Review

Cell PRESS

Protein folding and *de novo* protein design for biotechnological applications

George A. Khoury, James Smadbeck, Chris A. Kieslich, and Christodoulos A. Floudas

Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ 08544, USA

In the postgenomic era, the medical/biological fields are advancing faster than ever. However, before the power of full-genome sequencing can be fully realized, the connection between amino acid sequence and protein structure, known as the protein folding problem, needs to be elucidated. The protein folding problem remains elusive, with significant difficulties still arising when modeling amino acid sequences lacking an identifiable template. Understanding protein folding will allow for unforeseen advances in protein design; often referred to as the inverse protein folding problem. Despite challenges in protein folding, *de novo* protein design has recently demonstrated significant success via computational techniques. We review advances and challenges in protein structure prediction and *de novo* protein design, and highlight their interplay in successful biotechnological applications.

Protein folding and design are two sides of the same coin

Proteins are polymeric chains of amino acids that organisms and cells rely on for signaling, pathogen clearing, mobility, catalysis, recognition, shape, ordering, and stability. The precise ordering of the amino acids in a protein sequence determines how the protein folds into a 3D structure, and thus its biological function. As our knowledge of the connection between sequence, structure, and function has advanced, interest has grown in designing proteins on a sequence level to produce novel folds and function. Brute-force experimental approaches to resolving protein structures and designing protein sequences for new functions remain time consuming and expensive, and add little to our understanding of the physical principles required for both problems [1].

Protein structure prediction aims to determine accurately the full 3D structure of a protein given only its amino acid sequence. Structure prediction is challenging if only

Corresponding author: Floudas, C.A. (floudas@titan.princeton.edu).

Keywords: protein structure prediction; *de novo* protein design; computational biology; biotechnology.

0167-7799/$ – see front matter

 http://dx.doi.org/10.1016/j.tibtech.2013.10.008

Glossary

Cartesian minimization: refers to a process operating on the variables as 3D vectors of x, y, and z coordinates in order to reduce the potential energy of a conformer.
EC50: a metric for the concentration of compound at half the maximal value on a dose-response curve. The curve is usually sigmoidal.
Generalized Orientation-Dependent All-Atom Statistical Potential (GOAP): a distance-dependent statistical potential that scores models to aid in selecting near-native conformations of a target protein. It utilizes information about the relative plane orientation of interacting pairs of atoms.
Global Distance Test Total Score (GDT_TS): this is a metric that approximately represents the percentage of residues located in the correct position after structural alignment. This is a more robust metric than RMSD.
Hot spot: key interactions at the interface of a protein–protein complex. Many hot spots include salt bridges where oppositely charged side chains attract, hydrogen bonds, and/or ideal van der Waals interactions subject to shape complementarity.
IC50: A metric for the half-maximal inhibitory concentration in a competitive binding assay. The curve is usually sigmoidal.
Iterative Threading Assembly Refinement (I-TASSER): structure prediction method using multiple threading alignments to templates and fragment assembly.
Local Meta-Threading Server (LOMETS): generates structure predictions using high scoring alignments of a target sequence to a template using information from ten threading programs.
MODELLER: protein structure homology modeling program that generates structures satisfying spatial constraints.
Molecular dynamics (MD): an algorithm for solving the equations of motion iteratively over time and used to sample conformational space in a physically meaningful way.
Monte Carlo (MC): an algorithm reliant on randomly sampling the sequence or structural space according to a probability distribution.
Non-deterministic polynomial time complete (NP-complete): a difficult class of decision problems that have not been proven to be solvable with an algorithm within polynomial time – $O(n^k)$.
QUARK: a protein structure prediction program that assembles fragments without any global template information.
REMO: program that constructs a full protein model using only α-carbon traces.
Root-mean squared deviation (RMSD): this is a metric that measures the average distance between two structurally aligned sets of atoms. It is often used a metric for the quality of a prediction, and often computed with the α carbon atoms. A predicted structure with RMSD to the native of $\lesssim$3 Å is considered to be good enough to perform subsequent computational studies.
Rotamer: statistically abundant side-chain conformation.
SP^3: fold recognition method that combines structural information through sequence profiles of structure fragments, secondary structure predictions, and dynamic programming to generate an alignment of a target sequence to a template.
Torsion minimization: refers to a process operating on a reduced set of variables representing torsion angles that control the distance between the first and fourth atom in a series of four atoms in order to reduce the potential energy of a conformer.
Z score: computed as $\frac{Value-Mean}{StdDev}$, it is a metric denoting the separation of a value from counterparts. It is useful for assessing the significance of top structure predictions compared to the entire population of predictions from other methods.

low homology templates exist. *De novo* protein design is the inverse problem [2,3]; given a rigid or flexible backbone structure, one aims to determine a sequence that will fold into that structure. Different sequences can fold into the same structure, so there is degeneracy in the protein design space. The existence and accuracy of protein structures as templates for protein design can have a significant impact on potential success. For this reason, the ability to produce viable protein templates through protein structure prediction is important for protein design, and for advancement in biotechnology and drug discovery.

In this review, we describe advances and challenges in the fields of protein structure prediction and *de novo* protein design focusing on the interplay necessary for success. Figure 1 schematically shows the roadmap and key challenges in protein structure prediction and *de novo* protein design. The past few years have shown impressive applications of computational structure prediction and design to biotechnology, spanning peptide or antibody therapeutics, novel biocatalysts, and self-assembling nanomaterials.

State-of-the-art advances and challenges in protein structure prediction and refinement

The consistent determination of structure from sequence is one of the greatest unsolved problems in nature and has recently passed the 50-year milestone [4]. Accurately predicting the 3D structure of a protein involves a series of steps performed on a sequence of amino acids: secondary structure prediction (identifying whether local segments are helical, beta-strand, or loop), structural alignment to candidate template structures, conformational sampling, and selection (Figure 2A and Box 1). A predicted structure may then undergo refinement, in an attempt to improve the accuracy of that structure [5]. Historically, most refinement methods degrade rather than improve the accuracy of the predicted structure, making protein structure refinement a substantial unsolved problem in its own right [5,6]. We review recent progress and challenges and refer you to the reviews by Zhang [7] and Floudas [8] for prior advances.

The Critical Assessment of Techniques for Protein Structure Prediction (CASP) [9] occurs biennially and recently completed its tenth experiment. For the CASP

