

Review: Protein Design—Where We Were, Where We Are, Where We're Going

Navin Pokala and Tracy M. Handel

Department of Molecular and Cell Biology, 229 Stanley Hall, University of California, Berkeley, California 94720

Received December 21, 2000, and in revised form March 27, 2001; published online June 13, 2001

Protein design has become a powerful approach for understanding the relationship between amino acid sequence and 3-dimensional structure. In the past 5 years, there have been many breakthroughs in the development of computational methods that allow the selection of novel sequences given the structure of a protein backbone. Successful design of protein scaffolds has now paved the way for new endeavors to design function. The ability to design sequences compatible with a fold may also be useful in structural and functional genomics by expanding the range of proteins used for fold recognition and for the identification of functionally important domains from multiple sequence alignments. © 2001

Academic Press

Key Words: protein design; optimization methods; energy functions.

INTRODUCTION

Protein design has emerged as a powerful method for understanding the underlying physical principles that dictate protein folding and function. Substantial progress in scaffold design has been made in the past few years, largely because of the development of energy functions describing the forces that determine protein structure and optimization methods that allow an enormous number of sequence possibilities to be evaluated. Most design efforts have targeted small protein scaffolds, with the goal of understanding the balance of forces that stabilize protein structure. In some cases, the designed proteins have been tested experimentally to validate and parameterize energy functions. More recently, other computational metrics have been introduced to evaluate and improve the methods (see below).

This review focuses on recent advances in computational protein design. First we discuss the optimization methods and energy functions used to design protein scaffolds and the problems that have yet to be solved. More recent forays involving design of

function will then be discussed along with the computational challenges that these loftier goals require.

OPTIMIZATION METHODS

Rotamer Libraries and the Fixed Backbone Approximation

One of the challenges in protein design is selecting an optimal solution from the enormous number of sequence and structure possibilities. Ponder and Richards (1987) addressed this in a seminal study, where they proposed that protein design calculations could be simplified greatly if one assumed a fixed backbone and that side chains exclusively adopt statistically preferred conformations or rotamers. These assumptions make the search space discrete and more computationally feasible by effectively casting the design problem as one of identifying the lowest energy combination of side-chain rotamers for a backbone template. Furthermore, by simply selecting the most favorable rotamers for a given sequence, ~70% of the χ_1 rotamers can be predicted correctly (Dunbrack and Cohen, 1997).

Nevertheless, the sequence–structure space is still large. For a 100-residue protein in which all 20 amino acids are permitted at every position, with only two rotatable bonds and five conformations each, there are 500^{100} sequence–structure solutions. Ponder and Richards (1987) considered only a handful of residues and a hard-sphere van der Waals energy function and were able to perform an exhaustive search for compatible sequences, but this is not practical for total sequence design of large proteins. Improved optimization methods and energy functions have since been developed, expanding the range of tractable problems to the computational design of entire proteins. However, the use of discrete rotamer libraries and fixed backbones remains an important simplifying assumption in these new methods, as described below.

Search Algorithms

An excellent overview of the strengths and weaknesses of various search algorithms was recently reported (Desjarlais and Clarke, 1998), and implementations of these algorithms were evaluated quantitatively (Voigt *et al.*, 2000). They fall into two broad categories. Stochastic algorithms (including Monte Carlo methods and Genetic algorithms) semi-randomly sample sequence-structure space and move toward lower energy solutions, while deterministic algorithms (e.g., Dead End Elimination and Mean Field methods) perform semiexhaustive searches.

Monte Carlo (MC) methods and variations thereof are the simplest stochastic methods (Metropolis *et al.*, 1953). In the context of design, a starting structure is perturbed by a random change in residue type or rotamer at some position. If the change decreases the energy of the structure, it is accepted. Otherwise, the Metropolis criterion is used to accept or reject the change; in this case, a Boltzmann weighted probability calculated from the temperature and the old and new energies is compared to a random number. If the random number is lower than the Boltzmann probability of the move, it is accepted. This permits energetically uphill moves and escape of local minima. The temperature can also be adjusted to allow energy barriers to be overcome or slowly annealed to decrease the probability that higher energy changes will be accepted (Lee and Subbiah, 1991). Other similar methods have also been described (Wernisch *et al.*, 2000).

Genetic algorithms (GAs) evolve populations of solutions by cycles of operations based loosely on concepts in genetics and evolution (Holland, 1993). A population of random sequences and rotamers are generated for the target backbone. Structures with lower energies mate with one another, exchanging sequences and rotamers to form hybrids that often have better energy and therefore get selected. Mutations that alter an amino acid identity or rotamer conformation add diversity to the gene pool. High-energy structures are eliminated from the population. This process of recombination, mutation, and selection is repeated iteratively until convergence is achieved (Tuffery *et al.*, 1991). Hybrid algorithms using GAs followed by MC have also been reported (Desjarlais and Handel, 1999).

The advantage of stochastic methods is that they can deal with problems of significant combinatorial complexity because they do not require an exhaustive search. GAs are particularly effective in exploring a large sequence space and surmounting energy barriers because of the nature of the moves. The disadvantage is that there is no guarantee that

these methods will converge to the global minimum energy solution or even the same solution when run multiple times (Voigt *et al.*, 2000).

In contrast, deterministic methods, such as Self-Consistent Mean Field (SCMF) optimization and Dead End Elimination (DEE), always converge on the same solution. SCMF uses a multicopy representation (or ensemble) of side chains such that many amino acid types/rotamers are placed at each residue position of the template. Each rotamer in the ensemble is initially assigned a uniform Boltzmann probability. The rotamer probabilities at all other positions are used to calculate a weighted-average energy for each rotamer at each position. From these energies, a new Boltzmann probability is calculated for each rotamer, thus generating a new ensemble. This process is repeated iteratively until the rotamer probabilities converge (e.g., self-consistency is achieved). However, SCMF is not guaranteed to converge to the global minimum energy solution (Lee, 1994).

Being a quasi-exhaustive method, DEE is guaranteed to converge on the global minimum energy solution (Desmet *et al.*, 1992). The effectiveness of DEE for a combinatorial search is due to the systematic elimination or pruning of high-energy rotamers or rotamer combinations. A requirement is that the energy function must be written as the sum of individual and pairwise terms. Additionally, for extremely complex problems, DEE may fail to converge. To address these limitations, recent improvements to the DEE algorithm have been developed that dramatically increase efficiency and the size of problems that can be addressed (Goldstein, 1994; Gordon and Mayo, 1999; Wernisch *et al.*, 2000). Thus, as demonstrated in a quantitative comparison of these methods, DEE currently seems to be the most powerful method for finding the global minimum energy solution (Voigt *et al.*, 2000). However, as the complexity of problems increases, a combination of stochastic and deterministic methods may ultimately prove to be the best compromise between accuracy and speed.

ENERGY FUNCTIONS

Describing the interactions in a protein accurately is the second key element to protein design and probably the most difficult. Energy functions for protein design must be fast and accurate, yet not oversensitive to the fixed backbone approximations and discreteness of the rotamer library (reviewed in Gordon *et al.* (1999)). While structure-based pairwise potentials are fast and appear to have utility for fold prediction, they lack sensitivity to local structure at an atomic level that is required for design. There-

fore, most current protein design efforts use atomic-level molecular mechanics force fields as a foundation (Gordon *et al.*, 1999). In this section we discuss the different terms in energy functions used for design.

