Advances in computational protein design
Sheldon Park, Xi Yang and Jeffery G Saven*

Computational protein design continues to experience a variety of methodological advances. Several improvements have been suggested for the objective functions used to quantify sequence/structure compatibility. Disparate design strategies based upon dead-end elimination, simulated annealing and statistical design have each recently yielded striking successes involving de novo designed proteins with sizes on the order of 100 residues or greater. Such methods may be used to design new proteins, as well as to redesign natural proteins to facilitate structural and biophysical studies.

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

Current Opinion in Structural Biology 2004, 14:487–494
This review comes from a themed issue on Engineering and design  
Edited by Lars Baltzer and William F DeGrado
Available online 10th July 2004
0959-440X/5 – see front matter  
© 2004 Elsevier Ltd. All rights reserved.
DOI 10.1016/j.sbi.2004.06.002

Abbreviations
CPD computational protein design  
DEE dead-end elimination  
FDPB finite-difference Poisson–Boltzmann  
GA genetic algorithm  
MC Monte Carlo  
sCADs statistical computationally assisted design strategy

Introduction
Designing protein molecules with desired structural and functional characteristics critically assesses our understanding of protein sequence/structure relationships. Predictive tools for protein design also have a variety of applications in the development of new biotechnological therapeutics, materials and nanoscale systems. Because proteins are complex and numerous factors contribute to their stability and folding, sequence design efforts often face tough challenges. Computational protein design (CPD) can address large numbers of degrees of freedom and synthesize information from a variety of sources, such as structural databases and molecular potentials. To date, there are several notable examples that prove the effectiveness of CPD in designing novel protein sequences [1–3].

Difficult challenges for computation-driven sequence design remain, but many are being successfully addressed. The development of potential functions that are both fast and accurate is a topic of continued research. Modeling the role of solvent in protein stability is an important challenge, as it is usually impractical to treat solvent explicitly. Proper treatments of the electrostatic interactions and hydrogen bonds are important for designing sequences that resemble naturally occurring proteins. Further developments of sequence identification methods are necessary to efficiently sample and explore large sequence/rotamer spaces, where ‘rotamer’ refers to a particular sidechain conformation [4]. Variants of many of the algorithms that have been effective in solving other multiple minima problems, such as simulated annealing, genetic algorithms (GAs) and dead-end elimination (DEE), have seen success in designing sequences of large proteins [5]. An alternative to such sequence search methods involves statistical or probabilistic design, whereby the site-specific probabilities of the amino acids are calculated for a target structure and used to guide design [6]. Finally, several proteins have been designed by applying the principles of negative design, some of which have a smooth funnel-shaped free energy landscape [7].

Designing for foldability
Requisite properties of designed proteins are thermodynamic stability and kinetic foldability, that is, the ability to fold reversibly on a timescale comparable to that of natural proteins (microseconds to minutes). Although folding kinetics can certainly be investigated using carefully designed mutations, information concerning the kinetic details of the folding process need not be included explicitly in the design to arrive at fast folding sequences. Indeed, many proteins have been computationally designed by focusing on the consistency of interactions within only the target structure. Such de novo designed proteins are some of the fastest folding proteins known, with folding times on the order of 1–50 μs [8,9,10]. Nonetheless, even using just the free energy of folding to quantitatively design sequences is computationally prohibitive unless simplified models are used or the sequence diversity is severely limited. Hence, other simplified energy-based criteria are used to guide the design process.

Negative design
Folding implies the ability to distinguish between native and non-native structures. The folding of a protein to a unique, stable structure requires a sufficiently large free energy difference between native and non-native states. In particular, this difference must exceed the energy fluctuations observed among non-native states. However,
free energies are difficult to calculate computationally due to the extensive sampling required for accurate estimates of the unfolded state free energies. In order for the folded state of a particular sequence to be unique and thermodynamically stable, that structure must be separated energetically from other competing structures. Including information about such competing structures explicitly in the sequence design process is referred to as negative design. These concepts form the basis of the free energy landscape theory of protein folding, which postulates that naturally occurring proteins have a smooth funnel-like energy landscape to guide the folding of a protein to a well-defined minimum energy conformation [11].