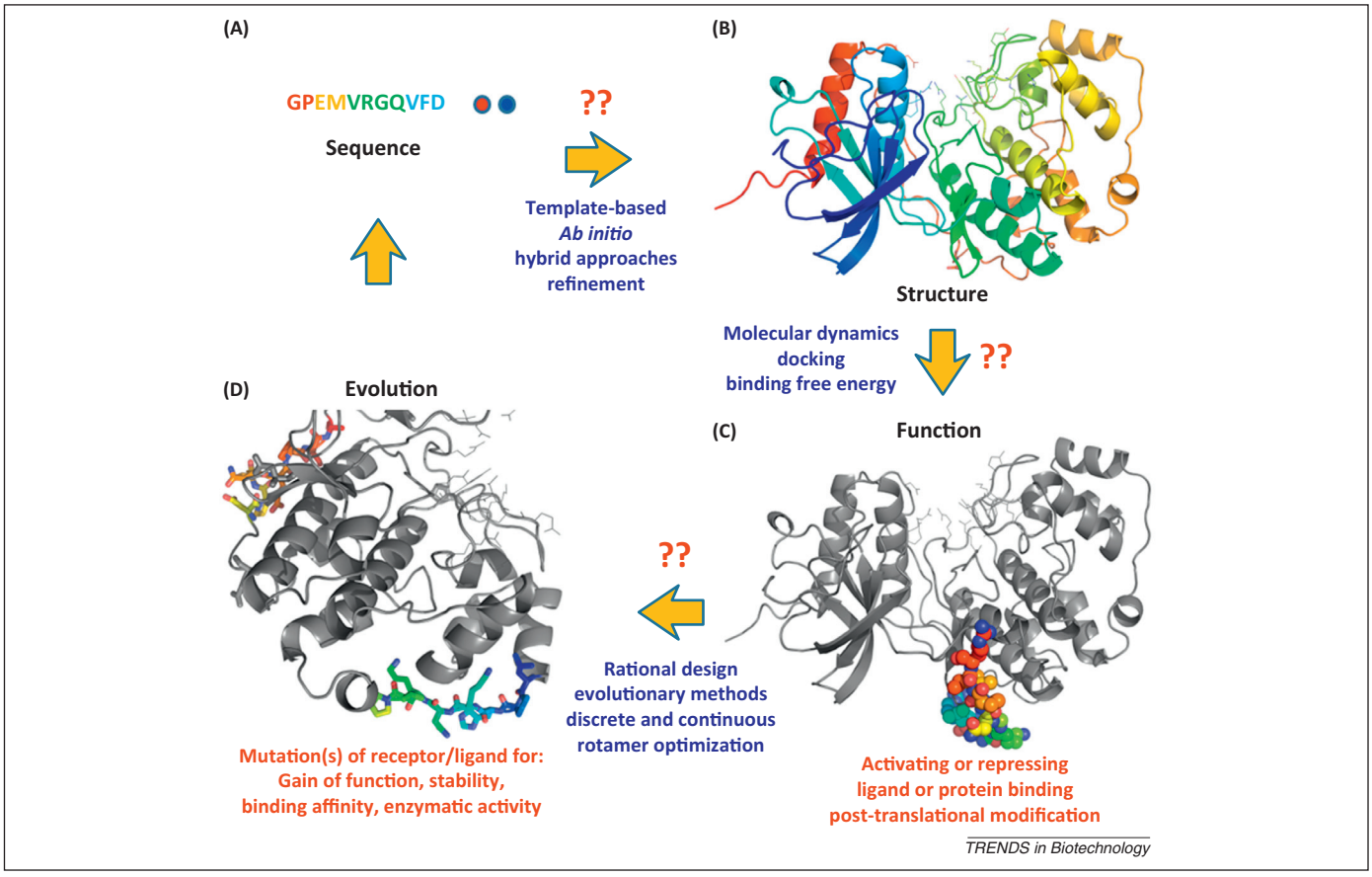

Figure 1. Roadmap of key challenges in understanding how to predict protein sequence to structure to function and design. Structure prediction begins with a primary amino acid sequence **(A)** and aims to predict the full 3D structure **(B)** of that sequence. **(C)** Other proteins, peptides, small molecules, or cofactors may form critical interactions with the protein structure critical to its function. Docking with or without binding free energy calculations may be required to find the most probable conformation for a ligand bound to a receptor protein. Understanding how structure leads to function remains a challenge. The protein structure may be subsequently post-translationally modified, and as most methods have focused in predicting the structures of canonical-amino-acid-containing proteins, the literature is lacking in the ability accurately represent post-translationally modified protein structures. The solution or accurate prediction of the 3D structure of a protein allows it to be used in a design context. **(D)** Biotechnological applications of protein design shown in the literature include designing/redesigning the receptor protein via site-specific mutations to change its binding affinity toward a ligand, change its fold, increase its stability, and create new or alternative enzymatic activity. The ligand of a peptide can be amenable to similar design strategies to design new sequences to bind more strongly to the receptor and compete with its native binding partner (antagonism) or to bind to and activate through a series of specific interactions with the receptor a particular downstream function (agonism). Upon design of the receptor or ligand peptide with new sequences, the cycle begins again as even a few mutations can cause structural conformation and topology changes. The structure shown in the figure is the mitogen-activated protein (MAP) kinase extracellular signal-regulated kinase (ERK)2. The ligand bound is the kinase interaction motif of MAP kinase phosphatase (MKP)3.

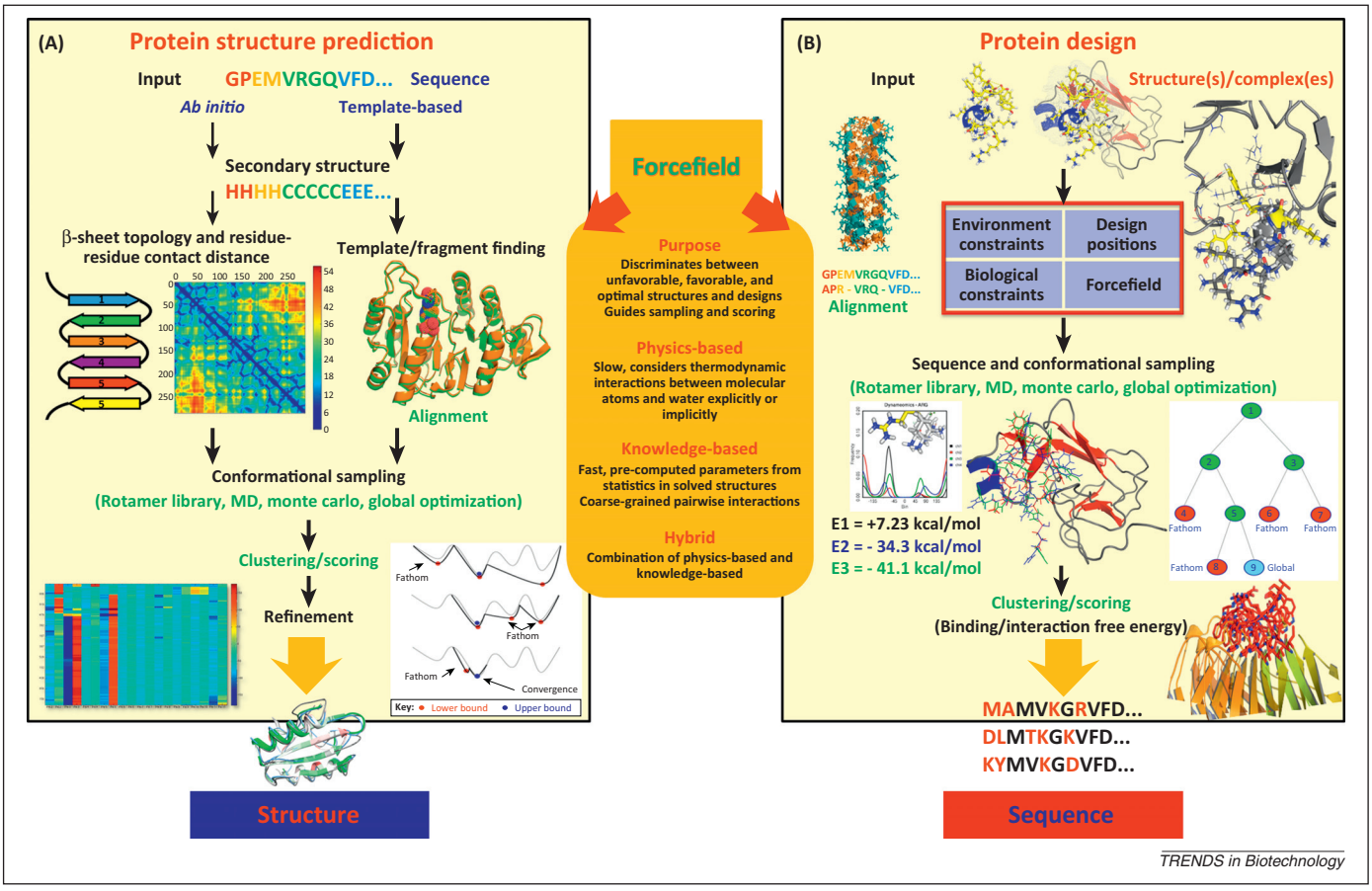

Figure 2. Detailed view of connections and differences between **(A)** protein structure prediction and **(B)** protein design. Dynaneomics image used with permission.

