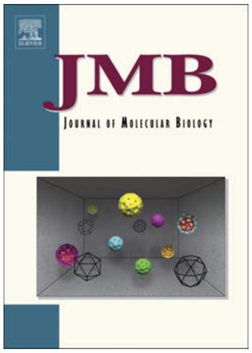


Using molecular dynamics simulations as an aid in the prediction of domain swapping of computationally designed protein variants

Yun Mou, Po-Ssu Huang, Leonard M. Thomas, Stephen L. Mayo

PII: S0022-2836(15)00346-0
DOI: doi: 10.1016/j.jmb.2015.06.006
Reference: YJMBI 64775

To appear in: *Journal of Molecular Biology*

Received date: 11 March 2015
Revised date: 11 June 2015
Accepted date: 16 June 2015

Using molecular dynamics simulations as an aid in the prediction of domain swapping of computationally designed protein variants

Yun Mou[a], Po-Ssu Huang[b,1], Leonard M. Thomas[c], and Stephen L. Mayo[a,d,2]

[a]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125; [b]Biochemistry and Molecular Biophysics Option, California Institute of Technology, Pasadena, CA 91125; [c]Department of Chemistry and Biochemistry, University of Oklahoma; and [d]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA 91125

[1]Present address: Department of Biochemistry, University of Washington, Seattle, WA 98195

[2]To whom correspondence should be addressed:

Stephen L. Mayo

Division of Biology and Biological Engineering, MC 114-96, California Institute of Technology,

1200 East California Blvd., Pasadena, CA 91125

phone: +1-626-395-6408

fax: +1-626-568-0934

e-mail: steve@mayo.caltech.edu

Keywords: molecular dynamics, computational protein design, *in silico* screening, homodimer, domain-swapped dimer

Abstract

In standard implementations of computational protein design, a positive-design approach is used to predict sequences that will be stable on a given backbone structure. Possible competing states are typically not considered, primarily because appropriate structural models are not available. One potential competing state, the domain-swapped dimer, is especially compelling because it is often nearly identical to its monomeric counterpart, differing by just a few mutations in a hinge region. Molecular dynamics (MD) simulations provide a computational method to sample different conformational states of a structure. Here, we tested whether MD simulations could be used as a post-design screening tool to identify sequence mutations leading to domain-swapped dimers. We hypothesized that a successful computationally-designed sequence would have backbone structure and dynamics characteristics similar to that of the input structure, and that in contrast, domain-swapped dimers would exhibit increased backbone flexibility and/or altered structure in the hinge-loop region to accommodate the large conformational change required for domain swapping. While attempting to engineer a homodimer from a 51 amino acid fragment of the monomeric protein engrailed homeodomain (ENH), we had instead generated a domain-swapped dimer (ENH_DsD). MD simulations on these proteins showed increased MD simulation derived B factors in the hinge loop of the ENH_DsD domain-swapped dimer relative to monomeric ENH. Two point mutants of ENH_DsD designed to recover the monomeric fold were then tested with an MD simulation protocol. The MD simulations suggested that one of these mutants would adopt the target monomeric structure, which was subsequently confirmed by X-ray crystallography.

Introduction

Computational protein design (CPD) provides *in silico* tools that facilitate the identification of amino acid sequences with desired properties. Most CPD algorithms sample an enormous number of amino acid types and side-chain conformations to find the most energy-favored sequences in the context of a single, fixed, backbone structure [1, 2]. Leveraging the speed of modern computers, CPD can effectively reduce the vast sequence space to an affordable number of sequences for experimental examination. CPD is particularly useful when combined with medium to high-throughput experimental screening and has led to successful designs for a variety of protein engineering problems [1, 3-10]. The utility of CPD, however, can be limited in applications where our understanding of the engineering problem is incomplete or an appropriate high-throughput experimental screening method does not exist. Another problem that can occur is that the designed sequence doesn’t fold into the desired structure but instead takes on the structure of a competing state, including unfolded or aggregated states. For example, Fleishman et al. [5] recently showed that only 2 of 88 CPD-designed variants from different protein scaffolds bound to the target molecule, influenza hemagglutinin. A community-wide assessment of this study suggested that many of the failed designs do not adopt the target fold [11]. In addition, only half express solubly [12], likely due to poor stability. The ability to predict whether a designed protein sequence will be correctly folded and stable prior to evaluating it experimentally would be extremely beneficial, as it would filter out “poor sequences” so that time-consuming and expensive experimental validations need only be done on sequences that are more likely to have the desired properties.

Although the conformational population of a protein depends on the relative energetic contributions of all possible states, most CPD methods evaluate designed sequences based on only one desired state. Consequently, even though the sequences obtained from these single-state designs may have acceptable CPD scores, this does not ensure that the desired fold dominates the population because the designed sequences may have better scores on structures other than the target state. Unfortunately, explicitly modeling alternative states is challenging because the structure of these states are typically unknown.

The stability, specificity, and activity of a protein often depend not only on the protein's structure but also on its dynamic properties. Protein variants with altered dynamics may lead to undesired outcomes such as amyloidogenesis [13, 14]. The goal of many protein engineering projects is therefore to maintain the basic structure and dynamics of the protein while improving a desired property (e.g., catalytic activity [15], thermostability [16], substrate specificity [17], ligand binding [18], and molecular transport [19]). However, protein dynamics is typically not directly modeled in CPD calculations. Molecular dynamics (MD) simulations provide a powerful tool for exploring local conformational ensembles of the native state and thus provide an opportunity to improve CPD by including an aspect of protein dynamics in the design process. Indeed, Allen et al. showed that MD generated structure ensembles could be successfully used for computational multi-state protein design [20].

MD simulations can also serve as a complementary tool to evaluate the dynamic properties of CPD-generated proteins. Both Kiss et al. [21] and Privett et al. [22] used MD simulations as a post-design screening method in the *de novo* design of an enzyme to catalyze the Kemp elimination reaction. In these studies, the dynamics of the substrate in the designed active site was monitored using MD. The population of competing states (i.e., substrate bound vs. substrate free) was calculated in the MD trajectories of putative enzyme variants and used as a filter to identify those likely to exhibit Kemp elimination activity. This approach proved successful and led to the development of a catalytically efficient computationally designed enzyme for the Kemp elimination [22] that was further optimized through directed evolution to yield an enzyme with kinetic parameters comparable to naturally evolved enzymes [15]. In addition, MD simulations have also been used to investigate the mechanisms of enzyme catalysis [23, 24].

Among the competing alternate structural states of proteins, domain-swapped dimers are common because they are often nearly structurally identical to their monomeric counterparts [25]. Altering one or two amino acids in a hinge-loop region can promote the conformational change needed for domain-swapped dimerization while keeping the rest of the protein structurally unchanged. Studies have shown that mutating these critical residues can significantly affect domain-swapping tendency [26, 27].

