FEBS Letters 587 (2013) 1147–1154

journal homepage: www.FEBSLetters.org

Review

Emerging themes in the computational design of novel enzymes and protein–protein interfaces

Sagar D. Khare [a,*], Sarel J. Fleishman [b,*]

[a] Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, NJ 08854, USA
[b] Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel

ARTICLE INFO

Article history:
Received 19 November 2012
Revised 10 December 2012
Accepted 10 December 2012
Available online 19 December 2012

Edited by Wilhelm Just

Keywords:
Computational design
Novel protein function
Rosetta
ORBIT
Negative design
Enzyme
Protein interaction
Energy function

ABSTRACT

Recent years have seen the first applications of computational protein design to generate novel catalysts, binding pairs of proteins, protein inhibitors, and large oligomeric assemblies. At their core these methods rely on a similar hybrid energy function, composed of physics-based and database-derived terms, while different sequence and conformational sampling approaches are used for each design category. Although these are first steps for the computational design of novel function, crystal structures and biochemical characterization already point out where success and failure are likely in the application of protein design. Contrasting failed and successful design attempts has been used to diagnose deficiencies in the approaches and in the underlying hybrid energy function. In this manner, design provides an inherent mechanism by which crucial information is obtained on pressing areas where focused efforts to improve methods are needed. Of the successful designs, many feature pre-organized sites that are poised to perform their intended function, and improvements often result from disfavoring alternative functionally suboptimal states. These rapid developments and fundamental insights obtained thus far promise to make computational design of novel molecular function general, robust, and routine.

1. Introduction

What are the design principles that underlie the complex, sophisticated and beautiful protein systems that are at the heart of all life processes? Nature provides an inspiring number of different functional classes of proteins, from signaling molecules that display exquisitely fine-tuned molecular recognition, through regulated membrane channels and pumps, to enzymes that catalyze essential reactions at specificities and efficiencies that are unmatched by human invention. Biochemical and theoretical work has long been used to characterize how function is encoded in these systems, often by studying the impact of mutations on natural proteins. However, a myriad of evolutionary forces operating over countless generations has shaped extant natural systems, confounding the inference of key design principles. Recent advances in computation and high-throughput experimental analysis have opened the way to generating molecular function from the bottom up. By controlling all inputs into the process, computational protein design of novel function offers an intriguing route to uncover fundamental principles that explain existing molecular functions and, by extrapolation, allows construction of functional systems with no known natural counterparts.

Recent years have seen the first steps made by computational protein designers to produce novel catalysts, binding proteins, inhibitors, and oligomeric assemblies. Their approaches all rely on the inverse-folding paradigm [1], where the target state (a protein bound to its partner, be it another protein or a transition-state model) is modeled in atomic detail and the designed protein's sequence is chosen to form energetically favorable interactions with its target. The choice of the sequence is guided by (1) the technique used to consider different candidate sequences, or the sampling method, e.g., simulated annealing or dead-end elimination, and (2) the energy (scoring) function used to compare these candidate sequences. Energy functions used in design are usually "hybrid" (Fig. 1) – they feature terms that are physics-based (e.g., the Lennard–Jones potential for atomic repulsion and dispersion forces) and terms that are derived from known three-dimensional structures of proteins (e.g., amino acid sidechain conformational preferences observed in the Protein DataBank (PDB)). A crucial early insight was that both the energy function and sampling techniques used should be general and independent of the particular design or modeling problem [2–4]. In such a framework, the design process provides a powerful mechanism to diagnose the state of our understanding of protein energetics; improvements in the energy

* Corresponding authors.
E-mail addresses: sagar.khare@rutgers.edu (S.D. Khare), sarel@weizmann.ac.il (S.J. Fleishman).

0014-5793/$36.00
http://dx.doi.org/10.1016/j.febslet.2012.12.009

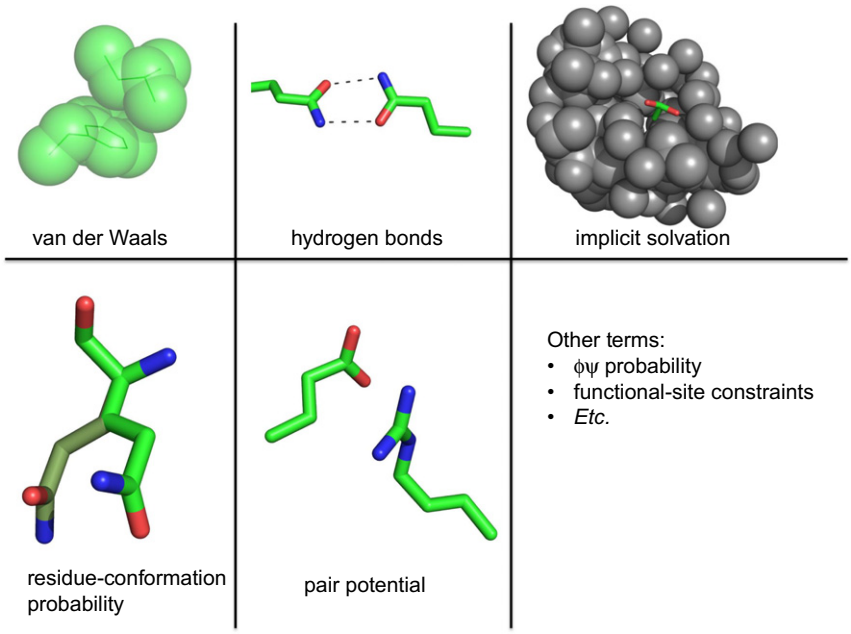

Fig. 1. A schematic description of hybrid all-atom energy functions used in design calculations. The energy functions used by several protein design software suites such as ORBIT and Rosetta are dominated by the following contributions (left-to-right, and top-to-bottom): van der Waals interactions accounting for both the attractive interactions among apposing molecular surfaces and the repulsive interactions due to steric overlap; hydrogen bonds between acceptor atoms (e.g., carbonyls), and donating polar hydrogens [62]. Hydrogen-bond strengths are determined by distance, orientation, and the polarity of the acceptor and donor; polar atoms are stabilized by interacting with water in their vicinity. Sequestering polar groups in the protein reduces some of this stabilization. Macromolecular forcefields used in design do not explicitly model the interactions with water molecules, rather use an implicit solvation model in which the total volume of excluded water in the vicinity of polar atoms is assessed by counting the number of protein atoms in a shell surrounding the atom [63,64]; residue sidechains are observed to reside in a limited set of preferred conformations, known as rotamers due to the nature of the chemical bonds within the residue and to dependencies on the local backbone conformations [64]. These probabilities are converted to pseudo-energies and used to bias conformations to the most likely ones; certain residue pairs are observed to cluster more often than others, for instance, due to the formation of stabilizing electrostatic interactions. A pair potential is derived from these propensities and used as a pseudo-energy term [65,66]. Other pseudo-energy terms are derived from the structures in the PDB and based on the desired molecular function (e.g., catalytic constraints). Increasing the reliability of macromolecular energy functions is an active area of research; major aspects that lack accuracy are the effects of solvation and electrostatics [55,67]. For a detailed treatment of the energy terms used in Rosetta we refer the reader to Ref. [49] and in ORBIT to Refs. [3,5]. All molecular figures were generated using PyMol [68].

function can then be fed back to improve all protein design, and more generally, protein modeling, efforts.