Molecular mechanics force fields, such as AMBER, OPLS, DREIDING, ECEPP/2, and CHARMM, are usually composed of van der Waals, electrostatics, dihedral angle, bond-angle, and bond stretching terms (Brooks *et al.*, 1983; Mayo *et al.*, 1990; Nemethy *et al.*, 1992; Cornell *et al.*, 1996; Jorgensen *et al.*, 1996). Since ideal geometry is always assumed for protein design calculations, the last two terms are not considered. The parameters for van der Waals terms are determined from small-molecule crystal structures. Partial charges and dihedral angle parameters are derived from electron distributions from quantum theory. These parameters are further adjusted by simulations that attempt to reproduce experimental data, such as small molecular crystal structures and heats of vaporization.

While these models work reasonably well for all-atom molecular dynamics simulations, they require considerable modification for protein design calculations. Energies must be adjusted to reduce artifacts resulting from the use of discrete rotamers and fixed backbones. Energy terms that describe solvation must be added. A reference state needs to be defined, since the relevant value for protein design is the difference in energy between the folded and the unfolded states (Koehl and Levitt, 1999a; Wernisch *et al.*, 2000). Finally, these terms must all be weighted appropriately. Since DEE-type search methods require the decomposition of energies into the sum of pairwise interactions, approximations that allow nonpairwise energies to be cast as pairwise terms are also important. Even methods that are not restricted by pairwise calculations can be sped up tremendously by such approximations.

van der Waals

Despite the overall requirement for accurate energy functions, fairly simple functions have worked well for designing protein cores (Hurley *et al.*, 1992; Hellinga and Richards, 1991; Lee, 1994; Kono *et al.*, 1998; Jiang *et al.*, 2000). This is because cores are the easiest part of the protein to describe energetically if one restricts the composition to hydrophobic residues.

In some recent studies aimed at probing the role of packing on structural specificity, core variants of 434 cro, ubiquitin, GCN4, and G β 1 were designed (Desjarlais and Handel, 1995; Dahiyat and Mayo, 1996, 1997; Lazar *et al.*, 1997). This was done with a backbone-dependent rotamer library for a subset of hydrophobic amino acids and a simple van der

Waals potential to score acceptable sequences. Remarkably, stable well-structured proteins were predicted with just this term. Subsequent structural characterization of select designs demonstrated impressive resemblance to the predicted structures, and reasonable correlations were observed between predicted and experimental energies (Dahiyat and Mayo, 1996; Lazar *et al.*, 1997, 1999; Johnson *et al.*, 1999). Part of the reason that this worked was because polar amino acids were excluded from the core. Nevertheless, it is clear that van der Waals potential is a dominant term that must be included in the force field, with all-atom representations of protein to accurately reflect local structure.

Van der Waals energies are typically calculated with a Lennard–Jones 6–12 potential, which has a mild attractive term and a strong repulsive term. The repulsive term causes the energy to become highly unfavorable as atoms approach each other. In molecular dynamics simulations, this has the important effect of preventing atoms from overlapping. In contrast, for protein design, the repulsive term must be softened to avoid overpenalizing slight overlaps that inevitably result from using discrete rotamers and a fixed backbone. One strategy that has been used is to uniformly scale the van der Waals radii down by 5–10% (Desjarlais and Handel, 1995; Dahiyat and Mayo, 1997b; Kuhlman and Baker, 2000). Radii scaling softens the effect of the repulsive component of the van der Waals term and implicitly allows for some level of flexibility in fixed backbone templates, although it also adversely affects the attractive component of the energy. In practice, reduced radii can lead to a slight overpacking of the hydrophobic core, but if carefully calibrated can lead to proteins with higher stability, perhaps because of increased buried hydrophobic surface area (Dahiyat and Mayo, 1997a). Another strategy that avoids altering the Lennard–Jones potential is expansion of the rotamer library near the database values (Lazar *et al.*, 1997; Desjarlais and Handel, 1999). This has the unfortunate consequence of expanding the search space, limiting its use to small systems. A related strategy is to minimize the structures of clashing rotamer pairs. The minimized energy is then used as the energy for that rotamer pair (Mendes *et al.*, 1999; Wernisch *et al.*, 2000).

Typically, one calculates van der Waals interactions between side chain and backbone, side-chain pairs, and intra-side-chain atoms. Intra-side-chain interactions are sometimes excluded. This can be justified by using only highly energetically favorable rotamers or supplementation with a torsion angle term.

Solvation

While van der Waals energies are described quite well by molecular mechanics force fields, electrostatics, hydrogen bonding, and solvation in the context of a protein in water are not (Edinger *et al.*, 1997). This is not surprising since the force fields are usually calibrated with *in vacuo* simulations and experiments. In addition, water itself is an extremely complicated substance to model and parameterize because of its polarizability, interactions with polar groups, and entropic contributions due to the hydrophobic effect. Inclusion of explicit waters in calculations is often done in molecular mechanism simulations, but this is prohibitively expensive for design; thus other approximations must be used (Wesson and Eisenberg, 1992).

The hydrophobic effect is typically modeled by assuming that the penalty for exposing nonpolar groups to water is dependent on the surface area (Eisenberg and McLachlan, 1986; Ooi *et al.*, 1987). The solvation energy is calculated from the change in solvent accessible surface area, multiplied by an atom-dependent atomic solvation parameter; these solvation parameters are typically derived from transfer free energies of amino acids between water and vacuum or some organic solvent (Wolfenden *et al.*, 1981; Fauchere *et al.*, 1988). The addition of solvation terms to the energy function has been shown to greatly improve correlations between predicted and experimental energies for core design (Dahiyat and Mayo, 1996) and has been used to predict native-like sequences as assessed by profile scores (Koehl and Delarue, 1994; Raha *et al.*, 2000).

In their traditional form, solvation parameters penalize exposure of hydrophobic groups and reward exposure of polar groups. More recently, solvation parameters were derived from fitting experimental protein stabilities to a solvation expression that includes an additional term favoring hydrophobic burial (Eriksson *et al.*, 1992; Dahiyat and Mayo, 1996; Street and Mayo, 1998). Other terms have also been added that penalize polar burial (Dahiyat *et al.*, 1997; Raha *et al.*, 2000) (see below).

One of the main issues in estimating solvation energies is that surface area calculations have been computationally intensive, because pairwise calculations overestimate buried areas. The problem is that the same area of a given atom may be masked by several atoms. However, it is possible to avoid overcounting by using empirical scaling factors that underestimate atom-atom overlap. Wodak and Janin (1980) described such a method almost 20 years ago for models that represented side chains as large pseudo-atoms. Recently, Street and Mayo (1998) have made considerable improvements to this strat-

egy using actual atoms. We have developed a hybrid approach that uses a combination of pseudo-atoms and actual atoms to calculate atomic surface areas additively (Pokala and Handel, in preparation).

Electrostatics

Electrostatic interactions such as hydrogen bonding and salt bridges are important for defining protein structural specificity and function, but are extremely complicated to model. One problem is that the strength of interactions depends on the local environment, which can range from completely buried to solvent exposed, making uniform treatments inaccurate. Water molecules also interact directly with polar atoms in a protein, competing with other polar groups for interactions. The interaction of water with itself is effected by charges in a protein. Thus, solvation and electrostatics are inextricably linked (reviewed by Bashford and Case (2000)). Individual charges placed in a protein affect the protein stability in a manner very sensitive to the local microenvironment (self-energy). Burial of charges is usually unfavorable; in fact, favorable interaction energies from a buried charged pair are more than offset by unfavorable desolvation energies (Hendsch and Tidor, 1994; Waldburger *et al.*, 1995; Hendsch *et al.*, 1996).

