Journal of Structural Biology xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Combining Rosetta with molecular dynamics (MD): A benchmark of the MD-based ensemble protein design

Jan Ludwiczak[a,b], Adam Jarmula[b], Stanislaw Dunin-Horkawicz[a,*]

[a] Laboratory of Structural Bioinformatics, Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland
[b] Laboratory of Bioinformatics, Nencki Institute of Experimental Biology, Pasteura 3, 02-093 Warsaw, Poland

ARTICLE INFO

Keywords:
Protein design
Backbone flexibility
Molecular dynamics
Rosetta

ABSTRACT

Computational protein design is a set of procedures for computing amino acid sequences that will fold into a specified structure. Rosetta Design, a commonly used software for protein design, allows for the effective identification of sequences compatible with a given backbone structure, while molecular dynamics (MD) simulations can thoroughly sample near-native conformations. We benchmarked a procedure in which Rosetta design is started on MD-derived structural ensembles and showed that such a combined approach generates 20–30% more diverse sequences than currently available methods with only a slight increase in computation time. Importantly, the increase in diversity is achieved without a loss in the quality of the designed sequences assessed by their resemblance to natural sequences. We demonstrate that the MD-based procedure is also applicable to *de novo* design tasks started from backbone structures without any sequence information. In addition, we implemented a protocol that can be used to assess the stability of designed models and to select the best candidates for experimental validation. In sum our results demonstrate that the MD ensemble-based flexible backbone design can be a viable method for protein design, especially for tasks that require a large pool of diverse sequences.

1. Introduction

Computational protein design is a fast-growing field of structural bioinformatics (Huang et al., 2016; Lupas, 2014), including not only the creation of entirely new protein structures (Bhardwaj et al., 2016), but also new functions (Burton et al., 2016; Joh et al., 2014). Two key aspects of design are: 1) defining initial backbones and 2) calculating sequences that will fold to their structures. The backbones can be obtained from known structures, parametric equations (Boyken et al., 2016; Lupas et al., 2017) or assemblies of short peptide fragments. Given the essentially infinite number of amino acids combinations in a protein chain of typical length, the problem of sequence calculation is usually solved with the aid of heuristic algorithms, such as Monte Carlo optimization implemented in Rosetta Design. In a typical scenario, design procedure is started for a thousand of times on a single backbone structure to calculate a diverse set of suboptimal sequences. The main criterion in selecting the best designs for further experimental validation is the score provided by the Rosetta energy function. Models can subsequently be filtered according to the packing quality (Sheffler and Baker, 2009), hydrogen bonding patterns (Boyken et al., 2016), foldability (Bhardwaj et al., 2016; Murphy et al., 2012) and stability in molecular dynamics simulations (Bhardwaj et al., 2016).

Considering the dynamic nature of a protein structure is essential for a successful design as it allows for the accommodation of the side-chains that would be otherwise rejected by the energy function, e.g. due to the steric hindrances. Within the Rosetta environment, backbone flexibility can be achieved by iteratively repeating sequence design on a fixed backbone and full-atom relaxation (design and relax protocol; D&R). Alternatively, flexibility can be simulated through multiple parallel design simulations on an ensemble of relatively similar (1–2 Å RMSD) initial backbones. Such backbone ensembles can be generated based on simulations of full-atom or backbone structures (Rosetta Backrub protocol (Davis et al., 2006) or KIC (Mandell et al., 2009), respectively), parametric equations (Boyken et al., 2016), or the homologous structures available in the PDB database (Sun and Kim, 2017). Flexible backbone design approaches provide better sampling of the sequence space and generate more diverse sequences thus increasing the chances of identifying highly-designable backbones, for which a substantial number of high-scoring sequences can be designed. For this reason, flexible backbone design is routinely applied to various design tasks such as *de novo* design, re-design of a known folds (Murphy et al., 2012), probing the designability (Szczepaniak et al., 2014), or testing

* Corresponding author.
E-mail address: s.dunin-horkawicz@cent.uw.edu.pl (S. Dunin-Horkawicz).

https://doi.org/10.1016/j.jsb.2018.02.004
Received 5 November 2017; Received in revised form 25 January 2018; Accepted 13 February 2018

the mutational robustness (Humphris-Narayanan et al., 2012).

Given the number of various flexible and fixed backbone design approaches, it is essential to define a benchmark procedure to compare them in an accurate way. A straightforward bioinformatics method to assess the quality of the designs was proposed by Kortemme and co-workers (Ó Conchúir et al., 2015; Ollikainen and Kortemme, 2013) in which each design method was initialized on a representative set of experimentally-determined protein structures and the resulting designed sequences were compared to natural sequences (homologs of a given structure). The comparisons were made in terms of sequence profile similarity, which measures the agreement between the positional probabilities of occurrence of the amino acids and covariation similarity, where the emphasis is put on the overlap between the sequence covariation patterns. The covariation similarity coefficient is an especially important metric, as it reflects sequence features resulting from the interactions critical for structural stability. Another important parameter is the sequence entropy, which quantifies the diversity of the sequences. Comparison of designed and natural sequences allows to check whether sequence constraints imposed by the input backbone are fulfilled in a similar way by the computational design and the natural selection. Design procedures that yield sequences with the most natural-like properties are assumed to be best-performing.