Given that domain-swapped dimer structures are usually not explicitly modeled in CPD calculations, it is not surprising that designed sequences that could assume these folds would be among those predicted, especially if the design involves alteration of loop residues. For example, O'Neill et al. used CPD to design the IgG-binding domain of protein L and found that one of the point mutants (G55A) led to a weak domain-swapped dimer with a dissociation constant (K_d) of ~30 μM [26] that was further stabilized by two additional mutations (A52V and D53P) resulting in an obligate domain-swapped dimer with a K_d of ~0.7 nM [27]. Similarly, while attempting to design a homodimer from a 51 amino acid fragment of the monomeric protein engrailed homeodomain (ENH), we generated a high affinity dimer, ENH_DsD (Table S1), with a K_d of ~40 nM that also proved to be domain-swapped when examined by X-ray crystallography. Comparison of the crystal structures of ENH and the domain-swapped dimer, ENH_DsD, suggested that domain swapping might be accommodated by opening of a putative hinge loop between the first and second helices. We hypothesized that ENH_DsD's ability to form a domain-swapped dimer would be revealed in higher backbone flexibility and/or altered local structure along this hinge loop in the corresponding monomeric state, and that the wild-type protein (ENH), which does not adopt a domain-swapped structure, would have lower flexibility and stably adopt the structure in this loop. We anticipated that these differences in loop flexibility and/or structure might be observable in MD simulations of the two proteins and set out to explore this possibility. As anticipated, short 20 ns MD simulations revealed greater flexibility in the hinge loop for ENH_DsD than for wild-type ENH (although the gross structure in the hinge-loop region remained essentially the same). Similarly, we reasoned that mutations to ENH_DsD that caused the protein to revert to the native ENH fold would also show wild-type-like hinge-loop flexibility and structure. This proved to be the case – an ENH_DsD point mutant that showed wild-type-like hinge-loop flexibility and structure was confirmed by X-ray crystallography to assume the wild-type native fold. To assess the general applicability of this MD simulation protocol, we also investigated domain-swapped oligomer variants of the IgG-binding domain of protein L and the IgG-binding domain of protein G and found that their hinge-loop properties correlated with the strength of domain-swapped oligomerization.

Results and Discussion

***De Novo* Homodimer Design Produced the Domain-Swapped Dimer ENH_DsD.** Our original goal was to use computational tools to engineer a *de novo* homodimer from a monomeric protein. We docked 51 amino acid fragments of two *Drosophila melanogaster* engrailed homeodomain monomers (PDB ID: 1ENH) [28] *in silico* to form a C2-symmetric homodimer [29] then used CPD to redesign the formed protein-protein interface in order to create affinity between the two identical chains. The oligomeric state(s) of the designed variants was determined using size exclusion chromatography and fluorescence polarization techniques.

To assist in these assays, a yellow fluorescent protein (YFP) [30] was fused to the C terminus of each of the designed sequences. Size exclusion chromatography revealed that one of the designed sequences, ENH_DsD, elutes primarily as a dimer (Fig. 1*A*). Fluorescence polarization was determined using a Förster resonance energy transfer (FRET) assay, which yielded a K_d of ~40 nM (Fig. 1*B*). The decreased polarization at higher concentrations is caused by the dimerization of ENH_DsD-YFP in which two nearby YFPs transfer energy to each other (homo-FRET). Note that these assays cannot distinguish between the target dimer configuration (where each chain retains the native ENH fold) and a domain-swapped dimer. X-ray crystallography was used to determine the detailed molecular architecture of ENH_DsD-YFP, resulting in a structure at 1.9 Å resolution (Fig. 2*A*; Table S2). Analysis revealed a domain-swapped dimer formed between two ENH_DsD sequences (Fig. 2*B*). In the domain-swapped structure, the first helix, the hinge loop, and the second helix form a single extended helix such that the first helix on one chain interacts with the second and third helices on the second chain (Fig. 2*C*). The rest of the structure essentially remains intact; excluding the hinge-loop region, superposition of ENH_DsD and wild-type ENH gives a C_α RMSD of 0.6 Å (Fig. 2*B*). Note that ENH_DsD shares only 49% sequence identity with wild-type ENH (25 out of 51 residues). To reduce the concern that the YFP fusion may play a role in causing the domain-swapping, we also solved the X-ray crystal structure of a closely related variant that lacks the YFP fusion and found that it also forms a domain-swapped dimer structure similar to ENH_DsD (Fig. S1 and Table S4).

Molecular Dynamics Simulations Suggest Increased Hinge-Loop B Factors are Associated with ENH_DsD Domain-Swapping. We hypothesized that flexibility and/or altered structure in the hinge-loop region could be observed with B factors computed from MD simulation, and that the B factors of the ENH_DsD sequence threaded onto the ENH native structure would show high flexibility for the hinge-loop region relative to wild-type ENH. We ran multiple 20 ns MD simulations on wild-type ENH and on the ENH_DsD sequence threaded onto the native ENH structure, and for each trajectory, calculated the backbone root mean square fluctuation (rmsf) for each C_α over the entire trajectory and converted these values to B factors. Three trajectories were run for each sequence and the average B factors were calculated and used as a measure of both altered structure and structural flexibility. As seen in Fig. 3*A*, nearly all the residues in wild-type ENH have low computed B factors with the exception of residues close to the N and C termini. The B factors for the ENH_DsD sequence threaded onto the native ENH structure (Fig. 3*B*) are generally larger than those of ENH – the average B factor for residues 4 through 48 are 17 Å^2 and 27 Å^2 for ENH and ENH_DsD, respectively (Table 1). Notably ENH_DsD residue 24 in the hinge-loop region has a B factor more than three times larger than the equivalent position in ENH, supporting the hypothesis that higher B factors in the hinge-loop region is associated with domain swapping.

ENH_DsD Mutant Designed to Recover Native Fold Reverts to Wild-Type Hinge-Loop Flexibility. Introduction of proline in the hinge-loop region was pursued as an approach to break the extended helical secondary structure seen in the ENH_DsD X-ray structure (Fig. 2*A*) and to recover the native ENH fold. Steric interference from the pyrrolidine ring in proline precludes the residue directly N-terminal to proline from adopting a helical structure. In addition, since prolines often occupy the first or second position of α-helices [31], we substituted a proline into position 23 (E23P) and separately position 24 (E24P) of ENH_DsD and examined the B factors of these variants using the MD simulation protocol described above. As can be seen in Fig. 3*C* for the E23P variant, the average of the B factors for the non-terminal residue positions for E23P, 18 Å^2, is nearly identical to wild-type ENH, 17 Å^2, and well below that of ENH_DsD at 27 Å^2

(Table 1). Moreover, the maximum B factor value in the hinge-loop region of E23P, 33 Å^2, is almost half that of ENH_DsD, 71 Å^2, and approaches that of wild-type ENH, 21 Å^2 (Table 1). In contrast, E24P (Fig. 3*D*) has characteristics that resemble ENH_DsD: the average overall B factor, 35 Å^2, is nearly twice that of wild-type ENH; and, the maximum hinge-loop B factor, 107 Å^2, is more than five times that of wild-type ENH (Table 1). These results suggest that the E23P sequence is stable in the ENH native fold and less likely to fold to an alternative structure, while the E24P is less stable in the ENH native fold and more likely to adopt an alternative fold in a manner similar to ENH_DsD.

We also performed 100-ns MD simulations on the ENH variants, which show trends qualitatively similar to the 20-ns simulations but are less distinctive (Fig. S2). For example, the B-factor enhancement of E24P at the hinge loop, while still apparent, is less obvious compared to the 20-ns simulation (Fig. 3D). A longer MD simulation might relax the structure more and find alternative conformations with less flexibility. The crystallographic input structure is clearly important for the MD protocol – the later time windows appear to lose important information present in the starting structures.

Biophysical Analyses Support MD Simulation Predictions. We constructed the E23P-YFP and E24P-YFP proteins and examined their dimerization properties using a homo-FRET assay. E23P-YFP forms a tight dimer with a K_d of ~10 nM (Fig. S3). However, E24P-YFP expresses poorly and shows no detectable dimerization at concentrations as high as 5 μM (anisotropy >300 mA; data not shown).