There are two classes of electrostatics/solvation models. Semiempirical models are dependent on solvation parameters fit directly from experimental data to approximate the self-energy. The simplest model is the use of a distance-dependent dielectric to screen coulombic interactions, coupled with a surface area-dependent term that favors the exposure of polar groups and penalizes their burial, as described above (Koehl and Delarue, 1994). These types of models provide, at best, qualitative solutions since they do not represent the environment-dependent nature of electrostatic interactions (Hendsch and Tidor, 1999). It has been shown that the distance-dependent dielectric/surface area model gives poor correlations with the more accurate energies calculated by Poisson-Boltzmann methods (Edinger *et al.*, 1997) (see below). However, better parameterizations can improve these models significantly (Mehler, 1996).

Despite their shortcomings, simple models have been successfully applied in protein design and it may generally be the case that they are perfectly adequate for design of protein scaffolds (Dahiyat *et al.*, 1997; Raha *et al.*, 2000). These models may be accurate enough to penalize unduly unfavorable interactions, which may be all that is necessary. As Warshel and Papazyan (1998) point out, many simple models can work well at the protein surface, if they use a large effective dielectric constant. How-

ever, it is likely that more accurate environment-dependent models will result in even better designs and will be required for accurate prediction of stabilities (see below). Furthermore, protein–protein interactions often involve buried polars that help to impart specificity usually at the expense of stability, and enzymes require buried polar residues as part of their catalytic mechanism. As protein design efforts address more complex problems such as binding and function, more accurate electrostatic models will be necessary. Since these will be an important addition to the next generation of design algorithms, we discuss them here.

An environment-dependent model was recently described by Lazaridis and Karplus (1999). It is based on the hydration shell model, which assumes that the self-energy of a group is related to the number of waters of hydration that are excluded by other atoms in the molecule. Their model is parameterized with experimental small-molecule heat capacity data (Lazaridis and Karplus, 1999). A novel feature of this method is that it does not require surface area calculations for either polar or nonpolar atoms. While there has been no direct experimental verification, the successful design of native-like sequences using this model as part of an energy function has been reported (Kuhlman and Baker, 2000).

Another class of more accurate models depends on analytical approximations to the Poisson–Boltzmann (PB) continuum dielectric model (reviewed in Honig *et al.* (1993). This model assumes that the protein and water can be treated as the classical physics problem of a low-dielectric charged object (protein) in a high-dielectric medium (solvent). The finite-difference (FDPB) numerical implementation has been remarkably successful in quantitatively predicting solvation energies and p*K*_a shifts of side chains in proteins from their canonical values in free amino acids (Bashford and Karplus, 1990; Antosiewicz *et al.*, 1994), as well as the effects of charge substitutions on protein stability (Spector *et al.*, 2000). For these reasons, it is considered the current “gold standard” model for protein electrostatics. Unfortunately, it is also computationally expensive, making it impractical for total protein design. Nevertheless, it provides a useful benchmark to calibrate and compare faster approximate methods.

One such method is the Tanford–Kirkwood model. The solvent-excluded volume of a protein is approximated as a sphere, and protein charges are mapped to the sphere, individually or in pairs (Tanford and Kirkwood, 1957). The energy of charges in a low-dielectric sphere embedded in a high-dielectric medium can then be calculated analytically. While simple, this model was recently used by Makhatadze and colleagues to identify electrostatically strained

sites in ubiquitin, which were then stabilized by generating appropriate mutations (see Loladze *et al.* (1999)).

Approaches that are even more sensitive to the local geometry have been developed. Solutions to the Poisson equation may be approximated with an image charge approach (Moult and James, 1986). Briefly, a charge buried at some depth in a low-dielectric sphere can be shown to have a self-energy equivalent to the coulombic interaction energy it would have with an image charge placed at a specific distance in the external medium. Shielding energies can be calculated in a similar manner. Both the image charge and its location in the external medium can be calculated from the radius of the sphere and the depth of the real charge in the sphere. These calculations are fast enough to include in dynamics simulations and *ab initio* structure prediction of peptide and protein structures (Abagyan and Totrov, 1994). Further improvements have been achieved by the use of shell charges, rather than point charges. These enable accurate prediction of p*K*_a shifts in proteins (Havranek and Harbury, 1999).

Another continuum dielectric approximation is the Generalized Born (GB) model. Briefly, this model uses a Born radius for each atom to confer environment-dependence on self-energies and charge pair energies. The Born radius is essentially equivalent to the ionic radius of an atom. In the original formulation used for molecular dynamics of small molecules, Born radii were first precalculated for each atom by FDPB, making it computationally intensive. Methods to analytically calculate Born radii directly from structure have been developed recently (Hawkins *et al.*, 1995; Qiu *et al.*, 1997; Dominy and Brooks, 1999; Onufriev *et al.*, 2000). In these implementations, the Born radii are dependent exclusively on the volumes of, and distances to, neighboring atoms. The agreement between energies calculated in this manner and with the FDPB is quite good. In addition, small-molecule solvation energies can be predicted with good accuracy. Our group and others have developed modifications that reduce the error of the analytical method for calculating solvation energies of proteins quickly and accurately. It has been shown that the analytical Born model is capable of calculating p*K*_a shifts in proteins to an accuracy similar to that of FDPB (Onufriev *et al.*, 2000).

One problem with the environment-dependent electrostatics models is that they are, unlike the simpler coulombic models, not readily and exactly decomposed into sums of rotamer pair energies, since they are dependent on all the atoms in the molecule. To address this, we have developed an

approximation that estimates Born radii for all atoms in all rotamers in a precalculation step (Pokala and Handel, in preparation). While not exact, it does approximate the energies well and extends the utility of this model to DEE-type search methods that require rotamer pair energies.

Hydrogen Bonding

Hydrogen bonding interactions play an important role in stabilizing secondary structures and imparting specificity to proteins. Because there is an optimum distance between donor and acceptor groups, the simplest expression resembles a van der Waals term and has both an attractive and a repulsive component. Explicit hydrogen bonding terms are included in some force fields, such as DREIDING. In this case, the strength of the interaction is dependent not only on the distances between the donor and the acceptor groups, but also on the orientation of the bond vectors. Other force fields such as AMBER, OPLS, and CHARMM treat hydrogen bonding implicitly, through the electrostatics and van der Waals energies. Explicit inclusion of a hydrogen bonding term led to the design of coiled coils with increased thermal stability (Dahiyat *et al.*, 1997).

Other Terms

Secondary structure propensities have also been used as constraints for sequence design. Dahiyat *et al.* (1997) found that the addition of a secondary structure term was helpful for predicting stability changes due to mutations at helical surface positions. In contrast, it appears that beta-sheet propensities are highly context dependent so this strategy may not be useful for designing beta sheets (Minor and Kim, 1994, 1996). Related to helical propensities are helix dipole and helix capping effects. Due to favorable electrostatic interactions with the helix dipole as well as capping interactions with the backbone, basic residues near the C-terminus of helices and acidic residues near the N-terminus appear to be stabilizing (Blaber *et al.*, 1993; Chakrabarty *et al.*, 1994; Doig *et al.*, 1994; Doig and Baldwin, 1995). These observations have been cited by Mayo and colleagues to rationalize the improved stabilities of some of their designed variants (see Dahiyat *et al.* (1997); Strop *et al.* (2000)). While empirical secondary structure propensities and helix dipole constraints may have practical utility in design, it should be noted that these effects, in principle, can be accounted for with the van der Waals and electrostatic energy terms (Srinivasan and Rose, 1999).