competition, prediction targets are categorized into two groups depending on the availability of structural templates: (i) template-based modeling, in cases for which templates are available; and (ii) free modeling, in cases for which templates are not available. Table 1 shows the top five structure prediction servers in the template-based modeling (TBM) and free modeling (FM) categories (www.predictioncenter.org/casp10/). The average of the top five server methods in the FM category represents approximately half the accuracy of the models produced in TBM. These servers may not apply the same prediction protocol for all targets and instead may perform different pipelines based on the predicted difficulty of the target [10]. The Zhang Server was assessed to be the best overall server in CASP10 in both TBM and FM. Additionally, Table 1 lists the top five protein structure refinement methods assessed in CASP10. The methods listed in Table 1 represent the current state-of-the-art in protein structure prediction and protein structure refinement.

Kryshtafovych *et al*. constructed a difficulty scale [4] based on the similarity of the sequence and structure of the target to that of the closest template available in the Protein Data Bank (PDB) during CASPs 1–9 [11]. Easy targets typically corresponded to those which there are directly identifiable templates through sequence homology. Hard targets may have excellent structural templates in the PDB, but their sequences are often so dissimilar to the target protein that it is nearly impossible to identify them. Additionally, hard targets may have no template at all and may represent a new fold. Figure 3 highlights CASP performance for easy and hard targets over the past 18 years, and several top FM predictions in the past three CASPs. Despite the progress attained for easy targets with identifiable templates, predictors face challenges accurately predicting structures for sequences with difficult to identify templates [12].

Template-based modeling

TBM has served as a reliable prediction method given an appropriate template structure. This approach utilizes the input sequence and attempts to identify structures whose sequences can be aligned with the target sequence to infer information about secondary and tertiary structure (including topology and residue-residue contacts).

The top-performing servers in the TBM category in CASP10 are exhibited in Table 1 and described below. Zhang Server utilizes a combination of Local Meta-Threading Server (LOMETS; see Glossary) [13], Iterative Threading Assembly Refinement (I-TASSER) [14–16], and QUARK [17]. I-TASSER identifies templates via LOMETS, performs fragment assembly via replica-exchange Monte Carlo (MC) simulations, and refinement using REMO [18] and fragment-guided molecular dynamics (FG-MD) [19]. The Protein Modeling System (PMS) uses conformational space annealing (CSA) with Lorentzian energetic restraints in MODELLER, combining physical and knowledge-based energy terms [20]. HHpred-thread is fast and accurate, and includes improvements with three statistical scores to

Box 1. Protein structure prediction and *de novo* protein design are related problems

Protein structure prediction (Figure 2A) begins with a sequence and produces a structure. Two paths are often followed: *ab initio* and template-based. *Ab initio* methods attempt to predict the structure from first principles without a template. Some methods utilize secondary structure and contact predictions as constraints. The most expensive step is the conformational sampling in the presence or absence of constraints. Template-based methods begin with a sequence, predict the secondary structure, and attempt to find a template structure and/or fragments from existing structures in the PDB that will fold similar to the target sequence. These methods rely on the ability to identify suitable templates and then align the target sequence properly to the template sequence. Both methods use advanced sampling techniques such as MD, rotamer optimization, MC, and global optimization. After sampling, both methods may cluster or rescore the structures, and may subject them to a refinement stage to increase prediction accuracy. For sequences of ~30% identity or more to a template, one can expect that the predicted structure is a reasonable estimate of the topology. Below 30%, accurate prediction is more challenging.

Protein design (Figure 2B) begins with a structure or complex and produces new sequences. Design positions are chosen to be mutated. Next, the sequence may be aligned to other homologous sequences to produce biological constraints on the sequence space. The solvent accessible surface area (SASA) of each residue being designed can be taken into consideration to constrain further the design space. Sequence design is then performed and can be done using a single state or multiple states. In this step, the structure being designed can remain fixed, with only side-chain rotamers changing, or may be completely flexible. The algorithms for sampling come from the same classes of techniques used in protein folding. Designed sequences may then be clustered and evaluated with a more detailed scoring function. Design produces one or many sequences that are predicted to fold into the input structure, often with enhanced biophysical characteristics.

Forcefields are the glue connecting structure prediction and design. They describe the interactions between atoms in a system, guide sequence and structural search, and discriminate between optimal and suboptimal solutions. An improved description of atomistic interactions in forcefields benefits both areas. Improving our ability to predict structures will improve our ability to model complexes of druggable targets and design new sequences.

compare the sequence profile of the target with template structure and sequence profiles [21]. RaptorX-YZ is an enhancement to RaptorX [22] using machine learning to predict contacts between residues for use as restraints. BAKER–ROSETTASERVER aligns the candidate sequence to multiple templates, assembles fragments using coarse-grained insertion, utilizes MC search for both coarse-grained and all-atom sampling of favorable backbone and rotameric states, and energy minimization in both torsion and Cartesian space. Top-scoring models are relaxed according to the Rosetta all-atom force field [23]. Commonalities in these top-performing methods are that both PMS and RaptorX utilize MODELLER in model building, and both BAKER–ROSETTASERVER and Zhang Server utilize MC fragment assembly to aid in sampling.

TASSER-VMT was introduced by Zhou and Skolnick [24], and uses the improved SP^3 alternative target-template alignment combined with other alignment methods as input to TASSER simulations. They introduced Generalized Orientation-Dependent All-Atom Statistical Potential (GOAP), a statistical potential with orientation-dependent correction terms for evaluating model quality, recognizing 226 native structures of 278 targets stemming from 11 commonly-used decoy sets [25].

Free modeling

FM is the prediction of structures for sequences that have no distinguishable template in the PDB. These predictions are considered to be hard and success on this front remains limited (Figure 3A) and represents the 'holy grail' of protein folding. In discussing the challenges in FM, it is important to point out the difference between indistinguishable and nonexistent templates. The PDB contains >92 000 resolved structures that offer a wide variety of templates and often several candidate templates for a target sequence. Zhang and Skolnick showed for a set of nonhomologous proteins that they can always find similar folds to the native with an average root-mean squared deviation (RMSD) of 2.5 Å [26].

The ability to predict such difficult targets relies on the ability to select the proper template from structures contained in the PDB and this remains challenging, as demonstrated by the low average Global Distance Test Total Score (GDT_TS) of even the top predictors in CASP10 (Table 1) and overall in CASPs 1–9 (Figure 3A) [12]. Interestingly, none of the best FM methods use strictly *ab initio* methods; all utilize template information. Also, Zhang Server, using an interplay of I-TASSER (which uses templates) and QUARK [17] (which is denoted as first principles) outperforms QUARK alone.

