Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Using natural sequences and modularity to design common and novel protein topologies

Aron Broom, Kyle Trainor, Duncan WS MacKenzie and Elizabeth M Meiering

Protein design is still a challenging undertaking, often requiring multiple attempts or iterations for success. Typically, the source of failure is unclear, and scoring metrics appear similar between successful and failed cases. Nevertheless, the use of sequence statistics, modularity and symmetry from natural proteins, combined with computational design both at the coarse-grained and atomistic levels is propelling a new wave of design efforts to success. Here we highlight recent examples of design, showing how the wealth of natural protein sequence and topology data may be leveraged to reduce the search space and increase the likelihood of achieving desired outcomes.

Address
Department of Chemistry, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

Corresponding author: Meiering, Elizabeth M (meiering@uwaterloo.ca)

Current Opinion in Structural Biology 2016, **38**:26–36

This review comes from a themed issue on **Sequences and topology**

Edited by **L Aravind** and **Elizabeth Meiering**

http://dx.doi.org/10.1016/j.sbi.2016.05.007

Introduction

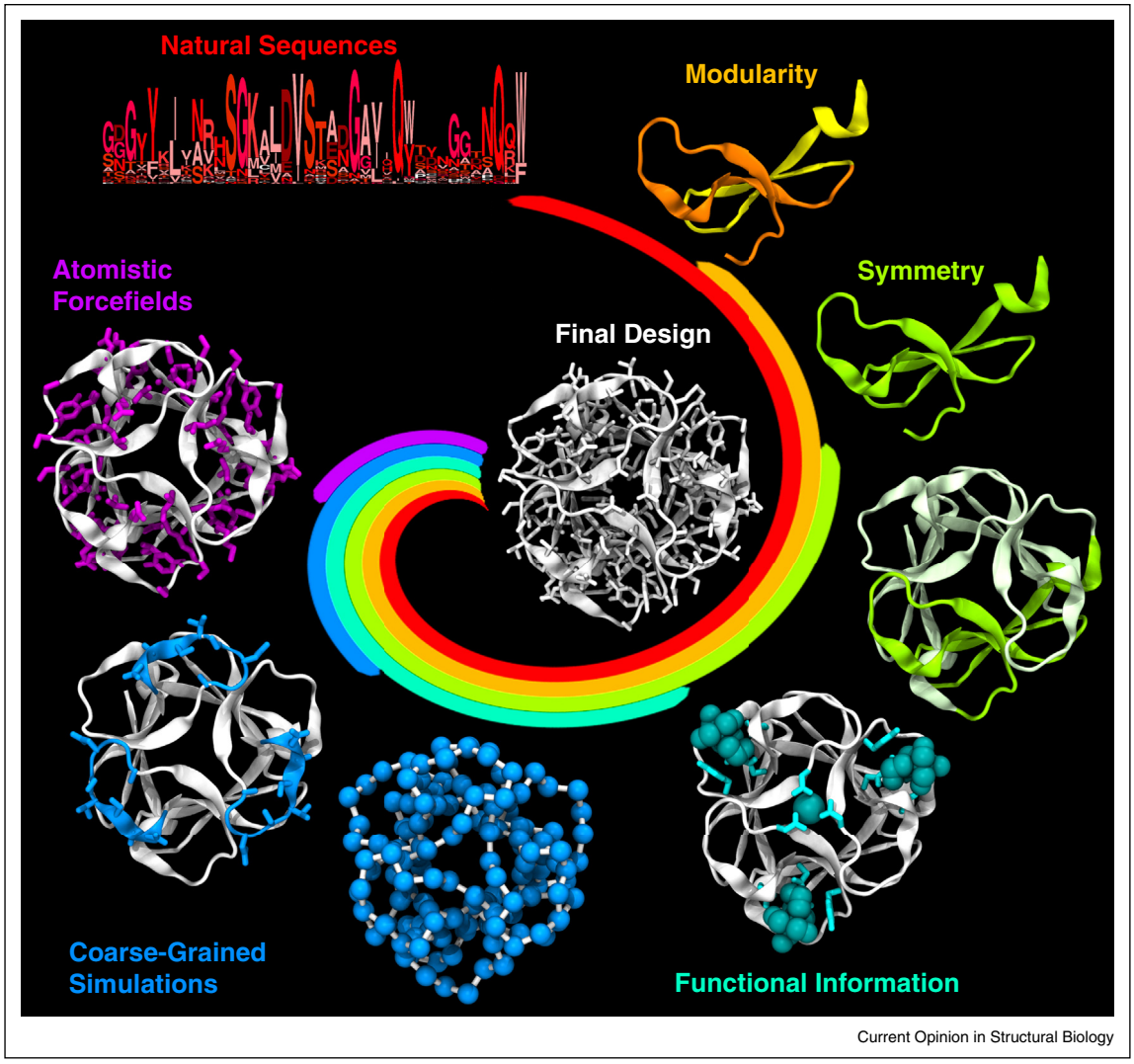
Protein design has advanced tremendously over the last several decades; yet, the reliable design of a stable, well-folded, and soluble protein with the intended structure remains far from routine, often requiring multiple attempts, iterative improvements, and substantial resources [1,2]. On the other hand, nature has successfully explored a great diversity of sequences and topologies [3], offering large and rapidly growing repositories of information that are increasingly leveraged in design. The advent of computational protein design (CPD) enabled the exploration of fully *de novo* sequences for natural topologies [4,5], and more recently the design of *de novo* topologies (also using natural structural information) [6,7•]. Although nature's existing sequences and topologies offer a solid foundation for protein design, recent breakthroughs generating natural and novel topologies, often with very high stability, demonstrate that many forms are possible [6,8,9••,10••,11,12].