Putting It All Together

For molecular mechanics simulations, the individual energy terms are typically added together to obtain the energy of a protein. Unfortunately, the approximations required for protein design do not permit this, and the energy terms must be appropriately parameterized and scaled with respect to one another. In addition, the unfolded state must be considered implicitly, if not explicitly.

Optimization of the energy function has been addressed in few different ways. One approach involves finding a potential function that maximizes the Z-score (number of standard deviations from the mean) of the energy of the wild-type sequence relative to an ensemble of random sequences (Chiu and Goldstein, 1998). Optimizing the Z-score amounts to increasing the energy gap between the wild-type sequence and other possible sequences and has been suggested as a method for ensuring specificity. Using this approach, Street *et al.* (2000) optimized an energy function and used it to redesign six to seven surface β -sheet positions, in one case generating a variant that was slightly more stable than the natural protein.

A related strategy described by Kuhlman and Baker (2000) involved placing, one by one, all amino acid types at all positions in several protein structures, while keeping all other side chains and conformations fixed. The weighting factors on terms in the energy function were then optimized to maximize the Boltzmann probability of the natural residue type at each position. With this approach, the authors were able to optimize their energy function and design sequences that had significant identity with natural sequences and similar profile characteristics (Kuhlman and Baker, 2000).

Despite the successes of these knowledge-based optimization approaches, it is possible that the energy function could be biased. These methods assume that the wild-type sequence is the most stable, but it is well established that naturally occurring sequences are not optimized for stability (Shoichet *et al.*, 1995). In addition, for the purposes of functional design, it may be more important to predict a sequence of a specified (but not necessarily optimal) stability. Accordingly, another approach is to optimize the weighting factors to predict the energies of mutant proteins. An early example of this strategy involved modifying the substrate specificity of α -lytic protease. In this study, the basis and test sets were composed of K_i 's of mutant boronic acid-based inhibitors. The van der Waals, coulombic, electrostatics, and torsion terms were treated together and scaled with surface-area-dependent solvation energies. The resulting function was successfully used to

redesign the active site of the protease such that its substrate specificity was changed (Wilson *et al.*, 1991) (see below).

The best successes so far for predicting the energies of designed proteins has been for hydrophobic core repacking. Our group has predicted the relative stabilities of several ubiquitin, 434 cro, and T4 lysozyme core variants with fair accuracy, using just a van der Waals function and a rotamer library with fine sampling (Lazar *et al.*, 1997; Desjarlais and Handel, 1999). Mayo and colleagues were able to predict the relative T_m 's of several GCN4 core mutants with the addition of surface-area dependent energy terms (see Dahiyat and Mayo (1996)). We have recently parameterized a more general energy function for both core and surface design that can predict the relative stabilities of ~ 200 mutants from seven different proteins to an error of ~ 1 kcal/mol (Pokala and Handel, in preparation).

Calculation of stabilities requires a reference baseline state or an explicit unfolded state. For the solvation and entropic energies, it is not the absolute energy that is important, but rather the change in energy upon folding. For other energies, one must account for the fact that residues with more atoms will necessarily contribute larger (either favorable or unfavorable) absolute energies than smaller residues.

Most reference state models assume that the unfolded state occupies a large ensemble of essentially random structures, which implies that the energy is composition dependent but not sequence dependent. It should be noted, however, that there are some examples of proteins that have residual structure in the unfolded state (Shortle and Meeker, 1986). A common and simple structural model for the unfolded state is an extended tripeptide group (Dahiyat and Mayo, 1996; Koehl and Levitt, 1999a; Wernisch *et al.*, 2000). Usually, a conformational ensemble of a residue type is scored with the same energy function as the folded protein and averaged. One can also place a residue at all positions in several proteins and calculate the average energy associated with that residue type (Raha *et al.*, 2000). Alternatively, Kuhlman and Baker (2000) derived reference values for residue types by empirically fitting these parameters in concert with the energy function scaling coefficients, as discussed above. We also note that their reference state energies scale quite well with water/vacuum transfer free energies of side chain analogs, although the significance of this is not clear. Finally, Harbury *et al.* (1998) described an energy of permutation derived from host-guest stability data for the special case of designed

coiled-coils of different structure, but identical compositions.

REMAINING CHALLENGES IN SCAFFOLD DESIGN

Backbone Flexibility

Most protein design efforts to date assume that the backbone is fixed; after all, the goal has been to design a sequence that will adopt a fold as close to the target as possible. With this as a simplification, it seems that the currently available methods are up to the task of designing medium-sized proteins. Although there have only been a few examples of full *de novo* designs that have been experimentally validated (Dahiyat and Mayo, 1997; Bryson *et al.*, 1998; Walsh *et al.*, 1999), the success of localized core, surface, and interface designs suggests that it is not unreasonable to expect that the design of stable proteins on the order of a few hundred residues will soon be routine.

What then are the next hurdles? While the fixed backbone assumption has been useful for scaffold design, and may be adequate for certain applications such as generating hyperstable proteins, it does have limitations. Similar to problems associated with using discrete rotamers, a fixed backbone can cause rejection of what would otherwise be acceptable sequence solutions. Yet in reality, proteins are tolerant to mutations that would be nonpermissible if the backbone was rigid because they can adjust to relieve the strain (Alber *et al.*, 1988; Baldwin *et al.*, 1993; Lim *et al.*, 1994). More optimal sequence solutions would be predicted if backbone flexibility could be incorporated into design algorithms.

However, this is not a trivial problem. Backbone flexibility massively increases the search space. Second, one must develop energy functions that can quantitatively rank backbone conformations. Finally, it is not sufficient to predict sequences that can be accommodated by slightly perturbed backbone conformations; ultimately it will be important to predict the backbone and side-chain structures. This is significantly more challenging than side-chain prediction in the context of a fixed backbone. Nevertheless, a few attempts to design proteins with backbone flexibility have been described.

One approach is to generate an ensemble of related backbones, design a sequence for each of them using fixed-backbone algorithms, and then select the best backbone-sequence combination (or the converse: design a sequence for a particular backbone and then find the backbone variant it best fits).

For symmetric protein structures, such as coiled-coils and TIM barrels, one can describe the backbone structure by parametric equations (Crick, 1953; Murzin *et al.*, 1994a,b). This greatly reduces the

number of main-chain conformations and ensures that there will be reasonable backbone conformations within the ensemble. Using this type of approach, Harbury *et al.* (1998) designed two-, three-, and four-stranded right-handed coiled-coils, folds not observed in natural proteins at the time. Even more impressive was the fact that the predicted structure of the tetramer matched the crystal structure in atomic detail.

Unfortunately, the parametric approach has not yet been generalized to the vast majority of folds, which are usually nonsymmetric. For nonsymmetric proteins, one approach is to treat secondary structure elements as rigid bodies that can move with respect to one another (Su and Mayo, 1997), as has been observed in natural proteins (Baldwin *et al.*, 1993; Lim *et al.*, 1994). Using this strategy, Su and Mayo (1997) redesigned core sequences for several backbone variants of G β 1. They found that for small but significant backbone perturbations, the optimal core sequences were identical to that predicted for the unperturbed backbone, suggesting that perhaps current energy functions are not sensitive to small changes. However, slightly different sequences were predicted for larger perturbations (Su and Mayo, 1997).

