# 7 Computational protein design and discovery

**Sheldon Park, Xiaoran Fu Stowell, Wei Wang, Xi Yang and Jeffery G. Saven***

*Makineni Theoretical Laboratories, Department of Chemistry, University of Pennsylvania, 231 South 34th Street, Philadelphia, Pennsylvania, 19104, USA. E-mail: saven@sas.upenn.edu*

Protein design has traditionally relied on an expert's ability to assimilate a myriad of factors that together influence the stability and uniqueness of a protein structure. As many of these forces are subtle and their simultaneous optimization is a problem of great complexity, sophisticated sequence prediction algorithms have been developed to aid in the design of novel proteins by providing quantitative analysis of the sequence–structure relationship. This review discusses some of the major developments in computational protein design, focusing on common inputs to the calculation and several often used search methods. We also highlight accomplishments in computational protein design, ranging from simple core redesign of an existing protein, to the design of new functionalities (catalytic or ligand-binding), and finally to a large-scale design of *de novo* proteins.

## Introduction

The genome of an organism contains the complete biochemical blueprint for its makeup and to a large extent determines the type of chemical processes that take place inside the organism. The availability of fully sequenced genomes, therefore, is an important step towards elucidating the fundamental chemical and biomolecular events that are the underpinnings of life. Unfortunately, the information encoded in the genome defies easy interpretation, as it is proteins not nucleic acids that perform most of critical cellular functions. Proteins are evolutionarily engineered nanomachines that carry out catalytic, structural and signaling functions essential to life. Gaining a predictive molecular understanding of their functions is a goal pursued by many, including molecular biologists, chemists, engineers and material scientists. Because the function of a protein may be largely dictated by its structure, structure determination is critical for fully understanding its function. While the primary sequence of a protein is easily determined from the nucleic acid sequence, it is notoriously difficult to accurately predict the three dimensional structure of a protein based on its sequence alone. While great advances are being made in the field of structure prediction, the current inability to reliably predict the three-dimensional structure of a protein based on its sequence implies that the structure of individual

proteins must still be determined experimentally, which limits structural elucidation only to those proteins that can be crystallized or are soluble enough for NMR studies.

A related yet different challenge that has seen many breakthroughs in recent years is protein design, which is the topic of the present review. Whereas structure prediction starts with the primary sequence and tries to deduce the corresponding three-dimensional structure, protein design starts with a target structure and searches for protein sequences that are compatible with the fold. This ''inverse protein folding''[1] problem can indirectly test many of the same theoretical and computational inputs used in studying protein folding, including the physical potential used to evaluate the sequence–structure compatibility. Predictions from a protein design study are tested experimentally by synthesizing the proposed sequences and verifying that they adopt the target fold. Protein design teaches us much about the molecular nature of interactions between amino acids by testing the assumptions and techniques used during the prediction–synthesis–analysis cycle. The successful design of a new protein can help us critically evaluate the quality of algorithms used and hone our skills to design novel proteins with atomic precision. The insights learned from these efforts will help elucidate how simple protein modules come together to create complex protein assemblies.

At the same time, protein design is considered an easier task than protein folding because of the redundancy in solution. Since more than one protein sequence can fold to a given structure, the odds of finding a sequence compatible with the structure is thought to be accordingly higher. This degeneracy, however, also complicates protein design since sequences folding to very similar structures may be broadly distributed in sequence space. As a result, proteins having essentially the same structure may have little or no sequence similarity. In addition, part of what makes protein structure prediction difficult is the covalent connectivity of the peptide backbone. As the protein folds, distant residues are brought into proximity to form hydrogen bonds or to participate in hydrophobic interactions. However, the covalent connectivity of the backbone requires that residues close in sequence also become close in physical space, and this constraint may lead to a situation where not all noncovalent interactions are simultaneously satisfied, an effect referred to as 'frustration'.[2] However, protein design has inherently greater latitude than protein folding, since such noncovalent frustration may often be resolved by changing the sequence which is at the discretion of the protein scientist.

The practical aspect of protein engineering is an important driver for the discipline. Novel protein molecules may have therapeutic potentials that are absent in natural proteins or satisfy demands unmet by wild type proteins. New proteins may be engineered for stability, expression, and function. A stable therapeutic molecule may be overexpressed in large quantities in the laboratory, whereas wild type protein may resist such overproduction due to undesirable chemical properties, *e.g.*, thermal instability or tendency to aggregate. Targeted expression of an engineered protein may reverse the harmful effects of a mutated protein—a notion central to gene therapy. Some proteins in nature are remarkable catalysts, performing difficult chemical reactions with ease, *e.g.*, stereospecific bond formations during cholesterol biosynthesis.[3] It is hoped that through artificial molecular evolution and rational design, novel protein catalysts may be discovered to help with chemical reactions that are currently difficult to achieve in the laboratory. There are also efforts to design

novel protein-based molecular sensors[4,5] and biomaterials to serve as a foundation for the next-generation electronics.[6] The field of protein design and protein engineering bridges science and engineering with potential breakthroughs for both.

This review surveys the current state of quantitative protein design. We start with an overview of some of the major accomplishments in heuristic protein design to place the recent interest in computational protein design in perspective. We summarize some of the key discoveries made regarding protein structures, which have critically contributed to the success of heuristic protein design as well as to the later development of sequence prediction algorithms. Even as the field of computational protein design is gaining momentum, many of these simple rules continue to play an important role in the planning, execution and analysis of all protein design projects. Next, we describe the inputs to computational design including the force field and rotamer model of side chains. Computational techniques frequently used in protein design are discussed, which are the workhorses of quantitative protein engineering. This is followed by success stories in quantitative protein design, where designed proteins have been experimentally tested against prediction. We conclude with potential applications of some of the same tools used in computational design to the design of non-biological folding polymers.[7]

## Knowledge-based protein design

Early works in protein design sought to apply the intuitive knowledge gleaned from biochemical experiments and structural databases to the construction of small protein motifs or modules. While many noncovalent interactions are important to protein folding, the burial of hydrophobic residues is thought to be one of the most important driving forces.[8] As such, the binary patterning of a protein sequence designed to hide hydrophobic side chains in the protein core and expose hydrophilic side chains on the protein surface (''hydrophobic in, polar out'') is a heuristic rule often followed during protein design. Using a knowledge-based top-down approach to protein design, researchers have built both simple α-helices and more complex α-helical bundles, sometimes introducing novel functionalities into the fold as well. The most impressive are those designs where the majority of the residues in a sequence have been engineered *en masse*, thus critically pushing the boundaries of our knowledge. Due to the largely cooperative nature of protein folding, it is unlikely that one would succeed in designing a large stably folding sequence unless the designing principles are at least qualitatively accurate. These empirical studies in turn have further fine-tuned our understanding of the sequence–structure relationship, and have laid the foundation for the more ambitious quantitative protein design that is discussed in later sections.

Most natural proteins have abundant secondary structures (helices, sheets and ordered loops), whose mutual interactions determine the tertiary structure of the protein. As secondary structures are smaller and ostensibly easier to design than the whole protein, a reductionist approach to protein design may suggest that we first identify a set of sequences that are individually expected to form distinct structural motifs, and then stitch them together to make a whole protein. Also referred to as hierarchical protein design, this modular approach to protein design was instrumental in achieving early successes in *de novo* protein engineering.[11]

Secondary structure design usually proceeds using statistical information inferred from known protein structures. For example, Eisenberg *et al.* applied simple heuristic rules regarding α-helix formation to design minimal 12 and 16 residue amphiphilic α-helices composed of Leu, Glu and Lys,[9] of which the 12 residue peptide was later shown to self-associate to form both tetramers and hexamers.[9,10] While oligomerization was expected given that the helix had separate hydrophobic and hydrophilic surfaces, the structural heterogeneity of the oligomer had not been predicted. To force the formation of a unique structure, four identical helical sequences were strung together with interhelical loops, resulting in a single tetrameric structure with a hydrophobic core.[11] A shorter peptide with two of the same helical elements connected together (helix–turn–helix) similarly formed a tetrameric helical bundle through antiparallel dimeric association.[12] Despite their stability, these early designed proteins and another peptide where some of the Leu side chains have been replaced with Ile exhibited much mobility in the interior, readily undergoing a thermal transition to a molten globule-like state.[13]

The α-helical coiled-coil is commonly observed in natural proteins, including DNA binding proteins. In order to study the formation of native-like helix bundles, a 29-residue peptide coil-Ser was designed and characterized.[14] Although the peptide has a similar seven residue repeat pattern as observed in GCN4, which forms a parallel coiled-coil,[15,16] the peptide instead formed an unexpected antiparallel coiled-coil trimer with the helices running up–up–down. Both attractive and repulsive interhelical electrostatic interactions are observed in this arrangement, suggesting that electrostatic interactions appear to play a minor role in determining the topology. On the other hand, the antiparallel orientation of the third helix does promote a mutually favorable arrangement of induced helix macrodipoles. More importantly, the stoichiometry seems to be influenced by the Leu residue at the **a** and **d** positions in the heptad repeat, since they adopt more favorable conformations in the observed structure than they would in a modeled parallel trimer. This view was later validated by the discovery of a coil-Ser variant with Val at the **a** position that forms a parallel three-helix bundle,[17] as well as GCN4 mutants that form trimers and tetramers.[18]

To estimate the relative importance of the individual amino acids in specifying the protein fold, Creamer and Rose put forth the "Paracelsus challenge": design two proteins that share 50% or greater sequence identity yet have different protein folds.[19] Responding to the challenge, the Regan group successfully transformed the B1 domain of Streptococcal IgG-binding protein G, which is a predominantly β-sheet protein, to a four-helix bundle protein by mutating 50% of the wild type amino acids.[20] To achieve this feat, they observed that the B1 domain contains both residues that have high α-helix forming propensities and those that have high β-sheet forming propensities. Hence, the β-sheet forming residues were selectively mutated to those that promote α-helix formation. The selection of residues to be replaced was made based on the expected α-helix hydrophobicity pattern. The resulting sequence, dubbed Janus, satisfied the intra-monomer salt bridge and surface charge distribution required to form a Rop-like struture and was shown to indeed form a four-helix bundle as expected. They also tested whether other variants of Janus with even higher sequence identity to the B1 domain can be designed. The mutants Janus-55, Janus-61, Janus-66 and Janus-86 respectively have 55, 61, 66 and 86% sequence identity with the

B1 domain. While Janus-55 and Janus-61 both maintained helical folds (although with decreased stability), Janus-66 formed aggregates of β-sheet and Janus-86 had a β-sheet fold similar to the B1 domain. The success in meeting the Paracelsus challenge illustrated that the stability and fold of a protein may be modulated through careful manipulation of a limited number of key amino acids.

A *de novo* protein with a custom-made function was designed by Schafmeister *et al.*, who constructed a novel amphiphilic α-helix composed of just five amino acid types that solubilized membrane proteins by shielding their exposed hydrophobic transmembrane domains.[21] This 24-residue minimal helix called "peptitergent" has a flat surface made of Ala's and Leu's for hydrophobic interaction on one side, and a hydrophilic surface consisting mostly of Gln on the other. When mixed with proteins containing transmembrane domains, *e.g.*, bacteriorhodopsin and rhodopsin, the peptide helped the proteins remain in solution over two days. The crystal structure of the peptide shows that the peptide forms a monomeric four-helix bundle with well-folded structure.[22] The design of a *de novo* protein to address a practical problem was also reported by Kim and coworkers, whose therapeutic five-helix protein retarded the HIV-1 infection of human T cells.[23] To that end, they noted that viral entry requires membrane fusion mediated by the gp41 envelope protein. The fusion active state of the gp41 ectodomain consists of a six-helix bundle comprised of a coiled-coil trimer of heterodimers.[24] Each heterodimer consists of the N-terminal and C-terminal helices of a polypeptide which folds on itself. Based on this information, they designed a so-called "5-helix" that binds to the C-terminal region of one of the heterodimers through an exposed hydrophobic cleft and competes with its association with the amino-terminus. The designed protein was soluble, maintained requisite helicity, and showed efficacy against HIV-1 in a culture study. Therefore, simple design ideas combined with an understanding of how helices are formed can lead to *de novo* proteins with useful functions.

As metal ions can play important roles both in catalysis and structure stabilization, the introduction of metal-binding activity to an existing protein has been the subject of many investigations (see ref. 25 for review). As a result, novel proteins with affinity towards zinc,[26] iron,[27] calcium,[28] copper,[29] cadmium[30] and mercury[31] have been engineered on templates without intrinsic affinity towards the metal. The protein scaffolds used in these studies included both designed helix bundles[32] and natural proteins such as protein G.[33] When used in a designed protein, metal ions can impose strong constraints on the coordinating residues, thus helping to organize the protein fold[34] and increase the native-like character of the core.[35] Some of the engineered metal-binding proteins have shown catalytic activity, performing such reactions as the hydrolysis of plasmid DNA.[36] Emulating nature in its use of metal-containing cofactors, several investigators have successfully designed *de novo* heme-binding proteins.[37] The heme cofactor appears in a variety of proteins, *e.g.*, myoglobin, hemoglobin, catalase, peroxidase, cytochrome P450, and participates in electron transfer and charge separation in respiration and photosynthesis. Using a binary patterned library, Moffet *et al.* isolated several heme binding proteins with peroxidase activity that was only a few fold less than the natural enzyme horseradish peroxidase.[38] Rau *et al.* reported the design of a heme binding four helix bundle with a ruthenium tris(bipyridine) complex covalently attached to the exterior hydrophilic surface. The complex exhibited laser-induced long-range electron transfer,[39,40] raising

a hope that novel proteins may be engineered to work as electron transfer agents. In an encouraging discovery, a 16 amino acid designed peptide was shown to efficiently incorporate an iron–sulfur cluster as a tetramer and exhibit redox properties typical of natural bacterial ferredoxins.[41]

The combinatorial library approach to protein design allows rapid examination of a large number of sequences by generating and screening libraries of targeted mutants. Usually, the library of mutants is constructed by using degenerate oligonucleotides during gene assembly,[42] by performing the polymerase chain reaction (PCR) under mutagenic conditions,[43] or by using DNA shuffling.[44,45] The library is transformed into a cell where desired mutants are identified based on biological screening. Studying the factors that affect protein stability and folding by combinatorial mutagenesis, Lim and Sauer constructed a library of lambda repressor mutants whose core residues were randomly mutated to other hydrophobic residues.[46] The high percentage of functional mutants in the library showed that there are many ways to repack the core and supported the hypothesis that hydrophobicity alone is the key determinant of whether a mutant core sequence is compatible with the wild type fold. In another combinatorial study designed to measure the cumulative effects of mutations, Gregoret and Sauer assembled a set of 2048 mutants containing either the wild type residue or Ala at eleven positions in the lambda repressor.[47] The group observed that roughly 25% of these mutants, many of which contained multiple mutations, were functional. By comparing the frequencies of pairwise mutations to those of single mutations, they concluded that the effects of multiple substitutions are largely additive but there are also residue pairs that are distant in the three-dimensional structure yet display statistically significant nonadditive effects. When they expressed a random library of 80- to 100-residue proteins mainly composed of Gln, Leu, and Arg in *E. coli* to study whether fine-tuning of sequence is required to achieve a stable three-dimensional structure, they noted that 5% of the mutants were readily expressed in soluble form.[48] Furthermore, three mutants that were examined biophysically had significant α-helical content and resisted proteolytic degradation, while one mutant exhibited highly cooperative unfolding.[49] Therefore, these early studies with a binary patterned peptide library suggested that relatively little sequence information is required to adopt a folded structure, and a significant fraction of sequences may be capable of folding to a stable structure.

