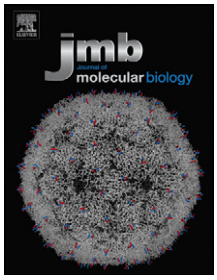

Computational Redesign of Thioredoxin Is Hypersensitive toward Minor Conformational Changes in the Backbone Template

Kristoffer E. Johansson[1,†], Nicolai Tidemand Johansen[1,†], Signe Christensen[1], Scott Horowitz[2], James C.A. Bardwell[2], Johan G. Olsen[1], Martin Willemoës[1], Kresten Lindorff-Larsen[1], Jesper Ferkinghoff-Borg[3], Thomas Hamelryck[4] and Jakob R. Winther[1]

*1 - **Linderstrøm-Lang Centre for Protein Science,** Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, Copenhagen DK-2200, Denmark*
*2 - **Howard Hughes Medical Institute,** Department of Molecular, Cellular and Developmental Biology, University of Michigan, 109 Zina Pitcher Place, Ann Arbor, MI 48109, USA*
*3 - **Biotech Research and Innovation Centre,** University of Copenhagen, Ole Maaløes Vej 5, Copenhagen DK-2200, Denmark*
*4 - **Section for Computational and RNA Biology,** Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, Copenhagen DK-2200, Denmark*

***Correspondence to Jakob R. Winther:** JRWinther@bio.ku.dk*
http://dx.doi.org/10.1016/j.jmb.2016.09.013
Edited by Arne Skerra

Abstract

Despite the development of powerful computational tools, the full-sequence design of proteins still remains a challenging task. To investigate the limits and capabilities of computational tools, we conducted a study of the ability of the program Rosetta to predict sequences that recreate the authentic fold of thioredoxin. Focusing on the influence of conformational details in the template structures, we based our study on 8 experimentally determined template structures and generated 120 designs from each. For experimental evaluation, we chose six sequences from each of the eight templates by objective criteria. The 48 selected sequences were evaluated based on their progressive ability to (1) produce soluble protein in *Escherichia coli* and (2) yield stable monomeric protein, and (3) on the ability of the stable, soluble proteins to adopt the target fold. Of the 48 designs, we were able to synthesize 32, 20 of which resulted in soluble protein. Of these, only two were sufficiently stable to be purified. An X-ray crystal structure was solved for one of the designs, revealing a close resemblance to the target structure. We found a significant difference among the eight template structures to realize the above three criteria despite their high structural similarity. Thus, in order to improve the success rate of computational full-sequence design methods, we recommend that multiple template structures are used. Furthermore, this study shows that special care should be taken when optimizing the geometry of a structure prior to computational design when using a method that is based on rigid conformations.

Introduction

The ability to routinely design new functional proteins and protein-based systems will significantly impact the development of novel technologies and medicinal products and also our basic understanding of proteins. One of the major challenges in this regard is the ability to rationally design an entire amino acid sequence that will adopt a given three-dimensional structure. To handle the vast complexity of full-sequence design, computational methods are particularly interesting. Analytical or non-computational approaches have successfully been applied to the full-sequence design of α-helical structures [1–5], for example, by using heptad repeats [6,7]. Also, small and less regular structures have been designed by non-computational consensus approaches and fragment assembly [8,9]. However, designing larger (>70 aa) globular αβ proteins with irregular contact patterns is a highly complex task and has only been achieved by employing

computational methods [10–14]. In addition to this unique achievement, computational methods have been employed to full-sequence design of a variety of protein structures including early mini-proteins [15–17], tandem repeats [18,19], and ligand binders [20–21]. Despite much effort, however, the total number of full-sequence designed proteins for which an atomic resolution structure has been solved still remains low; to our knowledge, less than 10 larger globular αβ proteins have been reported in the literature [10,12–14,22]. Among the available computational methods used here, Rosetta is by far the best validated, and thus, we base this study on the Rosetta software.

This relatively low number of successful designs highlights the need for further development of computational protein design methods. To our knowledge, all computational methods capable of optimizing an entire amino acid sequence of more than 100 residues approximate side-chain degrees of freedom by a discrete, typically small number of rigid conformers referred to as rotamers [23], and the backbone is kept completely fixed during sequence optimization. We will refer to this setup as the use of rigid conformations. Today, most design protocols optimize sequence and conformation iteratively. However, in the sequence optimization step, the backbone and rotamer conformations are always fixed. With a vast number of sequence combinations to be explored, the use of rigid conformations greatly reduces the complexity of the sequence optimization.

While being a key enabling factor in terms of computational time, the use of rigid conformations is also considered to be the main factor limiting the accuracy, and in practice, it limits the application to relatively rigid proteins [24–27]. In particular, the appearance of molten globule characteristics in non-successful designs have been associated with a lack of tight packing in the hydrophobic core caused by the use of rigid conformations [10,28]. To achieve a more accurate, comparative computational evaluation of structures, it is necessary to optimize the geometry using all degrees of freedom [29].

When based on a single template structure, design methods based on rigid conformations are known to converge to a narrow distribution of sequences [11]. In contrast, using more templates that display minor conformational differences increases the sequence variation of the output drastically [30,31]. Together, this indicates that computationally designed sequences based on a single rigid backbone template will only result in a small subset of the sequence solution space, as defined by the applied energy function, while another template of the same fold will yield another subset of solutions even assuming the same energy function.

To investigate this, we have conducted a full computational design study in which designed sequences based on several template structures were experimentally evaluated in an unbiased fashion. The thioredoxin fold was chosen as a design target because this fold is highly conserved throughout evolution and is also realized by a large variety of sequences in nature. Thus, we expect the thioredoxin fold to have a large sequence solution space and be highly designable in the sense that many sequences should be able to assume its fold. Furthermore, thioredoxin is a relatively rigid protein that is composed almost completely of segments with defined secondary structure (>90%) and has previously been shown to behave well in engineering contexts [32–34]. With a diversity of native sequences available, we tested templates with minor conformational changes (C_α RMSD = <2 Å), representing both a natural variation resulting from different wild-type sequences and a generated conformational variation resulting from computational geometry optimization. Starting with eight experimental template structures of the thioredoxin fold, we found a significant difference between the sequence outputs and, interestingly, also in the performance in experimental evaluations from template to template. In line with previous studies, we attribute these differences in template performance to the use of rigid conformations and show that these effects are enhanced by conducting thorough geometry optimization prior to design.

Results and Discussion

Design templates

To find a suitable set of templates to represent most of the natural thioredoxin sequence space, we searched the Protein Data Bank (PDB) for structures of the thioredoxin fold that shared high structural similarity despite low similarity in amino acid sequence. To enable a direct comparison of equivalent sequence positions in the resulting designs, we considered only the sets with a gap-free alignment. The search resulted in eight structures, which were truncated to the common, most structured 104 residues (Table 1). The structures are highly similar in backbone structure (Fig. 1) with an average C_α RMSD of 1.2 Å (0.7–1.8 Å) but are diverse in amino acid sequence with an average pairwise identity of 33% (15–61%; Fig. 2). Although it should not matter in principle, we note that all structures have been determined from protein expressed in *Escherichia coli*, the same host that we used for expression here.

Prior to computational design of an experimental structure, the Rosetta manual recommends preparing a structure that includes geometry optimization in an effort to remove disagreements between the experimental structure and the energy function, both of which potentially contain inaccuracies[‡]. The motivation is that a disagreement between the structure and the energy function, for example, an atomic overlap,