Much progress has been made concurrently using natural mechanisms to improve or guide design by selection and directed evolution approaches [13]. Here we focus predominantly on rational design of protein topology, highlighting recent developments that leverage sequence and topology databases in addition to CPD using both atomistic energy functions and coarse-grained simulations. Recent results demonstrate the numerous and increasingly sophisticated strategies, combining multiple approaches, which are being developed and validated; these promise to further advance fundamental understanding of the interplay between sequence, folding and topology and to improve the success of protein design in practical applications (Figure 1). The ongoing improvements may be likened to the development of refined and powerful machine tools at the onset of the industrial revolution, and may usher in a similarly transformative period.

Natural sequence statistics in design

Sequence data, without explicit consideration of structure, are now widely employed in designing stable and functional proteins. Numerous studies have shown that consensus sequence-based design — choosing the most common amino acid at each position of a multiple sequence alignment (MSA) — can be useful for increasing protein stability and may also aid in diversifying function. Recently, for example, the consensus design of FN3 domains using only a handful of closely-related sequences resulted in significantly increased stability [14], while using thousands of more distantly-related sequences produced an extraordinarily stable variant [15]. In the case of proteins with a catalytic function, large and diverse MSAs similarly resulted in improved stability but often with concurrent loss of catalytic activity and increased substrate promiscuity [16,17•]. A similar loss of specificity may occur when consensus designing proteins with ligand-binding, transport, and other functions. Reconstructing ancestral sequences from the same MSAs used for consensus design, Risso *et al.* found a similar trend in activity, though the ancestral reconstructions in this case were all more stable than the consensus designs [17•]. A cogent review of ancestral sequence reconstruction is included in this issue [18]. Sequence-based protein design may be more successful for highly populated folds which provide larger alignments, and such folds may be

Figure 1



Overview of approaches for protein design. The central colored spirals depict the contributions of different design methods, which may vary from one approach to another, and together generate the final design. For each method, the corresponding colored elements of structure are designed, and may then be fixed in later design stages (white). A wide range of protein topologies have been realized, using approaches based on: information derived from natural sequences, such as consensus (red); small subdomain-sized structural modules (yellow, orange); repetition of structural modules, that is, symmetry (green); consideration of functional sites (cyan, ligands are in space fill representation); coarse-grained simulations (blue spheres represent individual residues in simulations which inform the design of certain portions of the backbone or sidechains); and atomistic simulations using force fields (purple, modeled sidechains shown as sticks). Recent successful designs typically incorporate multiple methods. For instance, incorporating structural modularity (yellow, orange) with symmetry (green) can greatly reduce the size and complexity of the design problem. Including functional constraints (cyan) and/or coarse-grained folding simulations of many or selected residues (blue) can help retain function or optimize folding. The design methods are illustrated using ThreeFoil (PDB: 3PG0).

highly populated precisely because they are more amenable to functional diversification or engineering [19,20].

The success of consensus design is proposed to arise from the tendency for natural sequences to drift over time while the amino acids critical for stability and folding are retained owing to evolutionary pressure [21]. This phenomenon can be utilized to improve the folding and stability of natural and designed proteins by making specific 'consensus mutations' [22]. Further, the covariation or correlations between amino acids at different positions in an MSA may be useful for structure prediction and identifying functional residues [23,24]. Accounting for covariation when choosing consensus mutations may improve stability while also reducing potentially negative impacts on function [22,25]. Entirely sequence-based design methods have been successfully applied and continue to be developed; in more complex