Probing the importance of binary patterning in attaining a native-like protein structure, Kamtekar *et al.* assembled a degenerate library using an expanded binary code containing a total of eleven residues (Val, Ile, Met, Leu and Phe for the 24 hydrophobic positions; Asn, Asp, His, Gln, Glu, and Lys for the 32 hydrophilic positions).[50] Their library was patterned according the expected distribution of hydrophobic and hydrophilic side chains in a target four-helix bundle, *i.e.*, hydrophobic amino acids at interior buried positions and hydrophilic amino acids at solvent exposed positions. If binary patterning is the main determinant for protein folding, as suggested by previous experiments,[46,48] then a large percentage of the sequences in the library should fold to compact native-like structures. Of the $4.7 \times 10^{41}$ possible amino acid sequences, 48 were randomly sampled and expressed in *E. coli*. Roughly ~60% of the tested folded to proteins that were soluble and resistant to intracellular degradation.[50] Further analysis of three of the designed proteins by circular dichroism (CD), chemical denaturation, and size-exclusion chromatography

showed they were monomeric four-helical bundles. However, stability does not necessarily imply a unique structure. To examine whether a well-folded interior can be designed by a binary code strategy, Wei *et al.* lengthened the helices and constructed a new library patterned to fold into a 102-residue four-helix bundle.[51] When five proteins were randomly selected and characterized by NMR, all but one showed cross-peaks characteristic of tertiary interactions and well-ordered structures. Furthermore, the solution structure of one of them showed an antiparallel four-helix bundle with well-ordered side chains, demonstrating that a native-like structure can be designed using a binary patterned library.[52]

On the other hand, when a combinatorial library patterned with alternating hydrophobic and hydrophilic side chains was constructed to study novel β-sheets, the resulting proteins were found to self-assemble into large oligomers, such as amyloid-like fibrils.[53] A search through a database of 250,514 protein sequences revealed that alternating patterns of polar and nonpolar amino acids occur less often than other patterns with similar compositions.[54] Together, these results suggest that sequences of alternating hydrophobicity are inherently amyloidogenic and may have been disfavored by evolutionary selection. These observations also highlight the difficulty of designing novel β-sheets. Whereas most of the hydrogen bonds within a helix are satisfied locally, the hydrogen bonds required to stabilize a β-sheet are formed between residues distant in sequence, making the design process inherently more global. The tendency of some β-sheets to form aggregates and to precipitate out of solution is a consequence of the fact that a β-strand can interact with its neighbors both through backbone hydrogen bonds as well as through hydrophobic and hydrophilic interactions between side chains.[55]

Novel proteins with engineered properties and functionalities may be discovered by screening a large, randomly generated peptide library. Library-based protein engineering, in combination with molecular evolution or directed evolution, uses the iterative mutation–selection–enrichment cycle to engineer new protein molecules. The use of sequence libraries and directed evolution in protein engineering has been spurred by the availability of various library platforms and concurrent development of high throughput assays. Phage display libraries are a popular platform for protein engineering[56] but others have also used bacterial,[57] yeast,[58] and ribosomal display.[59] In a molecular evolution study, Braisted and Wells tested whether the third helix of the IgG binding domain of protein A, which does not contact IgG directly, may be removed without affecting the binding affinity.[60] They constructed and screened libraries of the first two helices to stabilize the truncated domain while maintaining high binding affinity ($K_d \sim 20$ nM). The stability of a mesophilic esterase was improved in a 96-well plate based parallel assay, resulting in a mutant with higher $T_m$ by 14 °C.[61] Molecular evolution was also used to evolve an RNA polymerase from a DNA polymerase;[62] to endow an antibody with catalytic activity;[63–65] and to speed up the maturation of a red fluorescent protein.[66] In a highly sensitive functional assay, Hilvert and co-workers converted dimeric *E. coli* chorismate mutase to a monomeric four-helix-bundle protein with near native activity (Fig. 1).[67] Their study also demonstrated that when residues in an interhelical turn are involved in long-range tertiary interactions, the fraction of acceptable turn sequences is substantially lower ($<0.05\%$) than previous studies on other four-helix-bundle proteins had suggested.[68] Despite these well-known successes, protein engineering by molecular evolution often

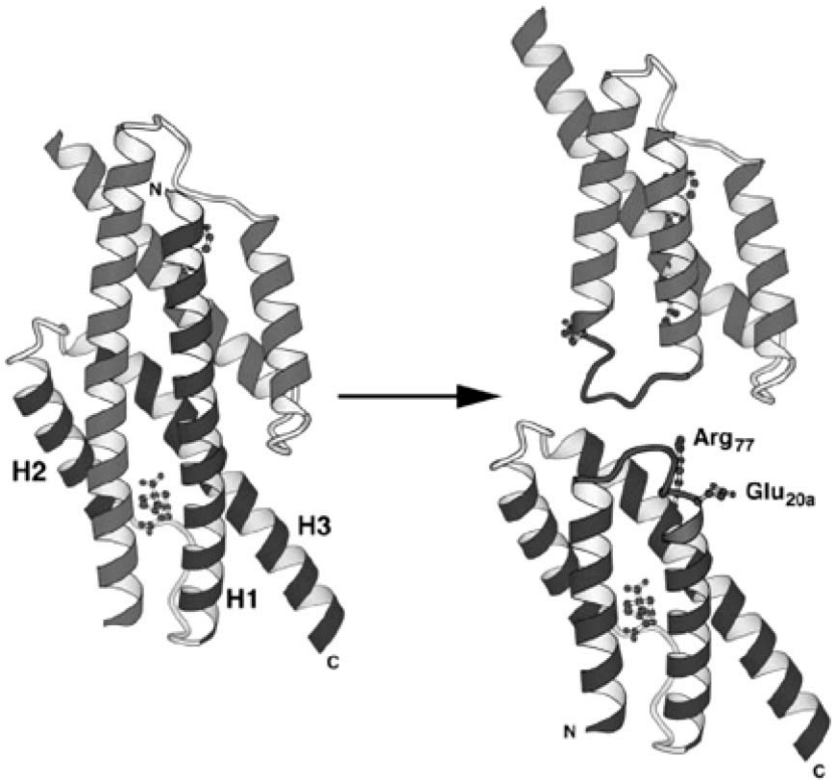


**Fig. 1** Redesigning the topology of *E. coli* chorismate mutase from an obligate dimer to a monomer through molecular evolution. (Re-printed with permission from ref. 67.)

does not explicitly incorporate knowledge of protein structure. Nor does the successful engineering of a novel protein necessarily help us design other interesting proteins in a predictable way. On the other hand, a large-scale library screening may be an efficient way to validate the predictions of rational protein design.

## Challenges in *de novo* protein design

While heuristic protein design has been successful in the discovery of numerous interesting molecules, protein designers working on *de novo* protein engineering continue to face difficult challenges. For one, many of the previously designed proteins including a well-characterized helix bundle[69] and a β sheet protein[70] fall short of behaving truly like natural proteins.[71] Their thermal denaturation is not entirely cooperative and often requires the presence of chaotropic chemicals. For example, in the case of the Richardson β-sheet protein, despite its sharp $^{1}$H NMR peaks, the protein binds hydrophobic dyes such as ANS to a degree expected for a poorly folded protein.[70] The difficulty in designing well-defined tertiary structures is

in part due to imperfect consideration of the physical and chemical interactions important for stabilizing particular structures. Furthermore, while numerous factors contribute to the design of a native-like protein, hierarchical, knowledge-based protein design is unable to capture all the subtleties of protein folding in an atomically detailed and systematic way that simultaneously optimizes the stability and uniqueness of a designed structure.

Unable to consider all the degrees of freedom available to a system, hierarchical design usually models only the most salient features such as hydrophobicity and optimizes design through an iterative cycle of construction, analysis and improvement. This sort of 'divide and conquer' approach disregards the fact that the local backbone structure is often contingent on its context within the larger tertiary structure of the protein, and thus it is preferable to consider the properties of the sequences as a whole. Nevertheless, a hierarchical approach is necessitated by the huge number of sequence degrees of freedom. Unfortunately, the exponentially large number of possible sequences (*e.g.*, more than $10^{130}$ sequences for a 100-residue protein) impedes direct rational sequence design. Furthermore, since the protein is not considered globally in the context of the target fold, such an approach may obscure the underlying physical principles of structural organization.

The information required for a polypeptide to fold to a unique three-dimensional structure is encoded in its primary structure, *i.e.*, its amino acid sequence. In addition to being energetically consistent with the native structure, the sequence must also be incompatible with other alternative structures. For example, if a sequence that folds to a monomeric β-sheet protein also folds to an amyloid fiber, its natural function would be severely compromised. Such ambiguity may have been eliminated during evolution to ensure structural uniqueness. In general, one would believe that two similar sequences in nature would not fold to two very different structures. However, as the works of Regan and coworkers have shown, a large degree of sequence similarity alone does not guarantee structural similarity.[20,72] Recently, a double mutant Arc protein ('switch Arc') was discovered that further demonstrates the subtle sequence–structure relationship. In 'switch Arc,' hydrophilic Asn11 and hydrophobic Leu12 have been switched with each other, with the concomitant change in the binary patterning from one that is consistent with a β-strand to one that promotes α-helix.[73] As a result, the region surrounding the residues 11 and 12 undergoes a transition from a β-strand to an α-helix. Surprisingly, the resulting mutant can still homodimerize and bind DNA. A single mutant Arc-N11L exhibits an even more ambiguous structural identity, as the region near residue 11 exists both as a β-strand and an α-helix at room temperature.[74] The relative distribution between the two structures depends on the ambient temperature. Thus, this mutant can be considered an evolutionary intermediate between the wild type protein and 'switch Arc.'

The sequence–structure relationship is highly context-dependent, making both protein folding and protein design subject to error. While the secondary structure propensities of amino acids are important determinants of protein folding, the local conformational preferences themselves can be influenced through tertiary interactions. The context-dependent secondary structure formation is shown by a variant of B1 domain of protein G, where an identical stretch of 11 amino acids ('chameleon sequence') folds to an α-helix when placed in one position but to a β-sheet when moved to a different location.[75] So far, all the subtle rules of sequence–structure

relationships have not been captured in a set of simple rules, making it difficult to consistently incorporate them during protein design. These difficulties have motivated the search for new ways of designing novel proteins that is founded on detailed quantitative analyses. A quantitative approach to protein design would help ensure that what we know heuristically about protein design is systematically incorporated in every design project. The development of algorithms to quantitatively evaluate the quality of a proposed design would also help identify gaps in our knowledge and point to ways to improve them. The remainder of the review discusses computation-driven protein design. A theoretical overview is presented first, followed by discussions of major experimental accomplishments.

## Computational approach to protein design

Heuristic protein design relies on human intuition to optimize the sequence–structure relationship, and its success depends critically on the designer's ability to assimilate numerous pieces of information consistently and coherently. Unfortunately, many of the factors that contribute to protein folding are subtle and cannot be easily visualized or addressed without detailed consideration of sequence and structure. Noncovalent interactions, *e.g.*, van der Waals forces, hydrogen bonds, and electrostatic interactions, are some of the most difficult quantities to estimate accurately, as the strength of each of these depends critically on the distance of separation and geometry of interaction in the presence of solvent. In addition, one must also consider a large number of amino acid sequences during protein design. The combinatorial possibilities for all but the simplest design project, therefore, far exceed what a person can meaningfully inspect by hand and evaluate. As a result, computational algorithms that can rapidly screen or characterize a large search space are needed in order to guide protein design with atomic resolution. Powerful optimization methods can efficiently search through or screen an astronomically large number of unique sequences *in silico* before any one of them is actually designed and tested in the laboratory. Potentially, the lessons learned from computational protein design can be applied to design other macromolecules. Since the same set of physical forces that govern the behavior of polypeptides also govern other small and large biomolecules, models based on physical and chemical interactions may be developed to assist the design of proteins and other folding heteropolymers. This offers a distinct advantage when experimental data to guide the heuristic design of such novel molecules is lacking. Common inputs to computational protein design are discussed below.

### Inputs to calculation

**Target structure.** The target protein structure is often obtained from an existing high-resolution structure, although it can also be modeled *de novo*. It may also be a fold based on an existing structure with additional design requirements modeled in afterwards, as for example, when optimizing a turn in a protein.[76] While fixing the backbone geometry greatly reduces the computational complexity by decreasing the total degrees of freedom, it also prevents the mainchain from making adjustments to

accommodate sequence variations. Several authors, hence, have introduced backbone flexibility in their design by exploring alternative conformations during the search by examining closely related structures.[77–79] The target structure is specified at atomic resolution unlike in heuristic protein design where the overall topology of the protein alone is the primary concern. The selection of a backbone from a known protein structure guarantees that the target protein is in fact designable. Designability is an important feature since not all possible protein folds may have sequences that fold uniquely to them.[80] At the same time, the choice of an existing structure as a design template may still permit a wide range of biological functionalities. Nearly identical structures have been used by proteins in nature that share no sequence or functional homology. The TIM barrel topology, for example, is found in 21 unrelated protein superfamilies,[81] suggesting that other novel functionalities may be successfully introduced to existing folds. Recent theoretical studies by the Shakhnovich group suggest that designability is correlated with the so-called contact trace, which is a measure of the fold's tertiary topology.[82,83] As of yet, most protein design usually proceeds from a fold found in nature.

**Residue degrees of freedom.** The total number of degrees of freedom per residue for a given target tertiary structure is specified by the amino acids and by the amino acid side chain conformations permitted at each position in the sequence. Both qualitative and quantitative methods of reducing the allowed degrees of freedom have been investigated and are commonly used during protein design.

*Amino acids.* The amino acid degrees of freedom refer to the number of different amino acid states allowed at each randomized position. The state of an amino acid is determined by both its identity and by its side chain conformation. The side chain conformations, rotamer states, are usually those inferred from a structural database and are usually consistent with the bond and torsional angles present in a molecular potential.[84] The simplest amino acids (Ala and Gly) are considered to have just one rotamer state, whereas amino acids with larger side chains may have as many as 80–100 different rotamer states. Typically there are on the order of 100's of such rotamer states, when summed over the 20 amino acids. While the inclusion of all 20 amino acids allows the entire sequence space to be searched, studies have suggested a variety of ways of reducing the amino acid degrees of freedom without compromising the quality of design. This can greatly influence the success or failure of a project since one must potentially optimize $m^N$ degrees of freedom ($m$ is the number of side chain conformations per residue, $N$ is the number of residues). When designing a medium-size protein of 100 residues with all 20 amino acids, this number is far greater than can be achieved either computationally or experimentally. A straightforward approach is to reduce the number of states per residue in a site specific manner. The number of possible sequence states $\prod_i m_i$, then, may be significantly smaller than $m^N$. Experiments have shown that not all twenty amino acid types are required to construct a functional protein.[85] For example, a de novo 108 residue protein composed of seven residue types can fold to a native-like four-helix bundle;[22] a five-letter amino acid alphabet can reconstruct 95% of an SH3 domain;[86] and a functional enzyme can be constructed using a total of 13 amino acid types.[87] Therefore, a reduction in the

amino acid alphabet can significantly speed up calculation without limiting the design scope. Most significantly, targeting amino acid variability in hydrophobic patterning can dramatically reduce the sequence search space and simultaneously drive the formation of desired secondary structures. In a validation of the use of a reduced alphabet during protein design, Marshall and Mayo automated the selection of amino acids based on the expected local hydrophobicity, and successfully designed a monomeric and well-folded variant of engrailed homeodomain using a limited set of side chains.[88] As protein design methods continue to improve, one may even consider monomers other than the naturally occurring amino acids. Though an expanded monomer set may increase complexity of the problem, nonnatural amino acids can potentially allow a greater range of functionalities in the designed protein.[89,90]

*Rotamer library.* During atomically detailed design, side chains are free to explore different conformations. While the side chain dihedral angles ($\chi$) may take on values from $-180$ to $+180^{\circ}$, in naturally occurring proteins, they usually adopt discrete staggered dihedral angles near the torsional energy minima. A side chain conformation corresponding to a local minimum energy is referred to as rotamer. Such rotamer states may be determined *via* the minimization of a molecular potential, or more often *via* analysis of the side chain conformations observed in high-resolution protein structures. As the bond angles and bond lengths in a side chain are usually well determined, a set of dihedral angles is sufficient to uniquely describe each rotamer. The rotamers for all amino acids together make up a rotamer library (reviewed in ref. 84). The use of a rotamer library significantly reduces the complexity of calculation by discretizing the search space. Although this discretization represents only an approximate representation of the full conformational flexibility of the side chains, it is a useful approximation that is rooted in statistical observation. There are a variety of rotamer libraries available for protein modeling and protein design, including backbone-independent, secondary-structure-dependent or backbone-dependent libraries.[91–103] Backbone-independent rotamer libraries are calculated from all available side chains of each amino acid regardless of the local context. Secondary-structure-dependent libraries provide separate sets of side chain dihedral angles for α-helix, β-sheet or coil secondary structures, while the side chain conformations of backbone-dependent rotamer libraries present them as functions of the local backbone conformation, *i.e.*, the backbone dihedral angles $\phi$ and $\psi$. The choice of a rotamer library constitutes an important part of computational protein design, since it can affect the calculation both qualitatively and quantitatively.[104]

Rotamer libraries are usually constructed from a statistical analysis of the side chain conformations in known protein structures by clustering observed conformations, or by dividing dihedral angles into bins and associating an average conformation in each bin.[84] Although the rotamers in a rotamer library usually correspond to local minima of a potential energy, broad ranges of sidechain dihedral angles are also observed for some side chains such as Asn and Gln. As different libraries are constructed using different statistical methods, fluctuations about a particular rotamer state have also been considered. Ponder and Richards provided means and standard deviations for their backbone-independent rotamer library.[96]

Dunbrack and Cohen used a Bayesian treatment to estimate the variance for each dihedral angle,[103] and Lovell *et al.* examined skew in the distribution by computing different half-widths on each side of the mode.[105] The existence of nonrotameric side chains with highly unusual dihedral angles has also received attention.[106,107] The prevailing theories to account for their existence include: interactions with the backbone, which force the side chain to adopt strained angles; stabilization by highly favorable interactions, *e.g.*, hydrogen bonding; improper fitting of the side chains to the observed electron density; and multiple conformations that coexist in equilibrium. Resolving these ambiguities may require excluding from the analysis those side chains with high *B*-factors or steric conflicts in the presence of predicted hydrogen atom locations, as well as limiting the dataset to high-resolution structures only (*e.g.*, better than 1.7 Å resolution). Nonetheless, the observation of nonstandard rotamers in the structural database highlights the limitations of rotamer libraries, in that the range of available side chain conformations may be overly restricted.