MD-driven folding

Duan and Kollman folded Villin headpiece starting from an unfolded state using MD without the simulation having knowledge of the native contacts [27]. Since that seminal result, several studies have reported the ability to simulate the folding of small proteins.

Scheraga, Liwo, and coworkers, using their developed UNRES coarse-grained molecular dynamics package, recently summarized notable first principles predictions made during CASP10 [28]. They were able to predict the correct packing symmetry for a target with a new fold. Recent advances in implementations, extensions, and applications of UNRES are reviewed by Liwo *et al*. [29]. Shaw and coworkers used equilibrium MD simulations to study the general folding landscape of 12 fast-folding small proteins [30]. In eight out of the 12 studied, a structure within 2 Å of the native was observed. Shaw and coworkers were able to fold ubiquitin [31]; a 76-residue-long protein contained in most eukaryotic organisms having a folding time on the millisecond timescale. These successes have required adjustments in force fields (CHARMM22* [32]), total simulation time on the order of milliseconds, explicit treatment of solvent, and specialized hardware (Anton [33]).

Conformational sampling has been suggested as a major limitation to predicting high-resolution structures [34], whereas it has been recently claimed that sampling is not the main issue, but instead it is forcefield inaccuracy that needs improvement [35]. At this point, there is limited evidence for the recent claim, and as the conformational search in even the simplest biophysical model is

Table 1. Top-performing protein structure prediction servers in TBM and FM categories and methods in the refinement category independently assessed in CASP10

TBM[a,c]				FM[b,c]			Refinement[d]	
Rank	*Server*	*Average GDT_TS*	*URL*	*Server*	*Average GDT_TS*	*URL*	*Group (Method)*	*URL*
1	Zhang Server	53.90	http://zhanglab.ccmb.med.umich.edu/I-TASSER/	Zhang Server	26.78	http://zhanglab.ccmb.med.umich.edu/I-TASSER/	FEIG (MD)	This method is not available as a server
2	PMS	48.78	Server to be released	BAKER-ROSETTA SERVER	24.17	http://robetta.bakerlab.org/	Seok (GalaxyRefine)	http://galaxy.seoklab.org/refine/
3	HHpred-thread	49.33	http://toolkit.tuebingen.mpg.de/hhpred	PMS	23.66	Server to be released	Mufold	http://mufold.org/
4	RaptorX-YZ	50.78	http://raptorx.uchicago.edu/	TASSER-VMT	23.77	http://cssb.biology.gatech.edu/skolnick/webservice/TASSER-VMT/index.html	KnowMIN (KobaMIN)	http://csb.stanford.edu/kobamin/
5	BAKER-ROSETTA SERVER	47.78	http://robetta.bakerlab.org/	Pcons-net (Metaserver)	24.55	http://pcons.net	FLOUDAS (Princeton_TIGRESS)	http://atlas.princeton.edu/refinement/

[a]TBM rankings were taken from the slides of the TBM assessor from the CASP10 website (www.predictioncenter.org/casp10/).

[b]FM rankings were taken based on best model FM rankings based on SUM Z-score.

[c]If multiple servers from the same laboratory are represented in the top five, the highest ranking server is chosen.

[d]Refinement rankings are taken from the CASP10 website and consider both the top human and server groups (http://predictioncenter.org/casp10/doc/presentations/ranking_CASP10_refinement_DJ.pdf).

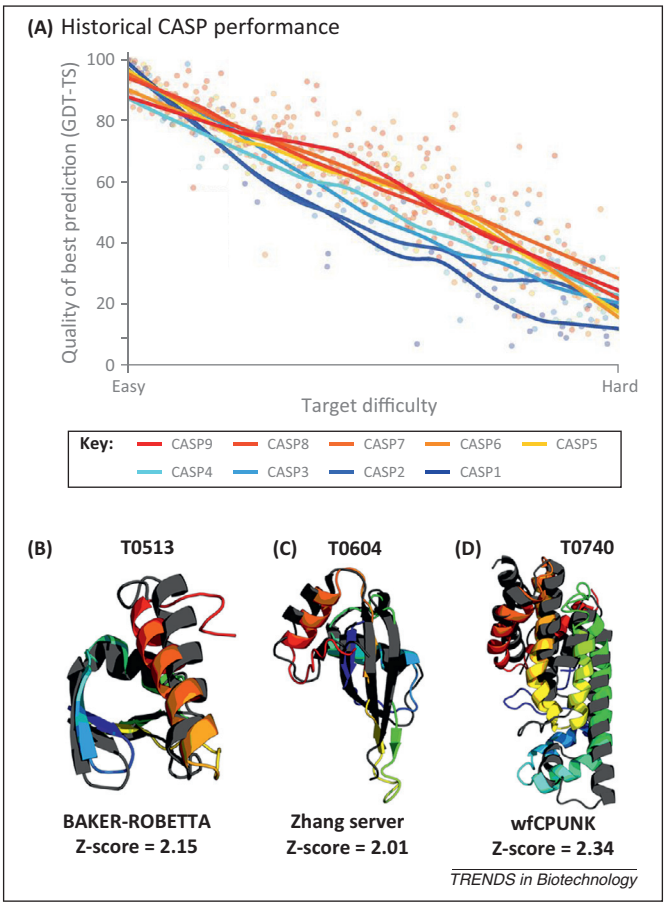

Figure 3. **(A)** Historical performance of the prediction with the best Global Distance Test Total Score (GDT_TS) versus target difficulty over the past 18 years in Critical Assessment of Techniques for Protein Structure Prediction (CASPs)1–9. Target difficulty accounts for both sequence and structural similarity of the target to known template structures available at the time of prediction [12]. Evidently, the quality of the best easy target predictions has been consistently accurate. Conversely, hard targets, most lacking identifiable templates, have not advanced significantly in this time period and still remain the biggest challenge (Adapted, with permission, from [4,12]). High-ranking blind free modeling predictions submitted to CASP8, 9, and 10 for targets **(B)** T0513 by BAKER-ROBETTA, **(C)** T0604 by Zhang Server, and **(D)** T0740 by wfCPUNK. The native structure is shown in dark grey and the prediction as a rainbow, with the N terminus in blue and the C terminus in red. GDT_TS Z scores are reported for Model 1 for T0513 and T0604 and for all models for T0740, with larger values indicating a larger separation from the rest of the predictions. The wfCPUNK prediction resulted from a collaboration between the Floudas, Liwo, and Scheraga laboratories as part of the collaborative folding experiment WeFold (http://www.wefold.org). The predictions shown in (B) and (C) used template information, whereas the prediction in (D) was strictly *ab initio*. The targets shown were among the most difficult in the respective CASP experiments.

NP-complete [36], we view that both conformational sampling and forcefield development are key limitations.

Contact and β-sheet topology prediction

Jones and coworkers introduced the contact prediction method PSICOV, which utilizes sparse inverse covariance estimation to predict contacts yielding a long-range L/5 contact precision $\geq$0.5 [37]. The approach has the limitation that it requires at least 500 sequences in the multiple sequence alignments for convergence. Marks and Sander introduced EVFold, which predicts contacts based on maximum entropy and coevolutionary couplings for contact predictions, but a similar challenge is faced in that 1000 sequences are required in the multiple sequence alignments to produce accurate contacts [38].

Success in contact prediction can substantially influence conformational search. Optimization-driven methods based on first principles were developed for the prediction of interhelical contacts in α-helical proteins [39] and both α/β and α+β proteins [40]. After input to the global-optimization framework ASTRO-FOLD [41,42], the contacts reduced the RMSD range of the sampled conformers by one-half [40]. Subramani and Floudas introduced BeST, for the prediction of β-sheet topologies with high precision and recall [43]. Baker and coworkers demonstrated in CASP10 with Rosetta-based methods [23] that given correct contacts for ~1 in 12 residues, this enabled the search for and construction of the correct topology [44], implying that if one can predict contacts with high positive predictive value, one can construct accurate topologies.