To explore the MD simulation suggestion that E23P is a regular dimer and not a domain-swapped dimer, we determined the X-ray crystal structure of E23P-YFP to 2.3 Å resolution (Fig. 4*A* and Table S3). Superposition of the wild-type ENH and E23P-YFP structures (Fig. 4*B*) shows that the single proline mutation at position 23 rescues the native fold. The C_α RMSD for E23P vs. ENH is 0.4 Å for all backbone atoms and 0.6 Å for the backbone atoms in the hinge-loop region. The X-ray structure thus confirms that the dimerization observed by homo-FRET results from a regular dimer (as opposed to a domain-swapped dimer). Notably, the dimer that is formed does not match the intended CPD homodimer design target (Fig. S4). Instead of a single four-helix bundle interface formed by association of the designed surfaces on the second and third helices, two interfaces are formed between each of the designed surfaces and a patch on YFP (Fig. 4*A*). This finding emphasizes the challenge of avoiding unwanted structural states in CPD and highlights the need for improved negative design approaches.

Protein L and Protein G Variants Test Applicability of MD Simulation Protocol. We also investigated variants of the B1 domain of protein L from *Peptostreptococcus magnus* generated by Baker and coworkers [26, 27] and variants of the B1 domain of streptococcal protein G generated by Gronenborn and coworkers [32, 33]. O'Neill et al. showed that the point mutant G55A of protein L is a weak domain-swapped dimer with a K_d of ~30 μM [26]. Kuhlman et al. used CPD to identify mutations that stabilize the domain-swapped dimer structure while destabilizing the monomer structure, creating a triple mutant (A52V/N53P/G55A) obligate dimer with a K_d of ~700 pM [27]. The proline substitution at position 53 is not compatible with the structure of the monomer and is expected, at a minimum, to alter the local structure. In protein G, Frank et al. showed that a quintuple mutant (L5V/A26F/F30V/Y33F/A34F) forms an intertwined tetramer [32]. Byeon et al. subsequently showed that the quadruple mutant (lacking the A26F mutation) forms a domain-swapping dimer [33]. We applied the MD simulation method described above to three protein L sequences (wild type, G55A, and the triple mutant) and three protein G sequences (wild type, the quadruple mutant, and the quintuple mutant) to evaluate whether the B factors of the hinge loop reflects the proteins' domain-swapping tendencies.

The two protein L and two protein G variant sequences were threaded onto their respective wild-type native monomeric structures (PDB IDs: 1HZ5 and 1PGB) [34] and the MD simulation protocol was applied. The torsion-angle strain caused by the N53P mutation in the protein L monomer structure was fully relaxed by energy minimization at the initial stage of the MD simulation protocol. Compared to wild-type ENH, wild-type protein L has larger B factors in the hinge-loop region (Fig. 5*A*) with a maximum B factor in that region almost four times that of ENH (Table 1), which emphasizes that relative B factors within a series of related variants will be more important than the absolute B factors of any individual protein. As can be seen in Fig. 5*B*, the hinge-loop B factors of G55A are reduced relative to wild-type protein L, suggesting that the MD simulation protocol is not able to detect variants that are only weakly domain-swapped. In contrast, the obligate domain-swapped dimer triple mutant shows a significant increase in hinge-loop B factors relative to wild-type protein L (Fig. 5*C*) with a maximum hinge-loop B factor greater than two times that of protein L (Table 1). Note that in addition to the hinge-loop region (residues 51-56), residues 41-45 of this triple mutant also display enhanced B factors compared to wild-type protein L. This points out a limitation of the MD protocol. Although the protocol can predict domain swapping, it may not unambiguously identify the hinge-loop region and thus require additional information in order to focus on the relevant protein segment.

In the case of protein G, the quadruple and the quintuple mutants show approximately two-fold and four-fold increases in the maximum hinge-loop B factor (at position 38) compared to wild-type protein G, respectively (Fig. 5D, 5E, 5F and Table 1). Note that although the quadruple mutant of protein G is a weak domain-swapped dimer with a K_d (~93 μM) similar to that of the G55A mutant of protein L (~30 μM), the MD simulation protocol can only identify the B factor increase for the protein G case. This might be explained by the extent of the mutations. The G55A mutant of protein L only changes the structure by a single methyl group, while the mutations introduced in the quadruple mutant of protein G are much more extensive (L5V/F30V/Y33F/A34F), suggesting that the G55A mutation might be too subtle for the MD simulation protocol used in this study.

Direct Computation of Monomer and Domain-Swapped Oligomer Stability. The availability of both monomeric and domain-swapped oligomer structures for ENH, protein L, and protein G variants allows the direct computation of CPD energies for the various sequences threaded onto different structural states. Table 1 summarizes the computed monomer and domain-swapped oligomer energies for the three protein series. With the exception of the E24P ENH variant, the monomer/domain-swapped oligomer fold preferences derived from computed stability match the MD simulation flexibility trends that are based on the monomer structure alone. The MD simulation results for the E24P variant suggest that the sequence is not stable on the native monomer fold, while the computed stabilities clearly indicate a preference for the native monomer fold over the domain-swapped dimer fold. Experimentally, the E24P variant is poorly behaved and shows no sign of dimer formation at concentrations as high as 5 μM (data not shown). Both the B factor analysis and CPD energy analysis incorrectly predict the behavior of the weakly dimeric G55A protein L variant (K_d of ~30 μM). However, the difference in computed monomer/dimer energies for wild-type protein L and the G55A variant, –32 and –8 Rosetta units, respectively, suggest that the G55A variant is less stable on the native monomer fold and more stable on the domain-swapped dimer fold relative to wild-type protein L. Overall, in the absence of explicit models for alternative structural states, MD simulation based B factor analysis appears to be a good surrogate for predicting the domain-swapping behavior of CPD-derived amino acid sequences.

MD Simulation as a Post-CPD Screening Method. MD simulations can serve as a complement to CPD by mitigating common method limitations such as the use of discrete side-chain rotamers, fixed protein backbone, and absence of explicit water. The continuous structure space of MD simulations allows for a much more accurate description of protein "softness." The explicit inclusion of water is also a key feature for accurate protein modeling, especially for surface/pocket designs, loop designs, binder designs, and enzyme designs. Several groups have thus used MD simulations as a post-CPD filter for various engineering applications. For example, Kiss et al. [21] used MD simulations to evaluate and re-rank variants obtained from *de novo* enzyme designs for the Kemp elimination reaction. Active designs could be clearly distinguished from inactive ones in 20 of the 23 cases tested, a significant improvement over the success rate of CPD alone, which had 8 active out of 59 tested designs [3]. Privett et al. used a similar MD simulation protocol to suggest a point mutant that resulted in a 3-fold improvement in Kemp elimination activity [22].

Liang et al. also reported studies in which MD simulations were used to validate designs [35]. They showed that the stability of a designed complex containing a computationally-grafted binding epitope could be evaluated by MD simulation. Although their root means square (rms) analysis requires a much longer MD simulation time (~400 ns), it is conceptually similar to our rmsf B factor protocol. The Liang et al. study emphasized the importance of protein dynamics—whereas MD analysis was able to distinguish stable protein complexes from unstable ones, monomer-dimer ΔG values calculated from static structures could not.