Other approaches for generating backbone families include using coordinates from molecular dynamics simulations of natural proteins, individual structures from NMR ensembles, or structures from different crystal forms. Regan and co-workers used structures from an NMR ensemble to design rubredoxin-like metal binding sites in G β 1 (see Farinas and Regan (1998)). They identified viable sites that would have been missed if the average coordinates of the ensemble had been used. An advantage of these backbone-family methodologies is that they permit the use of fixed-backbone algorithms and energy functions.

The most general approach that has been described allows explicit backbone flexibility at all positions in the backbone. Obviously, this expands the search space tremendously so deterministic optimization algorithms are unlikely to be practical. However, explicit backbone flexibility is well within the reach of Genetic and Monte Carlo algorithms. Using explicit backbone flexibility, our group designed two stable core variants of 434 cro that were predicted to be unstable in the context of a fixed-backbone model. However, in contrast to the reasonably good prediction of energies for sequences designed with a fixed backbone, prediction of stabilities with the flexible backbone was poor. This underscores the inherent difficulties of scoring backbone structures with standard force fields and led to the use of an empirical constraining potential that emphasizes the preser-

vation of local geometry (Desjarlais and Handel, 1999).

Loops and Turns

A simpler problem than total protein or core design with respect to backbone flexibility may be the design of loops and turns (Thanki *et al.*, 1997). Furthermore, important functional residues reside in loops. It is possible that once a stable fold is achieved, loops can be engineered to have different functions, as is the case for natural proteins (e.g., antibodies and TIM barrels). Although much smaller in magnitude than total design, sufficient conformational sampling and accuracy of the energy function remain important issues. A particularly important development for automated loop design has been the construction of an exhaustive classification of loop and turn structures observed in proteins and their sequence preferences (Oliva *et al.*, 1997). Such loop libraries may be used in a manner analogous to how rotamer libraries are used for side chains, reducing the number of backbone conformations that need to be searched.

Outlook and Utility of Scaffold Design

While the major goal of scaffold design has been to understand and analytically describe the forces that determine the structure and stability of proteins, scaffold design does have some practical applications. One very important industrial application is the design of thermostable proteins or those with better solubility characteristics. Mayo and colleagues have reported several examples of proteins that are more stable than the wild-type parent because of modifications to buried, surface, or interfacial residues (Malakauskas and Mayo, 1998).

Sequence design may also become a useful genomics tool for fold recognition by virtue of the ability to predict diverse sequences compatible with a fold (Koehl and Levitt, 1999b; Kuhlman and Baker, 2000; Raha *et al.*, 2000). Artificial sequences that are compatible with a given fold can be used to supplement the database of natural sequences that adopt that fold, in order to further refine profile definitions for threading-based prediction methods. This may become particularly powerful once backbone flexibility can be dealt with.

Finally, it has also been suggested that analysis of families of designed sequences for a fold may aid in the identification of functional positions in proteins. In natural protein sequences, residues of both functional and structural importance are conserved. However, artificial sequences that are selected on the basis of predicted stability will conserve only structurally important residues. It has been ob-

served that positions that are conserved in natural sequences, but not in artificial sequences, are those known to be functionally important (Raha *et al.*, 2000).

BEYOND SCAFFOLD DESIGN— DESIGN AND FUNCTION

The ultimate goal of protein design is the creation of proteins with desired functions. While this currently seems somewhat of a fantasy, it was not that long ago that designing protein scaffolds also seemed inconceivable. Nevertheless, some progress has been made in this area.

In general, protein function requires binding another molecule. Proteins bind molecules to transduce signals and to maintain cell structure. Enzymes bind transition states to catalyze reactions. Designing ligand binding sites in proteins is therefore an important step toward designing functional proteins.

Design of Peptide-Protein Interactions

The design of proteins that can bind peptides has much in common with scaffold design and is accessible with current design tools. Along these lines DeGrado and co-workers designed a two-helix receptor that recognizes the calmodulin binding domain (CBD) of calcineurin, making a three-stranded coiled-coil complex (Ghirlanda *et al.*, 1998). This was initially done by a combination of a core repacking algorithm (Desjarlais and Handel, 1995) and manual placement of potential salt bridges. Unfortunately, the first-generation design oligomerized with itself and did not bind the target peptide. In a second round, charged residues were manually chosen to disfavor oligomerization of the receptor and promote a bimolecular interaction between the receptor and the CBD (Ghirlanda *et al.*, 1998). These data emphasize the important concept of “negative design” for imparting specificity (O’Shea *et al.*, 1992; Ghirlanda *et al.*, 1998; Street *et al.*, 2000).

Design of Alternative Conformations

Proteins often undergo conformational changes upon ligand binding, and these changes regulate protein function. Understanding how to control these changes will be an important goal of functional design. Mayo, Springer, and colleagues have shown that computational protein design can be used to predict sequence changes that lock a protein into specific conformations. Integrin I is a cell-surface adhesion receptor that binds the complement component iC3b. Two crystal forms of the integrin I domain showed that it adopts two conformations, termed the open state and the closed state, and

these were hypothesized to represent the active and inactive conformations, respectively. In order to test this, Shimaoka *et al.* (2000) redesigned the hydrophobic core of this domain, using either the open state structure or the closed state structure as templates. They showed that the affinity of core variants designed to mimic the open state bound ligand with higher affinities than variants designed to mimic the closed state.

Design of Substrate Specificity

Closer to enzyme design is the modification of substrate specificity. One noteworthy study focused on changing the substrate specificity of α -lytic protease. Peptides containing Leu or Ile at the P1 position are poor substrates. Using an energy function that was optimized as described above, the authors computationally screened active site mutants that were predicted to improve activity for the Leu substrate. This was accomplished with a fixed-backbone protein model, discrete side-chain rotamers, and an exhaustive search over 1.4×10^6 sequence-rotamer conformations. They then computationally re-screened the Leu-improved subset to identify those that preferred Leu over Ile. The best prediction was experimentally characterized and found to have an improved activity toward the Leu substrate that quantitatively matched the predicted improvement. In addition, the designed enzyme had a >200-fold preference for Leu over Ile, exactly the goal of the design (Wilson *et al.*, 1991).

Design of Metal Binding Sites

A somewhat more difficult problem is conferring binding or activity onto a protein lacking that function. One strategy is to identify catalytic groups in an enzyme known to perform a reaction of interest and transplant those groups to the appropriate positions in another protein. This is precisely the idea behind the DEZYMER program of Hellinga (Hellinga and Richards, 1991) and related approaches (Clarke and Yuan, 1995; Klemba *et al.*, 1995). Briefly, a set of residues and ligand in a predefined geometry are defined. A protein backbone is then interrogated for positions that can accommodate the desired side chains and ligand. The other peripheral residues are replaced and repacked, if necessary. This program has been used to design zinc, blue copper, and Fe-S metal sites in thioredoxin, a protein that does not naturally support metal binding (Hellinga *et al.*, 1991). Perhaps the most impressive series were variants designed to mimic the metal site in superoxide dismutase (SOD) (Benson *et al.*, 2000). These designed proteins were found to have catalytic activities whose efficiencies depended

strongly on the location and burial of the metal center in the protein, illustrating structural principles that can be discovered by design. Future studies of the designed and natural SODs are likely to reveal further insight into the role that the protein matrix plays in modulating redox activity.

A complementary approach has been described recently for generating artificial backbone templates around a desired ligand binding site. Metal binding sites found in natural proteins can be described and parameterized by a few geometric descriptors, such as the positions of ligand-chelating residues, orientation of backbone groups, and symmetry. These parameters can be used to restrain the location and orientation of the backbone. The backbone is built around the metal binding site and is used as a template for designing the nonchelating residues that stabilize the scaffold and position the binding residues. Using this process, Lombardi *et al.* (2000) reported the successful design of a di-iron site into an artificial four-helix bundle scaffold.

