Proteins are the molecules cells primarily rely on for catalysis, recognition, signaling, defense, locomotion, and structural integrity. Engineering proteins for improved function or new applications is a fast-growing segment of biotechnology and biomedicine. Experimental efforts based on the screening of large mutant libraries have led to many successes but they do not provide quantitative design principles and/or insight into the structural features that underpin the desired function. The computational de novo design of proteins promises to bridge this gap; however, it requires reliable structure prediction, provisions for protein stability, and accurate descriptions of inter-molecule interactions. Studies that successfully meet all these criteria are beginning to emerge including the design of an O2-binding protein and a novel enzyme that catalyzes a Diels–Alder reaction.

Introduction

Computationally designing proteins is a crosscutting challenge that directly impacts many scientific and engineering endeavors, ranging from improved catalytic activity, genetic circuits, biosensors, chiral separations, creation of gene switches, and signal transduction pathways. Although purely experimental design efforts relying on combinatorial library construction and screening have been widely successful, the lessons learned do not easily generalize to inform the redesign of other systems. Proteins have been previously computationally designed to bind new ligands [1], proteins [2], and nucleic acids [3], to improve protein stability [4,5], as well as to introduce novel enzymatic activity [6,7], demonstrating that the fundamental rudiments of molecular recognition and interactions can be adequately captured via computational design. Despite these successes, predictably changing or even improving a protein’s function in response to a performance target remains a formidable challenge. Successful de novo computational protein design requires accurate structure prediction, protein stability at the desired operating conditions, and correct modeling of the protein’s interactions with other molecules (e.g., substrates, ligands, and cofactors). As illustrated in Figure 1, this review will discuss advances reached over the past couple of years in addressing each of these design challenges as well as examples where all three have been brought to bear in de novo design efforts.

Modeling and predicting protein structure

Reliable protein structure prediction is paramount in protein design, as protein geometry and flexibility along with proper presentation of charges and molecular groups on the surface determine function (or lack thereof). The central dogma behind protein structure prediction is that the native structure reaches a conformation that achieves (near) global minimum energy. The bi-annual Critical Assessment for protein Structure Prediction (CASP) benchmarks the current state of the art in protein structure prediction, with the most recent round, CASP9, completed in the summer of 2010. Using a feature space representation Kim et al. [8] sought to understand why identification of the native state is so challenging and discussed how the magnitude of the sampling problem dictates whether the problem can be solved with extra computational resources or if improved algorithms must be developed beforehand.

The development of improved protein structure prediction algorithms has been the focus of a number of recent publications. McAllister and Floudas [9] developed improved bounding methods for the structure search problem. In contrast to trimming the search space, Hahn et al. [10] sought to search more rapidly using a cluster expansion technique, albeit at the cost of introducing a controllable error. A popular concept that reduces the conformational search space is the use of rotamers (short for rotational isomers) of the statistically preferred conformations of amino acid side chains dependent upon the protein backbone geometry. Berkholz et al. [11] discussed how the backbone geometry varies as a function of the backbone dihedral angles. Havranek and Baker [12] considered how to identify acceptable backbone changes that would allow rotamers to assume optimized orientations. Krivov et al. [13] developed SCWRL4 to more accurately and quickly predict side-chain conformations in proteins, while Shandler et al. [14] used foldamers to explore generation of better rotamer libraries. Blum et al. [15] developed a new method for de novo protein structure prediction that combines conformational space annealing and genetic algorithms that...
achieved significant improved over a standard Rosetta implementation.

While it is customary in protein design to assume a single, well-defined backbone geometry, this does not always hold true. Xue et al. [16] developed the meta-analysis tool PONDR-FIT to develop predictions for disordered regions of proteins. An alternative to using a single backbone structure is to model an ensemble of low-energy protein structures. Allen et al. [17] developed a multi-state design algorithm for modeling protein properties (e.g. stability, activity, and solubility) that are dependent upon backbone conformation variability. McAllister and Floudas [18] combined the αBB deterministic global optimization approach with conformational space annealing to predict lower energy protein structures (i.e., unique and ensembles thereof) and compared results with other methods. Allen and Mayo [19] developed MSD-FASTER and Subramani et al. [20] created ICON to generate and screen ensembles of low energy protein structures.