Conclusions

We presented a simple MD simulation protocol for evaluating the domain-swapping tendency of CPD-designed protein variants. The protocol consists of a 20 ns MD simulation of the designed sequence followed by rmsf analyses to generate B factors for backbone atoms. Applying this protocol to the domain-swapped dimer ENH_DsD, we found higher B factors in the hinge-loop compared to that of the monomeric wild-type protein. To recover the wild-type fold, two mutants, E23P and E24P, were made and examined using the MD simulation protocol. E23P exhibited wild-type-like backbone B factors and was therefore predicted to revert to the wild-type fold. Determination of the structure by X-ray crystallography confirmed the MD simulation prediction. We tested the applicability of our method on variants of protein L and protein G and similarly found that domain-swapping tendency correlated with elevated B factors in the hinge region. We anticipate that our MD protocol can be used as an *in silico* post-CPD filter, which may circumvent the need for time-consuming structural studies and may be most useful when experimental screening techniques are not adequate.

Materials and Methods

Construct Preparation, Expression, and Purification. Oligonucleotides (Integrated DNA Technologies) containing ~20 bp overlapping segments were assembled via a modified Stemmer polymerase chain reaction (PCR) method using KOD Hot Start Polymerase (Novagen) to generate the full-length designed sequence ENH_DsD (Table S1). The PCR product and YFP gene [30] with a C-terminal His_6 tag at the 3′ end were fused using overlap extension PCR. The ENH_DsD-YFP gene was then cloned into pET-11a using standard digestion/ligation methods. The E23P and E24P mutants were created by standard Quikchange protocols. The plasmids were transformed into *Escherichia coli* BL21(DE3) cells; colonies were picked, and the plasmids mini-prepped and sequenced. Sequence-verified constructs were expressed in standard Luria Broth at 37 °C using 1 mM isopropyl β-D-1-thiogalactopyanoside (IPTG) for 3 h. The cells were centrifuge harvested, sonication lysed, and Ni^{2+}-NTA (Qiagen) purified.

Size-Exclusion Chromatography. Size-exclusion chromatography was carried out at room temperature using an analytical Superdex-75 column (Amersham Pharmacia). After affinity column purification, each sample was loaded on an ÄKTA FPLC system with 0.5 mL sample volume and run at 0.5 mL/min flow rate with running buffer (100 mM NaCl and 20 mM Tris-HCl, pH 8.0). Absorbance at 520 nm was tracked for protein elution. Typically >10 mg protein/L cell culture was purified for each construct.

Fluorescence Polarization Assay. The fluorescence polarization was measured at room temperature with a Fluorolog-3 spectrofluorometer (HORIBA). ENH_DsD-YFP was serially diluted in buffer containing 100 mM NaCl and 20 mM TrisHCl at pH 8.0. The fluorescence anisotropy was measured for each sample and the G-factor was determined individually. The data were analyzed according to a simple monomer-dimer equilibrium model and fit with KaleidaGraph software. The anisotropy values (mA) for the completely monomeric and dimeric states were fit to be 330 and 260, respectively.

Crystallography. The ENH_DsD-YFP crystals were grown in 0.8 M monosodium phosphate, 1.2 M dipotassium phosphate, and 0.1 M sodium acetate at pH 4.5 using hanging-drop diffusion. The E23P-YFP crystals were grown in 0.1 M sodium chloride, 0.1 M 12% v/v/ 2-propanol, 0.1 M sodium acetate at pH 4.6 using hanging-drop diffusion. The Crystals were flash frozen in glycerol cryo-protectant and shipped to beamline 12-2 at Stanford Synchrotron Radiation Lightsource. Phases were obtained through molecular replacement using YFP as a model (PDB:

1MYW) [30]. Following molecular replacement, the ENH_DsD and E23P residues were built manually into the electron density maps using COOT [36]. Further refinement was performed using PHENIX [37]. Final coordinates were deposited in the Protein Data Bank with the codes 4NDJ for ENH_DsD-YFP and 4NDK for E23P-YFP. Data collection and refinement statistics are listed in Tables S2 and S3.

MD Simulations. The input structures for the MD simulations were prepared as follows. 1ENH, 1HZ5, and 1PGB were used as the template backbone structures for the engrailed homeodomain, IgG-binding domain of protein L, and IgG-binding domain of protein G sequences, respectively. The TRIAD computational protein design software was used for creating the mutant atomic coordinates. Sequences were first threaded onto the corresponding backbone structure and side-chain repacking optimization was applied on the full sequence. An improved version of FASTER was used for the side-chain repacking optimization [38] with a rotamer library based on the backbone-dependent library of Dunbrack and Karplus [39] and the ROSETTA scoring function [40]. The structures were then input to GROMACS 4.5.5 for energy minimization (GROMOS 43a1) with an explicit water box under periodic boundary conditions [41]. Charge neutrality was achieved by adding sodium or chloride ions. After energy minimization converged to Fmax < 1000 kJ/mol, a 20 ps position-restrained MD simulation was run for water relaxation at 300 K with 2 fs steps. Finally, a 20 ns unrestrained MD was run at 300 K under NPT conditions. The rmsf of C_α atoms was analyzed by g_rmsf built in GROMACS for the later 20 ns trajectory. A total of three trajectories initiated with different random seeds were run for each sequence and the averaged B factors ware calculated using the following equation: $B = (8\pi^2/3) \times (\text{rmsf})^2$, where the rmsf units are Å.

ACCESSION NUMBERS: Coordinates and structure factors have been deposited in the Protein Data Bank with accession number 4NDJ, 4NDK and 4ZN8.

ACKNOWLEDGMENTS. The authors are grateful for the use of beamline 12-2 at the Stanford Synchrotron Radiation Lightsource (SSRL) in Menlo Park, CA, operated by Stanford University and supported by the Department of Energy and the National Institutes of Health. We also thank Jens Kaiser and Pavle Nikolovski at the California Institute of Technology for their advice on crystallography. We thank the Gordon and Betty Moore Foundation for support of the Molecular Observatory at the California Institute of Technology. This work was supported by the Defense Advanced Research Projects Agency Protein Design Processes Program, a National Security Science and Engineering Faculty Fellowship (NSSEFF N00244-09-1-0011, N00244-09-1-0082) and the Gordon and Betty Moore Foundation through Grant GBMF2809 to the Caltech Programmable Molecular Technology Initiative. We are grateful to Marie Ary for assistance with the manuscript.

References

[1] Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. Science. 1997;278:82-7.

[2] Pantazes RJ, Grisewood MJ, Maranas CD. Recent advances in computational protein design. Curr Opin Struc Biol. 2011;21:467-72.

[3] Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, et al. Kemp elimination catalysts by computational enzyme design. Nature. 2008;453:190-5.

[4] Jiang L, Althoff EA, Clemente FR, Doyle L, Röthlisberger D, Zanghellini A, et al. De novo computational design of retro-aldol enzymes. Science. 2008;319:1387-91.

[5] Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch E-M, et al. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science. 2011;332:816-21.

[6] King NP, Sheffler W, Sawaya MR, Vollmar BS, Sumida JP, André I, et al. Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science. 2012;336:1171-4.

[7] Grigoryan G, Kim YH, Acharya R, Axelrod K, Jain RM, Willis L, et al. Computational design of virus-like protein assemblies on carbon nanotube surfaces. Science. 2011;332:1071-6.

[8] Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods. 2009;6:551-2.

[9] Stranges PB, Machius M, Miley MJ, Tripathy A, Kuhlman B. Computational design of a symmetric homodimer using β-strand assembly. Proc Natl Acad Sci USA. 2011;108:20562-7.