The earliest demonstrations that hybrid energy functions are useful for design came from the complete computational redesign of a zinc-finger protein by Dahiyat and Mayo [5], followed by the de novo design of a protein fold not observed in nature, by Kuhlman, Baker and co-workers [6], and more recently, a similar strategy led to the design of a novel protein loop [7]. Here, we limit ourselves to discussing the computational design of novel protein function – particularly novel enzymes and protein binders – that has been corroborated by experimental atomic structures but note that very exciting progress has been made in computational design of *altered* protein function, such as novel binding specificities [8,9] and allosteric regulation [10], and refer readers to a recent review [11].

2. Design of novel enzymes

Natural enzymes are amazingly proficient catalysts that can accelerate the rates of their cognate reactions by as much as 10^{23} fold [12]. The ability to de novo design an enzyme to catalyze any desired chemical reaction is a stringent test of our understanding of catalysis and will have significant practical applications in medicine and industry. Early computational design efforts focused on introducing metal-binding sites in proteins ([13,14]), and "nascent" metalloenzymes for redox chemistries were obtained by virtue of open metal co-ordination sites in the designed proteins ([15–17]). However, these studies did not include explicit computational models of the chemical transformation being catalyzed. A pioneering effort by Bolon and Mayo included atomistic details of the catalyzed reaction and introduced a nucleophilic histidine residue on the surface of a catalytically inert thioredoxin to obtain catalysts ("protozymes") for the hydrolysis of an activated ester substrate [18].

It is widely accepted that natural enzymes make two primary contributions to catalysis: they interact favorably with the reaction transition state [19] and they shield the chemical groups that aid catalysis from water [20], thereby increasing their reactivity; together these mechanisms lower the transition-state free energy in the active site microenvironment compared to the bulk solvent. To generate novel enzymes, design efforts in the framework of programs such as ORBIT and Rosetta have attempted to emulate these properties of natural enzymes. The process starts by modeling a so-called theozyme that is composed of a model of the chemical transition state(s) and key amino acid residues placed in orientations that are predicted to favor interactions with the transition-state model [21]. The transition-state structure cannot be experimentally determined due to its short lifespan (a few femtoseconds at room temperature [22]), so it is either adapted from crystal structures of transition-state analogue bound enzymes, or is based on quantum-chemical calculations. Constellations of backbones that can support the theozyme model are searched among hundreds of small-molecule binding pockets in crystallographic protein structures [23,34]. The sequence of residues in the putative catalytic pocket is then optimized to both favor maintenance of catalytic geometry and to provide additional stabilization to the transition state(s) [25]. Several candidate designs are synthesized in the laboratory and assayed for their programmed activity.

The first reports of successful computational de novo design of enzymes utilizing the approach outlined above came from efforts to catalyze the model reactions, Kemp elimination [26] and the retro-aldol reaction [27], for both of which no natural biocatalyst is known. In both cases small and relatively activated substrates were chosen for initial proof-of-principle efforts (Fig 2A and B). The catalytic efficiencies of designed enzymes were initially quite modest but could be improved through in vitro evolution, yielding in some instances >10^3-fold increases over the original design [28–30]. Subsequently solved crystallographic structures of enzymes bound to substrate analogues and mutational analysis showed that in some enzymes much of the catalytic machinery was in place, but also revealed that the design of polar interactions and water-mediated contacts was not accurate [31], and these elements were visualized in orientations that were not optimal for catalysis [32,33] (Fig 2C and D). Analysis by crystallographic structure determination and molecular dynamics (MD) simulation suggested that conformational plasticity underlies the lack of activity in another designed Kemp eliminase [34]. In the latter case, substitutions were subsequently introduced to disfavor hydration of the active-site environment and to reduce the flexibility of catalytic groups, leading to an active enzyme with a catalytic efficiency, k_{cat}/K_m, of 430 $M^{-1}s^{-1}$. These results demonstrate a potentially powerful strategy in which an assessment of the conformational stability of the designed active site from MD simulations is used iteratively to complement design algorithms that typically operate on static structures. Further, they suggest that improvements in designed enzymes can be made by restricting the conformational degrees of freedom of polar amino acid residues, possibly with additional contacts formed to other structural elements in the designed protein [35]. Other areas where improvements are needed are in understanding how to confine water molecules in the catalytically desired orientation and how to modulate the degree of hydration of key catalytic groups.

In a second generation of enzyme design, a possible route to address the challenges of designing polar interaction networks was tested by constructing much of the network using backbone carbonyl and amide groups rather than water molecules and sidechain atoms; backbone functional groups are on average less flexible than those of sidechains and water molecules [36]. To understand the minimal requirements for ester hydrolysis, a reaction carried out by a number of natural enzymes, cysteine-histidine dyads along with various conceptions of an oxyanion-hole were used as theozymes for design. Oxyanion holes contribute to catalysis by donating hydrogen bonds to polarize the substrate, therefore different theozyme conceptions featured sidechain, backbone amide and explicitly modeled water molecule donor groups, respectively. While designs featuring sidechain and explicitly modeled water molecules were inactive, utilizing backbone-mediated hydrogen bonds yielded designed enzymes with catalytic efficiencies k_{cat}/K_m ~400 $M^{-1}s^{-1}$. Crystal structures of the designed (apo-) enzymes showed that the backbone-mediated hydrogen bonds and the nucleophilic cysteine residues were in their designed conformations. However, the catalytic histidine residue, which was designed on a loop segment, was not in an appropriate position for catalysis. In contrast, in the natural enzymes that feature similar catalytic machinery, the histidine residue is highly pre-organized by a conserved hydrogen bond to a spatially proximal