Successful protein designs with biotechnological applications

Protein design is the inverse folding problem [2,3] (Figure 2B, Box 1). Given a target fold, can we design a sequence to fold into that structure? Several notable examples are highlighted in Table 2. We present an overview of recent computational protein designs with biotechnological applications and describe the interplay with structural modeling necessary for success of the designs as appropriate. We refer the reader to [45,46] for excellent recent reviews of methodological advances and applications in *de novo* protein design.

Design of proteins and peptides for therapeutic applications

Over 200 peptides, proteins, or antibody therapeutics have been marketed as of 2010 [47]. Computational approaches have recently been applied to design new proteins and peptides for therapeutic applications. Elucidation of the sequences, structures, and interaction patterns of several disease-related proteins have allowed for the application of computational approaches for peptide therapeutic design [48]. Craik *et al*. [49] predict that by 2020 we will see more prevalence of peptides as drugs, while outlining the challenges to meeting that outcome. Here, we review timely applications by target.

Cancer. Generally, therapeutic proteins/peptides can: (i) interfere with signal transduction cascades; (ii) arrest the cell cycle through modulation of cyclin-dependent kinase activity; or (iii) directly induce apoptosis by modulation of the proteins controlling apoptosis [48]. Cysteine-rich intestinal protein 1 (CRIP1) is an early biomarker for breast cancer. Hao *et al*. used phage display to identify peptide sequences that bound to CRIP1. Subsequently, they computationally redesigned the scaffold sequence to optimize the binding free energy to increase its affinity for CRIP1, finding experimentally that it improved the IC50 27.5× over the phage-displayed sequence [50]. Cosic and coworkers used the Resonant Recognition Model (RRM) to design a short therapeutic peptide with myxoma virus antitumor/cytotoxic activity [51]. RRM represents a protein sequence as a series of numbers that can be analyzed by Fourier transformation and converted into a discrete spectrum, where a significant correlation to biological activity has been identified [51].

HIV. Correia *et al*. developed a computational method using side-chain grafting and Rosetta to transplant a continuous structural epitope, 4E10, into scaffold proteins for conformational stabilization and immune presentation [52]. The method produces epitope-containing designs that bind stronger to monoclonal antibody (mAb) 4E10 than 4E10 alone, and inhibits neutralization by HIV^+ sera. Floudas and coworkers designed HIV-1 entry inhibitors starting from the structure of the C14linkmid peptide in complex with the hydrophobic core of gp41 [53]. C14linkmid is a crosslinked peptide derived from the C-terminal heptad repeat gp41. A global optimization-based sequence selection was performed with a distance-dependent forcefield originally developed for protein folding [54] to select candidate sequences from the vast combinatorial space. These sequences were reranked using fold-specificity calculations, which sample conformations near the template structure with substitutions dictated by the newly designed sequences. It aims to determine how favorably a new sequence folds into the fold of the design template. A subset of top-ranked sequences identified in the fold-specificity stage was evaluated using approximate binding-affinity calculations, which approximate the binding equilibrium constant. The best design had an IC50 between 29 and 253 μM for different HIV-1 donors and mutants. This *de novo* design approach was made into an interactive web interface, Protein WISDOM [55].

Alzheimer's disease. Eisenberg and colleagues performed computationally guided design to predict and experimentally validate peptide inhibitors of fibril formation by the τ protein associated with Alzheimer's disease, as well as an amyloid promoting the sexual transmission of HIV [56]. The designs bind to the end of the steric zipper and inhibit elongation. Focusing on the τ protein inhibitor methodology, for a rotameric, fixed-backbone sequence optimization, they inverted the chirality of the design target to enable use of the Rosetta suite of tools. They designed L-amino acid sequences that favorably interact with a fixed-atom D version of the scaffold. Subsequently, the scaffold was reverted to its native L- form, and D-amino-acid-containing peptides were used as inhibitors experimentally. The designed D-peptides were then verified for shape complementarity, noting that D-Leu2 of the peptide was designed to clash with the target VQIVYK on the opposite sheet, and upon alanine substitution, inhibitory activity ceased. Introducing a tight-binding interface and clashes destroying the ability of a cascade of amyloid-forming sequences to propagate is effective for inhibition. Pande and coworkers, guided by observations made in simulations of $A\beta_{42}$, designed a noncanonical and D-amino-acid-containing peptide that organizes $A\beta_{42}$ into stable oligomers [57].

Antibody therapeutics. Gray and coworkers utilized Rosetta to introduce a noncanonical amino acid (NCAA) as an oxidizable crosslinker into an antibody complementarity determining region (CDR), with the best design experimentally crosslinking 52% of the available antigen [58]. Ellington and coworkers developed a supercharging protocol to substitute multiple surface residues with charged amino acids into proteins, using it to design an

Table 2. Summary of recent successful computational *de novo* designed and redesigned systems and their biotechnological applications.

Biotechnological application	Summary	Structure modeling	Forcefield(s)	*De novo* design method	Discriminating prediction metrics	Refs
Design of disease therapeutics	Computationally redesigned peptide sequence to bind to breast cancer biomarker	Cyclic peptides docked to NMR-derived conformers of target biomarker CRIP1	Physics	Eris	Change in binding free energy ΔΔG	[50]
	Bioactive peptide cytotoxic to tumor cells	Design based on primary structure alone	Physics	Resonant recognition model (RRM)	Electron–ion interaction potential	[51]
	Grafting 4E10 HIV epitope to new protein scaffolds	Matching 4E10 to putative protein scaffolds and docking to mAb 4E10	Hybrid	RosettaDesign	Rosetta energy	[52]
	HIV-1 entry inhibitors targeting gp41	Designed peptides and docked poses using TINKER and Rosetta *ab initio*	Physics, knowledge-based and hybrid	Components and subcomponents of the Protein WISDOM protocol	Distance-dependent energy followed by fold specificity and approximate binding affinity (*K**)	[53]
	Non-natural amino acid inhibitors of amyloid fibril formation by capping fibril ends	D-Amino acid containing sequence designed with Rosetta	Hybrid	RosettaDesign	Shape complementarity, total binding energy between inhibitor and scaffold, solubility	[56]
	Engineered peptides that stabilize amyloid-β oligomers	MD generated oligomer structure	Physics	Rational	Contact map and secondary structure analysis of MD	[57]
	Engineered crosslinking antibody with NCAA	NCAA mutations	Hybrid	Rational	Rosetta interface score	[58]
	Supercharged thermally resistant antibodies	Homology model of anti-MS2 antibody and supercharged surface mutations	Hybrid	RosettaDesign	Rosetta energy	[59]
	Design of antibodies targeting a peptide from hepatitis C, fluorescein, and VEGF	Combinations of backbones of CDR structures	Physics	OptCDR and IPRO	Mixed-integer linear optimization formulation to select structures followed by interaction energy between designs	[60]
	Inhibitors of hemagglutinin from the 1918 H1N1 pandemic virus that bind to and inhibit multiple other subtypes	Helical protein binders using Rosetta with hot-spot residue identification and shape-complementarity identification	Hybrid	RosettaDesign	Shape complementarity, electrostatic complementarity, RosettaDock, Rosetta energy	[80]
	Complement C3aR receptor agonists and antagonists	C3a-derived structures using TINKER	Physics, knowledge-based	Components and subcomponents of the Protein WISDOM protocol	Distance-dependent energy followed by fold specificity	[81]
	Complement system inhibitors of the compstatin-family targeting C3	Peptide inhibitors of C3	Physics, knowledge, and hybrid	Components and subcomponents of the Protein WISDOM protocol	Distance-dependent energy followed by fold specificity and *K**	[82,83]
Design of self-assembling proteins/peptides	Computational design of a protein crystal	Mathematically created idealized homotrimeric coiled-coiled protein	Physics	Site-specific amino acid probabilities calculated using a statistical thermodynamic method	AMBER energy function	[72]
	Design of a symmetric β-strand mediated homodimer	Computationally identified protein scaffold with symmetric protein–protein docking with Rosetta	Hybrid	Symmetric design with side-chain/backbone minimization using Rosetta	Rosetta-based energy metrics with visual inspection	[73]
	Self-assembling 24- and 12-subunit nanomaterials with different symmetries	Quaternary structures and interfaces of designed subunits through Rosetta with symmetric docking	Hybrid	RosettaDesign	Interface shape complementarity and energy	[74]