[10] Bolon DN, Mayo SL. Enzyme-like proteins by computational design. Proc Natl Acad Sci USA. 2001;98:14274-9.

[11] Fleishman SJ, Whitehead TA, Strauch E-M, Corn JE, Qin S, Zhou H-X, et al. Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol. 2011;414:289-302.

[12] Whitehead TA, Baker D, Fleishman SJ. Computational design of novel protein binders and experimental affinity maturation. Method Enzymol. 2013;523:1-19.

[13] Lim KH, Dyson HJ, Kelly JW, Wright PE. Localized structural fluctuations promote amyloidogenic conformations in transthyretin. J Mol Biol. 2013;425:977-88.

[14] Shukla UJ, Marino H, Huang P-S, Mayo SL, Love JJ. A designed protein interface that blocks fibril formation. J Am Chem Soc. 2004;126:13914-5.

[15] Blomberg R, Kries H, Pinkas DM, Mittl PRE, Grutter MG, Privett HK, et al. Precision is essential for efficient catalysis in an evolved Kemp eliminase. Nature. 2013;503:418-21.

[16] Malakauskas SM, Mayo SL. Design, structure and stability of a hyperthermophilic protein variant. Nat Struct Biol. 1998;5:470-5.

[17] Havranek JJ, Harbury PB. Automated design of specificity in molecular recognition. Nat Struct Biol. 2003;10:45-52.

[18] Tinberg CE, Khare SD, Dou JY, Doyle L, Nelson JW, Schena A, et al. Computational design of ligand-binding proteins with high affinity and selectivity. Nature. 2013;501:212-6.

[19] Koder RL, Anderson JLR, Solomon LA, Reddy KS, Moser CC, Dutton PL. Design and engineering of an O-2 transport protein. Nature. 2009;458:305-9.

[20] Allen BD, Nisthal A, Mayo SL. Experimental library screening demonstrates the successful application of computational protein design to large structural ensembles. Proc Natl Acad Sci USA. 2010;107:19838-43.

[21] Kiss G, Röthlisberger D, Baker D, Houk KN. Evaluation and ranking of enzyme designs. Protein Sci. 2010;19:1760-73.

[22] Privett HK, Kiss G, Lee TM, Blomberg R, Chica RA, Thomas LM, et al. Iterative approach to computational enzyme design. Proc Natl Acad Sci USA. 2012;109:3790-5.

[23] Jimenez-Oses G, Osuna S, Gao X, Sawaya MR, Gilson L, Collier SJ, et al. The role of distant mutations and allosteric regulation on LovD active site dynamics. Nat Chem Biol. 2014;10:431-6.

[24] Hanoian P, Liu CT, Hammes-Schiffer S, Benkovic S. Perspectives on electrostatics and conformational motions in enzyme catalysis. Accounts of chemical research. 2015;48:482-9.

[25] Liu Y, Eisenberg D. 3D domain swapping: As domains continue to swap. Protein Sci. 2002;11:1285-99.

[26] O'Neill JW, Kim DE, Johnsen K, Baker D, Zhang KY. Single-site mutations induce 3D domain swapping in the B1 domain of protein L from Peptostreptococcus magnus. Structure. 2001;9:1017-27.

[27] Kuhlman B, O'Neill JW, Kim DE, Zhang KY, Baker D. Conversion of monomeric protein L to an obligate dimer by computational protein design. Proc Natl Acad Sci USA. 2001;98:10687-91.

[28] Clarke ND, Kissinger CR, Desjarlais J, Gilliland GL, Pabo CO. Structural studies of the engrailed homeodomain. Protein Sci. 1994;3:1779-87.

[29] Huang P-S, Love JJ, Mayo SL. Adaptation of a fast Fourier transform-based docking algorithm for protein design. J Comput Chem. 2005;26:1222-32.

[30] Rekas A, Alattia JR, Nagai T, Miyawaki A, Ikura M. Crystal structure of Venus, a yellow fluorescent protein with improved maturation and reduced environmental sensitivity. J Biol Chem. 2002;277:50573-8.

[31] Choi EJ, Mayo SL. Generation and analysis of proline mutants in protein G. Protein Eng Des Sel. 2006;19:285-9.

[32] Kirsten Frank M, Dyda F, Dobrodumov A, Gronenborn AM. Core mutations switch monomeric protein GB1 into an intertwined tetramer. Nat Struct Biol. 2002;9:877-85.

[33] Byeon IJ, Louis JM, Gronenborn AM. A protein contortionist: core mutations of GB1 that induce dimerization and domain swapping. J Mol Biol. 2003;333:141-52.

[34] O'Neill JW, Kim DE, Baker D, Zhang KYJ. Structures of the B1 domain of protein L from Peptostreptococcus magnus with a tyrosine to tryptophan substitution. Acta Crystallogr D. 2001;57:480-7.

[35] Liang S, Li L, Hsu WL, Pilcher MN, Uversky V, Zhou Y, et al. Exploring the molecular design of protein interaction sites with molecular dynamics simulations and free energy calculations. Biochemistry. 2008;48:399-414.

[36] Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D. 2004;60:2126-32.

[37] Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D. 2010;66:213-21.

[38] Allen BD, Mayo SL. Dramatic performance enhancements for the FASTER optimization algorithm. J Comput Chem. 2006;27:1071-5.

[39] Dunbrack RL, Jr., Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. Journal of molecular biology. 1993;230:543-74.

[40] Das R, Baker D. Macromolecular modeling with Rosetta. Annu Rev Biochem. 2008;77:363-82.

[41] Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. J Chem Theory Comput. 2008;4:435-47.

[42] Shah PS, Hom GK, Ross SA, Lassila JK, Crowhurst KA, Mayo SL. Full-sequence computational design and solution structure of a thermostable protein variant. Journal of molecular biology. 2007;372:1-6.

Figure Legends

Fig. 1. (*A*) Size-exclusion chromatography for ENH_DsD-YFP (black) and a monomeric control, UVF-YFP (gray). UVF is a computationally designed 39-fold mutant of ENH whose NMR solution structure matches the wild-type (monomeric) fold [42]. (*B*) Fluorescence polarization of ENH_DsD-YFP.

Fig. 2. (*A*) X-ray crystal structure of ENH_DsD-YFP resolved to 1.9 Å. The two chains in ENH_DsD are shown in green and cyan and the YFP sequences are shown in gray. (*B*) Superposition of ENH_DsD portion of ENH_DsD-YFP with wild-type ENH structure (PDB ID: 1ENH) (gray). (*C*) The hinge loop rearrangement (arrow) between the first and second helices in the wild-type ENH structure (gray) and ENH_DsD-YFP (cyan). This change leads to an ~180º rearrangement of helix 2 such that helix 1, the hinge, and helix 2 form a single long helix in ENH_DsD.

Fig. 3. MD simulation-based B factor analyses for wild-type ENH (*A*), ENH_DsD (*B*), E23P (*C*), and E24P (*D*). Hinge residues are indicated with a gray bar in each panel.

Fig. 4. (*A*) X-ray crystal structure of E23P-YFP resolved to 2.3 Å. The E23P sequence is shown in green and cyan for each of the two chains and the YFP sequences are shown in gray. (*B*) Superposition of E23P portion of E23P-YFP with the wild-type ENH structure (gray).