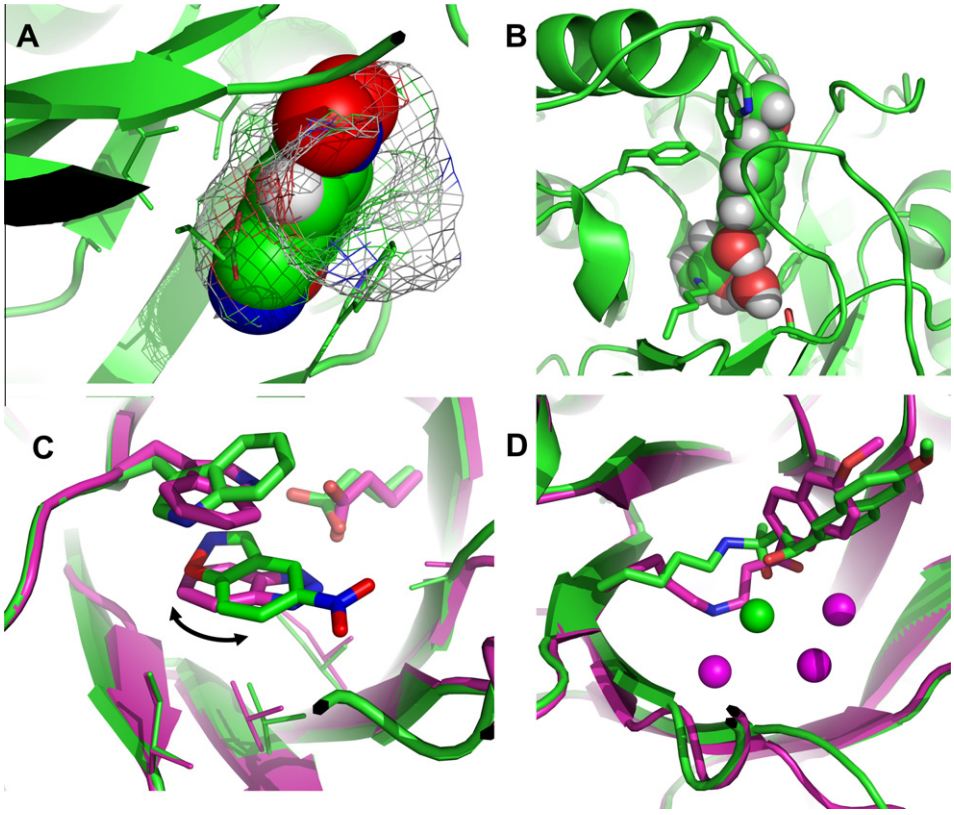

Fig. 2. Models and crystal structures of computationally designed enzymes provide high-resolution mechanistic insights and highlight challenges for modulating conformational plasticity in design. (A) and (B) Computational models of the designed Kemp eliminase, KE59, and Retro-aldolase, RA34, respectively. The TS models employed in the calculations are shown in spheres, and the catalytic and some of the binding residues are highlighted. The designs feature highly shape complementary interfaces. (C) Comparison of the KE59 design model (green) with a crystal structure of the protein bound to the substrate analogue benzotriazole (magenta). The benzotriazole is in a flipped orientation relative to the TS model, and rotameric changes are observed in the active site. (D) Comparison of the RA34 design model (green) with a crystal structure of the protein bound to a diketone inhibitor, 1-(6-methoxynaphthalen-2-yl)butane-1,3-dione (magenta). The catalytic lysine residue adopts a different rotamer in the two structures. The modeled and crystallographic water molecules are shown as spheres.

asparate/glutamate sidechain. The low catalytic efficiency of designed enzymes could thus be attributed to a reorganization penalty for orienting the histidine in a catalytically productive conformation, in agreement with earlier calculations highlighting the importance of preorganization in natural enzymes [37].

Another design approach that takes advantage of pre-organized polar functional groups is to computationally repurpose the machinery of existing enzymes for catalyzing new, non-cognate reactions. A set of mononuclear zinc-containing metalloenzymes was used to design catalysts for organophosphate (OP) hydrolysis, and a redesigned adenosine deaminase that catalyzed OP hydrolysis with a catalytic efficiency of $k_{cat}/K_m \sim 10^4\ M^{-1}s^{-1}$ after directed evolution was obtained [38]. The wild type deaminase had no detectable OP hydrolysis activity at comparable enzyme and substrate concentrations. While both OP hydrolysis and deamination reactions are hydrolyses, their respective transition-state geometry, leaving-group character, and substrate electrophilicity are quite distinct. The most active designed OP hydrolase variant featured 11 substitutions in the wild type deaminase, and analysis of the mutational landscape of activity switching indicated that 4 designed substitutions were required simultaneously for obtaining detectable OP hydrolysis activity. Obtaining these epistatic mutations using directed evolution would require generating and screening an impractically large library. Thus, success with this computational approach offers a general route to exploring the untapped catalytic prowess of natural enzymes for novel reactivities, and suggests that once highly activated, solvent-shielded catalytic machinery is obtained, it can be channeled effectively to carry out novel transformations.

Significant improvements in the catalytic efficiencies of designed enzymes have been made using directed evolution techniques in which substitutions both spatially proximal to and distal from the designed active site are screened for activity improvements [30,33,36]. A general theme that has emerged from these studies is that substitutions that do not directly contact the substrate can have substantial effects on catalytic efficiencies (k_{cat}/K_m), and most of the improvements that result through evolution occur in the catalytic rate (increases in k_{cat}) as opposed to substrate binding (decreases in K_m). Rationalizing these observed improvements using comparative structural analyses ([33,34,36]) as well as detailed quantum-chemical simulations ([39–41]) of the designed active sites will be useful feedback for improving the energy functions used in design. Taken together, recent efforts in these directions suggest that the observed improvements are likely the result of one or more of the following: (a) increased pre-organization of catalytic residues induced by the destabilization of non-productive alternative states, (b) better activation of the catalytic groups (e.g., pK_a modulation) by modulation of their electrostatic environment, and (c) better alignment of the substrate with respect to the catalytic machinery. These insights make a strong case for improvements in modeling long range, electrostatic effects of amino acid substitutions on both the active site environment and the entire free energy landscape of the reaction (as opposed to a single transition state). Such improvements will be crucial to obtain the next generation of highly active computationally designed catalysts and bridge the currently large efficiency gap between designed and natural enzymes [42].

3. Design of novel protein interactions

Protein–protein interactions underlie all life processes. The ability to design protein interactions that are not seen in nature is a stringent test of our understanding of the physical basis for biomolecular recognition, and provides a future route to novel molecules with therapeutic, technological, and research utility. Initial successes were demonstrated in designing low-affinity binding pairs through computational docking followed by design to generate a heterodimer comprising two variants of a single natural protein [43] and a designed protein that bound a natural target protein [44]. Likely due to the low affinities ($K_d \geqslant 100\ \mu M$), experimentally determined molecular structures of the designed interactions have not been obtained. In the following we discuss designs where molecular structures were obtained, providing a clear test for the design strategy.