In arriving at a completely redesigned helical protein, Jin et al. [12] applied the free energy landscape theory to design a three-helix bundle. Using a protein model with simplified sidechains, they generated an ensemble of denatured decoy states by performing folding simulations. To apply the principles of negative design, they used weighted Z-scores of candidate sequences to guide sequence design, where the Z-score is the energy difference between the target and denatured states relative to the size of the energy fluctuations among denatured states. Sequences were designed through Monte Carlo (MC)-based simulated annealing, using the modified Z-score as the objective function. NMR and circular dichroism studies of one of the designed sequences are consistent with the expected α-helical content and a well-defined three-dimensional structure.

More commonly, atomistic energy functions are used to optimize the energy of a sequence in the target structure relative to its energies in competing misfolded conformations. Incorporating negative design is likely to be important when designing protein–protein interfaces, for which the subtlety of specific interactions presents a significant challenge, or when simplified representations of the amino acids are used. In designing an A8B8 helical heterotetrameric protein with a dinuclear metal center, Summa et al. [13] used negative design successfully to select for sequences with appropriately charge patterned external positions, such that the peptides do not form homotetramers or heterotetramers of the wrong orientation. Havranek and Harbury [14] have described an algorithm that uses explicit negative design to engineer coiled-coil interfaces that favor the formation of either homodimers or heterodimers. To design specificity, they computed and compared the energies of candidate sequences in four competing structures, including unfolded and aggregated states, and homodimeric and heterodimeric states.

Energy function optimization

Rather than address competing structures directly, many studies include structural information concerning only the target structure (or target structure ensemble). These methods typically identify low-energy sequences by minimizing a carefully weighted and parameterized energy function. The target structure is largely constrained. Standard terms associated with molecular mechanics potentials are usually included and database-derived propensities of the amino acids for inter-residue interactions or secondary structure formation may be included in the objective function. The energetic contribution due to unfolded states is crudely quantified by effective reference energies for each amino acid. These reference energies model, in an independent manner, the contributions of each amino acid to the free energy of the unfolded state and may also be used to bias the composition of the designed sequences to mirror their distribution among naturally occurring sequences. Using repeated MC sequence searches with a weighted energy function comprising atomistic and database-derived potential terms, Kuhlman et al. [15] have designed a protein with a novel tertiary structure. Shifman et al. [16] fine-tuned the ligand specificity of calmodulin by optimizing interactions with a single target ligand, with the assumption that such optimization would destabilize interactions with other ligands.

Energy functions

Though similar in many respects, the potential functions used during protein design are different from those used for protein folding and molecular dynamics studies, and represent a compromise between speed and accuracy. In their most common forms, they include pairwise interaction energies, such as van der Waals, electrostatic, hydrogen bonding and solvation terms. In practice, these individual contributions to the energy may be fine-tuned to best suit the design objectives. The van der Waals term is often used with scaled atom radii to compensate for the discreteness of sidechain conformations or rotamers [17]; hydrogen bonding and solvation terms are often ignored when core packing is the design goal [18]; and the dielectric constant in the Coulomb energy term may be given a distance-dependent form [6]. Scoring functions used in protein sequence design continue to be optimized, often using parameterization to a high-resolution data set [19,20]. The potential function used to evaluate the fitness of each candidate sequence may contain terms that may not have explicit physical interpretations or only approximately model the forces present in proteins. Coefficients derived from high-resolution protein structures may be used to differentially weight various terms in the physical potential [15,21]. Statistical (database-derived) potentials may be used on an equal footing with physical energy functions to account for the trends observed in nature’s proteins, such as helical propensities and residue–residue interactions.