Enzyme Design

Extension of computational methods to broader classes of enzyme function will be a significantly greater challenge. It is thought that all enzymes derive much of their catalytic power by binding to and stabilizing the transition state of a reaction (Pauling, 1948). Based on this premise, Jencks (1969) proposed that antibodies raised against transition state analogs should have catalytic activity. Many groups have shown that this is indeed the case (reviewed by Lerner *et al.* (1991)). To date, approximately 100 different reactions have been catalyzed by antibodies raised against transition-state analog haptens; these include activities never observed in nature, such as the Diels–Alder reaction (Gouverneur *et al.*, 1993; Romesberg *et al.*, 1998).

Despite the remarkable successes of catalytic antibodies, they do suffer from the fact that few have catalytic efficiencies comparable to those of natural enzymes, in part because transition state analogs only approximate the shape and charge distribution of the transition state (reviewed by Hilvert (2000)). Another problem is that design and synthesis of transition state analogs can be difficult and time-consuming. Finally, the scaffold is restricted to the antibody fold, limiting the scope of accessible reactions.

In principle, these limitations can be addressed by computational design of enzymes. However, it will require improvements in energy functions, especially for the treatment of charged and polar groups. For example, enzymes bury functionally important polar groups, but buried positions in most designs are restricted to hydrophobic residues. Enzymes

must be stable, yet flexible enough to undergo the structural fluctuations associated with substrate binding, catalysis, and product release, and this will require the incorporation of backbone flexibility in design algorithms. The resulting increase in combinatorial complexity will also require the development of even more efficient optimization methods. Despite these hurdles, based on the advances described herein, it appears that the field is ready to tackle functional design.

This work was supported by a National Science Foundation Young Investigator Award to T.M.H. and a National Institutes of Health training grant to N.P.

REFERENCES

- Abagyan, R., and Totrov, M. (1994) Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins, *J. Mol. Biol.* **235**, 983–1002.
- Alber, T., Bell, J. A., Sun, D. P., Nicholson, H., Wozniak, J. A., Cook, S., and Matthews, B. W. (1988) Replacements of Pro86 in phage T4 lysozyme extend an alpha-helix but do not alter protein stability, *Science* **239**, 631–635.
- Antosiewicz, J., McCammon, J. A., and Gilson, M. K. (1994) Prediction of pH-dependent properties of proteins, *J. Mol. Biol.* **238**, 415–436.
- Baldwin, E. P., Hajiseyedjavadi, O., Baase, W. A., and Matthews, B. W. (1993) The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme, *Science* **262**, 1715–1718.
- Bashford, D., and Case, D. A. (2000) Generalized Born models of macromolecular solvation effects, *Annu. Rev. Phys. Chem.* **51**, 129–152.
- Bashford, D., and Karplus, M. (1990) pKa's of ionizable groups in proteins: Atomic detail from a continuum electrostatic model, *Biochemistry* **29**, 10219–10225.
- Benson, D. E., Wisz, M. S., and Hellinga, H. W. (2000) Rational design of nascent metalloenzymes, *Proc. Natl. Acad. Sci. USA* **97**, 6292–6297.
- Blaber, M., Zhang, X., and Matthews, B. (1993) Structural basis of amino acid alpha helix propensity, *Science* **260**, 1637–1640.
- Brooks, B., Brucoleri, R., and Olafson, B., *et al.* (1983) CHARMM—A program for macromolecular energy, minimization, and dynamics calculations, *J. Comput. Chem.* **4**, 187–217.
- Bryson, J. W., Desjarlais, J. R., Handel, T. M., and DeGrado, W. F. (1998) From coiled coils to small globular proteins: Design of a native-like three-helix bundle, *Protein Sci.* **7**, 1404–1414.
- Chakrabarty, A., Kortemme, T., and Baldwin, R. (1994) Helix propensities of the amino-acids measured in alanine-based peptides without helix-stabilizing side-chain interactions, *Protein Sci.* **3**, 843–852.
- Chiu, T. L., and Goldstein, R. A. (1998) Optimizing potentials for the inverse protein folding problem, *Protein Eng.* **11**, 749–752.
- Clarke, N. D., and Yuan, S. M. (1995) Metal search: A computer program that helps design tetrahedral metal-binding sites, *Proteins* **23**, 256–263.
- Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J., and Kollman, P. (1996) A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J. Am. Chem. Soc.* **118**, 2309–2309.