**Energy function.** The physico-chemical potential for interatomic interactions is a key element of computational protein design, since it is used to quantify sequence–structure compatibility. The physical potential determines if a particular combination of side chains is energetically favorable for a particular backbone structure. There are several potential functions currently in use for large-scale protein design, *e.g.*, Amber,[108] CHARMm,[109] and Gromos,[110] that are based on an all-atom representation of amino acids. In some cases, a united atom representation in which hydrogens are subsumed to the heavy-atoms to which they are bonded can speed up the energy calculation without significantly affecting the results.[111] Most potential functions contain terms involving bond lengths, bond angles and dihedral angles; as well as two-body terms for interactions between directly contacting amino acids. The Amber force field, for example, is parameterized using six such terms:[108]

$$E_{tot} = \sum_{bonds} K_r\left(r - r_{eq}\right)^2 + \sum_{angles} K_\theta\left(\theta - \theta_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} + \frac{B_{ij}}{R_{ij}^{6}}\right] + \sum_{i<j}\frac{q_i q_j}{\varepsilon R_{ij}} + \sum_{H-bonds}\left[\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}}\right] \quad (1)$$

In protein design, solvent is usually not treated explicitly but can be addressed implicitly *via* pairwise hydrophobic interactions or a variable dielectric constant.[112] As an energy function is often used in combination with a rotamer library during protein design, not all parameterized terms in the potential need to be evaluated during the calculation. The rotamer library in effect freezes the bond length, bond angle, and dihedral angle terms in the potential (the first three terms), leaving just the van der Waals, electrostatic, and hydrogen bonding terms to contend with. A side effect of fixing the dihedral angles, *i.e.*, discrete rotamer states, is an overestimation of van der Waals energy at short distances due to the steep repulsive part of the nonbonded (van der Waals and hydrogen bonding) interactions. This is often compensated by softening the van der Waals term *via* a scaling of the van der Waals radii by a factor slightly less than unity.[113,114]

A limitation in most of the existing physical potential functions is that they are

based on pairwise interactions between directly contacting amino acids. The physical basis of coupling between distant residues is not well-understood and their proper representation may require high-order energy terms not included in the existing force fields. Since the inclusion of multibody interaction terms in the potential can rapidly make the computation intractable, they are frequently omitted during numerical studies of bio-macromolecules.

If the state of a sequence is denoted by ($\alpha_1$, $r(\alpha_1)$; $\alpha_2$, $r(\alpha_2)$; … ; $\alpha_N$, $r(\alpha_N)$), where $\alpha_i$, $r(\alpha_i)$ represent the type and the rotamer of a residue, then the energy of a particular set of amino acids and rotamer states is computed by summing over one- and two-residue interactions.

$$E(\alpha_1,r(\alpha_1);\alpha_2,r(\alpha_2);\ldots;\alpha_N,r(\alpha_N)) = \sum_{i=1}^{N} \varepsilon_i(\alpha_i,r(\alpha_i)) + \sum_{i=1}^{N}\sum_{j>i}^{N} \gamma_{ij}\left(\alpha_i,r(\alpha_i);\alpha_j,r\left(\alpha_j\right)\right) \quad (2)$$

The one-body energy $\varepsilon$ is calculated by summing over the interatomic interactions between the side chain atoms of amino acid $\alpha_i$ when in rotamer state $r(\alpha_i)$ with the backbone. Similarly, the two-body term $\gamma$, representing the rotamer-rotamer interaction, is calculated as the sum of all interatomic energies between the atoms in the side chains of amino acids $\alpha_i$ and $\alpha_j$ given that their rotamer states are $r(\alpha_i)$ and $r(\alpha_j)$.

Using well parameterized molecular energy functions in atomistic molecular simulations, we would expect to be able to recover (at least partially) some of the solvation and secondary structure preferences of the amino acids. In protein design calculations, however, the use of explicit solvent is usually precluded by the fact that large numbers of possible sequence changes must be evaluated. Since atomistic calculations involving each potential sequence are not feasible if large portions of the sequence space are to be searched and/or characterized, these solvation effects and local structural preferences are often addressed in an implicit manner without the use of explicit solvent. These terms may be included in design calculations as additional contributions to the effective energy function or additional constraints on sequence properties, some of which are discussed below.

**Solvation.** While protein-solvent interactions are critical to protein folding, an accurate quantitative analysis of the hydrophobic effect is difficult because of its fundamentally multibody nature. In addition, the use of explicit solvent models introduces additional degrees of freedom and can become impractical for protein design, where averaging over solvent degrees of freedom must be performed for every sequence considered in the design process. A widely used method of quantifying solvation effects involves expressing the solvation free energy as the product of the solvent accessibility of a buried atom ($A_i$) and its intrinsic atomic solvation parameter, or ASP ($\Delta\sigma_i$) in the form:[115]

$$\Delta G_i = \Delta\sigma_i A_i \quad (3)$$

A solvent-accessible surface area is defined as the area over which the center of a water molecule of radius 1.4 Å can move while maintaining unobstructed contact with the group. The ASP is fit to the free energies of transfer of amino acid analogs between a hydrophobic medium (octanol, vacuum) and water.[116] Such implicit

models of solvation may also include generalized Born terms, where solvent is treated as a polarizable dielectric.[117–119] The free energy of hydration based on experimental free energies of solvation for simple aliphatic and aromatic compounds has also been proposed as a way of including the effects of solvation on protein folding.[120,121] Koehl and Delarue have generalized similar models in order to take into account contributions from protein/protein interactions.[122] A revised environmental free energy term when used together with the van der Waals and electrostatic terms has helped improve the predictive power of an algorithm, resulting in a higher percentage of correctly predicted sequences in a controlled test.[123,124] A simpler alternative has also been proposed based on the pairwise sum of the surface area buried by neighbor atoms with the goal of reducing the total computation time.[125]

A different approach to estimating the hydrophobic force has been suggested by Takada *et al.*, who only considered interactions based on $C_\alpha$ and $C_\beta$ atoms.[126] Similarly, an environmental potential energy based on the $C_\beta$ density $\rho$ in the vicinity of each side chain has been parameterized from a set of soluble proteins.[127] Since the locations of $C_\beta$ atoms are invariant for a fixed backbone, the hydrophobic propensity effectively takes the form of a one-body energy. This choice of potential is a useful parameterization of the hydrophobic effect and was shown to correlate well with commonly used hydrophobicity scales.[127,128]

**Secondary structure preferences.**

*Helix.* The statistical analysis of known protein structures shows that amino acids appear with different frequency in various secondary structures, which prompted Chou and Fasman to parameterize their propensities based on 15 protein structures.[129] In a more detailed study, Richardson and Richardson studied 215 α-helices from 45 globular proteins and tabulated the distribution of all twenty amino acids on the α-helix.[130] They reported that Asn has a strong preference for the N-cap position, while Pro has a significantly above average frequency at the beginning of a helix where it serves as a helix initiator. While chemically similar, Gln cannot replace Asn at the N-cap position, where the side chain does not have the correct geometry to form a hydrogen bond with the main chain, and Gln is not statistically favored over other residues as a capping residue. At the opposite end of the helix, Gly was by far the most common C-cap residue, terminating 34% of all helices. They also noted that Ala has a relatively smooth and favorable distributions throughout the helix, in accord with physical measurements that show Ala is a strong helix maker.[131] The helix propensities of all side chains have since been measured in experimental peptides using host–guest systems.[131–133] Measurements of relative helix forming tendencies were also measured in a 17-residue Ala-based peptide where a specific position was systematically mutated to five other nonpolar amino acids,[134] but with a somewhat different result than from the host–guest system. In addition to Ala, Leu and Met have been found to be helix promoting, while branched side chains such as Ile, Val, and Thr are in general poor helix formers. The distribution of charged residues is uneven along the helix with acidic and basic residues clustered near the $NH_2$- and COOH-termini of the helix, respectively. This charge-dependent distribution is attributed to the resulting neutralization of the induced macrodipole moment that stabilizes the helix by $\sim$0.5 kcal mol$^{-1}$.[71] Acidic and basic charged residues on adjacent turns of a helix can increase the stability of a helix by another $\sim$0.5 kcal mol$^{-1}$

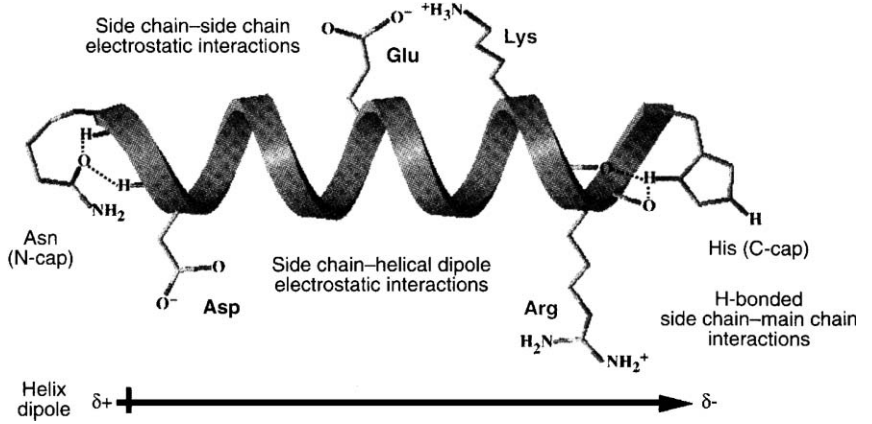


**Fig. 2** An idealized helix. See text for explanation of various stabilizing influences. (Re-printed with permission from ref. 71.)

through intrahelical salt-bridge formation.[71] An idealized helix with various built-in stabilizing interactions is shown in Fig. 2. Computational studies have attributed the physical source of the α-helix propensity to conformational entropy by correlating the side chain entropy loss upon helix formation with measured helix propensities.[135,136]

*β-Strand.* The β-strand propensities of amino acids have similarly been investigated by a number of groups, who measured their propensities by systematically substituting a solvent exposed β-strand residue in a host protein such as a zinc-finger protein or the B1 domain of protein G.[137–139] Studies show that β branched amino acids in general have the highest β-sheet propensities, regardless of the exact chemical nature of the side chain. Thr, Ile, Val, Tyr and Phe are among the best β-sheet formers, while charged residues Lys and Glu are among the poorest β-sheet formers. Furthermore, the β-sheet propensities have been shown to be context-dependent and vary depending on whether the amino acid is located on an edge strand with one neighboring β-strand, or on a central strand buffeted by β-strands on both sides.[140] Similar to the stabilizing side chain–side chain interactions observed in a helix, cross-strand pairs also interact with each other, either augmenting or opposing the β-sheet propensities of the neighboring residues. Smith and Regan studied the interaction energy between cross-strand pairs of side chains on an antiparallel β-sheet by creating double mutations on adjacent strands in the B1 domain.[141] They measured favorable interaction energy of nearly 1 kcal $mol^{-1}$ for the Lys–Glu pair and the Phe–Phe pair, and unfavorable interaction energy of 0.75 kcal $mol^{-1}$ for Thr–Ile pair. Their measurement was in good agreement with the statistical correlation observed in the protein structure database. To provide a theoretical foundation for the intrinsic β-sheet propensities, Street and Mayo derived the Ramachandran plots for each of the amino acids by modeling them in a dipeptide environment.[142] These data were then used to determine the changes in entropy ($\Delta S$) and Helmholtz free energy ($\Delta A$) on folding into a β-sheet. Both $\Delta S$ and $\Delta A$ correlated well with average normalized

experimental propensities, which led the authors to conclude that the β-sheet propensity arises from the steric interaction of an amino acid side chain with its local backbone.

*Loops and turns.* Interhelical loops and β-hairpins form another important class of structural motif. Loops can be understood as a combination of common helix C-cap and N-cap motifs joined back to back, and not surprisingly, loop residues often adopt conformations that are characteristic of residues found near helix termini.[143] Turns in β-hairpins (reviewed in detail in ref. 144) connect two antiparallel β-strands and may contribute $\sim$2–5 kcal mol$^{-1}$ to the overall stability, making the judicious selection of a turn a critical part of designing a β-sheet. Turns can be classified based on the number of residues involved and the handedness,[145] of which two residue turns with a left-handed twist, type I′ and II′ turns, are the most often found in β-hairpins.[146] The required main chain torsion angles in a hairpin put the turn residues in the left-handed region of the Ramachandran plot, and as a result turn sequences often include Gly, Asn, Asp and Pro that frequently adopt these conformations. In an *in vitro* evolution study, Zhou *et al.* introduced random turn sequences to host proteins of differing thermodynamic stabilities.[147] The percentage of active mutants was lower for hosts of lower stability and decreased further as the temperature was raised, demonstrating that optimized β-turns can affect the evolution of marginally stable proteins.

**Negative design.** Designing a *de novo* protein, one must ensure that the new sequence folds to a unique structure. In so doing, the protein avoids appreciably populating alternative structures that are significantly different from the target structure. In the language of the energy landscape theory of protein folding, the energy landscape must be appropriately ''funneled'' toward the native state. This aspect of protein design that includes bias against misfolded structures is referred to as ''negative design''.[148] An example from protein engineering that demonstrates the importance of negative design is an antiparallel three-helix bundle designed from three copies of a helix-forming sequence that were concatenated together. In the absence of explicit negative design, the resulting protein populated a mix of two competing topologies, with the third helix on either side of the plane formed by the first two helices.[69] To achieve structural specificity, therefore, the target structure must correspond to the energy ground state with a significant energy gap separating the target structure energetically from other possible competing structures.[149–152] Raising the energy of alternative conformations is a design strategy used both by protein engineers and nature alike. In Arc repressor, buried salt bridges confer conformational specificity at the expense of destabilizing the entire protein.[153] The statistical analysis of edge strands in naturally occurring β-sheet proteins shows that negative design elements such as β-bulges, prolines, inwardly directed charged residues, and very short edge strands protect free edge strands and help avoid the formation of aggregates.[154] The introduction of negative design features allowed Wang and Hecht to convert amyloid-like fibrils to a monomeric β-sheet protein.[155]

The notion that decreasing energy is correlated with improved foldability has been shown to be problematic, particularly for models involving reduced representations of the amino acids where oftentimes negative design is crucial.[156] For simple models

of proteins, foldability criteria that more accurately approximate the free energy of folding may be used as objective functions in sequence design. For atomically detailed representations of proteins, however, it is generally regarded that lower energy sequences are more likely to fold to the desired structure, and the minimization of interaction energy is at the heart of most design algorithms. This is not unreasonable, given that most such design algorithms yield tight packing of interior side chains, consistent with what is observed in naturally occurring structures. Such tightly packed sequences are likely specific to a particular backbone structure, and it is unlikely that the same interior packing could be observed in alternative conformations of the backbone. In a sense, use of explicit representations of the side chains increases the effective number of monomer types by associating a set of rotamers with each amino acid, and as the effective number of monomer types increases, it becomes more straightforward to encode a particular tertiary structure.

### Search methods

Protein folding is driven by the free energy difference between the denatured and native states. Due to the difficulty of computing the free energy differences accurately with atom-based models of proteins, however, the fitness of a designed sequence involves only the target structure and is often evaluated based on energies computed from one or more potential functions that quantify sequence–structure compatibility. There are well-tested algorithms of finding low-energy sequences, employing either a stochastic or deterministic approach to the search. Stochastic algorithms include the Monte Carlo method and its variants, simulated annealing, and genetic algorithms. Their individual differences notwithstanding, these methods systematically generate random sequences, evaluate their fitness, and iteratively improve on previous discoveries until a convergence is achieved. Other search methods include algorithms that start with initial search parameters and consistently drive toward low energy sequences. The elimination based methods yield global optima for pairwise additive potentials. Rather than specific sequences, some methods produce a site-specific probabilistic description of those sequences that are compatible with the target structure. The probability profile may then be used to construct a biased peptide library or to guide the selection of specific sequences.

**Monte Carlo.** The Monte Carlo (MC) method is one of the most widely used search methods when studying a system of great complexity. Energy minimization during protein design is one such area where MC and other similar methods based on the same principle have made significant contributions.[157,158] At each elementary step in a MC search, a test sequence is generated from the current amino acid sequence in a partially random manner, and its energy is computed using a physical potential. These elementary steps may involve changes in both rotamer state as well as amino acid identity. To bias the evolution of a sequence down the energy landscape towards the global minimum, the test sequence is either accepted or rejected according to the Metropolis criterion which depends on an effective temperature of the process.[159] The value of the effective temperature dictates the allowed range of energy changes that are available at each step in the MC process, with larger changes in energy being more

probable at higher effective temperatures. If the test sequence is accepted, then it is used during the next round to generate a new sequence, but if it is rejected, the test sequence is discarded and the original sequence is re-used to construct another test sequence. The iteration continues until some convergence criteria have been met.