Table 2 (*Continued*)

Biotechnological application	Summary	Structure modeling	Forcefield(s)	*De novo* design method	Discriminating prediction metrics	Refs
Design of novel enzymes	Design of enzymes for Retro–Aldol, Kemp elimination, Diels–Alder, and organophosphate hydrolysis reactions	Idealized catalytic sites for transition state stabilization hashed onto template protein structures using RosettaMatch	Hybrid	RosettaDesign	Catalytic geometry, computed transition-state-binding energy, and consistency in side-chain conformations	[64–66, 68]
	Increased Diels–Alderase activity in *de novo* designed enzyme	Active sites reshaped by human players using Foldit	Hybrid	Interactive puzzles that use human intuition and pattern recognition in the online multi-player game Foldit	High-scoring (low-energy) sequences	[67]
	Redesigned cofactor binding site of CbXR to switch enzymatic cofactor specificity from NADPH to NADH	Homology model of *Candida boidinii* xylose reductase	Physics	IPRO	Interaction energy between designed sequences of CbXR and cofactors	[69]
	Redesign of GrsA-PheA specificity to non-native amino acid substrates	Penultimate Rotamer Library for rotameric states	Physics	*K**	Statistical mechanics-based metric using rotameric ensembles to approximate the binding constant K_d	[70]

Full details are provided in the corresponding references.

antibody with enhanced thermal inactivation resistance and a 30-fold affinity improvement [59]. Pantazes and Maranas introduced OptCDR for the design of antibodies to bind a targeted antigen epitope [60]. They applied it to design antibodies targeting a peptide from the capsid of hepatitis C, fluorescein, and vascular endothelial growth factor (VEGF), and validated the approach with computational metrics and binding energies. They recently introduced the Modular Antibody Parts (MAPs) database [61]. MAPs works in the spirit of template-based modeling where the templates are prototype structures of the random variable (V), diversity (D), and joining (J) regions in the database, resulting in gene combinations with the fewest amino acid changes from the target. Using this database, they were able to predict antibody tertiary structures with an average all-atom RMSD of 1.9 Å on a testing set of 260 antibodies [61]. Upon successful prediction of a target structure, such antibodies can be computationally affinity matured using the Iterative Protein Redesign and Optimization (IPRO) framework [62]. IPRO is an iterative framework that optimizes side-chain substitutions in user-determined design positions using a mixed-integer optimization model in which subsequently the backbone of the protein being designed is allowed to adjust through local minimizations to the new side-chains.

Design or redesign of enzymes and biocatalysts

Baker has reviewed the challenges and utility in succeeding in this endeavor [63]. Jiang *et al*. have developed a computational method for constructing an active site for multistep reactions, designing 32 enzymes, spanning different protein folds and having detectable retro-aldolase activity for 4-hydroxy-4-(6-methoxy-2-naphthyl)-2-butanone, which is not found in biological systems [64]. The method designs active sites for these reactions with superimposed transition states of the reactions involved. Notably, designs identified using explicit water molecules were more successful, achieving enhancements of up to four orders of magnitude compared to the uncatalyzed reaction [64]. They also designed eight enzymes with different catalytic motifs for catalysis of Kemp elimination reactions [65].

Siegel *et al*. computationally designed stereoselective enzyme catalysts for the Diels–Alder reaction, prior to which none existed [66]. The most active design was confirmed to match the X-ray structure, noting that the success in design presumably was related to the success in modeling the designed catalytic site. Human game players, through the visual interface of the online multiplayer game, Foldit, were able to 'hands-on' remodel and redesign structures and side-chains of a 24-residue helix–turn–helix motif [67]. Based on this crowd-sourced design, an 18-fold increase in enzymatic activity over their previously designed enzyme was achieved. The players were guided in the design by scores, which were inversely proportional to the Rosetta energy, and visual intuition.

Khare *et al*. computationally redesigned mononuclear zinc metalloenzymes to catalyze non-native organophosphate hydrolysis activity with the experimental structures largely matching the designed ones [68]. Common themes from the catalyst design work were the correct modeling of the reaction transition states, which require quantum chemical calculations.

Maranas and coworkers computationally redesigned *Candida boidinii* xylose reductase (CbXR) to switch experimentally its cofactor specificity from NADPH to NADH with the IPRO procedure [69]. There is no experimentally resolved structure for CbXR, so they used a homology model to perform the design blind of the true native structure. A 10^4-fold specificity change was observed to NADH, due primarily to changes in hydrogen bond and local charge interactions. Seven of ten predictions had

significant xylose reductase activity utilizing NADH; the remaining two variants had dual cofactor specificity [69].

Donald and coworkers redesigned the specificity of nonribosomal peptide synthetase enzyme gramicidin S synthetase A (GrsA-PheA) from Phe to Leu, Arg, Glu, Lys, or Asp [70]. The computational redesign used physics-based energy evaluations of rotamerically sampled sequence space through the statistical mechanics based *K** algorithm to approximate the binding constant K_d for the different analogs [71]. This study suggested that structure-based computational design can identify different mutants than those that have evolved, and that the designs could be used for charged amino acid adenylation [70].

Self-assembling proteins/peptides

Controlling ordered (i.e., crystals) or disordered (i.e., hydrogels) self-assembly of proteins is a critical test of our understanding of both structure and interactions, having applications in biologically inspired materials. Lanci *et al*. computationally designed a protein crystal starting from an idealized homotrimeric parallel coiled-coil template and redesigned the interfaces [72]. They utilized strictly physics-based energy functions to discriminate favorable interfaces. Stranges *et al*. took the solvent-exposed β strands of two monomeric proteins and redesigned them to form an intermolecular β sheet symmetric homodimer with near atomic-level accuracy [73]. This design demonstrated the creation of unique stabilizing interactions at an interface. King *et al*. designed symmetric self-assembling complexes to atomic level accuracy [74]. They performed symmetric docking of subunits followed by redesign at the interfaces to design cage-like nanomaterials with tetrahedral or octahedral point group symmetry. The designed structures were confirmed experimentally by crystallography and electron microscopy to high agreement. The control over such self-assembling can be used to design advanced functional materials and molecular machines [74].