Hydrogen bonds

Hydrogen bonds are ubiquitous in proteins [22] and accurate modeling of these interactions is likely to be
important in the design of truly functional, native-like protein sequences. Kortemme et al. [23] introduced an orientation-dependent hydrogen bond potential based on the geometric characteristics of hydrogen bonds in high-resolution protein crystal structures. Treated as partially covalent interactions, the hydrogen bonds in their model are parameterized with three angles and one distance. They demonstrated that this potential is more effective than a purely distance-dependent Coulomb model of hydrogen bonds using a series of tests involving the prediction of sequences in proteins and protein–protein interfaces, and decay discrimination for both monomeric single-domain proteins and docked protein–protein complexes. To study the role of buried polar sidechains and hydrogen bonding networks in the protein core, Bolon et al. [24] generated simple hydrogen bond rules based on geometric hydrogen bond criteria to be used during protein design. According to these rules, all buried polar sidechains must form at least one hydrogen bond to mainchain atoms not involved in hydrogen bonds to other mainchain atoms, and a minimum number of total hydrogen bonds required for that amino acid type. A ‘prudent polar’ strategy that requires the minimum number of hydrogen bonds for buried polar residues improved thioredoxin core redesign relative to a ‘no polar’ strategy, and resulted in a sequence with improved thermodynamic stability while introducing a minimal increase in combinatorial complexity. Hellenga and co-workers [25] have developed methods for keeping a detailed inventory of hydrogen bonding interactions in designing proteins that bind with high specificity and affinity to a variety of small molecules.

Electrostatics

Although electrostatic interactions play a critical role in protein structure and function, a consensus has yet to be reached regarding how best to model these interactions for the purpose of protein design. The added complexity of sequence variation usually precludes the use of more detailed electrostatic treatments, such as Poisson–Boltzmann calculations. Simplifying approximations such as reduced or distance-dependent dielectric constants are often introduced [6,17]. An effective scoring function may be developed by optimizing the relative contributions from electrostatic and van der Waals interactions against proteins in a training set [19]. With their long range, electrostatic interactions may be disproportionately represented compared to other terms in the potential. Often, an energy cutoff is imposed on the maximum contribution from electrostatic interactions [26]. Marshall et al. have probed the importance of helix capping and helix macrodipole interactions, and their role in stabilizing redesigns of the engrailed homeodomain. They found that much more stable sequences are obtained when these delocalized interactions are accounted for in the design calculations [26]. Wisz and Hellenga [27] modeled the heterogeneous protein environment by introducing a dielectric constant that takes into account protein geometry, local structure and the types of interacting residues. Their model reproduces measured pKᵢ values as well as continuum solvent models. Because the electrostatic interactions are pairwise decomposable and can be rapidly calculated, they may be readily incorporated to model electrostatic interactions during protein design.

Solvation

Explicit solvent and even many implicit solvent models are impractical for protein design, wherein large variation in sequence must be considered. The solvation properties of amino acids arise from a myriad of interactions involving van der Waals forces, hydrogen bonding, electrostatic interactions and solvent degrees of freedom. Generally, globular proteins bury hydrophobic residues in their interiors and expose hydrophilic residues to solvent. A recent survey reveals, however, that 25–30% of core residues in globular proteins are polar [28]. On the one hand, it is believed that buried polar residues destabilize the protein. This view is supported by the observations that replacing buried polar residues with non-polar residues increases the stability of a protein [29] and that hydrophobic patterning has been a successful element of several de novo designed proteins [30,31**]. On the other hand, the substitution of five buried hydrophilic residues in thioredoxin with computationally optimized hydrophobic residues resulted in a mutant that had lower stability than the wild type, suggesting a possible role for buried polar residues in attaining thermodynamic stability [28]. Buried polar residues are probably most important for achieving conformational specificity rather than structural stability. In addition to providing directional interactions that can confer structural specificity, buried polar sidechains may stabilize the protein by satisfying the hydrogen bonding potential of the backbone amides, thus offsetting the energetic penalty of losing hydrogen bonds with solvent upon folding.