Fig. 5. MD simulation-based B factor analyses for protein L and protein G variants. (*A*) Wild-type protein L. (*B*) protein L single mutant G55A. (*C*) Protein L triple mutant A52V/D53P/G55A. (*D*) Wild-type protein G. (*E*) Protein G quadruple mutant L5V/F30V/Y33F/A34F. (*F*) Protein G Quintuple mutant L5V/A26F/F30V/Y33F/A34F. Hinge residues are indicated with a gray bar in each panel.

Table 1. Comparison of MD-derived hinge-loop B factors with domain swapping, K_d values, and CPD energies for ENH, protein L, and protein G variants.

	Domain swappe	K_d *	Average Protein B-factor	Maximum hinge-	Monomer CPD	Domain-swapped oligomer CPD
ENH (WT)	No	n/a	17.	21.1	-	-
ENH_DsD	Yes	40	27.	70.9	-	-
E23P	No	10	18.	32.7	-	160.
E24P	n/a	n/a	35.	106.	-	104.
Protein L (WT)	No	n/a	27.	78.4	-	-
G55A	Yes	30,00	21.	58.2	-	-
A52V/D53P/G55A	Yes	0.700	34.	179.	249.	-
Protein G (WT)	No	n/a	14.	21.9	-	-
L5V/F30V/Y33F/A34F	Yes	93,00	19.	41.8	-	-
L5V/A26F/F30V/Y33F/A3	Yes	n/a	29.	86.6	-8.1	-

*Dissociation constant for dimer if applicable.
§The domain-swapped dimer or tetramer energy divided by two or four for comparison with the monomer energy.
†Average B factors for residues 4 through 48, 4 through 61, and 4 through 53 for ENH, protein L, and protein G variants, respectively.
‡Maximum B factors for residues 24, 53, and 38 for ENH, protein L, and protein G variants, respectively.

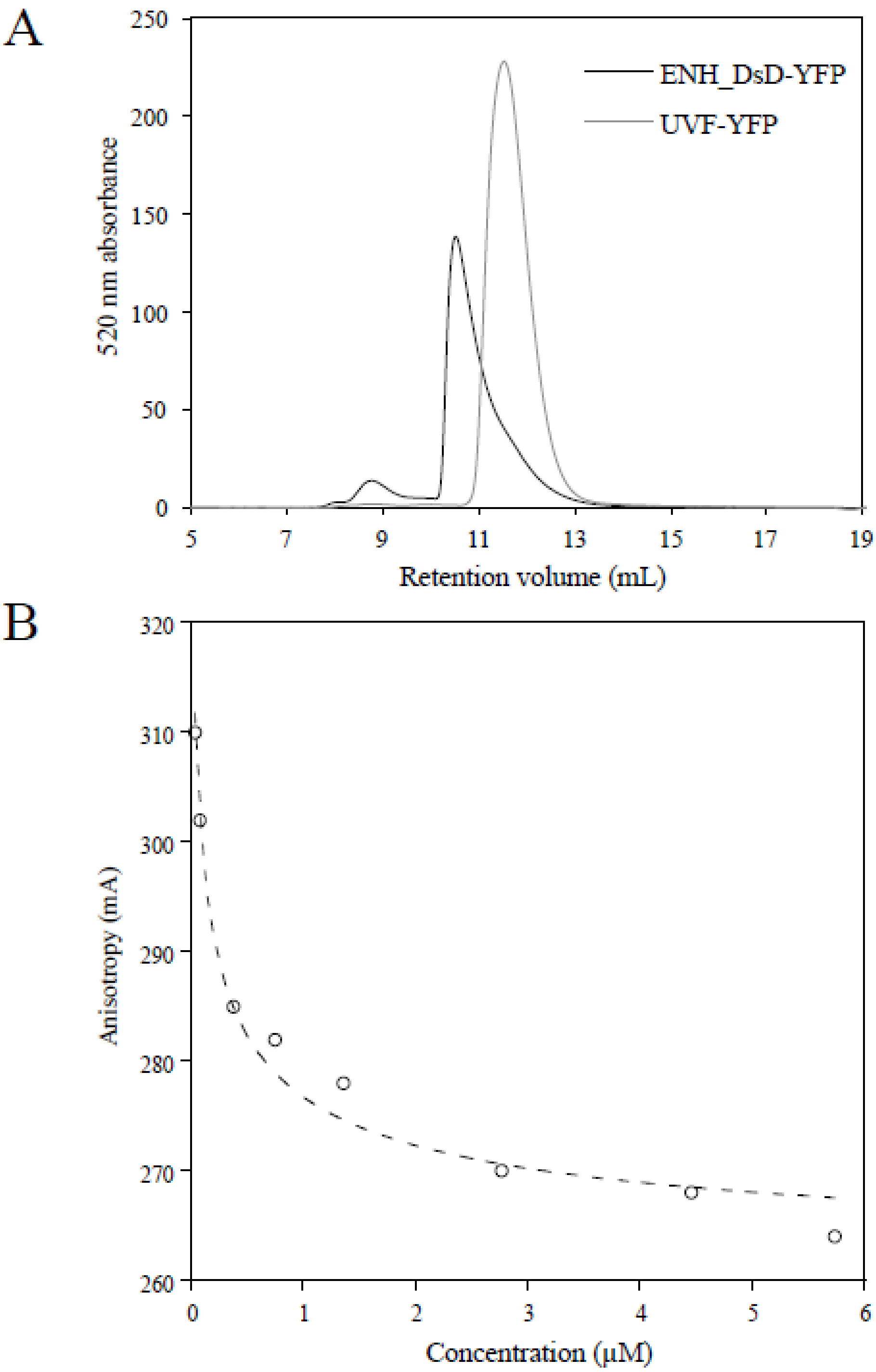

A
250
200
150
100
50
0
520 nm absorbance
ENH_DsD-YFP
UVF-YFP
5
7
9
11
13
15
17
19
Retention volume (mL)
B
320
310
300
290
280
270
260
Anisotropy (mA)
0
1
2
3
4
5
6
Concentration (µM)