There have been several publications in the last two years where authors have customized and deployed computational protein structure prediction systems to specific protein classes. Correia et al. [21] successfully designed protein scaffolds to present target epitopes recognized by antibodies. Luo et al. [22] used computations to model the allosteric changes of eight single-point mutations of αIIbβ3 to the integrin headpiece and observed conformational changes propagating from the headpiece to the legs of the integrin. In more general applications, Rosetta has been used to predict the structures of oligomers with
near atomic-level accuracy [23], which should be helpful in conjunction with NMR data to resolve structures and to model the allosteric changes of ligand-free proteins from their bound states [24].

**Designing stabilized proteins**

After an appropriate structure for a protein has been modeled, care must be taken to ensure that it will be stable at the desired pH and temperature. Although literature attention to this topic waned recently, it remains a critical factor in protein engineering. Belien et al. [25] used the pKD software to improve the low-pH stability of the B. subtilis endo-β-1,4-xylanase by making mutations that affected the local pKa of key residues. Heinzelman et al. [26] used SCHEMA to recombine several parent cellulases to design a library of thermostabilized proteins. Tian et al. [27] used computations to identify glycine to proline mutations to thermostabilize proteins by exploiting the fact that glycine has the highest conformational entropy of any amino acid whereas proline has the lowest. Joo et al. [28] used a more general computational approach to identify thermo-unstable residues and correcting mutations. Finally, Gribenkon et al. [29] and Gao et al. [30] used computations to identify thermostabilizing mutations while imposing active site geometry criteria to safeguard the activity of the redesigned proteins.

**Engineering proteins for molecular interactions**

Computational protein design for a given function relies on optimizing a complex choreography of interactions with other molecules. A significant number of recent studies have focused on engineering these inter-molecule contacts. An important class of protein interaction partners is in fact other proteins. Tuncbag et al. [31] developed a computational method to identify “hot-spot” residues that are most important in mediating protein–protein interactions. In a study aimed at redesigning the interactions between a high-affinity pair of proteins, (i.e., acetylcholinesterase and fasciculin), Sharabi et al. [32] found that changes in the interaction energy, rather than total energy, correlated well with the experimental changes in binding energy. Guntas et al. [33] used a joint computational and experimental approach to redesign the ubiquitin-ligase E6AP to act on the unnatural partner Ubɛ12 in an effort to demonstrate efficiency advantages computations can offer. Yosef et al. [34] used ORBIT to switch the specificity of Calmodulin between its two main-target interaction partners, demonstrating the plasticity of interactions in signaling networks.

On the opposite side of the size-scale for protein interaction partners are metal ions. Their small size allows for more computationally complex descriptions of molecular interactions. Hayik et al. [35] used a mixed QM/MM protocol to predict metal ion binding energies. Fazelinia et al. [36] developed the OptGraft method to identify the location in a target protein that can best accommodate a metal ion binding site along with beneficial mutations in the surrounding residues. Wang et al. [37] used a similar approach specific to zinc ions.

A class of protein interaction partners with increasing attention in the literature is nucleic acids. Ashworth and Baker [38] used computations to assess the degree of optimization in known protein–DNA interactions and identified the contribution of individual residues. Several groups used nucleic acid binding proteins as targets for specificity alterations: Liu et al. [39] increased the specificity of a nucleoside kinase for 3’-deoxythymidine, Lopes et al. [40] modified asparaginyl-tRNA synthetase to favor the binding of aspartyl-adenylate, and Murphy et al. [41] used loop remodeling to alter the specificity of a human guanininedeaminase for ammelide over guanine.

Many other studies aimed to engineer proteins for optimizing their interactions with a variety of target molecules. Yang et al. [42] used free-energy perturbation calculations on the free and transition-state butyrylcholinesterase to identify high-activity mutants for the hydrolysis of cocaine. Berroondo et al. [43] analyzed the structural and regulatory consequences of mutations in the N-terminus arm of AraC, which is a gene expression regulatory protein that promotes expression when bound to arabinose and suppresses it otherwise. Chaudhury and Gray [44] used computational docking techniques to identify residues in an HIV protease that were important for activity and found they were residues that tended to confer drug-resistance. Grigoryan et al. [45*] were successful in designing orthogonal interaction partners for specific members of the B-ZIP family of proteins in spite of their high sequence and structural homology. Finally, Khoury et al. [46] used IPRO to change the cofactor specificity of C. boidini xylose reductase from NADPH to NADH, while Chica et al. [47] destabilized the fluorophore ground state and stabilized the excited state to design improved red-fluorescent proteins.