Solvation plays an important role in determining the electrostatic potential energy of proteins, either directly by interacting with charged residues on the surface or indirectly by shielding local charges. Incorporating the protein–solvent interaction explicitly is, however, a difficult task, as sampling solvent configurations for each attempted sequence would be computationally intractable. A modified Tanford–Kirkwood continuum electrostatic approach models the effects of solvent polarization on charged atoms in proteins; it predicts the self-energies and interaction energies of 36 polar sidechains in bovine pancreatic trypsin inhibitor (BPTI) with an accuracy comparable to that of a finite-difference Poisson–Boltzmann (FDPB) calculation [32]. Recently, Pokala and Handel introduced a method of computing the solvent-accessible surface area (SASA), which is often used to estimate hydrophobic solvation energies, by mimicking sidechain atoms with enlarged backbone pseudo-atoms. Similarly, atomic Born
radii of the generalized Born family of continuum solvation models [33] can be computed using backbone pseudo-atoms and a scaling factor [34]. This model is six orders of magnitude faster than the FDPB continuum model, while successfully predicting the pKa of ionizable groups in five proteins [34]. A simple parameterization of the solvation preferences of amino acids based on the local density of β carbons has also been successfully applied in protein design and re-engineering [6]. Although implementation of the most accurate methods for calculating electrostatic energies is perhaps not possible during sequence design, the availability of rapid and accurate solvation models should help future efforts in CPD.

Search methods

A variety of sequence search methods are available for protein design. Pruning and DEE algorithms can provide very accurate estimates of global energy minima. MC and simulated annealing algorithms have the advantages of being rapid and straightforward to implement. Probabilistic methods provide estimates of site-specific amino acid variability and can be used to guide search algorithms.

The basic strategy of all DEE algorithms is the elimination of rotamers that cannot be members of the minimum energy sequence. The interaction energies are pairwise decomposed to include one-body (e.g., sidechain–backbone) and two-body (e.g., sidechain–sidechain) energies. In the original form of DEE, a rotamer is flagged if the lowest energy obtainable with it is higher than the highest energy obtainable with another rotamer at the same position [35]. This elimination criterion was later relaxed by Goldstein to also flag rotamers for which other rotamers can be found that have lower energy against an identical conformational background [36]. Generalized DEE, introduced by Looger and Hellinga [37], goes a step further by comparing clusters of rotamers instead of individual rotamers, thus increasing the convergence power of the algorithm. The new implementation extends the applicability of DEE to systems that are much larger than previously studied and has been effective in solving both sidechain modeling and protein design problems. Revised DEE elimination and flagging criteria were suggested by Gordon et al. [38], whose implementation includes an MC-based search to generate a reference energy used to eliminate high-energy rotamers. Although useful in accelerating the convergence process, the stochastic nature of reference energy calculation makes the algorithm no longer deterministic. In addition, they also proposed a ‘split flags’ method, in which the conformational space is partitioned into sub-conformer sets and rotamers are flagged as ‘elimination likely’ based on partial dominance by other rotamers. Because of the exponentially growing total computation time, the application of DEE methods had previously been limited to small to medium-size proteins, but with recent advances, greater complexity can be accommodated. For example, the method has been applied to the redesign of TIM-barrel proteins with 216 residues [39*].

Simulated annealing methods have also been increasingly applied to sequence design [15**]. In addition, predetermined site-specific amino acid probabilities can be used to improve the efficiency of these searches and the sampling of low-energy sequence/rotamer configurations [40]. Although convergence to a global optimum is not ensured, determining the global minimum energy sequence is probably not necessary for successful protein design.