Certain protein scaffolds recur through evolution as protein-binding modules, which have favorable characteristics for binding a wide range of molecular surfaces. For instance, the ankyrin domain has been studied extensively using in vitro evolution and crystallography, yielding many insights on the importance of certain positions for its stability and binding [45]. Following this lead from evolution, a design approach utilized these insights to design an ankyrin domain and a set of hyperthermophilic proteins to form binding pairs by constraining residue choices and interaction types between the two designed partners [46]. For instance, aspartate residues that were known to form important stabilizing interactions within the ankyrin repeat as well as cross-interface interactions were conserved in design calculations. On the hyperthermophilic partner, two aromatic residues, tyrosine and tryptophan, which are often seen to make important contributions to binding affinity in natural complexes [47], were introduced to form hydrophobic and hydrogen-bonding interactions across the interface. These aromatic residues were supported by a hydrophobic ring of residues, followed by polar and charged residues at the interface periphery, and designs were selected on the basis of how well they conformed to properties of naturally occurring protein–protein interactions, such as the content of polar groups at the interface. This design strategy yielded one medium-affinity designed pair ($K_d \sim 150$ nM) comprising approximately 20 substitutions on each partner protein relative to the wild-type pair and substitutions identified through in vitro evolution improved binding affinity by 3 orders of magnitude generating subnanomolar dissociation constants. Crystallographic analysis of this complex showed that the designed interfaces mediated the interaction, but the binding mode was inverted by 180°. The inverted binding mode may indicate inaccuracies in the design procedure or could be due to the substitutions accumulated during the affinity-maturation process favoring the inverted mode. Conformational plasticity may also have had a role in reducing the specificity of the designed surface to the target conformation. Indeed, the co-crystal structure reveals that several sidechains adopted conformations that were quite different from the designed ones and that the designed hyperthermophilic protein had increased temperature factors, indicative of conformational flexibility.

Interfacial contacts between backbone atoms provide a way to design polar interactions that are less prone to reconfiguring relative to the design conception. This inference led to the design of a symmetric homodimer mediated by a cross-interface β sheet [48]. The design protocol consisted of computationally scanning the PDB for exposed β strands and then subjecting these proteins to symmetric design and refinement to obtain designed homodimeric proteins that form tight binding interactions. One of these proteins was experimentally determined to dimerize with a dissociation constant, $K_d \sim 1\ \mu M$, and the dimer crystal structure was essentially identical to the original design conception (Fig. 3a). The designed interface is dominated by hydrophobic sidechain contacts and six backbone mediated hydrogen bonds. Three other designed proteins based on the same scaffold protein did not form homodimers in experiment and were noted to have more polar interfaces than the successful design. Taken together, these results highlight the challenges in the modeling and design of polar interaction networks mediated through amino acid sidechains. A comparison of

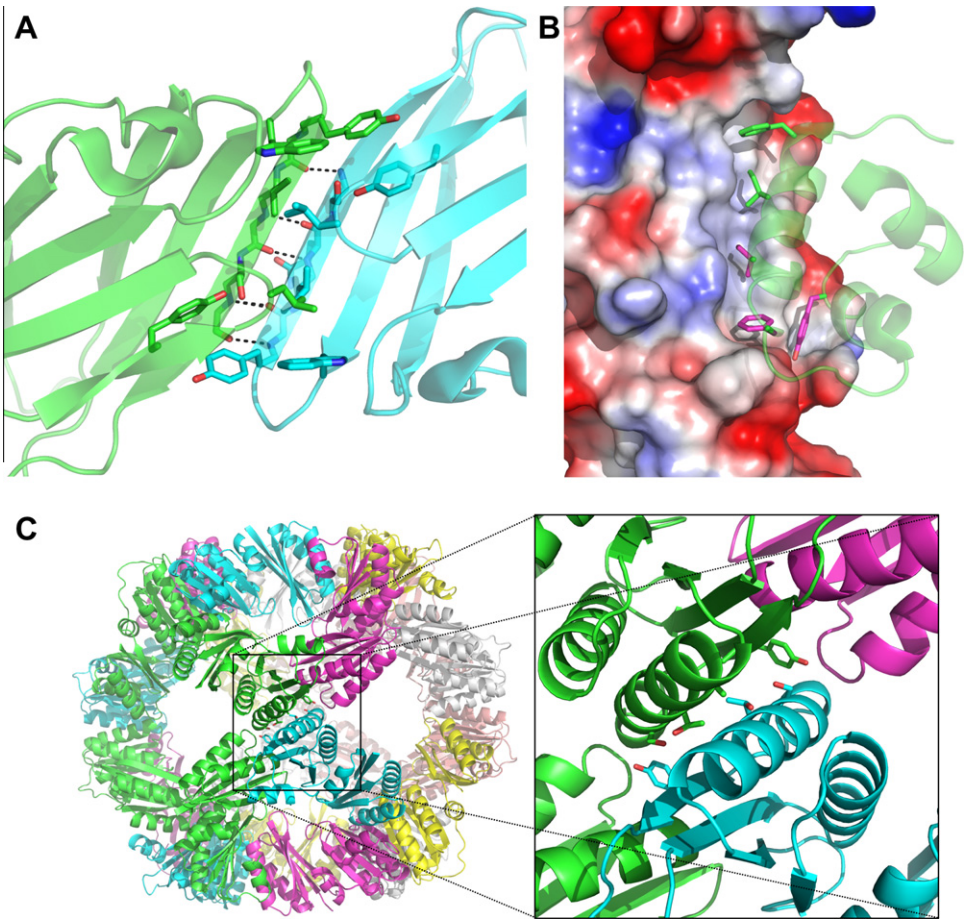

Fig. 3. Crystal structures of computationally designed protein interfaces using three different methods highlight strategies employed in protein design to limit conformational plasticity at the interface. (A) A computationally designed symmetric homodimer incorporates a cross-interface β sheet, which specifies the proteins' orientations [48]. (B) HB80 (green backbone) uses a dense interaction network comprising three residues (pink) to interact with the conserved surface of Group 1 influenza hemagglutinins (surface) [54,55]. (C) A large cage-like structure was designed by forming higher-order oligomers between symmetric homotrimers [57]. In all of these designs, the interfaces are dominated by hydrophobic contacts, and the designed surface comprises rigid secondary-structural elements. Success in designing flexible backbones and polar interaction networks in interfaces has been elusive [56].

failed and successful protein-interface design attempts has also highlighted the difficulty in accurate design of buried polar interaction networks mediated by sidechains [69].

