Protein Science (1996), 5:895-903. Cambridge University Press. Printed in the USA.

Protein design automation

BASSIL I. DAHIYAT[1] AND STEPHEN L. MAYO[2]

[1] Division of Chemistry and Chemical Engineering, and [2] Howard Hughes Medical Institute and Division of Biology, California Institute of Technology, Pasadena, California 91125

(RECEIVED December 5, 1995; ACCEPTED February 7, 1996)

Abstract

We have conceived and implemented a cyclical protein design strategy that couples theory, computation, and experimental testing. The combinatorially large number of possible sequences and the incomplete understanding of the factors that control protein structure are the primary obstacles in protein design. Our protein design automation algorithm objectively predicts protein sequences likely to achieve a desired fold. Using a rotamer description of the side chains, we implemented a fast discrete search algorithm based on the Dead-End Elimination Theorem to rapidly find the globally optimal sequence in its optimal geometry from the vast number of possible solutions. Rotamer sequences were scored for steric complementarity using a van der Waals potential. A Monte Carlo search was then executed, starting at the optimal sequence, in order to find other high-scoring sequences. As a test of the design methodology, high-scoring sequences were found for the buried hydrophobic residues of a homodimeric coiled coil based on GCN4-p1. The corresponding peptides were synthesized and characterized by CD spectroscopy and size-exclusion chromatography. All peptides were dimeric and nearly 100% helical at 1 °C, with melting temperatures ranging from 24 °C to 57 °C. A quantitative structure activity relation analysis was performed on the designed peptides, and a significant correlation was found with surface area burial. Incorporation of a buried surface area potential in the scoring of sequences greatly improved the correlation between predicted and measured stabilities and demonstrated experimental feedback in a complete design cycle.

Keywords: computational; dead-end elimination; packing; protein design; side chain

Efforts to design proteins rely on knowledge of the physical properties that determine protein structure, such as the patterns of hydrophobic and hydrophilic residues in the sequence, salt bridges and hydrogen bonds, and secondary structural preferences of amino acids. Various approaches to apply these principles have been attempted. For example, the construction of α-helical and β-sheet proteins with native-like sequences was attempted by individually selecting the residue required at every position in the target fold (Hecht et al., 1990; Quinn et al., 1994). Alternatively, a minimalist approach was used to design helical proteins, where the simplest possible sequence believed to be consistent with the folded structure was generated (Regan & DeGrado, 1988; DeGrado et al., 1989; Handel et al., 1993). An experimental method that relies on the hydrophobic and polar (HP) pattern of a sequence was developed where a library of sequences with the correct pattern for a four-helix bundle was generated by random mutagenesis (Kamtekar et al., 1993). Among non de novo approaches, domains of naturally occurring proteins have been modified or coupled together to achieve a desired tertiary organization (Pessi et al., 1993; Pomerantz et al., 1995).

Reprint requests to: Stephen L. Mayo, 147-75 Biology, California Institute of Technology, Pasadena California 91125; e-mail: steve@mayo.caltech.edu.

Although the correct secondary structure and overall tertiary organization seem to have been attained by several of the above techniques, many designed proteins appear to lack the structural specificity of native proteins. The complementary geometric arrangement of amino acids in the folded protein is the root of this specificity and is encoded in the sequence. However, few protein design methods to date have applied specific packing interactions systematically (Hurley et al., 1992; Hellinga & Richards, 1994; Jones, 1994; Kono & Doi, 1994; Desjarlais & Handel, 1995). In addition, the qualitative nature of many design approaches has hampered the development of improved, second-generation proteins, because there are no objective methods for learning from past design successes and failures.

We have conceived and implemented a cyclical design strategy that couples theory, computation, and experimental testing in order to address the problems of specificity and learning (Fig. 1). Our protein design automation (PDA) cycle is comprised of four components: a design paradigm, a simulation module, experimental testing, and data analysis. The design paradigm is based on the concept of inverse folding (Pabo, 1983; Bowie et al., 1991) and consists of the use of a fixed backbone onto which a sequence of side-chain rotamers can be placed, where rotamers are the allowed conformations of amino acid side chains (Ponder & Richards, 1987). Specific tertiary interactions based on the three-dimensional juxtaposition of atoms

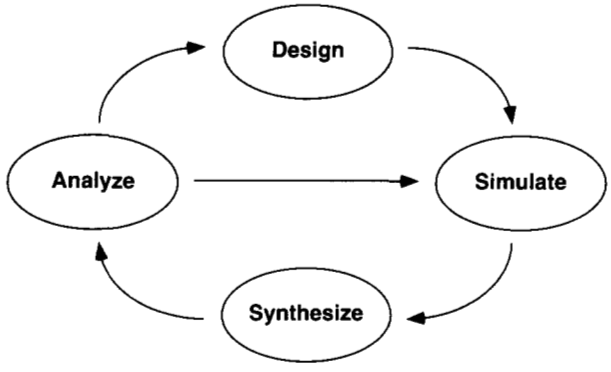

Fig. 1. Protein design automation cycle.