Backbone flexibility

For many CPD projects, a single fixed backbone from a high-resolution structure serves as the design template. Decoupling the backbone degrees of freedom can vastly reduce the required computation time. This is usually accomplished by fixing the backbone of a protein as a rigid scaffold. The lack of backbone flexibility, however, can significantly bias sequence selection. Some side-chains may be assigned high interaction energies and ruled unfavorable, whereas in reality a slight backbone adjustment might have corrected such problems. The importance of the interplay between the backbone and side-chains is showcased by the successful re-engineering of a coiled-coil dimer to a trimer and a tetramer using a series of parameterized backbones [41]. Still, the computational burden of managing multiple structures prevented the backbone degrees of freedom from being extensively explored in the past. Studies have shown that there are subtle yet significant discrepancies between modeled and actual backbone structures [42,43], thus motivating the use of multiple target structures during design studies [44**]. Regressive analysis of designed sequences against the correct structure shows that the quality of the designed sequences is often superior when the backbone flexibility is explicitly modeled [45].

Several recent efforts highlight the role of backbone flexibility. Studying the coupling of sidechain and backbone movements in the protein G β1 domain, Mayo and co-workers [46,47] initially reported that their sequence design algorithm (ORBIT) is robust against backbone adjustment. They showed that, even with a 15% shift in the backbone, ORBIT produces sequences with similar overall energies. A later study with T4 lysozyme, however, showed that protein stability is difficult to achieve upon core repacking without shifts in the backbone [45]. Desjarlais and Handel [48] explicitly modeled the backbone degrees of freedom using a GA algorithm and MC sampling. Because many backbone conformations near the target structure can be randomly sampled, an ensemble of related backbones can thus be easily generated even when the protein lacks apparent symmetry, such as that found in coiled coils. Kraemer-Pecore et al. [44**] used a sequence prediction algorithm and a GA-based sampling algorithm to design new sequences for the WW
domain. To weight the sequences obtained for a series of related backbones, an effective free energy was calculated for each rotamer by computing a rotamer-specific partition function averaged over all structures. Because each model contains a unique backbone structure and sidechain combination, each rotamer state is exposed to a wide range of environments. Two designed sequences had the expected secondary structure features, cooperative denaturation and NMR signatures consistent with the target structure. Baker and co-workers [15**] have used a design protocol based on iterative cycling between sequence selection and backbone optimization to design a novel α/β fold not observed in nature. First, a series of 93-residue backbone models was assembled, then the optimum sequence for each model was designed with an MC search algorithm. The lowest free energy backbone conformation was next identified using an MC minimization protocol. The sequence/structure pairs obtained after 15 rounds of sequence design and backbone optimization had energies comparable to those of naturally occurring proteins. The structure of one sequence, Top7, as determined by X-ray crystallography, was shown to have a backbone atom rmsd of 1.17 Å from the design model. No negative design was implemented explicitly. Interestingly, sequences designed without the backbone minimization step but with a damped repulsive term had somewhat molten cores, highlighting the importance of explicitly modeling interatomic interactions and the backbone flexibility.

Statistical design methods

Complementary to sequence search methods are statistical and probabilistic approaches to protein design [6,49–51]. Rather than identifying one or a few low-energy sequences, statistical design methods estimate the site-specific probabilities of the amino acids in sequences that are structurally consistent with the target structure. This structural consistency is quantified by atomistic energy functions that address rotamer–backbone, rotamer–rotamer and rotamer–solvent interactions. In addition, a variety of constraints may be introduced so as to pre-pattern sequences or to specify amino acid properties that are important for conferring function, such as the sidechain ligands that form a metal-binding site. The most likely set of site-specific amino acid probabilities is determined by maximizing an effective entropy function subject to any constraints that may be present on the sequences. Thus, characterization of sequence space is recast as a problem in statistical thermodynamics. The most common choice of entropy function yields an effective self-consistent field approximation upon optimization. As a result, large numbers of variable positions and great diversity of the amino acids may be considered in the calculations. The method has been extended so as to bias simulated annealing methods [40] and include symmetric quaternary structures [51]. For the purposes of protein design, a suitable sequence may be identified as that comprising the most probable amino acid at each position after one or more iterative rounds of calculation, where the identities of an increasing number of amino acids may be specified with each iteration.