Other applications

Hecht and coworkers designed *de novo* artificial sequences using a binary code strategy that encoded function and enabled cell growth after knocking out several naturally occurring genes required for cell viability [75]. The binary code strategy postulates that a simple code of alternating polar and nonpolar residues patterned in different ways can yield α helices or β strand structures. They used this strategy to design a series of helical bundles which rescued *Escherichia coli* cells with essential genes conditionally knocked out, and showed how a simplistic design strategy can produce proteins of novel function sufficient to sustain life. Piana *et al*. computationally designed the fastest folding β protein [76]. They noted that the prior fastest β protein, FiP35, was about an order of magnitude slower than its helical counterpart. The reduced folding time of the predicted design was experimentally confirmed to be ~3 times faster than the previous record holder.

Box 2. Outstanding questions

- How can we predict structures of sequences that are not homologous to any known protein?
- How can we accurately predict the β-sheet topology?
- How can we accurately predict medium and long-range tertiary contacts?
- How can we consistently and substantially refine predicted protein structures to be closer to the native?
- How can we predict structures of membrane proteins?
- How can we predict the effects of the many PTMs and NCAAs on the structures of proteins?
- How can we design soluble, passively permeable, metabolically stable peptides and proteins as therapeutics?
- How do we incorporate PTMs and NCAAs into design, and address the massive increase in combinatorial complexity?

Concluding remarks and future perspectives

One can be successful in accurately predicting protein structures from sequence alone if templates can be identified and properly aligned. However, there is no method yet that can consistently predict structures template free, possibly because conformational sampling and forcefields to guide sampling/selection are still imperfect [34,35]. Even if accurate conformations are sampled, no method exists to score accurately those models more favorably from other decoys. Interestingly, it has been suggested that all the puzzle pieces needed to construct any structure are available, despite the fact that no method is currently able to assemble them properly in a blind predictive capacity [26,77]. In our opinion, improvement in forcefields, the ability to predict accurately residue–residue contacts, β-sheet topologies, alignments to nonhomologous templates, and effective conformational sampling methods are the key elements to solving the protein folding problem (Box 2).

Transmembrane proteins are a class of targets that remain challenging for protein folding and design, despite being of significant interest to the pharmaceutical industry. These proteins are extremely difficult to solve experimentally due to their insoluble nature, and therefore few template structures exist for membrane proteins, although they account for the majority of current drug targets [78]. Further advances in the modeling of the membrane protein environment are needed to allow for improved structural models and evaluation of designed ligands.

In the *de novo* design paradigm, one has $20^{\#\text{DesignPositions}}$ sequences to evaluate. Doing this exhaustively computationally is largely impractical and even more so experimentally for even a few design positions. Proteins as potential therapeutics are hindered by proteolytic cleavage, poor solubility, and poor permeability. For these reasons, most have extracellular targets and often must be injected in order to be clinically successful. Using post-translational modifications (PTMs) and NCAAs can help with these challenges, because modified peptides are less likely to be recognized by proteases, and these peptide modifications can be selected to fine-tune bioavailability. Design of modified peptide sequences adds complexity, because by considering the over 400 known PTMs for design, the combinatorial problem increases significantly to $>420^{\#\text{DesignPositions}}$ combinations [79]. The methods to model PTMs and NCAAs are still at an early stage of development, and represent a challenge in protein structure prediction and *de novo* protein design. Looking forward, we have just touched the surface of the allowable chemical space of proteins and their potential biotechnological applications.

Acknowledgments

CAF acknowledges support from the National Institutes of Health grant number R01GM052032 and the National Science Foundation. GAK is grateful for support by a National Science Foundation Graduate Research Fellowship under grant number DGE-1148900. We thank members of the Computer-Aided Systems Laboratory for helpful discussions.

References

1 Pantazes, R.J. *et al.* (2011) Recent advances in computational protein design. *Curr. Opin. Struct. Biol.* 21, 467–472
2 Drexler, K.E. (1981) Molecular engineering: an approach to the development of general capabilities for molecular manipulation. *Proc. Natl. Acad. Sci. U.S.A.* 78, 5275–5278
3 Pabo, C. (1983) Molecular technology. Designing proteins and peptides. *Nature* 301, 200
4 Dill, K.A. and MacCallum, J.L. (2012) The protein-folding problem, 50 years on. *Science* 338, 1042–1046
5 MacCallum, J.L. *et al.* (2011) Assessment of protein structure refinement in CASP9. *Proteins* 79, 74–90
6 MacCallum, J.L. *et al.* (2009) Assessment of the protein-structure refinement category in CASP8. *Proteins* 77, 66–80
7 Zhang, Y. (2008) Progress and challenges in protein structure prediction. *Curr. Opin. Struct. Biol.* 18, 342–348
8 Floudas, C.A. (2007) Computational methods in protein structure prediction. *Biotechnol. Bioeng.* 97, 207–213
9 Moult, J. *et al.* (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins* 23, ii–iv
10 Zhang, Y. (2013) Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins* http://dx.doi.org/10.1002/prot.24341
11 Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242
12 Kryshtafovych, A. *et al.* (2011) CASP9 results compared to those of previous casp experiments. *Proteins* 79, 196–207
13 Wu, S. and Zhang, Y. (2007) LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.* 35, 3375–3382
14 Roy, A. *et al.* (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* 5, 725–738
15 Zhang, Y. (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9, 40
16 Roy, A. *et al.* (2011) A protocol for computer-based protein structure and function prediction. *J. Vis. Exp.* e3259
17 Xu, D. and Zhang, Y. (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* 80, 1715–1735
18 Li, Y. and Zhang, Y. (2009) REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins* 76, 665–676
19 Zhang, J. *et al.* (2011) Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. *Structure* 19, 1784–1795
20 Joo, K. *et al.* (2013) Protein structure modeling for CASP10 by multiple layers of global optimization. *Proteins* http://dx.doi.org/10.1002/prot.24397
21 Hildebrand, A. *et al.* (2009) Fast and accurate automatic structure prediction with HHpred. *Proteins* 77 (Suppl 9), 128–132
22 Peng, J. and Xu, J. (2011) RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins* 79 (Suppl 10), 161–171
23 Leaver-Fay, A. *et al.* (2011) ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* 487, 545–574
24 Zhou, H. and Skolnick, J. (2012) Template-based protein structure modeling using TASSERVMT. *Proteins* 80, 352–361
25 Zhou, H. and Skolnick, J. (2011) GOAP: A generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.* 101, 2043–2052
26 Zhang, Y. and Skolnick, J. (2005) The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1029–1034
27 Duan, Y. and Kollman, P.A. (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 282, 740–744
28 He, Y. *et al.* (2013) Lessons from application of the UNRES force field to predictions of structures of CASP10 targets. *Proc. Natl. Acad. Sci. U.S.A.* http://dx.doi.org/10.1073/pnas.1313316110
29 Liwo, A. *et al.* (2011) Coarse-grained force field: general folding theory. *Phys. Chem. Chem. Phys.* 13, 16890–16901
30 Lindorff-Larsen, K. *et al.* (2011) How fast-folding proteins fold. *Science* 334, 517–520
31 Piana, S. *et al.* (2013) Atomic-level description of ubiquitin folding. *Proc. Natl. Acad. Sci. U.S.A.* http://dx.doi.org/10.1073/pnas.1218321110
32 Piana, S. *et al.* (2011) How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* 100, L47–L49
33 Shaw, D.E. *et al.* (2009) Millisecond-scale molecular dynamics simulations on Anton. In *Proceedings of the Conference on High Performance Computing, Networking, Storage and Analysis*. pp. 1–11 ACM
34 Bradley, P. *et al.* (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309, 1868–1871
35 Raval, A. *et al.* (2012) Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins* 80, 2071–2079
36 Berger, B. and Leighton, T. (1998) Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *J. Comput. Biol.* 5, 27–40
37 Jones, D.T. *et al.* (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 28, 184–190
38 Marks, D.S. *et al.* (2011) Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* 6, e28766
39 Rajgaria, R. *et al.* (2009) Towards accurate residue–residue hydrophobic contact prediction for α helical proteins via integer linear optimization. *Proteins* 74, 929–947
40 Rajgaria, R. *et al.* (2010) Contact prediction for beta and alpha-beta proteins using integer linear optimization and its impact on the first principles 3D structure prediction method ASTRO-FOLD. *Proteins* 78, 1825–1846
41 Subramani, A. *et al.* (2012) ASTRO-FOLD 2.0: An enhanced framework for protein structure prediction. *AIChE J.* 58, 1619–1637
42 Klepeis, J.L. and Floudas, C.A. (2003) ASTRO-FOLD: a combinatorial and global optimization framework for Ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophys. J.* 85, 2119–2146
43 Subramani, A. and Floudas, C.A. (2012) β-sheet topology prediction with high precision and recall for β and mixed α/β proteins. *PLoS ONE* 7, e32461
44 Kim, D.E. *et al.* (2013) One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins* http://dx.doi.org/10.1002/prot.24374
45 Samish, I. *et al.* (2011) Theoretical and computational protein design. *Annu. Rev. Phys. Chem.* 62, 129–149
46 Fung, H.K. *et al.* (2008) Computational de novo peptide and protein design: rigid templates versus flexible templates. *Ind. Eng. Chem. Res.* 47, 993–1001
47 Vlieghe, P. *et al.* (2010) Synthetic therapeutic peptides: science and market. *Drug Discov. Today* 15, 40–56
48 Pirogova, E. and Istivan, T. (2013) Toward development of novel peptide-based cancer therapeutics: computational design and experimental evaluation. In *Bioinformatics of Human Proteomics*. pp. 103–126, Springer
49 Craik, D.J. *et al.* (2013) The future of peptide-based drugs. *Chem. Biol. Drug Des.* 81, 136–147
50 Hao, J. *et al.* (2008) Identification and rational redesign of peptide ligands to CRIP1, a novel biomarker for cancers. *PLoS Comput. Biol.* 4, e1000138
51 Istivan, T.S. *et al.* (2011) Biological effects of a de novo designed myxoma virus peptide analogue: evaluation of cytotoxicity on tumor cells. *PLoS ONE* 6, e24809
52 Correia, B.E. *et al.* (2010) Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic hiv vaccine epitope. *Structure* 18, 1116–1126
53 Bellows, M.L. *et al.* (2010) Discovery of entry inhibitors for HIV-1 via a new de novo protein design framework. *Biophys. J.* 99, 3445–3453
54 Rajgaria, R. *et al.* (2008) Distance dependent centroid to centroid force fields using high resolution decoys. *Proteins* 70, 950–970