Another approach for increasing the specificity of the designed binding mode utilized metal-mediated interactions, by designing a symmetric interaction incorporating a zinc ion binding site comprising residues on both partners [50]. The protein pair's affinity in the presence of Zn^{2+} was enhanced by 2 orders of magnitude compared to its affinity in the absence of Zn^{2+} and a crystal structure of the Zn^{2+}-bound complex showed that the binding mode was largely as designed. However, one of the designed histidines adopted a conformation that was not consistent with co-ordinating the zinc ion. Replacing this histidine residue with a designed glutamate yielded a design variant that ligated Zn^{2+} as conceived, demonstrating how iterative cycles of design and experimental testing can improve the design methodology. This strategy of incorporating metal-binding sites to induce interactions among non-interacting starting proteins has been extended to generate large and ordered protein arrays [51]. Here, a 3 histidine zinc binding site was designed on a monomeric protein to induce a C2 symmetric dimer. The fourth Zn-ligand was left open, such that aspartate residues from other monomers could form higher-order oligomers. Interestingly, the arrays could be assembled and disassembled with changes in pH and Zn concentration, mimicking the plasticity of biological assemblies.

Many natural protein interfaces show a hotspot region of high-affinity interactions mediated by 2–4 amino acids on either side of the interface, which make crucial contributions to binding affinity [47]. Computational design calculations have suggested that the clustering of these hotspot residues may limit conformational plasticity at the binding surface [52]. To emulate this design feature of natural binding surfaces a two-step procedure was developed: first, a hotspot region comprising clustered, high-affinity interactions between the target protein and amino acid sidechains was computed, and second, scaffold proteins capable of presenting the hotspot while forming high shape complementarity interfaces were identified and further sequence optimized for target binding [53]. This strategy was used to generate two initially low-affinity (K_d > 1 μM) binders of the conserved stem region of influenza hemagglutinin that were then refined through in vitro evolution to K_d ~10 nM [54]. Crystal structures showed that these proteins bound to their target in a mode that was almost identical with the original design conception providing atomic-level validation for this design strategy (Fig. 3b).

Given the high accuracy of this procedure, designs were further analyzed using two complementary approaches to identify methodological deficiencies. The first approach was informed by experimental data. To identify substitutions that can improve binding affinity, a novel combination of deep sequencing and affinity maturation was used, whereby a library encoding each single point

substitution on the two binders was generated and this library (comprising each of the 20 amino acids at each position on the two hemagglutinin binders) was subjected to weak selection for binding affinity to the target to obtain a selected library [55]. The relative proportion of all point substitutions in the original and selected libraries was assessed using deep sequencing, singling out substitutions that were enriched in the selected library on account of increasing the binding affinity. When substitutions that increase binding affinity were modeled on the initial designed structures it emerged that these substitutions could decrease the desolvation penalty upon binding and increase the long-range charge complementarity between the two proteins. Solvation and electrostatics are aspects that are poorly modeled by current methods, and guided by these data a Poisson–Boltzmann equation based electrostatic model was developed to aid interface design. The second approach analyzed the designs which completely failed to bind under the experimental conditions tested and compared them to a set of natural protein–protein complexes. Twenty-eight research groups specializing in structure prediction of protein complexes were asked to independently develop metrics that distinguish these failed designs from natural interfaces [56]. One key finding from this study was that the binding surfaces of a majority of natural protein–protein interfaces are conformationally more stable than those of the failed designs. Taken together, these results reinforce the conclusions from analysis of designed enzymes, that modeling the contributions from polar interactions needs to be improved and that conformational plasticity in designed interfaces is detrimental to their efficiency.

High avidity provides another route to interaction design that leverages small contributions from many interacting surfaces to drive the interaction. In two independent studies, large homo-oligomeric complexes were generated by (1) docking of homotrimeric structures into larger assemblies consistent with a target symmetry, and (2) symmetric design of the resulting interfaces between trimers to optimize their binding affinity [57,58]. In the former study, two proteins, one comprising 12 trimeric subunits and the other 24 trimeric subunits, were found to form the anticipated interaction stoichiometry in solution and crystallography showed very high fidelity to the original design concept (Fig. 3c), and in the latter study, a 3-helix coiled coil was designed to form a crystal lattice. This procedure could be used in the future to custom design drug delivery devices and other nanomaterials for biotechnological applications [59].

4. Emerging themes and future perspectives

Computational design of protein function is a very young field and success rates have so far been low – tens of failed designs for a few successes are the norm. At this early stage, we can learn much about the scope of our understanding of protein energetics and its relationship with molecular function from both the successes and the failures. A recurring theme is that the conformational stability of the active site cannot be taken for granted. Employing naturally occurring rigid backbone structures helps reduce uncertainty in the configuration of the designed site, yet backbone rigidity does not ensure that the sidechains do not reconfigure. Three strategies that have relied on limiting sidechain flexibility during design featured constellations of amino acid residues that were predicted to form a dense network of stabilizing interactions among themselves, as in the design of protein inhibitors [54], amino acids with short sidechains, as in the design of oligomeric structures [57], or focused on designing interactions that utilize backbone atoms, as in the design of homodimers with an interface β-sheet [48] and ester hydrolases featuring backbone groups for the catalytic machinery [36]. In cases where these routes were not taken, some flexibility at the sidechain level caused altered binding, either through minor sidechain rearrangements [30] or large binding mode differences [46]. The challenge in restricting the plasticity of the designed site is due to the fact that the energy gap that separates alternative sidechain conformations or substrate-binding modes is small unless the sidechains are explicitly designed to form uniquely stabilizing interactions. Such small energy gaps are difficult for design calculations to predict accurately [60]. The problem is compounded in the case of small and/or flexible substrates and in designing sidechain-mediated polar interaction networks, including through water molecules, where correct balancing of electrostatic, solvation, and hydrogen bonding is required [49].

A second emerging theme is that design of segments lacking secondary structural elements (loops) for function has so far seen no success (although design of a stable loop has been demonstrated [7]). In all cases mentioned here, where regions containing large loops were subjected to design, the resulting protein either did not interact with its target [56], or bound to it in a very different binding mode. Here again, the likely reason is that small energy gaps separate multiple different backbone configurations, confounding the energy functions on which design relies. There are two complementary routes that design efforts are undertaking to address these challenges: one is to improve the underlying forcefield to encompass a more accurate treatment of electrostatics and solvation [55]. These efforts should help identify the lowest-energy structure for a given sequence in design, and estimate the energy gap between the lowest-energy and alternative structures accurately. A second thrust is to encode large energy gaps by ruling out alternative states. For instance the hotspots used in inhibitor design are predicted to form dense interaction networks among themselves in order to disallow alternative configurations [53]. Another approach is to replace loop structures with larger, well-folded domains that likely will have access to a smaller number of alternative states. Significant improvements in a designed Diels–Alderase resulted from using such an approach [61]. These strategies could be generally beneficial for other enzyme design cases, particularly for those in which binding-site reconfiguration has been associated with the low efficiency of de novo designs. Natural evolution makes much use of the protein backbone degrees of freedom to encode molecular specificity, affinity, and regulation. Therefore, efforts should be directed to understand how backbone conformations that lack secondary structural elements are stabilized in nature and how conformational specificity is encoded in functional sites constructed from these flexible building blocks. Such insights will help account for and eventually programmatically encode subtle but important regulatory effects like protein dynamics and allostery in design.