Such a statistical computationally assisted design strategy (scads) has been applied to design water-soluble analogues of the membrane-spanning potassium channel KcsA (see Figure 1) [52**]. A version of the tetrameric membrane-bound protein with an engineered toxin-binding site was subjected to computational analysis. Exposed, transmembrane hydrophobic residues on the exterior of the protein were targeted for mutation. In addition to considering interatomic interactions, the value of a database-derived solvation energy (‘environmental energy’) [6] was constrained to the value expected for a soluble protein with the same number of residues as KcsA. For the tetrameric KcsA complex, 140 exposed residues were initially targeted for variation. A unique sequence was
selected from the computed probabilities by adopting the most probable amino acids at all sites, except at a few sites where alternative residues were selected to achieve greater helical propensities or polarity. A computationally redesigned variant, WSK-3, expresses in high yield, and shares structural and functional signature properties with its membrane-soluble counterpart: it is predominantly tetrameric in solution; it binds the toxin specifically with the same affinity and stoichiometry as wild-type KcsA; and toxin binding may be inhibited by a small molecule blocker, tetraethyl ammonium. This study reveals how protein design may be used to facilitate characterization of the structure and functional properties of membrane proteins, which are notoriously difficult to work with due to their low expression levels and poor solubility. In related work, sequence search algorithms have been used to identify solubilized variants of the integral membrane protein phospholamban [53**].

The statistical approach also played a central role in designing a 114-residue four-helix bundle (DFsc) with a dinuclear metal center (see Figure 2) [54**]. Two diiron proteins, a heterotetrameric A$_2$B$_2$ protein [13**] and a helix-loop-helix DF1 protein [55], had been previously designed (via symmetry-derived parameterization of helical structures) and characterized. The target template for a monomeric variant DFsc was generated by redesigning interhelical loops. The identities of residues that confer metal binding and substrate accessibility to the active site were constrained. Scads was used to determine the identities of the remaining 88 amino acids. Despite the challenge of burying six ionizable residues within the interior, the designed apo protein folds in the absence of metal ions. The protein stoichiometrically binds two equivalents of Fe(III), Co(II), Mn(II) and Zn(II), and has increased thermal stability upon metal binding. The diiron protein exhibits catalytic activity against known peroxidase substrates with rate enhancements of 10$^{10}$–$10^3$. This protein is a promising realization of the de novo design of protein structure, sequence and functionality.

**Conclusions**

A striking result of recent work in protein design is the fact that many disparate computational methods — simulated annealing, DEE and statistical design — have been successful in identifying sequences that fold into a target structure. These methods are mostly based on atomistic molecular information that includes interactions within the folded structure, but do not explicitly consider negative design. Why are such methods that focus only on native state structures successful? This may be due to several reasons. The energetic competition with unfolded states is in fact addressed, although in a crude and approximate manner, by incorporating reference energies, by constraining the amino acid composition, or by pre-patterning polar and hydrophobic residues. In addition, energy landscape theory suggests that increasing the number of amino acids facilitates the design of unique structures [11]. Atomistic modeling of the amino acids and their sidechain conformational states significantly increases the effective number of monomer types permitted at each position in a protein compared to reduced-residue representations, thus helping to create favorable complementary packing and inter-residue interactions. Sequences that are specifically tailored to a particular structure can be obtained.

