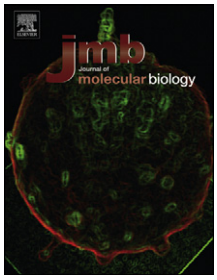

Using Molecular Dynamics Simulations as an Aid in the Prediction of Domain Swapping of Computationally Designed Protein Variants

Yun Mou [1], Po-Ssu Huang [2,1], Leonard M. Thomas [3] and Stephen L. Mayo [1,4]

*1 - **Division of Chemistry and Chemical Engineering,** California Institute of Technology, Pasadena, CA 91125, USA*
*2 - **Biochemistry and Molecular Biophysics Option,** California Institute of Technology, Pasadena, CA 91125, USA*
*3 - **Department of Chemistry and Biochemistry,** University of Oklahoma, Norman, OK 73019, USA*
*4 - **Division of Biology and Biological Engineering,** California Institute of Technology, Pasadena, CA 91125, USA*

***Correspondence to Stephen L. Mayo:** Fax: +1 626 568 0934. steve@mayo.caltech.edu*
http://dx.doi.org/10.1016/j.jmb.2015.06.006
Edited by G. Hummer

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 with 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 *B*-factors derived from MD simulation 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 does not 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.* recently showed that only 2 of 88 CPD-designed variants from different protein scaffolds bound to the target molecule, influenza hemagglutinin [5]. 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 performed 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 structures 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 multistate 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 *versus* 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 with 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