A final theme is the use of high-throughput experimental characterization to screen and improve computationally designed proteins. Many of the limitations of computational-design methodology, including forcefield inaccuracies and our limited understanding of the relationships between sequence, backbone conformation, and function, have been at least in part overcome by screening dozens of computationally designed proteins using sensitive experimental assays, identifying weakly active hits and subsequently improving their efficiencies using directed evolution techniques. Importantly, activities of computationally designed proteins can be improved to the point that, in the case of hemagglutinin binders, they rival known antibodies in affinity and specificity [54,55], and in the case of designed enzymes, they are comparable to some natural enzymes [33,38]. This combined computational–experimental strategy builds upon the strengths of these complementary methods and promises to make rapid gains in design applications, in the process improving our understanding of crucial aspects of protein conformation and energetics.

Incorporating these insights will help make computational design of novel function robust, routine and powerful.

Acknowledgments

We apologize that not all contributions to protein design could be reviewed here due to space limitations and the review's focus on the computational de novo design of function that has been corroborated by experimentally determined atomic structures. S.J.F. is supported by the Israel Science Foundation, the Human Frontier Science Program, the Marie Curie Reintegration Grant, an Alon fellowship, the Yeda-Sela Center, the Geffen Trust, and a charitable donation from Sam Switzer and family. S.D.K. is supported by a startup grant from Rutgers University.