MD simulations provide an opportunity to describe the heterogeneity of a protein structure, depicting it as an ensemble of the physically possible conformations rather than as a single, unique state. Various applications of MD simulations in protein design were recently reviewed (Childers and Daggett, 2017) and include, but are not limited to, the guidance of the stability improvement (Alfarano et al., 2012; Joo et al., 2011), design or re-design of proteins functional sites (Liang et al., 2009; Privett et al., 2012), and evaluation or ranking of the models (Bhardwaj et al., 2016; Kiss et al., 2013, 2010). Additionally, the iterative MD-Rosetta protocol was recently found to significantly improve the quality of Cryo-EM models (Leelananda and Lindert, 2017; Lindert and McCammon, 2015). Another straightforward application of MD simulations in protein design is modeling of the backbone flexibility by generation of the structural ensembles that can be used as a starting point for design simulations (Babor et al., 2011; Fu et al., 2007; Li et al., 2009; Schenkelberg and Bystroff, 2016; Sun et al., 2016). So far, to the best of the authors knowledge, the MD-based design procedures have not been systematically benchmarked. This has become increasingly important with the advent of GPU computing, as MD simulations have become more accessible to the community, providing performances comparable to large CPU clusters, achievable on a handful of graphic units (Nobile et al., 2016; Salomon-Ferrer et al., 2013).

We assessed the applicability of MD simulations in sequence design using Rosetta and benchmarked various design procedures (Fig. 1, Table 1.). It is to be expected that ensembles generated based on full-atom structures may tend to yield more native-like sequences when used for design. Therefore, we benchmarked design protocols starting either from full-atom structures (Fig. 1A) or from backbone structures (Fig. 1B). Since neither MD nor other ensemble generation methods such as Backrub can be executed directly on backbone structures, for the second scenario we implemented a two-step protocol in which a preliminary design is performed to generate full-atom structures that are subsequently used as an input to MD or Backrub. In our study, we adhered to the previously described design quality assessment procedure (Ó Conchúir et al., 2015) with some modifications described in the Methods section. Altogether, we tested 10 different design pipelines and found that the best results were obtained with the Rosetta design and relax protocol started on the backbone ensembles obtained by clustering of the MD trajectories. Moreover, we discuss the applicability of MD in the process of model selection.

2. Methods

2.1. Overview

To assess the applicability of MD-generated backbone conformational ensembles for protein design, we selected 12 Pfam domains together with their representative PDB structures and multiple sequence alignments (Table 2; the structures were selected to represent all major structural classes). The structural backbone ensembles were generated from 200 ns all-atom MD simulations of the 12 PDB structures and then extracting 500 representative backbone conformations from each trajectory (Fig. 1A). 500 independent Backrub simulations on the same set of PDB structures were also performed for comparison. Each of the ensembles (generated with either MD or with Backrub) served as a starting point for the design of 25,000 sequences (500 sequences per backbone conformation). In addition, for each of the 12 structures 25,000 sequences were designed using only the PDB structure (the ensemble generation step was omitted). The Rosetta sequence design was started in two modes: (a) a fixed backbone (FixBB) protocol (Kuhlman et al., 2003), which utilizes the simulated annealing approach to optimize the side-chain rotamers, while holding the backbone fixed and (b) a design and relax protocol (Murphy et al., 2012), where the cycles of the fixed backbone design and full-atom (side-chains and backbone) optimization are repeated iteratively. To benchmark the applicability of MD and Backrub for the design starting from backbone-only structures (*de novo* design) we included an additional sequence design step with a D&R protocol and selected 25 top-scoring models for the subsequent ensemble generation with either MD or Backrub (see Fig. 1B and the "Backbone ensemble generation from a single backbone structure" section). All relevant commands used to run Rosetta calculations are available in Supplementary File 3.

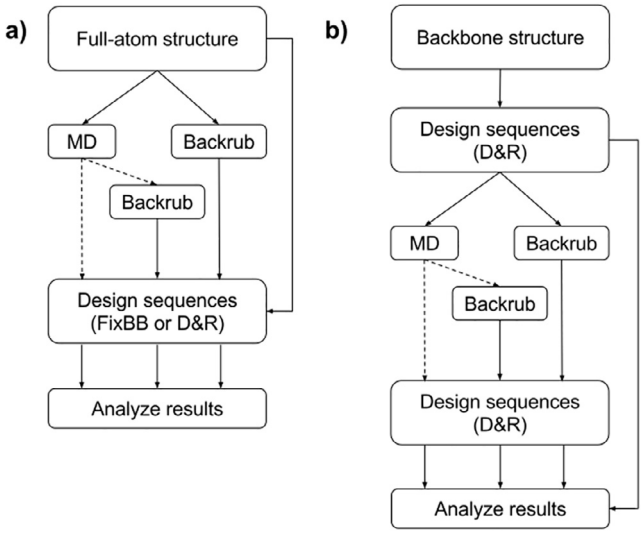

Fig. 1. Overview of the computational pipelines used in the study. The design was started either from (A) full-atom or (B) backbone structures. Dashed lines indicate the stages at which the combined clustering of MD trajectories is performed.

Table 1
Design protocols benchmarked in the study. IDs and names are used throughout the text to identify the individual protocols. Input, ensemble generation, and design protocol denote type of input structure, method for ensemble generation, and design method, respectively.

ID	Name	Input	Ensemble generation	Design protocol
F1	FixBB	Full-atom	None	FixBB
F2	FixBB + MD	Full-atom	MD	FixBB
F3	FixBB + BR	Full-atom	BR	FixBB
F4	D&R	Full-atom	None	D&R
F5	D&R + MD	Full-atom	MD	D&R
F6	D&R + BR	Full-atom	BR	D&R
F7	D&R + MD + BR	Full-atom	MD + BR	D&R
B5	D&R + MD	Backbone	MD	D&R
B6	D&R + BR	Backbone	DR	D&R
B7	D&R + MD + BR	Backbone	MD + BR	D&R