Journal of Structural Biology xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Structural Biology

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

Proteins of well-defined structures can be designed without backbone readjustment by a statistical model

Xiaoqun Zhou [a], Peng Xiong [a], Meng Wang [a], Rongsheng Ma [a], Jiahai Zhang [a], Quan Chen [a,*], Haiyan Liu [a,b,c,d,*]

[a] School of Life Sciences, China
[b] Hefei National Laboratory for Physical Sciences at the Microscales, China
[c] Collaborative Innovation Center of Chemistry for Life Sciences, University of Science and Technology of China, Hefei, Anhui 230027, China
[d] Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, Anhui 230031, China

ARTICLE INFO

Article history:
Received 21 April 2016
Received in revised form 26 July 2016
Accepted 2 August 2016
Available online xxxx

Keywords:
Computational protein design
Design accuracy
Statistical energy function
Protein foldability
NMR structure determination

ABSTRACT

We report that using mainly a statistical energy model, protein sequence design for designable backbones can be carried out with high confidence without considering backbone relaxation. A recently-developed statistical energy function for backbone-based protein sequence design has been rationally revised to improve its accuracy. As a demonstrative example, this revised model is applied to design a *de novo* protein for a target backbone for which the previous model had relied on after-design directed evolution to produce a well-folded protein. The actual backbone structure of the newly designed protein agrees excellently with the corresponding target. Besides presenting a new protein design protocol with experimentally verifications on different backbone types, our study implies that with an energy model of an appropriate resolution, proteins of well-defined structures instead of molten globules can be designed without the explicit consideration of backbone variations due to side chain changes, even if the side chain changes correspond to complete sequence redesigns.

1. Introduction

A fundamental tool to support protein design for different purposes (Regan et al., 2015) is the computational design of amino acid sequences that fold into a specific backbone. While the first automatically designed protein to fold as desired was reported almost two decades ago (Dahiyat and Mayo, 1997) and the first protein of a designed backbone more than ten years ago (Kuhlman et al., 2003), computational protein design (CPD) has so far produced only more than a dozen of automatically designed proteins with experimentally determined structures (Li et al., 2013). Apart from structures of particular types (Huang et al., 2016; Lin et al., 2015), the low success rates of automatic sequence design is constantly a major factor that limits CPD to realize its full potential. It has been suggested that the inaccuracy of current energy functions used by CPD is the major cause of low success rates (Li et al., 2013). For a task that involves the design of more than a handful of amino acid positions, small improvements of the energy function that reduce the chance of erroneous designs at a single position by even a tiny amount will have a significant effect on the final success rate. In addition, with current most successful energy functions such as RosettaDesign (Leaver-Fay et al., 2011), "perfect" packing needs to be emphasized through, e.g., iterative rounds of sequence selection and backbone relaxation (Kuhlman et al., 2003), to avoid the generation of molten globule structures. Although this approach has been able to produce designs of atomic precision (Kuhlman et al., 2003), the designed proteins may lack the sequence diversity and conformational plasticity commonly observed in natural proteins. While considering backbone flexibility during sequence optimization may increase sequence variability (Gainza et al., 2013; Ollikainen et al., 2015), introducing new energy functions that complement current ones can further expand the solution space of sequence design (Xiong et al., 2014). Another desired property of the energy function is to support the design of specific backbones with an intermediate structural resolution, so that the fluctuations of a backbone structure caused by tolerable sequence variations do not need to be differentiated explicitly. Then the resulting designs may encompass larger conformational plasticity, which should benefit subsequent functional adaptation (Murphy et al., 2016).

Abbreviations: ABACUS, acronym for A Backbone based Amino aCid Usage Survey; CPD, computational protein design; NMR, nuclear magnetic resonance; HSQC, heteronuclear single quantum coherence spectra; NOE, nuclear Overhauser enhancement; RDC, backbone residual dipolar coupling; SAI, solvent accessibility indices; RMSD, the root-mean-square-deviation; vdW, van der Waals.

* Corresponding authors at: School of Life Sciences, China (H. Liu).
E-mail addresses: chenquan@ustc.edu.cn (Q. Chen), hyliu@ustc.edu.cn (H. Liu).

http://dx.doi.org/10.1016/j.jsb.2016.08.002
1047-8477/© 2016 Elsevier Inc. All rights reserved.

Recently, we have reported a statistical energy function named ABACUS (acronym for A Backbone based Amino aCid Usage Survey, see also Computational Method) for backbone-specific protein design. We have reported experimental verifications of design results for three backbone structures taken from the protein data bank (PDB IDs 1ubq, 1cy5 and 1r26, respectively) (Xiong et al., 2014). Automatic design using the program itself led to a *de novo* sequence that fold into the 1ubq backbone with around 1.2 Å root mean square deviation (RMSD) of backbone atom positions according to the structure solved by solution nuclear magnetic resonance (NMR) . However, only after directed evolutions with a system to select and improve foldability (Foit et al., 2009), mutants of the sequences designed for the 1cy5 and 1r26 backbones were found to be well-folded as indicated by ^{1}H–^{15}N heteronuclear single quantum coherence spectra (HSQC) NMR spectra (Bax et al., 1990; Piotto et al., 1992). The structure of one mutant, D_1cy5_M1, was solved using solution NMR and it turned to be in agreement with the corresponding design target.