Simulated annealing, so called in reference to the gradual cooling of physical material to achieve internal order, is a MC algorithm with a gradually decreasing effective temperature which allows progressively lower energy configurations to be sampled. Simulated annealing has been used to solve NP-complete problems such as the traveling salesman problem[160] as well as such optimization problems as molecular docking,[161] side chain packing,[162] and sequence design.[163] Other extensions to the classical MC include MC with quenching (MCQ) and biased MC (BMC). MCQ provides periodic optimization at the end of each cycle by testing all possible rotamers of the amino acids. Hence, the rotamer combinations with the lowest interaction energy are identified before the next MC cycle.[164] The main feature that distinguishes BMC from the classical MC is that the trial moves are biased to increase the acceptance probability.[165] This is achieved by substituting the amino acid at each chosen site with a probability that is a function of the local energy surface. In a lattice model calculation, BMC and mean-field biased MC have provided more efficient sampling, faster permitted cooling rates, and better estimates of the lowest energy sequence for the target structure.[166]

**Genetic algorithms.** The concept of Darwinian evolution based on the survival of the fittest guided the invention of genetic algorithms as powerful stochastic methods for optimization.[167] Whereas MC and its variants maintain and repeatedly update one or a few trial configurations serially, genetic algorithms are inherently parallel algorithms that use a population of would-be solutions to arrive at optimal solutions through mutations, crossovers and natural selection. In a typical implementation of the genetic algorithm, a large number (a few tens) of randomly created configurations (sequence-rotamer states of the protein) are initially assembled to represent a population. Each configuration is referred to as chromosome in analogy to genetic information in biology. The obligate inputs to any genetic algorithm include the genetic representation of solution, objective function and definition of genetic operators. The chromosomal representation of a solution is flexible and can include trees, lists, arrays, in addition to the commonly used strings. For each representation, the genetic operators must be defined with full information regarding how two chromosomes would swap genetic information. The mutational and crossover rates must also be specified. An objective function is used to evaluate the fitness of a chromosome, and may be maximized or minimized during the optimization.

While there are differences in the implementation details, all genetic algorithms share the common computational framework. At each generation, the chromosomes are evaluated according to the objective function, and high-scoring chromosomes are allowed to mate with each other. The resulting offspring replace low-scoring chromosomes to improve the average fitness of the population while maintaining its total size. The exact choice of the chromosomes involved in crossovers varies just as there are a number of ways of selecting the members to be eliminated. Regardless, high-scoring chromosomes in general have a higher rate of mating and survival than

low-scoring ones, resulting in an improvement in fitness for the entire population over time. Once the population has been updated, it undergoes random mutations at a predetermined rate, which introduces variations in the genetic pool of the population, before it is subjected to another round of mating, selection, and mutation.

The peptide sequence search can be written as a genetic algorithm by letting each chromosome represent an amino acid sequence. The crossover step would then correspond to first identifying two peptide sequences with low energies, cutting them into two pieces by hydrolyzing the backbone, and ligating the N-terminal piece of one with the C-terminal piece of the other, and *vice versa*, to generate two new sequences. They would replace two other sequences with poor compatibility with the target structure. Finally, there would be a random substitution in the sequence to model point mutation. Similarly to the MC search, the strength of a genetic algorithm is in optimizing a solution on a rugged fitness landscape, as most free energy landscapes encountered during protein design are, where the algorithm can sample the search space without getting trapped in local minima. The success of a GA depends on choosing an appropriate size for the initial population as well as optimizing the rates of mutation and crossover, and hence the algorithm may have to be run multiple times before a desired solution is found.

**Dead end elimination (DEE).** A novel pruning method of identifying rotamers that correspond to the global minimum energy conformation (GMEC) was proposed by Desmet *et al.*,[168] which was since adopted by various groups for side chain modeling[123,169] and ligand docking.[170] Functionally equivalent to an exhaustive search, the algorithm guarantees to find the minimum energy solution of an energy function comprising at most two-body interactions. The DEE theorem can be succinctly stated as: for two rotamers $i_r$ and $i_t$ at position $i$, if the following inequality holds true

$$E(i_r) + \sum_{j \neq i}^{N} \min_s E(i_r, j_s) > E(i_t) + \sum_{j \neq i}^{N} \max_s E(i_t, j_s) \tag{4}$$

where $E(i_r)$ is side chain-backbone energy while $E(i_r,j_s)$ is side chain-side chain energy with other rotamers, then $i_r$ is incompatible with the GMEC and can be eliminated from further consideration. The pruning criterion was later relaxed by Goldstein to expedite the elimination of sub-optimal rotamers:[171]

$$E(i_r) - E(i_t) + \sum_{j \neq i}^{N} \min_s [E(i_r, j_s) - E(i_t, j_s)] > 0 \tag{5}$$

which is equivalent to eliminating rotamer $i_r$ if it has a higher energy than rotamer $i_t$ for all possible conformations at other sites. The process is repeated iteratively throughout the entire protein until no further amino acids or rotamer states may be eliminated. At that point an exhaustive search among the energies of the typically few remaining sequences identifies the global minimum. While a significant hurdle exists in applying DEE to protein design due to the exponentially growing computation time for large proteins,[164] the method has been successfully used in many design projects[172,173] and the algorithm continues to be refined.[174]

**Mean field theory.** While sampling methods are commonly used for optimization, the computational costs of such algorithms are high as it may take a long time to reach convergence. One alternative to stochastic sampling is the mean field approach. Rather than enumerating individual sequences, mean field calculations apply an optimization algorithm to an ensemble of sequences and try to determine the relative weights of the interactions that determine the local energy at a site (residue) that are consistent with a particular thermal average. The effective temperature may be gradually lowered so as to identify the properties of low energy sequences. Originally applied to study the Ising model of the spin–lattice and ferromagnetism,[175,176] mean field theory neglects fluctuations in the local energies of each site in a system and the interactions of a site with its neighbors are calculated as a weighted average. This results in a self-consistent set of equations determining the site-specific probabilities of particular states in a system. When describing proteins, these states represent the type and side-chain conformation of amino acids present at various sites. Mean field methods have been used both for side chain modeling and sequence design[101,177–180]

In the mean field approximation for protein design, the effective two-body potential is the weighted sum of all pairwise interactions. If the sequence of a protein of $N$ residues is described as a series of parameters ($\alpha_1$, $r(\alpha_1)$; $\alpha_2$, $r(\alpha_2)$; … ; $\alpha_N$, $r(\alpha_N)$) with $\alpha_i$, $r(\alpha_i)$ denoting the amino acid identity and side chain conformation, respectively, at site $i$, then the mean field approximation allows the average local energy $\varepsilon_i(\alpha_i,r(\alpha_i))$ at position $i$ to be written as:

$$\varepsilon_i(\alpha_i,r(\alpha_i)) = \sum_{j,\alpha_j} \sum_{r(\alpha_j)} w_j(\alpha_j,r(\alpha_j))\gamma_{ij}(\alpha_i,r(\alpha_i);\alpha_j,r(\alpha_j)) + \varepsilon_i^0(\alpha_i,r(\alpha_i)) \qquad (6)$$

where $\gamma_{ij}(\alpha_i, r(\alpha_i); \alpha_j,r(\alpha_j))$ is the two-body interaction energy between side chains $\alpha_i$ and $\alpha_j$, and $\varepsilon_i{}^0(\alpha_i,r(\alpha_i))$ is the one-body energy that results from side chain backbone interactions, or the structural propensities of the amino acids. The pair interactions of site $i$ with its neighbors are weighted with the normalized probabilities $w_j(\alpha_j,r(\alpha_j))$. The $w_j(\alpha_j,r(\alpha_j))$ are themselves Boltzmann functions of the local energies at the neighboring sites: $w_j(\alpha_j,r(\alpha_j)) \propto \exp(-\beta\varepsilon_j(\alpha_j,r(\alpha_j))$, where $\beta^{-1}$ is an effective inverse temperature. These equations are solved self-consistently for the local fields $\varepsilon_j(\alpha_j,r_j(\alpha_j))$, or equivalently the site-specific probabilities $w_j(\alpha_j,r(\alpha_j))$. The method may be used to investigate the properties of low energy sequences by solving for the site-specific probabilities for successively increasing values of the inverse temperature parameter $\beta$. The major advantage of mean-field methods over other algorithms is that by using the average energy to compute local interactions, an explicit enumeration or sampling of conformations can be avoided, greatly decreasing the total computation time. In a benchmarking study, Voigt *et al.* compared the performance of MC, GA, and self-consistent mean field algorithms in finding the global minimum as identified by dead-end elimination. Mean field algorithms performed well on hydrophobic core calculations (7% incorrect identification of amino acids) but only marginally for residues in boundary and surface positions (28 and 37% error, respectively).[164]

**Probabilistic approach to sequence design.** The use of site-specific amino acid probabilities, rather than specific sequences, is referred to as *probabilistic* protein design. Such a probabilistic approach is motivated by the complexity and uncertainty

associated with the process of identifying sequences that fold to a particular structure, *e.g.*, the energy functions used are approximate, side chain conformations are treated discretely, backbone atoms are often fixed, and solvation properties are treated using crude approximations. Probabilistic approaches are often used in science and engineering when we have only incomplete information about a system, as is certainly the case for protein folding. Nonetheless, probabilistic methods directly provide useful sequence information that can be used to guide design experiments and identify structurally important amino acids. The site specific amino acid probabilities can also highlight residues that are likely to tolerate mutation without adversely affecting structure. Such mutable sites can then be targeted for mutation in multiple rounds of protein design during the search for sequences that confer biological activity or other desired properties to a target tertiary structure.

An entropy based formalism to identify amino acid probabilities for a given backbone structure has been developed.[127,181] The theory borrows concepts from statistical mechanics to directly estimate the site-specific probabilities and addresses the whole space of available compositions, and the method is not limited to a small fraction that is accessible to experiment or to computational enumeration and sampling. Using this approach, the features of suboptimal sequences may also be readily examined. Large protein structures (more than 100 residues) can also be easily accommodated in such calculations. The effective entropy quantifies the variability of sequences consistent with the target structure. The number of possible sequences is reduced by decreasing the energy or imposing constraints on the system, which reduces the conformational entropy, thus diminishing the number of allowed sequences.

The notion of entropy maximization is central to this methodology, just as it is fundamental to statistical mechanics and information theory. There are an infinite number of possible sets of site-specific state probabilities, where the "state" of each residue is defined by both monomer identity and side chain conformation. The most probable set of probabilities is determined by optimizing an effective entropy function subject to any constraints imposed on the system. The method takes as input a target structure, energy functions, and constraints on amino acid properties. Both global considerations (*e.g.*, the overall energy of the sequences) and local features (*e.g.*, the allowed amino acids at particular sites) can be specified *via* constraints. With the judicious application of such constraints, the properties of sequences consistent with a particular tertiary structure and other desired properties may be readily identified.

Again with $w_i(\alpha_i,r(\alpha_i))$ denoting the amino acid and rotamer state probabilities at residue position $i$, the total sequence-conformational entropy $S_c$ (simply referred to as "conformational entropy") is written as

$$S_c = -\sum_{i,\alpha_i,r(\alpha_i)} w_i(\alpha_i,r(\alpha_i)) \ln w_i(\alpha_i,r(\alpha_i)) \qquad (7)$$

The sum extends over each sequence position $i$ and all available amino acids $\alpha$ at each position. Furthermore, for each amino acid, the sum is taken over each of the $k$ possible rotamer states $r_k(\alpha_i)$. Although writing the entropy $S_c$ in this manner implies a factorization approximation and seems to suggest that the site specific probabilities are independent. In fact, constraints on the sequences cause the probabilities to be

coupled. The $w_i(\alpha_i,r(\alpha_i))$ are determined as those that maximize $S_c$ subject to constraints $f_m$, which are themselves functions of the $w_i(\alpha_i,r(\alpha_i))$. In order to impose these constraints during maximization, a variational functional $V$ is defined using the method of Lagrange multipliers

$$V = S_c - \beta_1 f_1 - \beta_2 f_2 - \ldots \quad (8)$$

where the $m$th constraint function $f_m$ has a particular value $f_m^o = f_m(\{w_i(\alpha_i,r(\alpha_i))\})$ and the $\beta_m$ are Lagrange multipliers conjugate to the constraints. The functions $f_m$ may be used to specify a wide variety of properties on the sequences, including the overall energy of the structure, the patterning of residue properties, and effective energies that quantify solvation and/or secondary structure propensities of the amino acids. Different energy constraints each enter in a dimensionless manner in the variational functional $V$, obviating difficulties associated the relative weighting of physically-derived *vs.* database-derived energy terms. The set of equations to be solved simultaneously to determine the probabilities and the Lagrange multipliers take the form:

$$\partial V/\partial w_i(\alpha_i,r(\alpha_i)) = 0 \quad (9)$$

$$f_m(\{w_i(\alpha_i,r(\alpha_i))\}) = f_m^o \quad (10)$$

This large set of on the order of $10^4$ coupled, nonlinear equations is solved using constrained minimization or root finding methods.[182] If the only constraints imposed are those involving the atomistic energy and the normalization of the $w_i(\alpha_i,r(\alpha_i))$, this methodology reduces to the mean-field methods discussed in the previous section.

The probabilistic methods described may be used in several ways to guide protein design. First, a low energy consensus sequence may be identified as the sequence comprising the most probable amino acid at each position. Although the consensus sequence would not directly include correlations between residue identities, such correlation may be better addressed by an iterative series of calculations. For each iteration, an increasing number of residue identities may be constrained until a unique sequence is identified. Such an approach has been used in the design of a 114-residue four-helix metalloprotein.[183] The calculated probabilities may also be used to guide a search algorithm. For example, an efficient Monte Carlo (MC) based method has been reported that uses predetermined amino acid probabilities to bias the generation of trial sequences at each step of the Monte Carlo Markov trajectory.[166] Finally, probabilistic methods may be used to quantitatively guide the design of combinatorial libraries of proteins, in which an ensemble of sequences is generated in a manner that best reproduces the calculated site-specific amino acid probabilities.

## Efforts in quantitative protein design

In this section, we showcase some of the highlights in computational protein design, where proteins have been designed with the help of computation, and the experimentally realized proteins have been found to agree with the design specifications. In some cases, high-resolution structures of these proteins are available in addition to

biochemical and biophysical data, providing detailed information on the accuracy of computational prediction. The works reported in the literature include designs and redesigns of various structural elements such as helices, sheets, turns, and hydrophobic cores; as well as introduction of *de novo* functionalities such as ligand binding and catalysis. Computational protein design has advanced sufficiently to allow a medium size protein to be predicted efficiently, and to permit the complete design of novel proteins.

### Core packing and secondary structural elements

Some of the earliest computational work on quantitative protein design involved redesigning the hydrophobic core of globular proteins.[96,162,184,185] Several groups since then have studied the core-packing problem using both empirical and theoretical methods. Randomization studies of the hydrophobic core of lambda repressor, barnase and T4 lysozyme support the view that the protein core is closer to a "oil-drop" model than a "jigsaw" model, and the most important chemical properties of core side chains is their overall hydrophobicity.[46,186,187] Other studies, on the contrary, have suggested that good complementarity between contacting side chains is important to achieve stability and structural uniqueness.[188,189] To demonstrate the power of computational methods in the search of optimum core packing arrangements, Desjarlais and Handel developed a pair of algorithms that are together called Repacking of Cores (ROC).[77] The first program generates a custom rotamer library for the target structure, while the second uses a genetic algorithm to optimize core packing. ROC was applied to repack the core of two proteins 434 Cro[77] and ubiquitin.[190] While many of the 434 Cro mutants had stability comparable to that of the wild type, all the ubiquitin mutants studied were significantly destabilized compared to the wild type. As random mutants, which served as controls, had well-defined conformations independent of stability, they concluded that core packing affects protein stability but not the conformational specificity. Redesigning of the protein core was also reported by Jiang *et al.*, who used Metropolis-driven simulated annealing and low-temperature MC sampling in a new computer program CORE to find sequence and side chain conformations of hydrophobic core residues for hyperstable mutants of four naturally occurring proteins (B1 domain of protein G, 434 cro, myoglobin and methionine aminopeptidase).[191]

**Helix bundles.** Helices are some of the most common structural motifs in proteins, and structures comprising α-helices are fundamental targets of protein design. Most repacking algorithms assume fixed backbone coordinates. However, X-ray crystallography studies show that the backbone conformations often change in response to point mutations.[192,193] To introduce backbone flexibility, Harbury *et al.* parameterized parallel coiled-coil proteins using three parameters that represent the supercoil radius, supercoil frequency and the orientation angle of the **a** position in the helical heptad.[78] The application of a standard molecular force field to a series of parameterized backbone coordinates then successfully repacked the core of three parametrized structures (a dimer, a trimer, and a tetramer) with a RMSD of 0.6 Å or less from their crystal structures. Similarly, the same group designed a sequence to

fold to a tetramer with a right-handed superhelical twist by parameterizing the target structure to allow backbone flexibility, and by engineering the optimum main chain conformation and interior side chain rotamers through computational packing.[79] The designed protein agreed strikingly with the subsequently determined crystal structure in atomic detail.

As coiled-coils are often involved in mediating protein–protein interactions, a predictive method of designing new coiled-coil dimers would be invaluable in molding protein interaction interface. To study the role of buried hydrophobic residues in determining specificity, Keating *et al.* designed and estimated the stability of six heterodimeric coiled-coils derived from GCN4.[113] Their computational method combined extensive conformational sampling with molecular mechanics minimization to predict the stability and structure of designed heterodimers to high accuracy. Havranek and Harbury also investigated the computational design of coiled-coil structures, but with an emphasis on the incorporation of negative design to dictate specifity.[194] Rather than optimizing the sequence for the target structure alone, they maximized the free energy difference between the target structure and other competing states. As an example, for sequences designed to form homodimers, three other competing states were simultaneously considered, *i.e.*, heterodimers, aggregate state and unfolded state. Consideration of the full competitor set in general resulted in sequences that were superior in stability and specificity relative to those obtained from calculations with one or more of the competing states omitted.