Two recent papers build upon the pioneering de novo computational design of an enzyme that catalyzes the Kemp elimination reaction [7]. Khersonsky et al. [48] computationally generated and experimentally screened proposed beneficial mutations for this enzyme while Kiss et al. [49**] used computations to rank-order and evaluate active and inactive in silico designed enzymes for the Kemp elimination reaction, finding that molecular dynamics was most useful in explaining the experimental findings.

**Designing new proteins**

By bringing to bear structure elucidation, stability safeguards, and molecular interaction descriptions, a number of efforts achieved de novo design of novel proteins. One
particularly intriguing target is antibodies, because there are well-established rules governing their structures and their functions are limited to binding, not catalysis. RosettaAntibody [50] was recently developed for the homology modeling of antibody variable domains and SnugDock [51] can be used in conjunction to predict antibody–antigen complexes. Our group recently developed the Optimal Complementarity Determining Regions (OptCDR) method [52] for the de novo design of the binding portions of antibodies against any specified antigen epitope.

Other efforts include the work of Masica et al. [53] who used computations to de novo design peptides that can influence calcite binding. Fry et al. [54] designed a heterotetrameric protein that can selectively bind a chromophore whereas Koder et al. [55] designed an O₂ binding protein with properties similar to natural globin proteins with the key improvement of being able to bind O₂ better than CO. Finally, Siegel et al. [56] computationally de novo designed an enzyme to catalyze the Diels–Alder reaction, for which no naturally occurring enzyme was known beforehand.

Conclusions
Successful computational protein design depends on accurate structure modeling, ensuring protein stability, and optimizing inter-molecule interactions. Each of these major hurdles has received significant attention in the past two years and many de novo protein designs have been put forth as a result. However, the dream of efficiently, predictably and reliably computationally designing improved proteins remains beyond reach. Baker [57] eloquently reviewed in detail many of the unresolved challenges facing computational enzyme design. Biological systems are significantly more complicated than the idealized abstractions imposed by the assumptions used in computational protein engineering. It is increasingly realized that proteins rarely have unique functions, instead they participate in multiple interactions and processes in ways that may confound our ability to computationally assess their fitness. In addition, because of cost and time constraints experimentally resolved structures are only rarely obtained for successful designs. This limits our ability to learn from our successes and fairly assess which modeling predictions panned out. More importantly, failed designs are almost never reported or analyzed further. This implies that no quantitative guidelines are obtained to improve the modeling component of the computational system. The lack of communication of failures likely slows the overall rate of progress of this field. This explains why most of the recent successes have been system-specific and incremental in nature. Bold new steps are needed in integrating computational design methods, experimental screening protocols and structure identification techniques to achieve new milestones in protein design.

Acknowledgement
This work was funded by the National Science Foundation [CBET-0639962].

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest


The authors explored how to develop rotamers for systems that are less-well understood, an important step in expanding modeling capabilities.


Recent advances in computational protein design


The authors successfully designed interaction partners that were specific to their designed bZIP over other members of the bZIP family.


49. Kiss G, Rothlisberger D, Baker D, Houk KN: Evaluation and ranking of enzyme designs. Protein Sci 2010, 19:1760-1773. This paper used computations to examine and rank-order in silico designed enzymes for the Kemp elimination reaction, providing insight into why many of the designs did not work.


52. Pantazes RJ, Maranas CD: OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding. Protein Eng Des Sel 2010, 23:849-858. We have developed a computational method, OptCDR, to de novo design the binding portions of antibodies (i.e. the CDRs) to bind any specified antigen epitope.


54. Fry HC, Lehmann A, Saven JG, DeGrado WF, Therien MJ: Computational design and elaboration of a de novo heterotetrameric alpha-helical protein that selectively binds...
6 Engineering and design