55 Smadbeck, J. *et al.* (2013) Protein WISDOM: a workbench for in silico de novo design of biomolecules. *J. Vis. Exp.* e50476
56 Sievers, S.A. *et al.* (2011) Structure-based design of non-natural amino-acid inhibitors of amyloid fibril formation. *Nature* 475, 96–100
57 Rajadas, J. *et al.* (2011) rationally designed turn promoting mutation in the amyloid-β peptide sequence stabilizes oligomers in solution. *PLoS ONE* 6, e21776
58 Xu, J. *et al.* (2013) Structure-based non-canonical amino acid design to covalently crosslink an antibody–antigen complex. *J. Struct. Biol.* http://dx.doi.org/10.1016/j.jsb.2013.1005.1003
59 Miklos, A.E. *et al.* (2012) Structure-based design of supercharged, highly thermoresistant antibodies. *Chem. Biol.* 19, 449–455
60 Pantazes, R.J. and Maranas, C.D. (2010) OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding. *Protein Eng. Des. Sel.* 23, 849–858
61 Pantazes, R.J. and Maranas, C.D. (2013) MAPs: a database of modular antibody parts for predicting tertiary structures and designing affinity matured antibodies. *BMC Bioinformatics* 14, 168
62 Saraf, M.C. *et al.* (2006) IPRO: an iterative computational protein library redesign and optimization procedure. *Biophys. J.* 90, 4167–4180
63 Baker, D. (2010) An exciting but challenging road ahead for computational enzyme design. *Protein Sci.* 19, 1817–1819
64 Jiang, L. *et al.* (2008) De novo computational design of retro-aldol enzymes. *Science* 319, 1387–1391
65 Rothlisberger, D. *et al.* (2008) Kemp elimination catalysts by computational enzyme design. *Nature* 453, 190–195
66 Siegel, J.B. *et al.* (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. *Science* 329, 309–313
67 Eiben, C.B. *et al.* (2012) Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nat. Biotechnol.* 30, 190–192
68 Khare, S.D. *et al.* (2012) Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. *Nat. Chem. Biol.* 8, 294–300
69 Khoury, G.A. *et al.* (2009) Computational design of *Candida boidinii* xylose reductase for altered cofactor specificity. *Protein Sci.* 18, 2125–2138
70 Chen, C-Y. *et al.* (2009) Computational structure-based redesign of enzyme activity. *Proc. Natl. Acad. Sci. U.S.A* 106, 3764–3769
71 Lilien, R.H. *et al.* (2005) A novel ensemble-based scoring and search algorithm for protein redesign and its application to modify the substrate specificity of the gramicidin synthetase a phenylalanine adenylation enzyme. *J. Comput. Biol.* 12, 740–761
72 Lanci, C.J. *et al.* (2012) Computational design of a protein crystal. *Proc. Natl. Acad. Sci. U.S.A.* 109, 7304–7309
73 Stranges, P.B. *et al.* (2011) Computational design of a symmetric homodimer using β-strand assembly. *Proc. Natl. Acad. Sci. U.S.A.* 108, 20562–20567
74 King, N.P. *et al.* (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science* 336, 1171–1174
75 Fisher, M.A. *et al.* (2011) *De Novo* designed proteins from a library of artificial sequences function in *Escherichia coli* and enable cell growth. *PLoS ONE* 6, e15364
76 Piana, S. *et al.* (2011) Computational design and experimental testing of the fastest-folding β-sheet protein. *J. Mol. Biol.* 405, 43–48
77 Skolnick, J. *et al.* (2012) Further evidence for the likely completeness of the library of solved single domain protein structures. *J. Phys. Chem. B* 116, 6654–6664
78 Overington, J.P. *et al.* (2006) How many drug targets are there? *Nat. Rev. Drug Discov.* 5, 993–996
79 Khoury, G.A. *et al.* (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Rep.* 1, http://dx.doi.org/10.1038/srep00090
80 Fleishman, S.J. *et al.* (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science* 332, 816–821
81 Bellows-Peterson, M.L. *et al.* (2012) De novo peptide design with C3a receptor agonist and antagonist activities: theoretical predictions and experimental validation. *J. Med. Chem.* 55, 4159–4168
82 Bellows, M.L. *et al.* (2010) New compstatin variants through two de novo protein design frameworks. *Biophys. J.* 98, 2337–2346
83 Gorham, R.D., Jr *et al.* (2013) Novel compstatin family peptides inhibit complement activation by drusen-like deposits in human retinal pigmented epithelial cell cultures. *Exp. Eye Res.* 116, 96–108