In the current work, we introduce several important revisions of the ABACUS model to improve its accuracy. The revisions have been subjected to extensive computational tests, including systematic single-site-redesign (Kuhlman and Baker, 2000) for 40 natural proteins (Xiong et al., 2014) and heuristic manual inspections of completely redesigned sequences for different target backbones. The revised ABACUS allowed us to design well-folded sequences for different target backbones. Here as an example, we present the structure of a protein redesigned for the target backbone 1r26, for which the sequence produced by the previous ABACUS was foldable only after directed evolution. It turns out that the structure of the new protein is much more well-defined than the previously obtained mutant protein, and is in excellent agreement with the design target.

2. Materials and methods

2.1. Definition of the ABACUS energy function

The statistical components of the ABACUS energy function have the usual form of summations over single residue (or sequence position) terms and residue pair-wise terms,

$$E_{statistical}(s_1,s_2,\cdots,s_L) = \sum_{i=1}^{L} e_i(s_i) + \sum_{i=1}^{L-1}\sum_{j=i+1}^{L} e_{ij}(s_i,s_j) \qquad (1)$$

in which i or j = $1,\cdots,L$ stands for sequence positions along the peptide chain, $s_1,s_2,\cdots,s_L$ represent the amino-acid sequence (or more exactly, the rotamer (Dunbrack and Cohen, 1997) sequence), and e_i and e_{ij} are respectively single-residue and pair-wise energy terms depending on backbone structure. The single residue energy e_i depends on secondary structural type (SS_i), backbone torsional angles (φ and ψ_i) and solvent accessibility index (SAI_i) of sequence position *i*. The solvent accessibility index *SAI* has been defined as a relative rank of a position's solvent exposure (see computational details) normalized to between 0 and 1, with the value 1 corresponding to the most exposed (Xiong et al., 2014). The pair-wise energy e_{ij} depends on the above local features of both positions *i* and *j* as well as on the relative positioning of all backbone atoms at the two positions. Unlike other studies in which different structure features are considered using different statistical energy terms (Dantas et al., 2003; Dunbrack and Cohen, 1997; Poole and Ranganathan, 2006; Simons et al., 1999), the various structural features are considered jointly in one ABACUS energy term. More specifically, $e_i(s)$ is defined to be proportional to a single term like $-\ln P(s|SS_i,\varphi_i,\psi_i,SAI_I)$ instead of a summation like $-[\ln P(s|SS_i) + \ln P(s|\varphi_i,\psi_i) + \ln P(s|SAI_I)]$. Here *P* represent rotamer type distributions conditioned on structural features. In a similar spirit, the ABACUS pairwise term is defined as

$$e_{ij} \propto -\ln \frac{P(s,s'|\Theta_{ij},SS_i,\varphi_i,\psi_i,SAI_i,SS_j,\varphi_j,\psi_j,SAI_j)}{P(s|SS_i,\varphi_i,\psi_i,SAI_i)P(s'|SS_j,\varphi_j,\psi_j,SAI_j)} \qquad (2)$$

Here we use Θ_{ij} to formally represent the relative positioning of the two backbone positions (see below). For every position *i* (or position pair *i,j*) in a given target backbone, the distribution is estimated using training positions (or position pairs) selected from a general set of training protein structures (for benchmark purpose, training proteins sequentially or structurally similar to the target backbone have been excluded). To handle the multi-dimensional conditions effectively, ABACUS uses a neighbor filter (DeBartolo et al., 2012) with adaptive cutoffs to strike a balance between data relevance and data size (Xiong et al., 2014). Another important feature of ABACUS is that for the relative positioning of a pair of backbone positions (i.e.,Θ_{ij}), the root-mean-square-deviation (RMSD) of all backbone atoms is employed to construct the filter, thus Θ_{ij} is by itself of multiple dimensions, including not only distances but also relative orientations.

As previously reported (Xiong et al., 2014), the above pure statistical terms have been combined with simple van der Waals energy terms to give the total energy function for optimization,

$$E_{total} = E_{statistical}(s_1,s_2,\cdots,s_L) + \sum_{ij} w_{ij} e_{ij}^{vdw}(s_i,s_j) \qquad (3)$$

The van der Waals interactions are calculated as the sum of atomic van der Waals interactions computed using an adjusted Lennard-Jones form (Pokala and Handel, 2005). To calculate the van der Waals interactions, the coordinates of sidechain atoms for a given rotamer state have been generated according to "standard" internal coordinates (i.e., bond lengths, bond angles and torsional angles). The residue type-specific bond lengths and bond angles have been determined by taken averages over sidechains contained in the training PDB structures. The rotamer-specific torsional angles have been taken from the Dunbrack backbone independent rotamer library (Dunbrack and Cohen, 1997). For each of the amino acid residue types with aromatic sidechains, the number of rotamer types has been extended to 9 by including for each sidechain torsion two additional values, each of which is one standard deviation away from the respective central value given in the library. The way to calculate the van der Waals interactions has been the same as the EGAD method (Pokala and Handel, 2005), in which hydrogen atoms have been explicitly considered, but with their radius scaled by a factor of 0.9 from the original force field value. Other parameters in the ABACUS energy function, including the adaptive cutoffs of the neighbor filters and the weights of the van der Waals terms, have been optimized via single-site-redesign tests (Kuhlman and Baker, 2000), in which for a given set of native proteins, the residue at every position is one-by-one changed into different residue types with the residue types at all the other positions unchanged. Residue types are ranked according to respective calculated energies. The parameters were systematically varied to find optimal values for the averaged rank of native residue types to be as high as possible.

2.2. Revisions of the computational model

The total statistical energy has been rewritten with the local and non-local interactions separated and given different weights, namely,

Please cite this article in press as: Zhou, X., et al. Proteins of well-defined structures can be designed without backbone readjustment by a statistical model. J. Struct. Biol. (2016), http://dx.doi.org/10.1016/j.jsb.2016.08.002