A contiguous three helix bundle protein with unique topology and a native-like core (α3D) was designed by Bryson *et al.* in a hierarchical approach combined with computational optimization of the interior side chain conformations.[69,195] The modeling was initiated by shortening the previously studied Coil-Ser sequence[14] to form three helix turns rather than four, conjoining three copies of this sequence with hairpin loops, and arranging charged residues on the surface to stabilize antiparallel packing. Strong helix start/stop sequences "N–X–(X)–E" were then introduced to help define the topology of the bundle. The resulting protein α3B exhibited some characteristics of a molten globule, which were attributed to the 15 Leu residues in the core that can pack in many different conformations with roughly equal energies and to the coexistence of two alternative topologies involving clockwise and counterclockwise turning of the third helix with respect to the first two. To enforce the counterclockwise topology, they placed positive charged residues at the **e** and **g** positions of the first helix, negative charged residues at the **e** and **g** positions of the second helix, and negative and positive charged residues at the **e** and **g** positions, respectively, of the third helix. The 17 core residues were redesigned using the genetic algorithm ROC[77] in four rounds, systematically fixing eight, eleven, and sixteen residues at each round. The solution structure of the redesigned protein, after three conservative residue substitutions have been made to help with cross-peak identification, shows a counterclockwise bundle looking down the symmetry axis, with the backbone RMSD of 1.9 Å from the modeled structure. As in native proteins, the buried side chains were well packed, each largely populating a single predominant rotamer state, though in NMR experiments the side chains of some were observed to be more dynamic than those of natural proteins.[196] In a hydrogen-deuterium exchange experiment, nine amide protons showed protection factors within 0.5 kcal $mol^{-1}$ of that expected from its thermal stability.

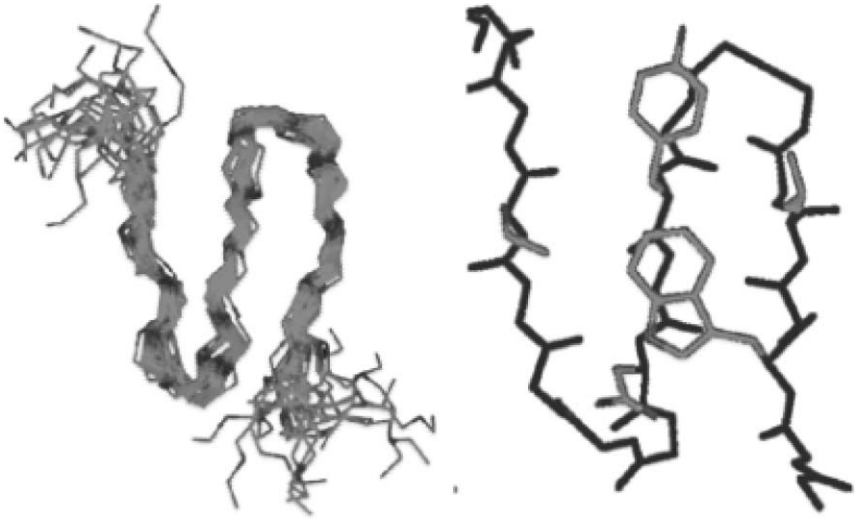

**Fig. 3** Three-stranded β-sheet, Betanova. (Left) Backbone traces of 20 NMR structures. (Right) Minimized average NMR structure with the residues $Trp^3$, $Val^5$, $Tyr^{10}$, $Asn^{12}$, $Thr^{17}$ forming the hydrophobic core. (Re-printed with permission from ref. 197.)

**β-Sheet.** A 20-residue three-strand β-sheet protein, Betanova, was designed by the Serrano group, and structurally characterized by NMR (Fig. 3).[197] The design of a monomeric β-sheet protein is difficult due to the tendency of isolated β-sheet secondary elements to aggregate. In order to simplify the design, they selected a template consisting of three strands with four residues per strand and took an iterative approach to improve their initial design. A sequence compatible with the target backbone structure was selected based on β-hairpin stability, amino acid β-sheet propensities, statistical preferences for inter-strand residue pairs, and side chain rotamer conformations. Also, van der Waals contacts were optimized through rotamer modeling, favorable inter-strand packing was achieved through a Glu–Lys ionic pair, and turn sequences were optimized for the canonical type I′ β-turns. As expected of a protein with native conformation, the resulting Betanova exhibited cooperative folding/unfolding transitions by CD and fluorescence spectroscopy. The successful design of a β-sheet protein through a combination of modeling and structurally stabilizing motifs shows that we now have some understanding of the principles guiding β-sheet formation.

To better understand the formation of stable β-sheets and proteins, single- or multiple residue mutants of Betanova were created by Lopez de la Paz *et al.* using an automatic design method PERLA (protein engineering rotamer library algorithm).[198] The algorithm evaluates the fitness of an amino acid sequence to the target structure with a scoring function that is based on the ECEPP/2 all-atom molecular force field[199] and a combination of statistical terms including entropy and solvation. Dead-end elimination is used to reduce the sequence space, while mean-field theory is used to weight different side chain conformers. In the end, candidate sequences are produced along with modeled structures. When the mutants predicted by the

algorithm were studied biophysically, some of them were more stable than the wild type by $\sim 1$ kcal mol$^{-1}$, in good agreement with the prediction. PERLA was also used to redesign other β-sheet-containing proteins including the SH3 domain of α-spectrin. Selected for designed mutagenesis were two solvent exposed four-residue clusters (Sheet I and II) and clusters of residues involved in two turns and the protein core. As before, more stable mutants were identified among the selected sequences.[200]

Recently, PERLA was used to study the formation of amyloid fibrils, thought to be responsible for the onset of several high profile diseases including Alzheimer's disease and spongiform encephalopathies.[201] On the assumption that propagation and stacking of preformed β-sheets would drive the assembly of amyloid fibrils, PERLA was used to design self-associating hexapeptides with a high propensity to form polymeric β-sheets. A six-stranded antiparallel β-sheet template with six residues per strand was constructed from the large single-layer β-sheet of the outer surface protein A, whose structure had been determined by NMR.[202] Sequences predicted to be amyloidogenic were synthesized and studied by CD and electron microscopy, which revealed that β-sheet formation is necessary but not sufficient for the formation of amyloid fibrils. Specific interactions appear to be crucial in the stabilization of fibril aggregates, since a single amino acid substitution (Ile→Leu) can completely block fibril formation. Similarly, Coulombic interactions can modulate the supramolecular organization of β-sheets, and fibril formation occurs only when the total net charge of the monomer is $\pm 1$. The presence of a large net charge on the peptide leads to the formation of amorphous aggregates in competition with the formation of ordered polymer. The rational design of a peptide-based model system for amyloidogenesis can facilitate the identification of sequences with a propensity to form fibril aggregates.

The successful redesign of a β-sheet protein using a genetic algorithm was reported by Desjarlais and coworkers.[203] Their design target is a small WW-domain protein involved in cell signaling, Pin-1, that is composed of three antiparallel β-strands and binds polyproline peptide ligands.[204] Unlike other design algorithms, the method used by these authors is unique in that an ensemble of closely related backbone structures are used as design templates, and the information from different backbone structures and design algorithm results are integrated using a novel sampling procedure. The backbone ensemble consisted of nondegenerate structures with the maximal RMSD of 0.30 Å from the PDB structure 1PIN, generated through a combination of MC perturbation and refinement algorithms. The amino acid probability was then calculated by evaluating the partition function of each rotamer across the structure set, thus exposing each rotamer state to a wide range of local environments with a unique configuration of backbone structure and side chain identities. In addition to rotamers, sub-rotamer states stochastically sample within 20° of canonical rotamers, bringing in extra degrees of freedom. Two sequences, each with 35 and 32% sequence homology to wild type, were obtained from the calculation. Despite only moderate sequence homology, NOESY data of one of the sequences showed a structure closely resembling that of the wild type, demonstrating that a β-sheet protein can be successfully designed with the inclusion of backbone flexibility.

**β-Turns.** In a quantitative study of β-turn sequences, the Baker group used a computational approach to redesign the second β-hairpin in protein L.[76] The native

four-residue turn contains three consecutive residues with positive $\phi$ angles, which presumably contribute to the low intrinsic stability in the region. The turn was therefore changed to a canonical β-turn by adding or subtracting two residues. First, the PDB was searched for alternative hairpins with termini that superimposed with the region of interest. A design algorithm using a MC search was then applied to identify amino acid rotamers that would form well-packed structures and result in low energy sequence–structure combinations. The sequences selected by the algorithm contained at the two turn positions amino acids commonly observed in the canonical turn types while differing considerably from wild type sequence. The accuracy of prediction was verified by determining the crystal structure of one of the redesigned proteins, which showed the designed turn and the *in silico* model had the all-atom RMSD of 1.4 Å from each other.

### Protein design automation

Mayo and coworkers have been pioneers of the computation-guided design of *de novo* proteins. Their design scheme ORBIT (optimization of rotamers by iterative techniques) applies dead-end elimination to find globally optimal sequences for a given structure.[123] During the final stage of the calculation, a MC search is conducted starting from the resulting optimum sequences to find other related high-scoring sequences. In an implementation that couples theory, computation, and experimental testing, the algorithm was applied to redesign the hydrophobic core of a homodimeric coiled-coil based on GCN4. A quantitative structure activity relationship (QSAR) analysis showed a significant correlation with surface area burial, which then prompted the inclusion of buried surface in the scoring function. Subsequently, the methodology led to the successful design of a novel protein entirely based on computation. This ββα-motif protein was modeled after the tertiary structure of a zinc finger DNA binding module of Zif268 but folds stably without the requisite $Zn^{2+}$ metal ion (Fig. 4).[172] The design procedure identified a single sequence from over

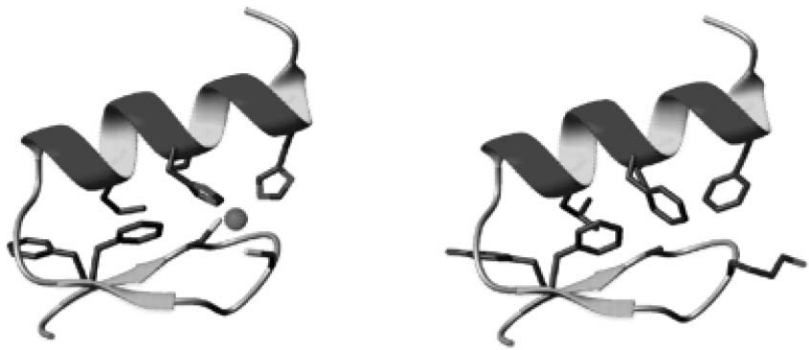

**Fig. 4** Automated sequence design of a ββα protein FSD-1. (Left) Zinc finger Zif268 with the zinc ion shown as sphere. (Right) Computed FSD-1 structure. (Re-printed with permission from ref. 172.)

$1.9 \times 10^{27}$ possible sequences by DEE. NMR analysis of the protein showed a compact, well-ordered structure with the predicted side chain conformations. A BLAST[205] search of a sequence database showed that the designed sequence had less than 40% homology to any other known sequence, showing that a novel sequence can be computationally designed completely from scratch.

ORBIT was also used to design a metal-free variant of thermophilic rubredoxin, a small protein ($\sim$6 kDa) naturally stabilized by a tetrahedrally coordinated iron.[206] The high-resolution structures of rubredoxin at resolutions of 0.95–1.8 Å served as design templates. In the calculation, the four conserved Cys at the iron binding site were optimized in the absence of metal. Two were classified as core residues and thus mutated to one of seven hydrophobic amino acids (Ala, Val, Leu, Ile, Phe, Tyr and Trp) whereas the other two were mutated to all amino acids except Gly, Pro, Cys and Met. Four other neighboring residues were allowed to change conformations to optimize core packing. The first round of calculation resulted in Thr at the two non-core Cys positions, while Leu and Ala were selected at the core Cys positions during the second round using van der Waals radii scaled to 0.7 of their nominal values. The resulting mutant was found to adopt a fold similar to the wild type based on the chemical shifts of the amide and $\alpha$ hydrogens, and undergoes cooperative reversible thermal denaturation at 82 °C.

While binary patterning is commonly used in protein design, the determination of a residue position as either polar or hydrophobic is not always obvious for globular proteins. In order to automate the binary patterning procedure during protein design, Marshall and Mayo developed an algorithm called Genclass, which classifies each position along a protein backbone to either exposed or buried based on solvent accessibility.[88] They first replaced all the naturally occurring side chains in the target structure with "generic" side chains, whose size and shape are similar to an average amino acid. A solvent accessible surface was then generated by applying the Connolly algorithm with the solvent radius of 1.4 Å.[207] Comparison with 29 proteins in the PDB shows that setting the minimum solvent accessible area for polar residues $SA_{cut}$ to 23.9 Å$^2$ yields the best agreement between prediction and the database binary pattern, *i.e.*, the highest percentage (76%) of hydrophobic residues whose generic surface area is less than $SA_{cut}$, and polar residues whose generic surface area is greater than $SA_{cut}$. The algorithm was then applied to generate a series of engrailed homeodomain variants, whose design quality was judged based on stability and conformational specificity. The proteins were experimentally studied using CD and $T_m$ measurements, dynamic light scattering, NMR and differential scanning calorimetry. The most successful variant B6 had $T_m$ of 114 °C, was stable at physiological pH, and exhibited well separated NMR peaks in the aromatic and amide region. The actual $SA_{cut}$ value of 43 Å used for the construction of B6 was, however, significantly larger than the predicted optimal value of 23.9 Å, and either raising or lowering the threshold resulted in suboptimal variants with reduced stability and tendency to aggregate. Hence, although the study demonstrates that well folded, stable proteins may be designed using an automated binary pattern prediction algorithm, it also raises a concern about the prepatterning of hydrophobic and polar residues, which must be fine-tuned in order to achieve stability and conformational specificity.

## Combinatorial design

Introduction of a new function to an existing protein can be achieved through molecular evolution. Starting from a protein library of targeted or random mutants, mutants with desired properties are screened in a customized functional assay. Some highly unusual properties may be successfully engineered if the diversity of the library is high enough, but the size of a library that can be assembled and screened in practice is far smaller than the full sequence diversity available to a moderately sized protein. As computational analysis has been shown in a number of cases to be effective in identifying low energy sequences among exponentially large numbers of potential sequences, one wonders whether computational analysis can be used to guide protein evolution in a library-based assay. Such was the idea behind the work of Hayes *et al.* who searched for sequences compatible with the active site of TEM-1 β-lactamase using DEE and simulated annealing.[208] Top scoring sequences were compared to calculate the amino acid probability at 19 positions within 5 Å of the active residues, which was then used to synthesize degenerate oligonucleotides needed to construct a library of ~200,000 mutants in *E. coli*. Screening the library for improved resistance to the antibiotic cefotaxime yielded novel sequences with multiple mutations, all different from those observed in mutagenesis studies, that conferred over three orders of magnitudes greater resistance against the antibiotic.

## Ligand binding

Computational protein design has been used to engineer new binding affinity as well as to modulate existing affinity. The availability of high-resolution structures of the insert domain from the α-chain of integrin αMβ2 (Mac-1) in both binding-active (open) and binding-inactive (closed) conformations allowed the Mayo and Springer groups to take a computational approach to design integrin mutants with enhanced affinity to the natural ligand.[209] They reasoned that mutations that stabilize the binding-active conformation would result in higher binding affinity. Using their DEE-based algorithm, they redesigned 40–45 core residues out of a total of 184 residues to bias the open structure over closed structure. The residues that may be directly involved with ligand binding or in $Mg^{2+}$ coordination were excluded from the calculation. The calculated energies of three selected mutant sequences were all lower than that of wild type in the open configuration and higher than that of wild type in the closed conformation. This result was independent of the solvation potential used in the calculation. When the designed mutants were investigated in a functional assay using cultured cells, they all showed increased binding to the protein ligand iC3b by 10- to 13-fold. This outcome compared favorably with other designed mutants (F302W, F302R, F302Y), of which only the F302W had a two-fold increase in activity.