References

[1] Bowie, J.U., Luthy, R. and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164–170.
[2] Kuhlman, B. and Baker, D. (2000) Native protein sequences are close to optimal for their structures. Proc. Natl. Acad. Sci. USA 97, 10383–10388.
[3] Dahiyat, B.I. and Mayo, S.L. (1996) Protein design automation. Protein Sci. 5, 895–903, http://dx.doi.org/10.1002/pro.5560050511.
[4] Havranek, J.J. and Harbury, P.B. (2003) Automated design of specificity in molecular recognition. Nat. Struct. Biol. 10, 45–52.
[5] Dahiyat, B.I. and Mayo, S.L. (1997) De novo protein design: fully automated sequence selection. Science 278, 82–87.
[6] Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L. and Baker, D. (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302, 1364–1368.
[7] Hu, X., Wang, H., Ke, H. and Kuhlman, B. (2007) High-resolution design of a protein loop. Proc. Natl. Acad. Sci. USA 104, 17668–17673. doi: 0707977104 [pii] 10.1073/pnas.0707977104.
[8] Kapp, G.T., Liu, S., Stein, A., Wong, D.T., Remenyi, A., Yeh, B.J., Fraser, J.S., Taunton, J., Lim, W.A. and Kortemme, T. (2012) Control of protein signaling using a computationally designed GTPase/GEF orthogonal pair. Proc. Natl. Acad. Sci. USA 109, 5277–5282, http://dx.doi.org/10.1073/pnas.1114487109.
[9] Sammond, D.W., Eletr, Z.M., Purbeck, C. and Kuhlman, B. (2010) Computational design of second-site suppressor mutations at protein–protein interfaces. Proteins 78, 1055–1065, http://dx.doi.org/10.1002/prot.22631.
[10] Deckert, K., Budiardjo, S.J., Brunner, L.C., Lovell, S. and Karanicolas, J. (2012) Designing allosteric control into enzymes by chemical rescue of structure. J. Am. Chem. Soc. 134, 10055–10060, http://dx.doi.org/10.1021/ja301409g.
[11] Samish, I., MacDermaid, C.M., Perez-Aguilar, J.M. and Saven, J.G. (2011) Theoretical and computational protein design. Annu. Rev. Phys. Chem. 62, 129–149, http://dx.doi.org/10.1146/annurev-physchem-032210-103509.
[12] Wolfenden, R. and Snider, M.J. (2001) The depth of chemical time and the power of enzymes as catalysts. Acc. Chem. Res. 34, 938–945.
[13] Hellinga, H.W. and Richards, F.M. (1991) Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry. J. Mol. Biol. 222, 763–785. doi: 0022-2836(91)90510-D [pii].
[14] Robertson, D.E., Farid, R.S., Moser, C.C., Urbauer, J.L., Mulholland, S.E., Pidikiti, R., Lear, J.D., Wand, A.J., DeGrado, W.F. and Dutton, P.L. (1994) Design and synthesis of multi-haem proteins. Nature 368, 425–432, http://dx.doi.org/10.1038/368425a0.
[15] Benson, D.E., Wisz, M.S. and Hellinga, H.W. (2000) Rational design of nascent metalloenzymes. Proc. Natl. Acad. Sci. USA 97, 6292–6297.
[16] Kaplan, J. and DeGrado, W.F. (2004) De novo design of catalytic proteins. Proc. Natl. Acad. Sci. USA 101, 11566–11570.
[17] Pinto, A.L., Hellinga, H.W. and Caradonna, J.P. (1997) Construction of a catalytically active iron superoxide dismutase by rational protein design. Proc. Natl. Acad. Sci. USA 94, 5562–5567.
[18] Bolon, D.N. and Mayo, S.L. (2001) Enzyme-like proteins by computational design. Proc. Natl. Acad. Sci. USA 98, 14274–14279.
[19] Wolfenden, R. (1976) Transition state analog inhibitors and enzyme catalysis. Annu. Rev. Biophys. Bioeng. 5, 271–306, http://dx.doi.org/10.1146/annurev.bb.05.060176.001415.
[20] Warshel, A., Sharma, P.K., Kato, M., Xiang, Y., Liu, H. and Olsson, M.H. (2006) Electrostatic basis for enzyme catalysis. Chem. Rev. 106, 3210–3235, http://dx.doi.org/10.1021/cr0503106.
[21] Tantillo, D.J., Chen, J. and Houk, K.N. (1998) Theozymes and compuzymes: theoretical models for biological catalysis. Curr. Opin. Chem. Biol. 2, 743–750.
[22] Schramm, V.L. (2011) Enzymatic transition states, transition-state analogs, dynamics, thermodynamics, and lifetimes. Annu. Rev. Biochem. 80, 703–732, http://dx.doi.org/10.1146/annurev-biochem-061809-100742.
[23] Zanghellini, A., Jiang, L., Wollacott, A.M., Cheng, G., Meiler, J., Althoff, E.A., Rothlisberger, D. and Baker, D. (2006) New algorithms and an in silico benchmark for computational enzyme design. Protein Sci. 15, 2785–2794. doi: 15/12/2785 [pii] 10.1110/ps.062353106.
[24] Lassila, J.K., Privett, H.K., Allen, B.D. and Mayo, S.L. (2006) Combinatorial methods for small-molecule placement in computational enzyme design. Proc. Natl. Acad. Sci. USA 103, 16710–16715, http://dx.doi.org/10.1073/pnas.0607691103.
[25] Richter, F., Leaver-Fay, A., Khare, S.D., Bjelic, S. and Baker, D. (2011) De novo enzyme design using Rosetta3. PLoS ONE 6, e19230, http://dx.doi.org/10.1371/journal.pone.0019230.
[26] Rothlisberger, D., Khersonsky, O., Wollacott, A.M., Jiang, L., DeChancie, J., Betker, J., Gallaher, J.L., Althoff, E.A., Zanghellini, A., Dym, O., et al. (2008) Kemp elimination catalysts by computational enzyme design. Nature 453, 190–195. doi: nature06879 [pii] 10.1038/nature06879.
[27] Jiang, L., Althoff, E.A., Clemente, F.R., Doyle, L., Rothlisberger, D., Zanghellini, A., Gallaher, J.L., Betker, J.L., Tanaka, F., Barbas 3rd, C.F., et al. (2008) De novo computational design of retro-aldol enzymes. Science 319, 1387–1391, http://dx.doi.org/10.1126/science.1152692.
[28] Khersonsky, O., Rothlisberger, D., Wollacott, A.M., Murphy, P., Dym, O., Albeck, S., Kiss, G., Houk, K.N., Baker, D. and Tawfik, D.S. (2011) Optimization of the in-silico-designed kemp eliminase KE70 by computational design and directed evolution. J. Mol. Biol. 407, 391–412. doi: S0022-2836(11)00084-2 [pii] 10.1016/j.jmb.2011.01.041.
[29] Althoff, E.A., Wang, L., Jiang, L., Giger, L., Lassila, J.K., Wang, Z., Smith, M., Hari, S., Kast, P., Herschlag, D., et al. (2012) Robust design and optimization of retroaldol enzymes. Protein Sci. 21, 717–726, http://dx.doi.org/10.1002/pro.2059.
[30] Khersonsky, O., Rothlisberger, D., Dym, O., Albeck, S., Jackson, C.J., Baker, D. and Tawfik, D.S. (2010) Evolutionary optimization of computationally designed enzymes: Kemp eliminases of the KE07 series. J. Mol. Biol. 396, 1025–1042, http://dx.doi.org/10.1016/j.jmb.2009.12.031.
[31] Lassila, J.K., Baker, D. and Herschlag, D. (2010) Origins of catalysis by computationally designed retroaldolase enzymes. Proc. Natl. Acad. Sci. USA 107, 4937–4942, http://dx.doi.org/10.1073/pnas.0913638107.
[32] Wang, L., Althoff, E.A., Bolduc, J., Jiang, L., Moody, J., Lassila, J.K., Giger, L., Hilvert, D., Stoddard, B. and Baker, D. (2011) Structural analyses of covalent enzyme–substrate analog complexes reveal the strengths and limitations of de novo enzyme design. J. Mol. Biol., http://dx.doi.org/10.1016/j.jmb.2011.10.043.
[33] Khersonsky, O., Kiss, G., Rothlisberger, D., Dym, O., Albeck, S., Houk, K.N., Baker, D. and Tawfik, D.S. (2012) Bridging the gaps in design methodologies by evolutionary optimization of the stability and proficiency of designed Kemp eliminase KE59. Proc. Natl. Acad. Sci. USA 109, 10358–10363, http://dx.doi.org/10.1073/pnas.1121063109.
[34] Privett, H.K., Kiss, G., Lee, T.M., Blomberg, R., Chica, R.A., Thomas, L.M., Hilvert, D., Houk, K.N. and Mayo, S.L. (2012) Iterative approach to computational enzyme design. Proc. Natl. Acad. Sci. USA 109, 3790–3795, http://dx.doi.org/10.1073/pnas.1118082108.
[35] Kiss, G., Rothlisberger, D., Baker, D. and Houk, K.N. (2010) Evaluation and ranking of enzyme designs. Protein Sci. 19, 1760–1773, http://dx.doi.org/10.1002/pro.462.
[36] Richter, F., Blomberg, R., Khare, S.D., Kiss, G., Kuzin, A.P., Smith, A.J., Gallaher, J., Pianowski, Z., Helgeson, R.C., Grjasnow, A., et al. (2012) Computational design of catalytic dyads and oxyanion holes for ester hydrolysis. J. Am. Chem. Soc. 134, 16197–16206, http://dx.doi.org/10.1021/ja3037367.
[37] Warshel, A. (1998) Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. J. Biol. Chem. 273, 27035–27038.
[38] Khare, S.D., Kipnis, Y., Greisen Jr., P., Takeuchi, R., Ashani, Y., Goldsmith, M., Song, Y., Gallaher, J.L., Silman, I., Leader, H., et al. (2012) Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis. Nat. Chem. Biol. 8, 294–300, http://dx.doi.org/10.1038/nchembio.777.
[39] Alexandrova, A.N., Rothlisberger, D., Baker, D. and Jorgensen, W.L. (2008) Catalytic mechanism and performance of computationally designed enzymes for Kemp elimination. J. Am. Chem. Soc. 130, 15907–15915, http://dx.doi.org/10.1021/ja804040s.
[40] Frushicheva, M.P., Cao, J., Chu, Z.T. and Warshel, A. (2010) Exploring challenges in rational enzyme design by simulating the catalysis in artificial kemp eliminase. Proc. Natl. Acad. Sci. USA 107, 16869–16874, http://dx.doi.org/10.1073/pnas.1010381107.
[41] Frushicheva, M.P., Cao, J. and Warshel, A. (2011) Challenges and advances in validating enzyme design proposals: the case of kemp eliminase catalysis. Biochemistry 50, 3849–3858, http://dx.doi.org/10.1021/bi200063a.
[42] Baker, D. (2010) An exciting but challenging road ahead for computational enzyme design. Protein Sci. 19, 1817–1819, http://dx.doi.org/10.1002/pro.481.
[43] Huang, P.S., Love, J.J. and Mayo, S.L. (2007) A de novo designed protein protein interface. Protein Sci. 16, 2770–2774. doi: 16/12/2770 [pii] 10.1110/ps.073125207.
[44] Jha, R.K., Leaver-Fay, A., Yin, S., Wu, Y., Butterfoss, G.L., Szyperski, T., Dokholyan, N.V. and Kuhlman, B. (2010) Computational design of a PAK1 binding protein. J. Mol. Biol. 400, 257–270. doi: S0022-2836(10)00472-9 [pii] 10.1016/j.jmb.2010.05.006.
[45] Binz, H.K., Amstutz, P., Kohl, A., Stumpp, M.T., Briand, C., Forrer, P., Grutter, M.G. and Pluckthun, A. (2004) High-affinity binders selected from designed ankyrin repeat protein libraries. Nat. Biotechnol. 22, 575–582.
[46] Karanicolas, J., Corn, J.E., Chen, I., Joachimiak, L.A., Dym, O., Peck, S.H., Albeck, S., Unger, T., Hu, W., Liu, G., et al. (2011) A de novo protein binding pair by computational design and directed evolution. Mol. Cell. 42, 250–260. doi: S1097-2765(11)00208-5 [pii] 10.1016/j.molcel.2011.03.010.