Figure 1

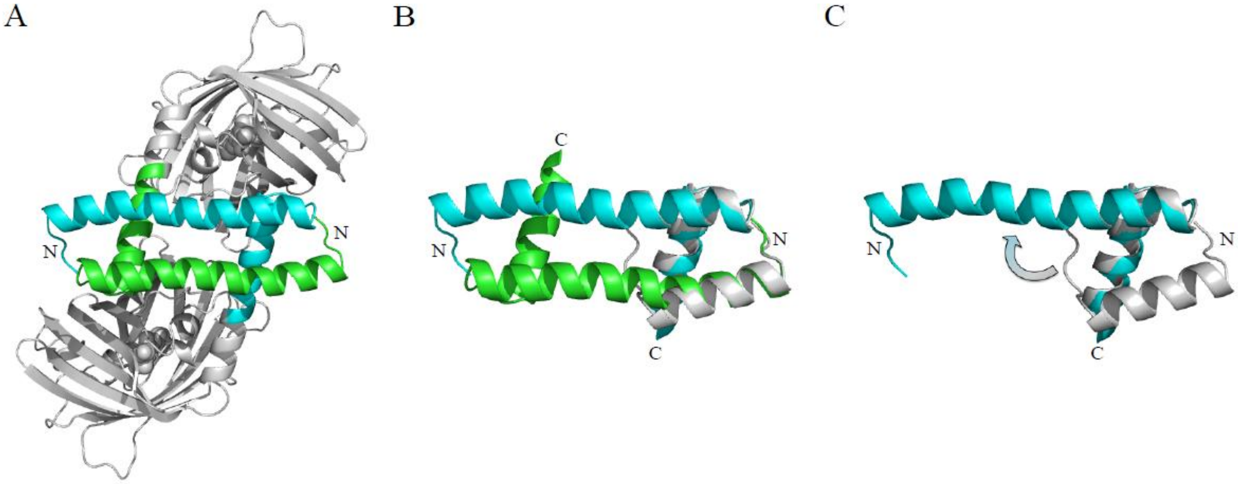


Figure 2

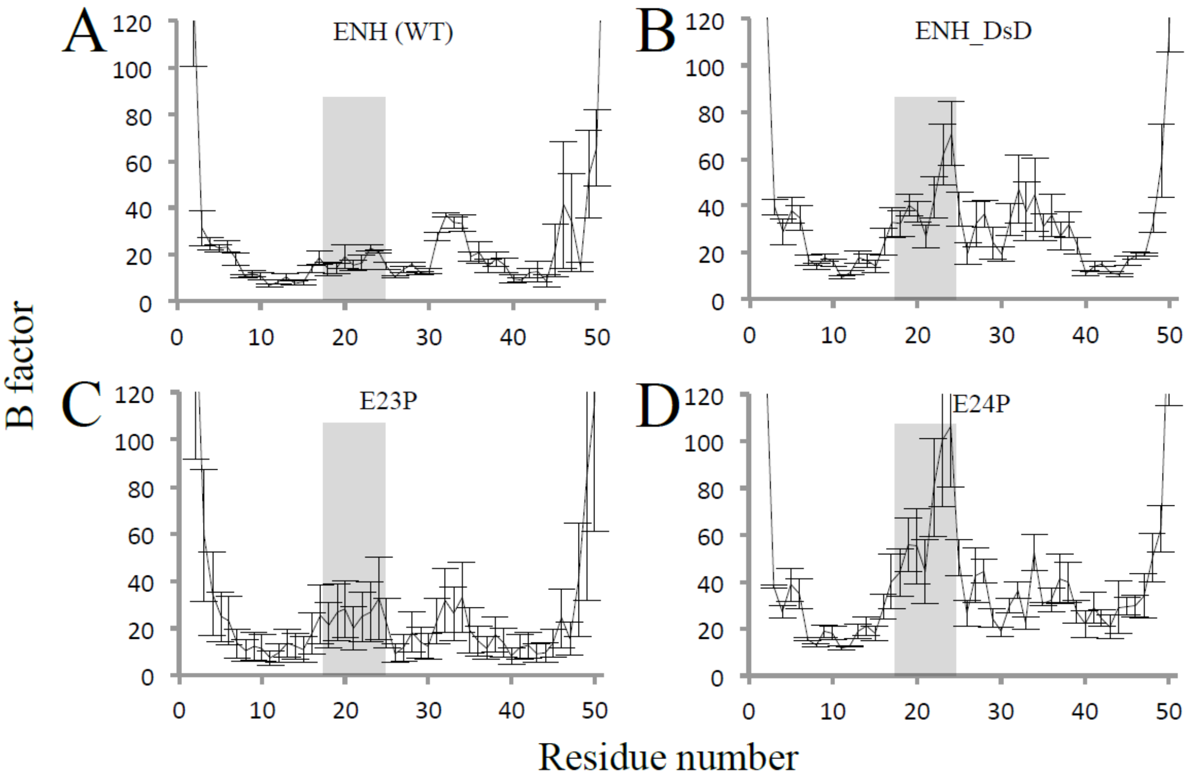

A
ENH (WT)
B
ENH_DsD
C
E23P
D
E24P
B factor
Residue number
0
20
40
60
80
100
120
0
10
20
30
40
50