The binding of calmodulin (CaM) has been modulated to improve specificity towards just one of its many natural targets.[210] CaM is an all α-helix protein with N-terminal and C-terminal domains connected by an α-helix of eight turns. $Ca^{2+}$ binding activates the protein by inducing a large conformational change that concurrently exposes hydrophobic residues. CaM binds its targets by wrapping its

two domains around helix ligands.[211–213] While the burial of a large amount of hydrophobic surface contributes to the high affinity of CaM for its targets ($K_d \leqslant 10^{-7}$ M), the flexibility of its two domains with respect to each other results in low binding specificity. To improve the specificity of CaM towards one of its targets, smooth muscle myosin light chain kinase (smMLCK), the CaM–smMLCK complex was optimized under the assumption that such an optimization would destabilize contacts between CaM and other targets. Twenty-four buried CaM residues within 4 Å of the target peptide were mutated to Ala, Val, Leu, Ile, Trp, Phe, Tyr, Met and Glu, while the ligand sequence was kept constant. One of the mutants containing eight substitutions was shown to have a substantially lower overall energy ($-508.4$ kcal mol$^{-1}$ compared to $-467.4$ kcal mol$^{-1}$ of wild type). The designed mutant, which had a superimposable CD spectrum as wild type, bound smMLCK with $K_d$ of $1.3 \pm 0.9$ nM, which is comparable to that of the wild type ($1.8 \pm 1.3$ nM), but the affinity to six other target peptides was lower by as much as 86-fold. The absence of explicit negative design in the study, however, leaves the question open whether a similar approach can be used to discriminate subtly different target sequences.

PDZ domains are small protein modules involved in signal transduction. They recognize the C-terminal 4–7 amino acids of target proteins. Of the three classes of PDZ domains, class I and class II share good structural homology whereas those of class III have a slightly displaced α-helix relative to the β-sheet when compared to either class I or II. Class I and class II proteins recognize different residues at position p($-2$) (*i.e.*, third residue from the C-terminus)—class I proteins require Thr/Ser at p($-2$) while class II proteins recognize an aliphatic residue at the same position. Reina *et al.* used computer-aided protein design to mutate PSD-95 from class I to recognize new target sequences.[214] Visual inspection was used to identify positions to be mutated, ranging from six to twelve different positions depending on the target ligand. The interaction between these mutant PDZ domains and their respective ligands were measured by fluorescence polarization assays, which showed that one mutant–ligand pair had an affinity two orders of magnitude greater than the (wild type PDZ)–(wild type peptide pair), while two other mutant PDZ–ligand pairs had similar affinities to that of wild type. Some of the residues predicted by the algorithm were the same as those identified from an independent study where PDZ domains with novel specificity were engineered by experimental screening.[215]

In a demonstration of computation-driven design of novel ligand affinity, Looger *et al.* engineered a series of new binding sites for trinitrotoluene (TNT), L-lactose and serotonin in five proteins from the *E. coli* periplasmic binding protein superfamily: glucose binding protein, ribose-binding protein, arabinose-binding protein, glutamine-binding protein, and histidine-binding protein.[216] The three target ligands have little resemblance to the cognate ligands of the wild type proteins and exhibit a wide range of chemical properties in terms of molecular shape (polar, aliphatic and aromatic), chirality, functional groups (nitro, hydroxyl and carboxylate), internal flexibility, charge, and water solubility. The semi-empirical potential function used for the design includes a Lennard-Jones potential, an explicit geometry-dependent hydrogen-bonding term and a continuum solvation term to represent the hydrophobic effect. Satisfying all potential hydrogen-bond donors and acceptors in the ligand was critical for high-affinity binding, as expected from the known importance

of hydrogen bonding in molecular recognition.[217] The specificity of ligand recognition was experimentally verified. All six designed receptors for TNT can distinguish the absence of a single nitro group and all ten lactate designs exhibited the desired chiral stereospecifity, selecting L-lactate over the D-lactate enantiomer, pyruvate, and the prochiral oxidized form of lactate. The serotonin design had a significantly lower affinity for two related molecules, tryptamine and tryptophan, both of which are missing a hydroxyl group and/or a carboxylate group. The observed binding affinities of many of the engineered proteins for their target ligands were in the same range as the wild type receptors for their cognate ligands ($K_d \sim 0.1$–$1.5$ μM), validating the computation-driven approach to the design of high affinity and high specificity binding sites.

## Towards catalysts and enzymes

Metal ions are key players at the active sites of many enzymes, and computational studies have led to the successful design of several metal binding proteins. In a series of studies using the *E. coli* protein thioredoxin, a protein that is naturally devoid of metal centers, Hellinga and coworkers used the rational protein design algorithm DEZYMER to introduce targeted mutations to bind metal ions and nonheme metal complexes. In an attempt to recreate one of the earliest evolved biological redox centers, they engineered a cuboidal [$Fe_4$–$S_4$] binding site in thioredoxin.[218] A mononuclear iron–sulfur center capable of reversible electron transfer was similarly introduced using DEZYMER by designing a tetrahedral tetrathiolate iron center, where the coordination is provided by Cys32 and Cys35, forming a disulfide bond in wild type thioredoxin, and by Cys28 and Cys75 that replace Trp and Ile, respectively.[219] The designed protein, which forms a 1:1 monomeric complex with Fe(III), undergoes successive cycles of oxidation and reduction, demonstrating that simple geometrical considerations can be sufficient to reproduce the dominant electronic structure and reactivity of a metal-based redox center. In a different study, the active site of nonheme iron superoxide dismutase (SOD) was successfully grafted into the hydrophobic interior of thioredoxin by re-creating the trigonal bipyramidal coordination.[220] This protein bound iron tightly and had an open coordination sphere capable of binding an exogenous ligand such as azide and fluoride. However, its SOD activity, $10^5$ $M^{-1}s^{-1}$, was $\sim 10^4$ lower than natural enzymes, which was attributed to suboptimal tuning of the redox potential of the metal center.

Thioredoxin was re-engineered by Bolon and Mayo to perform the hydrolysis of *p*-nitrophenyl acetate (PNPA).[221] The high stability of wild type thioredoxin allowed potentially destabilizing mutations required to build an active site to be introduced without resulting in an unfolded mutant. In order to achieve efficient catalysis, proximity and orientation of substrate molecules and transition-state stabilization must be carefully modeled. The need to destabilize the acylated enzyme intermediate relative to substrate suggested histidine as a potential nucleophile for catalysis. Therefore, a high-energy state involving a tetrahedral intermediate of histidine–PNPA was modeled as a series of side chain rotamers. The surrounding protein sequence was also optimized for binding to the high-energy state in order to reduce the activation energy and enhance catalytic turnover. Hydrophobic solvent accessible

surface area of the substrate atoms in the computed high-energy state was used as a measure of substrate recognition. A high-scoring sequence was synthesized and the rate of PNPA hydrolysis was measured. Impressively, the pH dependence of the reaction showed that His does indeed act as a catalytic nucleophile in the designed protein. The designed mutant falls short of exhibiting enzymatic efficiency, however, with only an order of magnitude enhancement in the turnover rate over 4-methylimidazole and catalytically inactive wild type thioredoxin, thus emphasizing the subtlety of computation-driven enzyme engineering.

## Large scale design and redesign

Most of the designs presented so far have involved engineering small parts of a protein, either to introduce a specific functionality or to enhance stability. A directed sequence design approach works well for studies of this nature since it can thoroughly examine the available sequence space in search of the optimum sequence. These methods may also be used for the complete design of a protein sequence, *e.g.*, a 27-residue fully designed protein by Dahiyat and Mayo.[172] Recently, the *de novo* design of a 216 residue β-barrel protein with idealized geometry was reported, demonstrating that a directed sequence search can also be applied to significantly larger proteins.[222] The preliminary results suggest a stable tertiary structure, yet the protein seems to behave unlike natural proteins. For example, there is increased binding to the hydrophobic dye ANS upon addition of the denaturant guanidinium hydrochloride (GndHCl) up to 1 M, but the ANS signal disappears when GndHCl concentration is increased to 2 M.

Using a coarse-grained protein model having simplified side-chain representations, Jin *et al* have included information about unfolded structures (negative design) in a stochastic search for a sequence with a "funneled conformational energy landscape".[223] Their design principle involves selecting for the global shape of a protein folding funnel, where sequences are identified having the target structure as the lowest energy state. A combination of Monte Carlo sequence search and repeated folding simulations ensures that the ground state is well separated from other structures sampled during simulation. The authors designed a three-helix-bundle topology and selected several of the designed sequences for synthesis, one of which had spectroscopic data (CD and NMR) consistent with a well-defined target structure. The work provides an elegant synthesis of energy landscape ideas and protein design methods.

Probabilistic approaches are well suited for large sale protein designs where large percentages of proteins are simultaneously engineered, since they can recover features of the ensemble of sequences compatible with the target fold. These methods can be used to study proteins that may be too large for direct sequence optimization. In a collaboration between the Saven and DeGrado groups, a statistical, computationally-assisted design strategy (SCADS) was used to engineer a *de novo* 114 residue DFsc with a diiron center (Fig. 5).[183] The protein forms a stable four-helix bundle with a melting temperature of $T_m = 53$ °C in the absence of metal, and remains unfolded at 98 °C in the presence of metal. A well-ordered interior is evident in the 1D- and 2D-NMR spectra. The tertiary template was modeled after a previously designed due

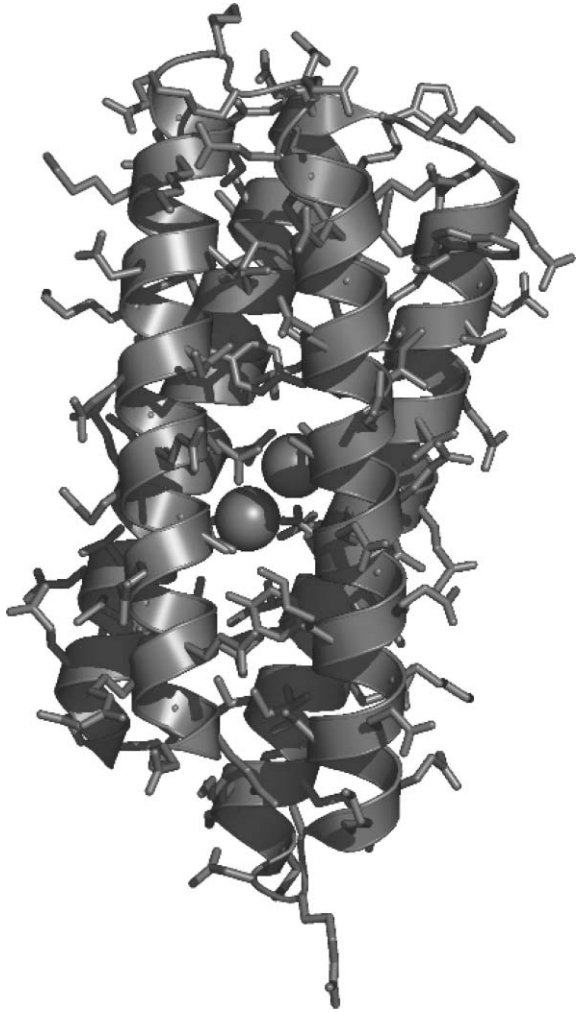

**Fig. 5** Rendering[241] of the putative structure of DFsc, a *de novo* designed 114-residue dinuclear metalloprotein.[183]

ferrin (DF1), an antiparallel dimer of two helix bundles whose structure was solved to high resolution by NMR and X-ray crystallography.[224] In order to generate a monomer starting from two antiparallel dimers, the helices were extended, two short loops were added and one of the original loops was deleted. The rearrangement of topology avoids the long intervening loop between the second and third helices that is observed in the natural dinuclear metalloproteins. During the calculation, the amino acid identities of a subset of 26 residues were constrained, where these residues participate in metal binding, facillitate access to the active site, initiate a helix, or form a turn sequence. The rest of the protein, a total of 88 residues, was designed using

SCADS to identify residues compatible with the structure. The calculation converged to a sequence that was ~30% homologous to the sequence of DF1 after two rounds of calculation. The number of residues that were subjected to simultaneous randomization was greater than in most other design studies discussed thus far. This was made possible by the use of an algorithm that computes the site-specific amino acid probabilities rather than optimizing individual sequences through explicit enumeration. In addition to being well-structured, the resulting protein also exhibits catalytic activity with regard to known peroxidase substrates.

In another application of a SCADS-based approach, the hydrophobic transmembrane domain of a potassium channel KcsA was re-engineered so as to construct a water-soluble version of the protein.[225] The known crystal structure of a potassium channel (PDB code 1K4C) comprising four transmembrane helical subunits packed as a tetramer was used as the template, and this structure comprises four transmembrane helical subunits packed as a tetramer. SCADS was then applied to redesign the exposed surface residues of the transmembrane domain to make the protein water-soluble. From the protein surface, 35 residues were selected for mutation. The computation was constrained using the environmental energy (a scale for the solvation propensities of the amino acids), which was fixed to a value that is representative of a water-soluble protein of comparable size. A sequence WSK-1 was identified. The protein is well expressed in soluble form in *E. coli*. When the protein was tested by analytical gel filtration, however, the protein was shown to have a tendency to form larger oligomers in addition to the expected tetramers. Hydrophobic patches found on the surface of the modeled protein and were suspected to cause the observed aggregation. Further redesign resulted in a protein containing 29 designed mutations in each of the four 104-residue protomers and removed the aggregation problem, leaving the protein mostly tetrameric. The newly designed protein was shown to have the functionally related toxin binding properties of the membrane soluble wild type, based on an assay with scorpion toxin that binds specifically to the extracellular domain of a potassium channel.

In an impressive achievement in protein design, a 97-residue α/β protein Top7 with a novel fold was successfully designed by Baker and coworkers (Fig. 6).[226] Thus a globular protein fold not found in nature is physically possible. With regard to the design and realization of novel, nonnatural protein structures, this extends the previous discovery of a right-handed helical coiled-coil[79] and now includes non-helical proteins. The design protocol consisted of cycling between sequence design and backbone optimization. At first, a rough two-dimensional diagram was created to describe the overall topology, and constraints were specified to define the topology. Three-dimensional models were then generated by assembling residue fragments from the PDB with secondary structures consistent with the desired topology. A sequence was designed using the RosettaDesign MC algorithm with a Lennard-Jones potential, an orientation-dependent hydrogen bonding term, and an implicit solvation model. The backbone optimization allowed the identification of the lowest free energy backbone conformation for a fixed amino acid sequence, again using a MC minimization algorithm. During the backbone optimization, low energy side chain conformations for a fixed sequence were also explored to replace high energy conformations caused by backbone adjustment. For each of the five initial structures, 15 cycles of sequence design and backbone optimization were used to

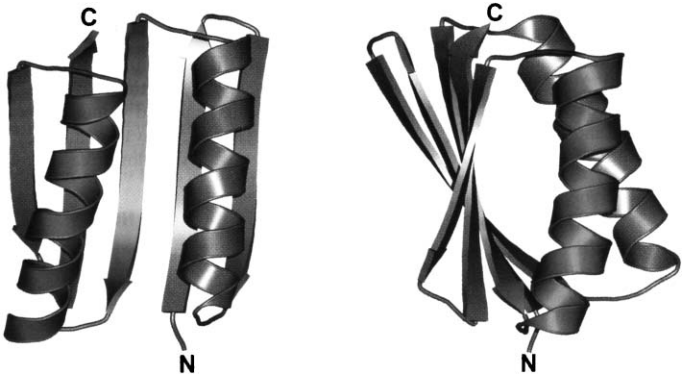


**Fig. 6** Designing a *de novo* protein Top7 with a new topology. Two views rotated by 90° from each other. (Re-printed with permission from ref. 226.)

obtain low energy sequence–structure pairs. Interestingly, they observed that a dampened Lennard-Jones repulsive term and a MC optimization without the minimization step resulted in a stable but molten globule-like core, questioning the practice of rescaling the atomic radii to soften van der Waals repulsion at short distances.[123] The crystal structure of Top7 at 2.5 Å resolution revealed that Top7 adopts the designed topology with 1.17 Å RMSD over all backbone atoms. The similarity between the designed and predicted structure is a validation of the utility and transferability of the energy function, which had been partially parameterized using known protein structures and sequences.

## Beyond peptides and proteins

Whereas proteins use a fixed set of 20 α-amino acids as building blocks, the potential monomers for non-biological folding polymers, foldamers,[7,227] are more numerous in numbers and diverse in chemical properties, making the three-dimensional structures of non-biological polymers harder to predict. A key problem in foldamer research is identifying those heterosequences that are likely to yield interesting, well-formed folded structures. Nevertheless, the lessons learned from studying proteins have been applied to the design of novel folding polymers. For example, the stabilization of protein backbone by a network of hydrogen bonds has inspired the creation of several foldamers that are similarly stabilized through a hydrogen bonding network, including vinylogous peptides,[228,229] α,α′-di-substituted-α amino acid peptides,[230] β-amino acid peptides,[231,232] γ-amino acid peptides (γ-peptides),[233] and trispyridylamide scaffold.[234] In addition to hydrogen bonding, aromatic stacking can also stabilize well-defined secondary structures.[234,235] Moore and coworkers have shown that (*m*-phenylene-ethynylene)$_n$ (oligo-PE) where $n > 8$, forms a helical structure in a variety of solvents.[236] The aromatic π–π stacking is thought to provide the main driving force for the helix formation (Fig. 7). Interestingly, another group has reported the formation of β-structure using a related backbone.[237] The functional

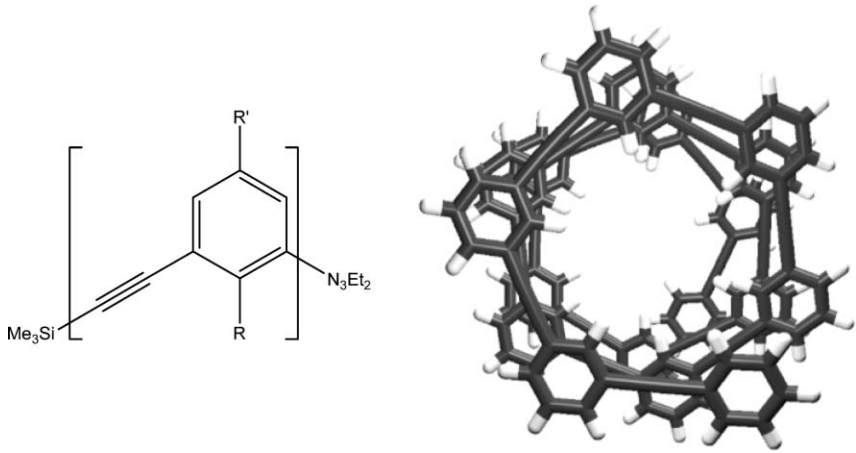


**Fig. 7** (Left) Chemical structure of oligo-phenylene-ethynylene (Oligo-PE). (Right) A view of oligo-PE helix with R = H, R′ = H from above.

potential of non-biological polymers have also been investigated by several groups[234,238] and the helical oligo-PE can bind small chiral molecules in its interior tubular cavity.[239]

The successful design of non-biological foldamers is a natural extension to quantitative protein design. A theoretical approach similar to one used to design proteins can be used to design non-biological polymers with novel properties. Non-biological foldamers may also be elaborated to yield new materials, *e.g.*, helical nanotubules.[240] The generalizability of computational protein design to other related disciplines will be amply demonstrated in the application of a common theoretical framework to study both biological and non-biological polymers.