[47] Bogan, A.A. and Thorn, K.S. (1998) Anatomy of hot spots in protein interfaces. J. Mol. Biol. 280, 1–9. doi: S0022-2836(98)91843-5 [pii] 10.1006/jmbi.1998.1843.
[48] Stranges, P.B., Machius, M., Miley, M.J., Tripathy, A. and Kuhlman, B. (2011) Computational design of a symmetric homodimer using beta-strand assembly. Proc. Natl. Acad. Sci. USA 108, 20562–20567, http://dx.doi.org/10.1073/pnas.1115124108.
[49] Rohl, C.A., Strauss, C.E., Misura, K.M. and Baker, D. (2004) Protein structure prediction using Rosetta. Methods Enzymol. 383, 66–93, http://dx.doi.org/10.1016/S0076-6879(04)83004-0.
[50] Der, B.S., Machius, M., Miley, M.J., Mills, J.L., Szyperski, T. and Kuhlman, B. (2012) Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer. J. Am. Chem. Soc. 134, 375–385, http://dx.doi.org/10.1021/ja208015j.
[51] Brodin, J.D., Ambroggio, X.I., Tang, C., Parent, K.N., Baker, T.S. and Tezcan, F.A. (2012) Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays. Nat. Chem. 4, 375–382, http://dx.doi.org/10.1038/nchem.1290.
[52] Fleishman, S.J., Khare, S.D., Koga, N. and Baker, D. (2011) Restricted sidechain plasticity in the structures of native proteins and complexes. Protein Sci. 20, 753–757.
[53] Fleishman, S.J., Corn, J.E., Strauch, E.M., Whitehead, T.A., Karanicolas, J. and Baker, D. (2011) Hotspot-centric de novo design of protein binders. J. Mol. Biol. 413, 1047–1062, http://dx.doi.org/10.1016/j.jmb.2011.09.001.
[54] Fleishman, S.J., Whitehead, T.A., Ekiert, D.C., Dreyfus, C., Corn, J.E., Strauch, E.-M., Wilson, I.A. and Baker, D. (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 332, 816–821.
[55] Whitehead, T.A., Chevalier, A., Song, Y., Dreyfus, C., Fleishman, S.J., De Mattos, C., Myers, C.A., Kamisetty, H., Blair, P., Wilson, I.A., et al. (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat. Biotechnol. 30, 543–548, http://dx.doi.org/10.1038/nbt.2214.
[56] Fleishman, S.J., Whitehead, T.A., Strauch, E.M., Corn, J.E., Qin, S., Zhou, H.X., Mitchell, J.C., Demerdash, O.N., Takeda-Shitaka, M., Terashi, G., et al. (2011) Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J. Mol. Biol. 414, 289–302, http://dx.doi.org/10.1016/j.jmb.2011.09.031.
[57] King, N.P., Sheffler, W., Sawaya, M.R., Vollmar, B.S., Sumida, J.P., Andre, I., Gonen, T., Yeates, T.O. and Baker, D. (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science 336, 1171–1174, http://dx.doi.org/10.1126/science.1219364.
[58] Lanci, C.J., MacDermaid, C.M., Kang, S.G., Acharya, R., North, B., Yang, X., Qiu, X.J., DeGrado, W.F. and Saven, J.G. (2012) Computational design of a protein crystal. Proc. Natl. Acad. Sci. USA 109, 7304–7309, http://dx.doi.org/10.1073/pnas.1112595109.
[59] Lai, Y.T., King, N.P. and Yeates, T.O. (2012) Principles for designing ordered protein assemblies. Trends Cell Biol., http://dx.doi.org/10.1016/j.tcb.2012.08.004.
[60] Fleishman, S.J. and Baker, D. (2012) Role of the biomolecular energy gap in protein design, structure, and evolution. Cell 149, 262–273, http://dx.doi.org/10.1016/j.cell.2012.03.016.
[61] Eiben, C.B., Siegel, J.B., Bale, J.B., Cooper, S., Khatib, F., Shen, B.W., Players, F., Stoddard, B.L., Popovic, Z. and Baker, D. (2012) Increased Diels–Alderase activity through backbone remodeling guided by Foldit players. Nat. Biotechnol. 30, 190–192, http://dx.doi.org/10.1038/nbt.2109.
[62] Kortemme, T., Morozov, A.V. and Baker, D. (2003) An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein–protein complexes. J. Mol. Biol. 326, 1239–1259.
[63] Lazaridis, T. and Karplus, M. (1999) Effective energy function for proteins in solution. Proteins 35, 133–152. doi: 10.1002/(SICI)1097-0134(19990501)35:2<133::AID-PROT1>3.0.CO;2-N [pii].
[64] Dunbrack Jr., R.L. and Karplus, M. (1994) Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nat. Struct. Biol. 1, 334–340.
[65] Munoz, V. and Serrano, L. (1994) Intrinsic secondary structure propensities of the amino acids, using statistical phi–psi matrices: comparison with experimental scales. Proteins 20, 301–311, http://dx.doi.org/10.1002/prot.340200403.
[66] Miyazawa, S. and Jernigan, R.L. (1993) A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. Method Enzymol. 6, 267–278.
[67] Song, Y., Tyka, M., Leaver-Fay, A., Thompson, J. and Baker, D. (2011) Structure-guided forcefield optimization. Proteins 79, 1898–1909, http://dx.doi.org/10.1002/prot.23013.
[68] DeLano WL (2002) The PyMol Molecular Graphics Systems. DeLano Scientific, Palo Alto, CA, USA.
[69] Stranges, P.B. and Kuhlman, B. (2013) A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. Protein Sci. 22, 74–82.