Figure 3

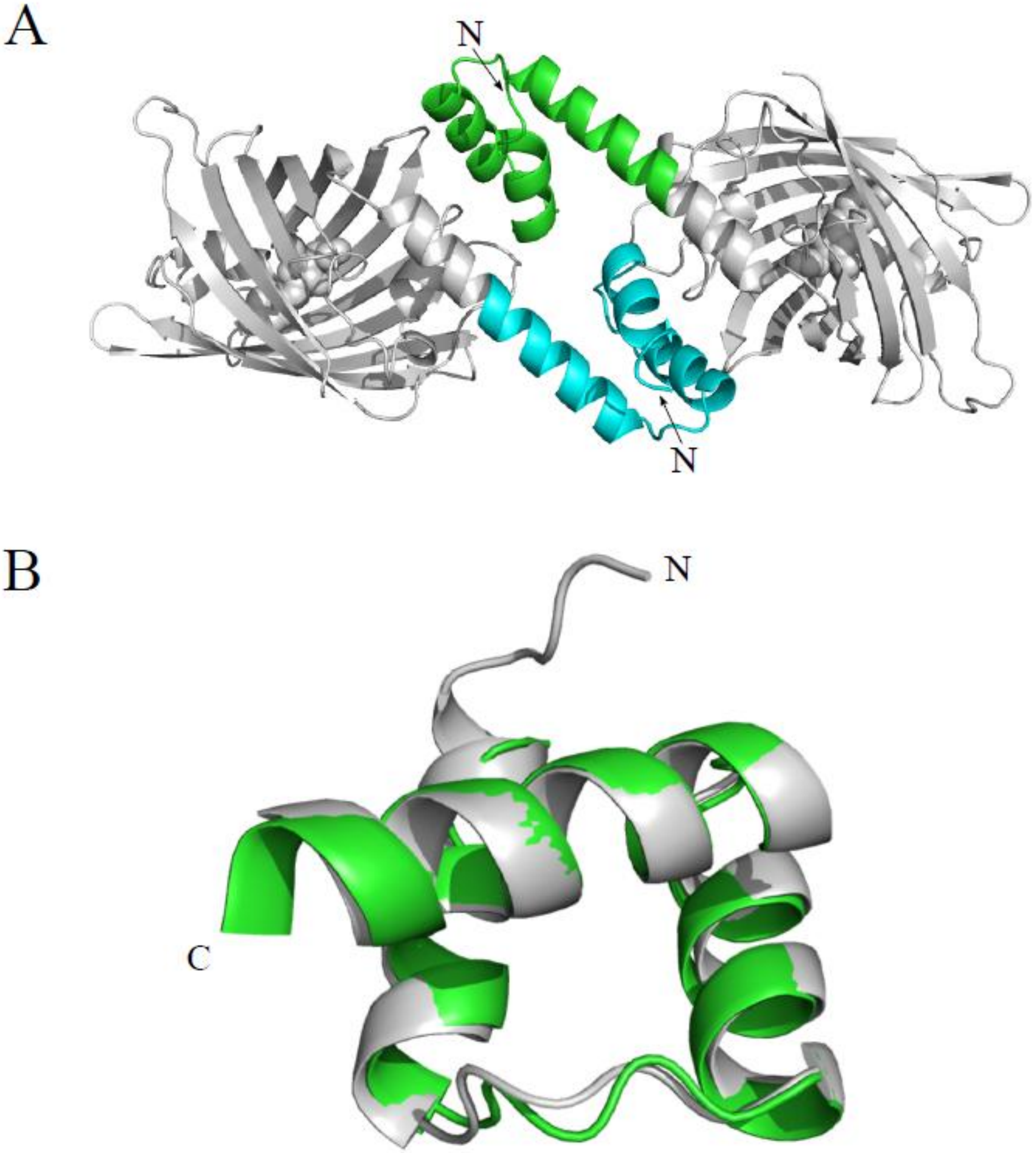


Figure 4

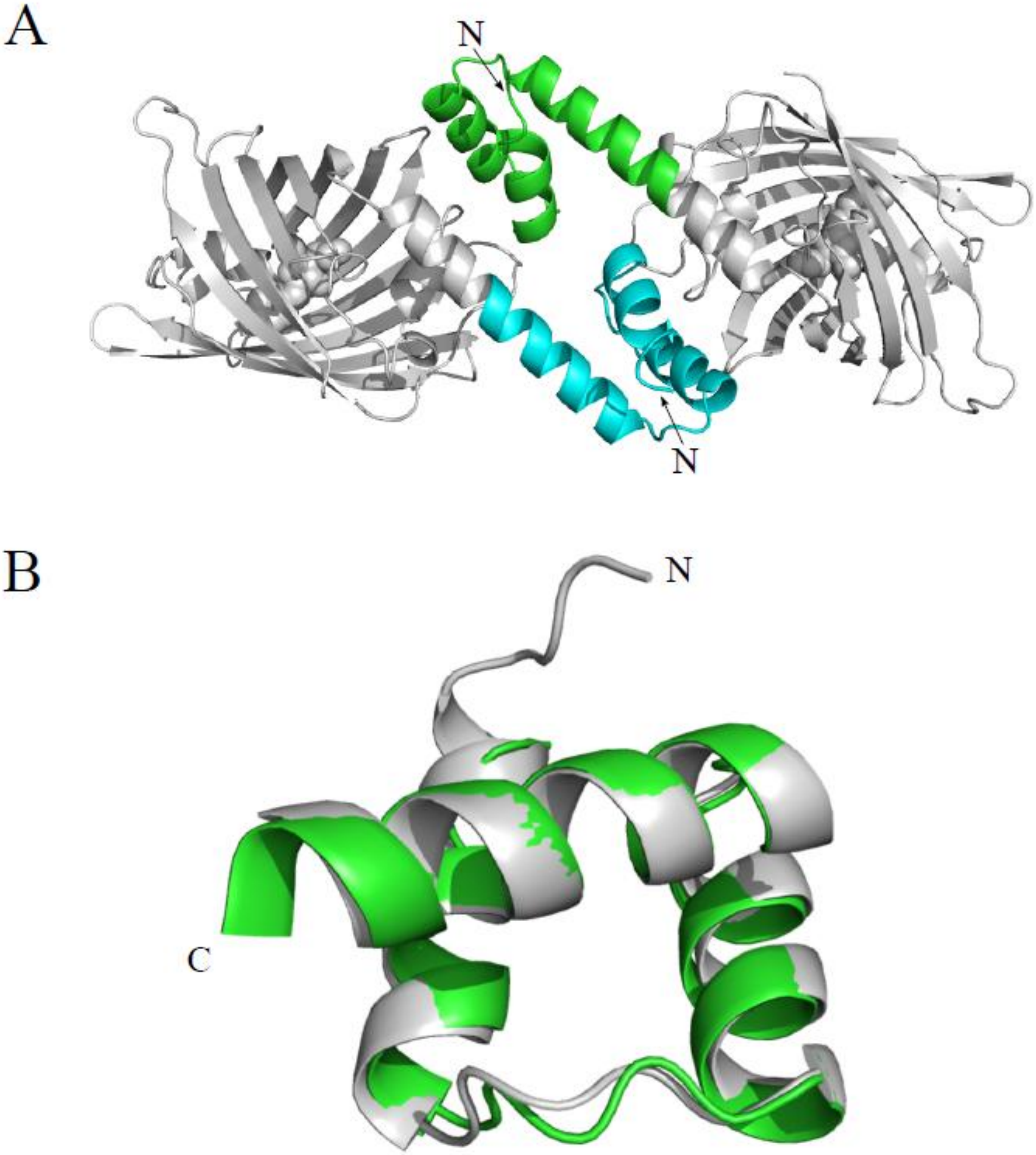
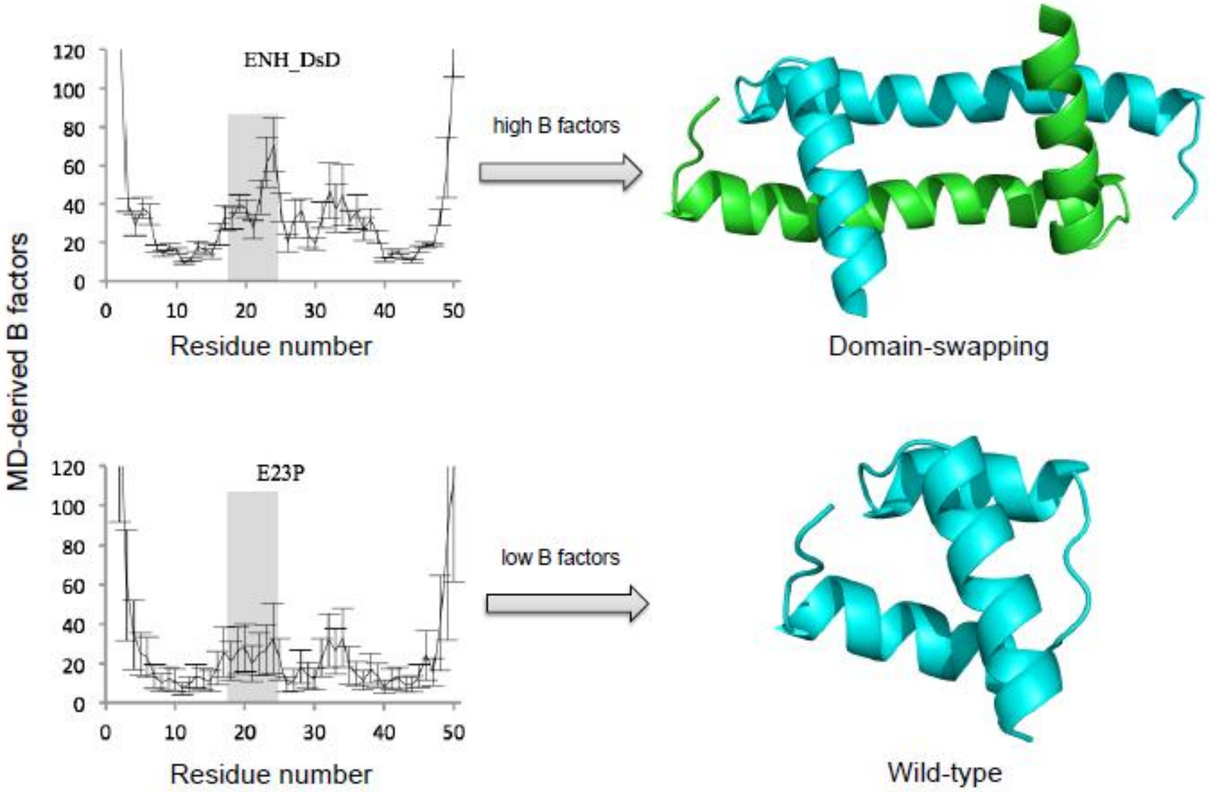


Figure 5



Graphical abstract

Highlights

- CPD calculations that do not consider competing states may lead to off-target folding.

- We developed a MD simulation protocol as a post-CPD screening tool.

- The MD protocol identifies mutations leading to undesired competing states.

- The MD protocol predicts mutations that favor the target fold.

- CPD combined with MD screening can greatly improve design success rates.