With sophisticated physical potentials and powerful search algorithms, CPD promises to dramatically advance protein engineering. With several alternative successful design
algorithms currently in use, protein designers may adopt one that best fits their needs. Negative design, for example, may be particularly important when working with low-resolution models (e.g. reduced or non-atomistic amino acid representations) or intermolecular complexes, because of the large number of potential competing structures. For problems where much is known quantitatively about the determinants of particular protein properties, sequence search methods appear to be appropriate, as a key feature of these methods is their ability to recover sequences with well-packed structures, and favorable electrostatic and hydrogen bonding interactions. Although statistical methods can be similarly used to design individual sequences, they are also appropriate for the partial design of ensembles of sequences. To arrive at proteins with the efficiency and selectivity seen in nature, a combination of de novo design methods and combinatorial or directed evolution methods may be necessary. Such approaches would be especially powerful when designing structures in the absence of complete information. Computational design will dramatically accelerate protein discovery. Lastly, by modulating solubility and intermolecular structure, CPD can facilitate the redesign of a wide variety of natural proteins to expedite the elucidation of their structures and biological activities.

Acknowledgements
The authors acknowledge support from the National Science Foundation (CHE 99-84752 and DMR 00-79909) and the National Institutes of Health (GM61267). JGS is a Cottrell Scholar of Research Corporation, and an Arnold and Mabel Beckman Foundation Young Investigator. The figures were rendered using PyMol (DeLano Scientific).

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

De novo computationally designed proteins are found to fold on timescales of 10-50 μs, suggesting that fast folding is consistent with high stability and close to optimal interatomic interactions in the folded state.


The authors studied the folding of a de novo designed three-helix-bundle protein using IR spectroscopy, temperature jump spectroscopy and molecular dynamics simulation. The protein is found to fold on a microsecond timescale and has a highly heterogeneous set of folding trajectories.


Statistical computational design was used to identify ‘frustrated’ sites in a 47-residue three-helix-bundle protein and to identify mutations that modulate hydrophobic content consistent with the folded structure. Several mutants were made and experimentally characterized. Correlation of the folding rate with tailored hydrophobic mutations is observed. One double mutant is among the fastest folding proteins known, with a folding time of 1 μs.


The authors designed a three-helix-bundle topology and selected several of the designed sequences for synthesis. The design principle involves selecting for the global shape of the protein folding funnel, whereby sequences are identified that have the target structure as the lowest energy state among all the structures sampled by many folding simulations. The authors use a coarse-grained protein model with simplified sidechain representations. NMR and CD spectroscopic data indicate that one designed sequence has a well-defined three-dimensional structure and helical content consistent with the target. This work provides an elegant synthesis of energy landscape ideas and protein design methods.


A simple parameterization of complementary electrostatic interactions in a four-helix bundle is used to design a four-helix heterotetramer. This represents one of the first computational efforts with explicit negative design and experimental verification.


A novel 93-residue α/β protein structure is identified and computationally designed. The crystallographic structure of one sequence is within 1.7 Å of the structure obtained from the calculation.


34. Alber T: Retrostructural analysis of metalloproteins: a first example of the computational redesign of a metalloprotein is presented. The protein binds a variety of metal ions with the appropriate stoichiometry. The diiron protein exhibits catalytic activity with regard to several peroxidase substrates. Proc Natl Acad Sci USA 2000, 97:6298-6305.


36. Kraemer-Pecore CM, Lecomte JT, Desjarlais JR: A de novo redesign of the WW domain. Protein Sci 2003, 12:2194-2205. A first example of the computational redesign of a β protein, the WW domain. The authors used their sequence prediction algorithm in combination with sampling fluctuations of the backbone in the vicinity of the target. CD and NMR measurements are consistent with the wild-type WW domain structure.


38. Slovic AM, Summa CM, Lear JD, DeGrado WF: Computational design and characterization of a monomeric helical dinuclear metalloprotein. J Mol Biol 2003, 334:1101-1115. The de novo design of structure, sequence and function for a 114-residue metalloprotein is presented. The protein binds a variety of metal ions with the appropriate stoichiometry. The diiron protein exhibits catalytic activity with regard to several peroxidase substrates.