- Crick, F. (1953) The packing of alpha-helices: Simple coiled-coils, *Acta Crystallogr.* **6**, 689–697.
- Dahiyat, B. I., Gordon, D. B., and Mayo, S. L. (1997) Automated design of the surface positions of protein helices, *Protein Sci.* **6**, 1333–1337.
- Dahiyat, B. I., and Mayo, S. L. (1996) Protein design automation, *Protein Sci.* **5**, 895–903.
- Dahiyat, B. I., and Mayo, S. L. (1997a) De novo protein design: Fully automated sequence selection, *Science* **278**, 82–87.
- Dahiyat, B. I., and Mayo, S. L. (1997b) Probing the role of packing specificity in protein design, *Proc. Natl. Acad. Sci. USA* **94**, 10172–10177.
- Desjarlais, J. R., and Clarke, N. D. (1998) Computer search algorithms in protein modification and design, *Curr. Opin. Struct. Biol.* **8**, 471–475.
- Desjarlais, J. R., and Handel, T. M. (1995) De novo design of the hydrophobic cores of proteins, *Protein Sci.* **4**, 2006–2018.
- Desjarlais, J. R., and Handel, T. M. (1999) Side-chain and backbone flexibility in protein core design, *J. Mol. Biol.* **290**, 305–318.
- Desmet, J., Maeyer, M., Hazes, B., and Lasters, I. (1992) The dead-end elimination theorem and its use in protein side-chain positioning, *Nature* **356**, 539–542.
- Doig, A. J., and Baldwin, R. L. (1995) N- and C-capping preferences for all 20 amino acids in alpha-helical peptides, *Protein Sci.* **4**, 1325–1336.
- Doig, A. J., Chakrabarty, A., Klingler, T. M., and Baldwin, R. L. (1994) Determination of free energies of N-capping in alpha-helices by modification of the Lifson–Roig helix-coil theory to include N- and C-capping, *Biochemistry* **33**, 3396–3403.
- Dominy, B., and Brooks, C. (1999) Development of a generalized born model parametrization for proteins and nucleic acids, *J. Phys. Chem. B* **103**, 3765–3773.
- Dunbrack, R. L., Jr., and Cohen, F. E. (1997) Bayesian statistical analysis of protein side-chain rotamer preferences, *Protein Sci.* **6**, 1661–1681.
- Edinger, S., Cortis, C., Shenkin, P., and Friesner, R. (1997) Solvation free energies of peptides: Comparison of approximate continuum solvation models with accurate solution of the Poisson–Boltzmann equation, *J. Phys. Chem. B* **101**, 1190–1197.
- Eisenberg, D., and McLachlan, A. D. (1986) Solvation energy in protein folding and binding, *Nature* **319**, 199–203.
- Eriksson, A. E., Baase, W. A., Zhang, X. J., Heinz, D. W., Blaber, M., Baldwin, E. P., and Matthews, B. W. (1992) Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect, *Science* **255**, 178–183.
- Farinas, E., and Regan, L. (1998) The de novo design of a rubredoxin-like Fe site, *Protein Sci.* **7**, 1939–1946.
- Fauchere, J. L., Charton, M., Kier, L. B., Verloop, A., and Pliska, V. (1988) Amino acid side chain parameters for correlation studies in biology and pharmacology, *Int. J. Pept. Protein Res.* **32**, 269–278.
- Ghirlanda, G., Lear, J. D., Lombardi, A., and DeGrado, W. F. (1998) From synthetic coiled coils to functional proteins: Automated design of a receptor for the calmodulin-binding domain of calcineurin, *J. Mol. Biol.* **281**, 379–391.
- Goldstein, R. F. (1994) Efficient rotamer elimination applied to protein side-chains and related spin glasses, *Biophys. J.* **66**, 1335–1340.
- Gordon, D. B., Marshall, S. A., and Mayo, S. L. (1999) Energy functions for protein design, *Curr. Opin. Struct. Biol.* **9**, 509–513.
- Gordon, D. B., and Mayo, S. L. (1999) Branch-and-terminate: A combinatorial optimization algorithm for protein design, *Structure Fold. Des.* **7**, 1089–1098.
- Gouverneur, V. E., Houk, K. N., de Pascual-Teresa, B., Beno, B., Janda, K. D., and Lerner, R. A. (1993) Control of the exo and endo pathways of the Diels–Alder reaction by antibody catalysis, *Science* **262**, 204–208.
- Harbury, P. B., Plecs, J. J., Tidor, B., Alber, T., and Kim, P. S. (1998) High-resolution protein design with backbone freedom, *Science* **282**, 1462–1467.
- Havranek, J. J., and Harbury, P. B. (1999) Tanford–Kirkwood electrostatics for protein modeling, *Proc. Natl. Acad. Sci. USA* **96**, 11145–11150.
- Hawkins, G., Cramer, C., and Truhlar, D. (1995) Pairwise solute descreening of solute charges from a dielectric medium, *Chem. Phys. Lett.* **246**, 122–129.
- Hellinga, H. W., Caradonna, J. P., and Richards, F. M. (1991) Construction of new ligand binding sites in proteins of known structure. II. Grafting of a buried transition metal binding site into *Escherichia coli* thioredoxin, *J. Mol. Biol.* **222**, 787–803.
- Hellinga, H. W., and Richards, F. M. (1991) Construction of new ligand binding sites in proteins of known structure. I. Computer-aided modeling of sites with pre-defined geometry, *J. Mol. Biol.* **222**, 763–785.
- Hensch, Z. S., Jonsson, T., Sauer, R. T., and Tidor, B. (1996) Protein stabilization by removal of unsatisfied polar groups: Computational approaches and experimental tests, *Biochemistry* **35**, 7621–7625.
- Hensch, Z. S., and Tidor, B. (1994) Do salt bridges stabilize proteins? A continuum electrostatic analysis, *Protein Sci.* **3**, 211–226.
- Hensch, Z. S., and Tidor, B. (1999) Electrostatic interactions in the GCN4 leucine zipper: Substantial contributions arise from intramolecular interactions enhanced on binding, *Protein Sci.* **8**, 1381–1392.
- Hilvert, D. (2000) Critical analysis of antibody catalysis, *Annu. Rev. Biochem.* **69**, 751–793.
- Holland, J. (1993) *Adaptation in Natural and Artificial Systems*, The MIT Press, Boston.
- Honig, B., Sharp, K., and Yang, A. (1993) Macroscopic models of aqueous solutions—Biological and chemical applications, *J. Phys. Chem.* **97**, 1101–1109.
- Hurley, J. H., Baase, W. A., and Matthews, B. H. (1992) Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme, *J. Mol. Biol.* **224**, 1143–1159.
- Jencks, W. (1969) *Catalysis in Chemistry and Enzymology*, McGraw-Hill, New York.
- Jiang, X., Farid, H., Pistor, E., and Farid, R. S. (2000) A new approach to the design of uniquely folded thermally stable proteins, *Protein Sci.* **9**, 403–416.
- Johnson, E. C., Lazar, G. A., Desjarlais, J. R., and Handel, T. M. (1999) Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin, *Structure Fold. Des.* **7**, 967–976.
- Jorgensen, W., Maxwell, D., and Tirado-Rives, J. (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids, *J. Am. Chem. Soc.* **118**, 11225–11236.
- Klemba, M., Gardner, K. H., Marino, S., Clarke, N. D., and Regan, L. (1995) Novel metal-binding proteins by design, *Nat. Struct. Biol.* **2**, 368–373.
- Koehl, P., and Delarue, M. (1994) Polar and nonpolar atomic