## Conclusion

Successful protein design poses many hurdles: the many degrees of freedom involving both sequence and local structure that lead to the combinatorial complexity of the search for viable sequences; the subtlety of the underlying physical forces that stabilize folded structures; and our incomplete understanding of the determinants of folding. These impediments pose challenges when designing novel proteins since subtle features of protein folding may be overlooked and result in a sequence that either fails to fold as expected or has other undesirable properties. Computational protein design seeks to remedy gaps in our intuition by codifying many fundamental rules governing protein folding and using efficient algorithms to search and characterize the range of possible sequences for a given target structure. In recent years, this quantitative approach to protein design has gained momentum with the development of a number of high quality sequence prediction algorithms, and in many cases, these efforts have led to milestone successes that have contributed to the design of new proteins with novel properties. With continued success, we may have a far better

understanding of proteins and the principles of protein folding in the near future, bringing within reach the creation of custom-made functional proteins and the re-engineering of natural proteins to facilitate detailed functional and structural studies.

## Acknowledgements

The authors acknowledge support from the US National Science Foundation (CHE 99-84752 and DMR 00-79909) and the National Institutes of Health (GM61267). J. G. S. is a Cottrell Scholar of Research Corporation and an Arnold and Mabel Beckman Foundation Young Investigator.

## References

1 C. Pabo, *Nature*, 1983, **301**, 200.
2 J. D. Bryngelson and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA*, 1987, **84**, 7524.
3 K. Bloch, *Science*, 1965, **150**, 19.
4 R. M. De Lorimier, J. J. Smith, M. A. Dwyer, L. L. Looger, K. M. Sali, C. D. Paavola, S. S. Rizk, S. Sadigov, D. W. Conrad, L. Loew and H. W. Hellinga, *Protein Sci.*, 2002, **11**, 2655.
5 R. T. Piervincenzi, W. M. Reichert and H. W. Hellinga, *Biosens. Bioelectron.*, 1998, **13**, 305.
6 http://www.lrsm.upenn.edu/research/.
7 S. H. Gellman, *Acc. Chem. Res.*, 1998, **31**, 173.
8 K. A. Dill, *Biochemistry*, 1990, **29**, 7133.
9 D. Eisenberg, W. Wilcox, S. M. Eshita, P. M. Pryciak, S. P. Ho and W. F. DeGrado, *Proteins*, 1986, **1**, 16.
10 C. P. Hill, D. H. Anderson, L. Wesson, W. F. DeGrado and D. Eisenberg, *Science*, 1990, **249**, 543.
11 L. Regan and W. F. DeGrado, *Science*, 1988, **241**, 976.
12 S. P. Ho and W. F. DeGrado, *J. Am. Chem. Soc.*, 1989, **107**, 1987.
13 D. P. Raleigh and W. F. Degrado, *J. Am. Chem. Soc.*, 1992, **114**, 10079.
14 B. Lovejoy, S. Choe, D. Cascio, D. K. McRorie, W. F. DeGrado and D. Eisenberg, *Science*, 1993, **259**, 1288.
15 E. K. O'Shea, R. Rutkowski and P. S. Kim, *Science*, 1989, **243**, 538.
16 E. K. O'Shea, J. D. Klemm, P. S. Kim and T. Alber, *Science*, 1991, **254**, 539.
17 J. A. Boice, G. R. Dieckmann, W. F. DeGrado and R. Fairman, *Biochemistry*, 1996, **35**, 14480.
18 P. B. Harbury, T. Zhang, P. S. Kim and T. Alber, *Science*, 1993, **262**, 1401.
19 G. D. Rose and T. P. Creamer, *Proteins*, 1994, **19**, 1.
20 S. Dalal, S. Balasubramanian and L. Regan, *Nat. Struct. Biol.*, 1997, **4**, 548.
21 C. E. Schafmeister, L. J. Miercke and R. M. Stroud, *Science*, 1993, **262**, 734.
22 C. E. Schafmeister, S. L. LaPorte, L. J. Miercke and R. M. Stroud, *Nat. Struct. Biol.*, 1997, **4**, 1039.
23 M. J. Root, M. S. Kay and P. S. Kim, *Science*, 2001, **291**, 884.
24 D. C. Chan, D. Fass, J. M. Berger and P. S. Kim, *Cell*, 1997, **89**, 263.
25 H. W. Hellinga, *Curr. Opin. Biotechnol.*, 1996, **7**, 437.
26 L. Regan and N. D. Clarke, *Biochemistry*, 1990, **29**, 10878.
27 E. Farinas and L. Regan, *Protein Sci.*, 1998, **7**, 1939.
28 W. Yang, L. M. Jones, L. Isley, Y. Ye, H. W. Lee, A. Wilkins, Z. R. Liu, H. W. Hellinga, R. Malchow, M. Ghazi and J. J. Yang, *J. Am. Chem. Soc.*, 2003, **125**, 6165.
29 R. Schnepf, P. Horth, E. Bill, K. Wieghardt, P. Hildebrandt and W. Haehnel, *J. Am. Chem. Soc.*, 2001, **123**, 2186.
30 X. Li, K. Suzuki, K. Kanaori, K. Tajima, A. Kashiwada, H. Hiroaki, D. Kohda and T. Tanaka, *Protein Sci.*, 2000, **9**, 1327.
31 G. R. Dieckmann, D. K. McRorie, D. L. Tierney, L. M. Utschig, C. P. Singer, T. V. Ohalloran, J. E. PennerHahn, W. F. DeGrado and V. L. Pecoraro, *J. Am. Chem. Soc.*, 1997, **119**, 6195.
32 A. Lombardi, C. M. Summa, S. Geremia, L. Randaccio, V. Pavone and W. F. DeGrado, *Proc. Natl. Acad. Sci. USA*, 2000, **97**, 6298.
33 M. Klemba, K. H. Gardner, S. Marino, N. D. Clarke and L. Regan, *Nat. Struct. Biol.*, 1995, **2**, 368.
34 W. D. Kohn, C. M. Kay, B. D. Sykes and R. S. Hodges, *J. Am. Chem. Soc.*, 1998, **120**, 1124.
35 T. M. Handel, S. A. Williams and W. F. DeGrado, *Science*, 1993, **261**, 879.
36 C. Sissi, P. Rossi, F. Felluga, F. Formaggio, M. Palumbo, P. Tecilla, C. Toniolo and P. Scrimin, *J. Am. Chem. Soc.*, 2001, **123**, 3169.

37 D. E. Robertson, R. S. Farid, C. C. Moser, J. L. Urbauer, S. E. Mulholland, R. Pidikiti, J. D. Lear, A. J. Wand, W. F. DeGrado and P. L. Dutton, *Nature*, 1994, **368**, 425.
38 D. A. Moffet, L. K. Certain, A. J. Smith, A. J. Kessel, K. A. Beckwith and M. H. Hecht, *J. Am. Chem. Soc.*, 2000, **122**, 7612.
39 H. K. Rau, N. DeJonge and W. Haehnel, *Angew. Chem., Int. Ed.*, 2000, **39**, 250.
40 H. K. Rau, N. DeJonge and W. Haehnel, *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 11526.
41 S. E. Mulholland, B. R. Gibney, F. Rabanal and P. L. Dutton, *Biochemistry*, 1999, **38**, 10442.
42 A. Knappik, L. M. Ge, A. Honegger, P. Pack, M. Fischer, G. Wellnhofer, A. Hoess, J. Wolle, A. Pluckthun and B. Virnekas, *J. Mol. Biol.*, 2000, **296**, 57.
43 M. Zaccolo, D. M. Williams, D. M. Brown and E. Gherardi, *J. Mol. Biol.*, 1996, **255**, 589.
44 W. P. Stemmer, *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 10747.
45 W. P. Stemmer, *Nature*, 1994, **370**, 389.
46 W. A. Lim and R. T. Sauer, *Nature*, 1989, **339**, 31.
47 L. M. Gregoret and R. T. Sauer, *Proc. Natl. Acad. Sci. USA*, 1993, **90**, 4246.
48 A. R. Davidson and R. T. Sauer, *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 2146.
49 A. R. Davidson, K. J. Lumb and R. T. Sauer, *Nat. Struct. Biol.*, 1995, **2**, 856.
50 S. Kamtekar, J. M. Schiffer, H. Y. Xiong, J. M. Babik and M. H. Hecht, *Science*, 1993, **262**, 1680.
51 Y. Wei, T. Liu, S. L. Sazinsky, D. A. Moffet, I. Pelczer and M. H. Hecht, *Protein Sci.*, 2003, **12**, 92.
52 Y. Wei, S. Kim, D. Fela, J. Baum and M. H. Hecht, *Proc. Natl. Acad. Sci. USA*, 2003, **100**, 13270.
53 M. W. West, W. Wang, J. Patterson, J. D. Mancias, J. R. Beasley and M. H. Hecht, *Proc. Natl. Acad. Sci. USA*, 1999, **96**, 11211.
54 B. M. Broome and M. H. Hecht, *J. Mol. Biol.*, 2000, **296**, 961.
55 D. A. Moffet and M. H. Hecht, *Chem. Rev.*, 2001, **101**, 3191.
56 R. H. Hoess, *Chem. Rev.*, 2001, **101**, 3205.
57 T. N. Nguyen, M. Hansson, S. Stahl, T. Bachi, A. Robert, W. Domzig, H. Binz and M. Uhlen, *Gene*, 1993, **128**, 89.
58 E. T. Boder and K. D. Wittrup, *Nat. Biotechnol.*, 1997, **15**, 553.
59 J. Hanes, C. Schaffitzel, A. Knappik and A. Pluckthun, *Nat. Biotechnol.*, 2000, **18**, 1287.
60 A. C. Braisted and J. A. Wells, *Proc. Natl. Acad. Sci. USA*, 1996, **93**, 5688.
61 L. Giver, A. Gershenson, P. O. Freskgard and F. H. Arnold, *Proc. Natl. Acad. Sci. USA*, 1998, **95**, 12809.
62 G. Xia, L. Chen, T. Sera, M. Fa, P. G. Schultz and F. E. Romesberg, *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 6597.
63 S. J. Pollack, J. W. Jacobs and P. G. Schultz, *Science*, 1986, **234**, 1570.
64 P. G. Schultz and R. A. Lerner, *Science*, 1995, **269**, 1835.
65 K. M. Shokat and P. G. Schultz, *Methods Enzymol.*, 1991, **203**, 327.
66 B. J. Bevis and B. S. Glick, *Nat. Biotechnol.*, 2002, **20**, 83.
67 G. MacBeath, P. Kast and D. Hilvert, *Science*, 1998, **279**, 1958.
68 A. P. Brunet, E. S. Huang, M. E. Huffine, J. E. Loeb, R. J. Weltman and M. H. Hecht, *Nature*, 1993, **364**, 355.
69 J. W. Bryson, J. R. Desjarlais, T. M. Handel and W. F. DeGrado, *Protein Sci.*, 1998, **7**, 1404.
70 T. P. Quinn, N. B. Tweedy, R. W. Williams, J. S. Richardson and D. C. Richardson, *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 8747.
71 J. W. Bryson, S. F. Betz, H. S. Lu, D. J. Suich, H. X. Zhou, K. T. O'Neil and W. F. DeGrado, *Science*, 1995, **270**, 935.
72 S. Dalal, S. Balasubramanian and L. Regan, *Fold. Des.*, 1997, **2**, R71.
73 M. H. Cordes, N. P. Walsh, C. J. McKnight and R. T. Sauer, *Science*, 1999, **284**, 325.
74 M. H. Cordes, R. E. Burton, N. P. Walsh, C. J. McKnight and R. T. Sauer, *Nat. Struct. Biol.*, 2000, **7**, 1129.
75 D. L. Minor, Jr. and P. S. Kim, *Nature*, 1996, **380**, 730.
76 B. Kuhlman, J. W. O'Neill, D. E. Kim, K. Y. Zhang and D. Baker, *J. Mol. Biol.*, 2002, **315**, 471.
77 J. R. Desjarlais and T. M. Handel, *Protein Sci.*, 1995, **4**, 2006.
78 P. B. Harbury, B. Tidor and P. S. Kim, *Proc. Natl. Acad. Sci. USA*, 1995, **92**, 8408.
79 P. B. Harbury, J. J. Plecs, B. Tidor, T. Alber and P. S. Kim, *Science*, 1998, **282**, 1462.
80 H. Li, R. Helling, C. Tang and N. Wingreen, *Science*, 1996, **273**, 666.
81 N. Nagano, C. A. Orengo and J. M. Thornton, *J. Mol. Biol.*, 2002, **321**, 741.
82 J. L. England, B. E. Shakhnovich and E. I. Shakhnovich, *Proc. Natl. Acad. Sci. USA*, 2003, **100**, 8727.
83 J. Miller, C. Zeng, N. S. Wingreen and C. Tang, *Proteins*, 2002, **47**, 506.
84 R. L. Dunbrack, Jr., *Curr. Opin. Struct. Biol.*, 2002, **12**, 431.
85 K. Fan and W. Wang, *J. Mol. Biol.*, 2003, **328**, 921.
86 D. S. Riddle, J. V. Santiago, S. T. Bray-Hall, N. Doshi, V. P. Grantcharova, Q. Yi and D. Baker, *Nat. Struct. Biol.*, 1997, **4**, 805.
87 S. Akanuma, T. Kigawa and S. Yokoyama, *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 13549.
88 S. A. Marshall and S. L. Mayo, *J. Mol. Biol.*, 2001, **305**, 619.