- environments in the protein core: Implications for folding and binding, *Proteins* **20**, 264–278.
- Koehl, P., and Levitt, M. (1999a) De novo protein design. I. In search of stability and specificity, *J. Mol. Biol.* **293**, 1161–1181.
- Koehl, P., and Levitt, M. (1999b) De novo protein design. II. Plasticity in sequence space, *J. Mol. Biol.* **293**, 1183–1193.
- Kono, H., Nishiyama, M., Tanokura, M., and Doi, J. (1998) Designing the hydrophobic core of thermus flavus malate dehydrogenase based on sidechain packing, *Protein Eng.* **11**, 47–52.
- Kuhlman, B., and Baker, D. (2000) Native protein sequences are close to optimal for their structures, *Proc. Natl. Acad. Sci. USA* **97**, 10383–10388.
- Lazar, G. A., Desjarlais, J. R., and Handel, T. M. (1997) De novo design of the hydrophobic core of ubiquitin, *Protein Sci.* **6**, 1167–1178.
- Lazar, G. A., Johnson, E. C., Desjarlais, J. R., and Handel, T. M. (1999) Rotamer strain as a determinant of protein structural specificity, *Protein Sci.* **8**, 2598–2610.
- Lazaridis, T., and Karplus, M. (1999) Effective energy function for proteins in solution, *Proteins* **35**, 133–152.
- Lee, C. (1994) Predicting protein mutant energetics by self-consistent ensemble optimization, *J. Mol. Biol.* **236**, 918–939.
- Lee, C., and Subbiah, S. (1991) Prediction of protein side-chain conformation by packing optimization, *J. Mol. Biol.* **217**, 373–388.
- Lerner, R., Benkovic, S., and Schultz, P. (1991) At the crossroads of chemistry and immunology: Catalytic antibodies, *Science* **252**, 659–667.
- Lim, W. A., Hodel, A., Sauer, R. T., and Richards, F. M. (1994) The crystal structure of a mutant protein with altered but improved hydrophobic core packing, *Proc. Natl. Acad. Sci. USA* **91**, 423–427.
- Loladze, V. V., Ibarra-Molero, B., Sanchez-Ruiz, J. M., and Makhatazde, G. I. (1999) Engineering a thermostable protein via optimization of charge–charge interactions on the protein surface, *Biochemistry* **38**, 16419–16423.
- Lombardi, A., Summa, C. M., Geremia, S., Randaccio, L., Pavone, V., and DeGrado, W. F. (2000) Inaugural article: Retrostructural analysis of metalloproteins: Application to the design of a minimal model for diiron proteins, *Proc. Natl. Acad. Sci. USA* **97**, 6298–6305.
- Malakauskas, S. M., and Mayo, S. L. (1998) Design, structure and stability of a hyperthermophilic protein variant, *Nat. Struct. Biol.* **5**, 470–475.
- Mayo, S., Olafson, B., and Goddard, W. (1990) Dreiding—A generic force-field for molecular simulations, *J. Phys. Chem.* **94**, 8897–8909.
- Mehler, E. (1996) Self-consistent, free energy based approximation to calculate pH dependent electrostatic effects in proteins, *J. Phys. Chem.* **100**, 16006–16018.
- Mendes, J., Baptista, A. M., Carrondo, M. A., and Soares, C. M. (1999) Improved modeling of side-chains in proteins with rotamer-based methods: A flexible rotamer model, *Proteins* **37**, 530–543.
- Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953) Equations of state calculations by fast computing machines, *J. Chem. Phys.* **21**, 1087–1092.
- Minor, D. L., Jr., and Kim, P. S. (1994) Context is a major determinant of beta-sheet propensity, *Nature* **371**, 264–267.
- Minor, D. L., Jr., and Kim, P. S. (1996) Context-dependent secondary structure formation of a designed protein sequence, *Nature* **380**, 730–734.
- Moult, J., and James, M. N. (1986) An algorithm for determining the conformation of polypeptide segments in proteins by systematic search, *Proteins* **1**, 146–163.
- Murzin, A. G., Lesk, A. M., and Chothia, C. (1994a) Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis, *J. Mol. Biol.* **236**, 1369–1381.
- Murzin, A. G., Lesk, A. M., and Chothia, C. (1994b) Principles determining the structure of beta-sheet barrels in proteins. II. The observed structures, *J. Mol. Biol.* **236**, 1382–1400.
- Nemethy, G., Gibson, K., Palmer, K., Yoon, C., Paterlini, G., Zagari, A., Rumsey, S., and Scheraga, H. (1992) Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides, *J. Phys. Chem.* **96**, 6472–6484.
- Oliva, B., Bates, P. A., Querol, E., Aviles, F. X., and Sternberg, M. J. (1997) An automated classification of the structure of protein loops, *J. Mol. Biol.* **266**, 814–830.
- Onufriev, A., Bashford, D., and Case, D. (2000) Modification of the generalized Born model suitable for macromolecules, *J. Phys. Chem. B* **104**, 3712–3720.
- Ooi, T., Oobatake, M., Nemethy, G., and Scheraga, H. A. (1987) Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides, *Proc. Natl. Acad. Sci. USA* **84**, 3086–3090.
- O'Shea, E. K., Rutkowski, R., and Kim, P. S. (1992) Mechanism of specificity in the Fos-Jun oncoprotein heterodimer, *Cell* **68**, 699–708.
- Pauling, L. (1948) Chemical Achievement and Hope, *Am. Sci.* **36**, 51–58.
- Ponder, J. W., and Richards, F. M. (1987) Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes, *J. Mol. Biol.* **193**, 775–791.
- Qiu, D., Shenkin, P. S., Hollinger, F. P., and Still, W. C. (1997) The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii, *J. Phys. Chem. A* **101**, 3005–3014.
- Raha, K., Wollacott, A. M., Italia, M. J., and Desjarlais, J. R. (2000) Prediction of amino acid sequence from structure, *Protein Sci.* **9**, 1106–1119.
- Romesberg, F. E., Spiller, B., Schultz, P. G., and Stevens, R. C. (1998) Immunological origins of binding and catalysis in a Diels–Alderase antibody, *Science* **279**, 1929–1933.
- Shimaoka, M., Shifman, J. M., Jing, H., Takagi, J., Mayo, S. L., and Springer, T. A. (2000) Computational design of an integrin I domain stabilized in the open high affinity conformation, *Nat. Struct. Biol.* **7**, 674–678.
- Shoichet, B. K., Baase, W. A., Kuroki, R., and Matthews, B. W. (1995) A relationship between protein stability and protein function, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 452–456.
- Shortle, D., and Meeker, A. K. (1986) Mutant forms of staphylococcal nuclease with altered patterns of guanidine hydrochloride and urea denaturation, *Proteins* **1**, 81–89.
- Spector, S., Wang, M., Carp, S. A., Robblee, J., Hendsch, Z. S., Fairman, R., Tidor, B., and Raleigh, D. P. (2000) Rational modification of protein stability by the mutation of charged surface residues, *Biochemistry* **39**, 872–879.
- Srinivasan, R., and Rose, G. D. (1999) A physical basis for protein secondary structure, *Proc. Natl. Acad. Sci. USA* **96**, 14258–14263.
- Street, A. G., Datta, D., Gordon, D. B., and Mayo, S. L. (2000) Designing protein beta-sheet surfaces by Z-score optimization, *Phys. Rev. Lett.* **84**, 5010–5013.

- Street, A. G., and Mayo, S. L. (1998) Pairwise calculation of protein solvent-accessible surface areas, *Fold. Des.* **3**, 253–258.
- Strop, P., Marinescu, A. M., and Mayo, S. L. (2000) Structure of a protein G helix variant suggests the importance of helix propensity and helix dipole interactions in protein design, *Protein Sci.* **9**, 1391–1394.
- Su, A., and Mayo, S. L. (1997) Coupling backbone flexibility and amino acid sequence selection in protein design, *Protein Sci.* **6**, 1701–1707.
- Tanford, C., and Kirkwood, J. (1957) Theory of protein titration. I. General equations for impenetrable spheres, *J. Am. Chem. Soc.* **79**, 5333–5339.
- Thanki, N., Zeelen, J. P., Mathieu, M., Jaenicke, R., Abagyan, R. A., Wierenga, R. K., and Schliebs, W. (1997) Protein engineering with monomeric triosephosphate isomerase (mono-TIM): The modelling and structure verification of a seven-residue loop, *Protein Eng.* **10**, 159–167.
- Tuffery, P., Etchebest, C., Hazout, S., and Lavery, R. (1991) A new approach to the rapid determination of protein side chain conformations, *J. Biomol. Struct. Dyn.* **8**, 1267–1289.
- Voigt, C. A., Gordon, D. B., and Mayo, S. L. (2000) Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design, *J. Mol. Biol.* **299**, 789–803.
- Waldburger, C. D., Schildbach, J. F., and Sauer, R. T. (1995) Are buried salt bridges important for protein stability and conformational specificity? *Nat. Struct. Biol.* **2**, 122–128.
- Walsh, S. T., Cheng, H., Bryson, J. W., Roder, H., and DeGrado, W. F. (1999) Solution structure and dynamics of a de novo designed three-helix bundle protein, *Proc. Natl. Acad. Sci. USA* **96**, 5486–5491.
- Warshel, A., and Papazyan, A. (1998) Electrostatic effects in macromolecules: Fundamental concepts and practical modeling, *Curr. Opin. Struct. Biol.* **8**, 211–217.
- Wernisch, L., Hery, S., and Wodak, S. J. (2000) Automatic protein design with all atom force-fields by exact and heuristic optimization, *J. Mol. Biol.* **301**, 713–736.
- Wesson, L., and Eisenberg, D. (1992) Atomic solvation parameters applied to molecular dynamics of proteins in solution, *Protein Sci.* **1**, 227–235.
- Wilson, C., Mace, J. E., and Agard, D. A. (1991) Computational method for the design of enzymes with altered substrate specificity, *J. Mol. Biol.* **220**, 495–506.
- Wodak, S., and Janin, J. (1980) Analytical approximation to the accessible surface-area of proteins, *Proc. Natl. Acad. Sci. USA* **77**, 1736–1740.
- Wolfenden, R., Andersson, L., Cullis, P., and Southgate, C. (1981) Affinities of amino acid side chains for solvent water, *Biochemistry* **20**, 849–855.