89 J. W. Chin, T. A. Cropp, J. C. Anderson, M. Mukherji, Z. Zhang and P. G. Schultz, *Science*, 2003, **301**, 964.
90 D. Datta, P. Wang, I. S. Carrico, S. L. Mayo and D. A. Tirrell, *J. Am. Chem. Soc.*, 2002, **124**, 5652.
91 R. Chandrasekaran and G. N. Ramachandran, *Int. J. Protein Res.*, 1970, **2**, 223.
92 J. Janin and S. Wodak, *J. Mol. Biol.*, 1978, **125**, 357.
93 T. N. Bhat, V. Sasisekharan and M. Vijayan, *Int. J. Pept. Protein Res.*, 1979, **13**, 170.
94 M. N. James and A. R. Sielecki, *J. Mol. Biol.*, 1983, **163**, 299.
95 E. Benedetti, G. Morelli, G. Nemethy and H. A. Scheraga, *Int. J. Pept. Protein Res.*, 1983, **22**, 1.
96 J. W. Ponder and F. M. Richards, *J. Mol. Biol.*, 1987, **193**, 775.
97 M. J. McGregor, S. A. Islam and M. J. Sternberg, *J. Mol. Biol.*, 1987, **198**, 295.
98 P. Tuffery, C. Etchebest, S. Hazout and R. Lavery, *J. Biomol. Struct. Dyn.*, 1991, **8**, 1267.
99 R. L. Dunbrack, Jr. and M. Karplus, *J. Mol. Biol.*, 1993, **230**, 543.
100 H. Schrauber, F. Eisenhaber and P. Argos, *J. Mol. Biol.*, 1993, **230**, 592.
101 H. Kono and J. Doi, *J. Comput. Chem.*, 1996, **17**, 1667.
102 M. De Maeyer, J. Desmet and I. Lasters, *Fold. Des.*, 1997, **2**, 53.
103 R. Dunbrack and F. E. Cohen, *Protein Sci.*, 1997, **6**, 1661.
104 Z. Xiang and B. Honig, *J. Mol. Biol.*, 2001, **311**, 421.
105 S. C. Lovell, J. M. Word, J. S. Richardson and D. C. Richardson, *Proteins*, 2000, **40**, 389.
106 J. Heringa and P. Argos, *Proteins*, 1999, **37**, 44.
107 J. Heringa and P. Argos, *Proteins*, 1999, **37**, 30.
108 S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta and P. Weiner, *J. Am. Chem. Soc.*, 1984, **106**, 765.
109 B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, *J. Comput. Chem.*, 1983, **4**, 187.
110 J. Hermans, H. J. C. Berendsen, W. F. Vangunsteren and J. P. M. Postma, *Biopolymers*, 1984, **23**, 1513.
111 N. Gibbs, A. R. Clarke and R. B. Sessions, *Proteins*, 2001, **43**, 186.
112 D. B. Gordon, S. A. Marshall and S. L. Mayo, *Curr. Opin. Struct. Biol.*, 1999, **9**, 509.
113 A. E. Keating, V. N. Malashkevich, B. Tidor and P. S. Kim, *Proc. Natl. Acad. Sci. USA*, 2001, **98**, 14825.
114 B. I. Dahiyat, D. B. Gordon and S. L. Mayo, *Protein Sci.*, 1997, **6**, 1333.
115 D. Eisenberg and A. McLachlan, *Nature*, 1986, **319**, 199.
116 K. A. Sharp, A. Nicholls, R. Friedman and B. Honig, *Biochemistry*, 1991, **30**, 9686.
117 M. S. Lee, M. Feig, F. R. Salsbury, Jr. and C. L. Brooks, *J. Comput. Chem.*, 2003, **24**, 1348.
118 W. Im, M. S. Lee and C. L. Brooks, *J. Comput. Chem.*, 2003, **24**, 1691.
119 O. Guvench, J. Weiser, P. Shenkin, I. Kolossvary and W. C. Still, *J. Comput. Chem.*, 2002, **23**, 214.
120 L. Wesson and D. Eisenberg, *Protein Sci.*, 1992, **1**, 227.
121 T. Ooi, M. Oobatake, G. Nemethy and H. A. Scheraga, *Proc. Natl. Acad. Sci. USA*, 1987, **84**, 3086.
122 P. Koehl and M. Delarue, *Proteins*, 1994, **20**, 264.
123 B. I. Dahiyat and S. L. Mayo, *Protein Sci.*, 1996, **5**, 895.
124 P. Koehl and M. Levitt, *J. Mol. Biol.*, 1999, **293**, 1161.
125 N. Kurochkina and B. Lee, *Protein Eng.*, 1995, **8**, 437.
126 S. Takada, Z. Luthey-Schulten and P. G. Wolynes, *J. Chem. Phys.*, 1999, **110**, 11616.
127 H. Kono and J. G. Saven, *J. Mol. Biol.*, 2001, **306**, 607.
128 J. L. Fauchere and V. Pliska, *Eur. J. Med. Chem.*, 1983, **18**, 369.
129 P. Y. Chou and G. D. Fasman, *Biochemistry*, 1974, **13**, 211.
130 J. S. Richardson and D. C. Richardson, *Science*, 1988, **240**, 1648.
131 K. T. O'Neil and W. F. Degrado, *Science*, 1990, **250**, 646.
132 A. Chakrabartty, T. Kortemme and R. L. Baldwin, *Protein Sci.*, 1994, **3**, 843.
133 P. C. Lyu, M. I. Liff, L. A. Marky and N. R. Kallenbach, *Science*, 1990, **250**, 669.
134 S. Padmanabhan, S. Marqusee, T. Ridgeway, T. M. Laue and R. L. Baldwin, *Nature*, 1990, **344**, 268.
135 T. P. Creamer and G. D. Rose, *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 5937.
136 T. P. Creamer and G. D. Rose, *Proteins*, 1994, **19**, 85.
137 C. K. Smith, J. M. Withka and L. Regan, *Biochemistry*, 1994, **33**, 5510.
138 C. A. Kim and J. M. Berg, *Nature*, 1993, **362**, 267.
139 D. L. Minor, Jr. and P. S. Kim, *Nature*, 1994, **367**, 660.
140 D. L. Minor, Jr. and P. S. Kim, *Nature*, 1994, **371**, 264.
141 C. K. Smith and L. Regan, *Science*, 1995, **270**, 980.
142 A. G. Street and S. L. Mayo, *Proc. Natl. Acad. Sci. USA*, 1999, **96**, 9074.
143 W. F. DeGrado, C. M. Summa, V. Pavone, F. Nastri and A. Lombardi, *Annu. Rev. Biochem.*, 1999, **68**, 779.
144 G. D. Rose, L. M. Gierasch and J. A. Smith, *Adv. Protein Chem.*, 1985, **37**, 1.
145 K. Gunasekaran, C. Ramakrishnan and P. Balaram, *Protein Eng.*, 1997, **10**, 1131.
146 B. L. Sibanda and J. M. Thornton, *Nature*, 1985, **316**, 170.
147 H. X. Zhou, R. H. Hoess and W. F. DeGrado, *Nat. Struct. Biol.*, 1996, **3**, 446.
148 M. H. Hecht, J. S. Richardson, D. C. Richardson and R. C. Ogden, *Science*, 1990, **249**, 884.

149 H. W. Hellinga, *Proc. Natl. Acad. Sci. USA*, 1997, **94**, 10015.
150 J. G. Saven, *Chem. Rev.*, 2001, **101**, 3113.
151 J. N. Onuchic, Z. Luthey-Schulten and P. G. Wolynes, *Annu. Rev. Phys. Chem.*, 1997, 48.
152 E. I. Shakhnovich, *Fold. Des.*, 1998, **3**, R45.
153 C. D. Waldburger, J. F. Schildbach and R. T. Sauer, *Nat. Struct. Biol.*, 1995, **2**, 122.
154 J. S. Richardson and D. C. Richardson, *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 2754.
155 W. Wang and M. H. Hecht, *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 2760.
156 K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich and K. A. Dill, *Proc. Natl. Acad. Sci. USA*, 1995, **92**, 325.
157 H. W. Hellinga and F. M. Richards, *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 5803.
158 B. Kuhlman and D. Baker, *Proc. Natl. Acad. Sci. USA*, 2000, **97**, 10383.
159 N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, *J. Chem. Phys.*, 1953, **21**, 1087.
160 S. Kirkpatrick, C. D. Gelatt and M. P. Vecchi, *Science*, 1983, **220**, 671.
161 D. R. Westhead, D. E. Clark and C. W. Murray, *J. Comput. Aided Mol. Des.*, 1997, **11**, 209.
162 C. Lee and S. Subbiah, *J. Mol. Biol.*, 1991, **217**, 373.
163 H. W. Hellinga and F. M. Richards, *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 5803.
164 C. A. Voigt, D. B. Gordon and S. L. Mayo, *J. Mol. Biol.*, 2000, **299**, 789.
165 A. P. Cootes, P. M. G. Curmi and A. E. Torda, *J. Chem. Phys.*, 2000, **113**, 2489.
166 J. Zou and J. G. Saven, *J. Chem. Phys.*, 2003, **118**, 3843.
167 J. H. Holland, *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*, University of Michigan Press, 1975.
168 J. Desmet, M. De Maeyer, B. Hazes and I. Lasters, *Nature*, 1992, **356**, 539.
169 S. Liang and N. V. Grishin, *Protein Sci.*, 2002, **11**, 322.
170 A. R. Leach, *J. Mol. Biol.*, 1994, **235**, 345.
171 R. F. Goldstein, *Biophys J.*, 1994, **66**, 1335.
172 B. I. Dahiyat and S. L. Mayo, *Science*, 1997, **278**, 82.
173 L. L. Looger and H. W. Hellinga, *J. Mol. Biol.*, 2001, **307**, 429.
174 N. A. Pierce, J. A. Spriet, J. Desmet and S. L. Mayo, *J. Comput. Chem.*, 2000, **21**, 999.
175 H. A. Bethe, *Proc. R. Soc. London, Ser. A*, 1935, **150**, 552.
176 W. L. Bragg and E. J. Williams, *Proc. R. Soc. London, Ser. A*, 1934, **145**, 699.
177 P. Koehl and M. Delarue, *J. Mol. Biol.*, 1994, **239**, 249.
178 C. Lee, *J. Mol. Biol.*, 1994, **236**, 918.
179 M. Vasquez, *Biopolymers*, 1995, **36**, 53.
180 J. Mendes, C. M. Soares and M. A. Carrondo, *Biopolymers*, 1999, **50**, 111.
181 J. M. Zou and J. G. Saven, *J. Mol. Biol.*, 2000, **296**, 281.
182 W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, *Numerical Recipes*, Cambridge University Press, Cambridge, 1992.
183 J. R. Calhoun, H. Kono, S. Lahr, W. Wang, W. F. DeGrado and J. G. Saven, *J. Mol. Biol.*, 2003, **334**, 1101.
184 P. E. Correa, *Proteins*, 1990, **7**, 366.
185 L. Holm and C. Sander, *J. Mol. Biol.*, 1991, **218**, 183.
186 D. D. Axe, N. W. Foster and A. R. Fersht, *Proc. Natl. Acad. Sci. USA*, 1996, **93**, 5590.
187 N. C. Gassner, W. A. Baase and B. W. Matthews, *Proc. Natl. Acad. Sci. USA*, 1996, **93**, 12155.
188 Y. Isogai, M. Ota, A. Ishii, M. Ishida and K. Nishikawa, *Protein Eng.*, 2002, **15**, 555.
189 M. Munson, S. Balasubramanian, K. G. Fleming, A. D. Nagi, R. O'Brien and L. Regan, *Protein Sci.*, 1996, **5**, 1584.
190 G. A. Lazar, J. R. Desjarlais and T. M. Handel, *Protein Sci.*, 1997, **6**, 1167.
191 X. Jiang, H. Farid, E. Pistor and R. S. Farid, *Protein Sci.*, 2000, **9**, 403.
192 J. H. Hurley, W. A. Baase and B. W. Matthews, *J. Mol. Biol.*, 1992, **224**, 1143.
193 E. P. Baldwin, O. Hajiseyedjavadi, W. A. Baase and B. W. Matthews, *Science*, 1993, **262**, 1715.
194 J. J. Havranek and P. B. Harbury, *Nat. Struct. Biol.*, 2003, **10**, 45.
195 S. T. R. Walsh, H. Cheng, J. W. Bryson, H. Roder and W. F. DeGrado, *Proc. Natl. Acad. Sci. USA*, 1999, **96**, 5486.
196 S. T. Walsh, A. L. Lee, W. F. DeGrado and A. J. Wand, *Biochemistry*, 2001, **40**, 9560.
197 T. Kortemme, M. Ramirez-Alvarado and L. Serrano, *Science*, 1998, **281**, 253.
198 M. Lopez de la Paz, E. Lacroix, M. Ramirez-Alvarado and L. Serrano, *J. Mol. Biol.*, 2001, **312**, 229.
199 G. Nemethy, M. S. Pottle and H. A. Scheraga, *J. Phys. Chem.*, 1983, **87**, 1883.
200 I. Angrand, L. Serrano and E. Lacroix, *Biomol. Eng.*, 2001, **18**, 125.
201 M. Lopez de la Paz, K. Goldie, J. Zurdo, E. Lacroix, C. M. Dobson, A. Hoenger and L. Serrano, *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 16052.
202 T. N. Pham and S. Koide, *J. Biomol. NMR*, 1998, **11**, 407.
203 C. M. Kraemer-Pecore, J. T. Lecomte and J. R. Desjarlais, *Protein Sci.*, 2003, **12**, 2194.
204 I. Landrieu, J. M. Wieruszeski, R. Wintjens, D. Inze and G. Lippens, *J. Mol. Biol.*, 2002, **320**, 321.
205 S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, *J. Mol. Biol.*, 1990, **215**, 403.

206 P. Strop and S. L. Mayo, *J. Am. Chem. Soc.*, 1999, **121**, 2341.
207 M. L. Connolly, *Science*, 1983, **221**, 709.
208 R. J. Hayes, J. Bentzien, M. L. Ary, M. Y. Hwang, J. M. Jacinto, J. Vielmetter, A. Kundu and B. I. Dahiyat, *Proc. Natl. Acad. Sci. USA*, 2002, **99**, 15926.
209 M. Shimaoka, J. M. Shifman, H. Jing, L. Takagi, S. L. Mayo and T. A. Springer, *Nat. Struct. Biol.*, 2000, **7**, 674.
210 J. M. Shifman and S. L. Mayo, *J. Mol. Biol.*, 2002, **323**, 417.
211 W. E. Meador, A. R. Means and F. A. Quiocho, *Science*, 1993, **262**, 1718.
212 W. E. Meador, A. R. Means and F. A. Quiocho, *Science*, 1992, **257**, 1251.
213 M. Ikura, G. M. Clore, A. M. Gronenborn, G. Zhu, C. B. Klee and A. Bax, *Science*, 1992, **256**, 632.
214 J. Reina, E. Lacroix, S. D. Hobson, G. Fernandez-Ballester, V. Rybin, M. S. Schwab, L. Serrano and C. Gonzalez, *Nat. Struct. Biol.*, 2002, **9**, 621.
215 S. Schneider, M. Buchert, O. Georgiev, B. Catimel, M. Halford, S. A. Stacker, T. Baechi, K. Moelling and C. M. Hovens, *Nat. Biotechnol.*, 1999, **17**, 170.
216 L. L. Looger, M. A. Dwyer, J. J. Smith and H. W. Hellinga, *Nature*, 2003, **423**, 185.
217 A. R. Fersht, *Structure and Mechanismm in Protein Science*, Freeman, New York, 1999.
218 C. D. Coldren, H. W. Hellinga and J. P. Caradonna, *Proc. Natl. Acad. Sci. USA*, 1997, **94**, 6635.
219 D. E. Benson, M. S. Wisz, W. T. Liu and H. W. Hellinga, *Biochemistry*, 1998, **37**, 7070.
220 A. L. Pinto, H. W. Hellinga and J. P. Caradonna, *Proc. Natl. Acad. Sci. USA*, 1997, **94**, 5562.
221 D. N. Bolon and S. L. Mayo, *Proc. Natl. Acad. Sci. USA*, 2001, **98**, 14274.
222 F. Offredi, F. Dubail, P. Kischel, K. Sarinski, A. S. Stern, C. Van de Weerdt, J. C. Hoch, C. Prosperi, J. M. Francois, S. L. Mayo and J. A. Martial, *J. Mol. Biol.*, 2003, **325**, 163.
223 W. Jin, O. Kambara, H. Sasakawa, A. Tamura and S. Takada, *Struct. Fold. Des.*, 2003, **11**, 581.
224 O. Maglio, F. Nastri, V. Pavone, A. Lombardi and W. F. DeGrado, *Proc. Natl. Acad. Sci. USA*, 2003, **100**, 3772.
225 A. M. Slovic, H. Kono, J. D. Lear, J. G. Saven and W. F. DeGrado, *Proc. Natl. Acad. Sci. USA*, 2004, **101**, 1828.
226 B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard and D. Baker, *Science*, 2003, **302**, 1364.
227 D. J. Hill, M. J. Mio, R. B. Prince, T. S. Hughes and J. S. Moore, *Chem. Rev.*, 2001, **101**, 3893.
228 M. Hagihara, N. J. Anthony, T. J. Stout, J. Clardy and S. L. Schreiber, *J. Am. Chem. Soc.*, 1992, **114**, 6568.
229 C. Gennari, B. Salom, D. Potenza and A. Williams, *Angew. Chem., Int. Ed. Engl.*, 1994, **33**, 2067.
230 M. Tanaka, *J. Synth. Org. Chem. Jpn.*, 2002, **60**, 125.
231 D. H. Appella, L. A. Christianson, D. A. Klein, D. R. Powell, X. Huang, J. J. Barchi and S. H. Gellman, *Nature*, 1997, **387**, 381.
232 R. P. Cheng and W. F. DeGrado, *J. Am. Chem. Soc.*, 2001, **123**, 5162.
233 G. P. Dado and S. H. Gellman, *J. Am. Chem. Soc.*, 1994, **116**, 1054.
234 J. T. Ernst, J. Becerril, H. S. Park, H. Yin and A. D. Hamilton, *Angew. Chem., Int. Ed.*, 2003, **42**, 535.
235 A. S. Shetty, J. Zhang and J. S. Moore, *J. Am. Chem. Soc.*, 1996, **118**, 1019.
236 J. C. Nelson, J. G. Saven, J. S. Moore and P. G. Wolynes, *Science*, 1997, **277**, 1793.
237 L. Arnt and G. N. Tew, *Langmuir*, 2003, **19**, 2404.
238 R. B. Prince, S. A. Barnes and J. S. Moore, *J. Am. Chem. Soc.*, 2000, **122**, 2758.
239 A. Tanatani, M. J. Mio and J. S. Moore, *J. Am. Chem. Soc.*, 2001, **123**, 1792.
240 M. J. Mio, R. B. Prince, J. S. Moore, C. Kuebel and D. C. Martin, *J. Am. Chem. Soc.*, 2000, **122**, 6134.
241 W. L. DeLano, The PyMOL Molecular Graphics System, DeLano Scientific, San Carlos, CA, USA, 2002, http://www.pymol.org.