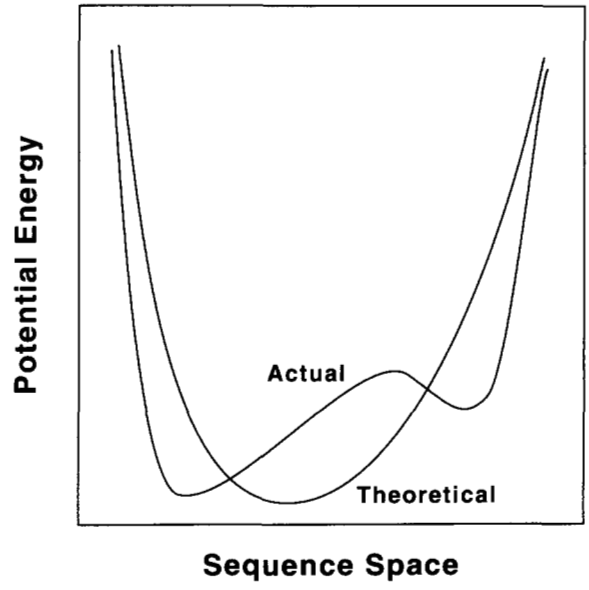

Fig. 2. Schematic of actual versus theoretical potential energy surfaces. The horizontal axis represents all of the possible solutions for the system (all sequences in all possible conformations) and the vertical axis represents the energy of the solutions. Note that the solution space is discrete; continuous lines are used for illustrative purposes only.

are used to determine the sequences that will potentially best adopt the target fold. Given a backbone geometry and the possible rotamers allowed for each residue position as input, the simulation must generate as output a rank-ordered list of solutions based on a cost function that explicitly considers the atom positions in the various rotamers. The principle obstacle is that a fixed backbone comprised of n residues and m possible rotamers per residue (all rotamers of all allowed amino acids) results in m^n possible arrangements of the system, an immense number for even small design problems. For example, to consider 50 rotamers at 15 positions results in more than 10^{25} sequences, which, at an evaluation rate of 10^9 sequences per second (far beyond current capabilities), would take 10^9 *years* to exhaustively search for the global minimum. The synthesis and characterization of a subset of amino acid sequences presented by the simulation module generates experimental data for the analysis module. The analysis section discovers correlations between calculable properties of the simulated structures and the experimental observables. The goal of the analysis is to suggest *quantitative* modifications to the simulation and in some cases to the guiding design paradigm. In other words, the cost function used in the simulation module describes a theoretical potential energy surface whose horizontal axis comprises all possible solutions to the problem at hand (Fig. 2). This potential energy surface is not guaranteed to match the actual potential energy surface that is determined from the experimental data. In this light, the goal of the analysis becomes the correction of the simulation cost function in order to create better agreement between the theoretical and actual potential energy surfaces. If such corrections can be found, then the output of subsequent simulations will be amino acid sequences that better achieve the target properties. This design cycle is generally applicable to any protein system and, by removing the subjective human component, allows a largely unbiased approach to protein design, i.e., protein design automation.

Results and discussion

Design paradigm

The PDA side-chain selection algorithm requires as input a backbone structure defining the desired fold. The task of designing a sequence that takes this fold can be viewed as finding an optimal arrangement of amino acid side chains relative to the given backbone. It is not sufficient to consider *only* the identity of an amino acid when evaluating sequences. In order to correctly account for the geometric specificity of side-chain placement, all possible conformations of each side chain must also be examined. Statistical surveys of the protein structure database (Ponder & Richards, 1987) have defined a discrete set of allowed conformations, called rotamers, for each amino acid side chain. We use a rotamer library based on the Ponder and Richards library to define allowed conformations for the side chains in PDA.

Using a rotamer description of side chains, an optimal sequence for a backbone can be found by screening all possible sequences of rotamers, where each backbone position can be occupied by each amino acid in all its possible rotameric states. The discrete nature of rotamer sets allows a simple calculation of the number of rotamer sequences to be tested. A backbone of length n with m possible rotamers per position will have m^n possible rotamer sequences. The size of the search space grows exponentially with sequence length, which, for typical values of n and m, render intractable an exhaustive search. This combinatorial "explosion" is the primary obstacle to be overcome in the simulation phase of PDA.

Simulation algorithm

We use an extension of the Dead-End Elimination (DEE) theorem (Desmet et al., 1992, 1994; Goldstein, 1994) to solve the combinatorial search problem. The DEE theorem is the basis for a very fast discrete search algorithm that was designed to pack protein side chains on a fixed backbone with a known sequence. Side chains are described by rotamers and an atomistic force field is used to score rotamer arrangements. The DEE theorem guarantees that, if the algorithm converges, the *global*

optimum packing is found. The DEE method is readily extended to our inverse folding design paradigm by simply releasing the constraint that a position is limited to the rotamers of a single amino acid. This extension of DEE greatly increases the number of rotamers at each position and requires a significantly modified implementation to ensure convergence (B.I. Dahiyat & S.L. Mayo, unpubl. results). The guarantee that only the global optimum will be found is still valid, and in our extension means that the globally optimal sequence is found in its optimal conformation. The initial scoring function for sequence arrangements used in the search was an atomic van der Waals potential. The van der Waals potential reflects excluded volume and steric packing interactions, which are important determinants of the specific three-dimensional arrangement of protein side chains.

Following DEE optimization, a rank-ordered list of sequences is generated by a Monte Carlo search in the neighborhood of the DEE solution. This list of sequences is necessary because of possible differences between the theoretical and actual potential surfaces (Fig. 2). Starting at the DEE solution, random positions are changed to other rotamers, and the new sequence energy is calculated. If the new sequence meets the Boltzmann criteria for acceptance, it is used as the starting point for another jump (Metropolis et al., 1953). After a predetermined number of jumps, the best scoring sequences are output as a rank-ordered list. Starting at the global optimum is critical for the Monte Carlo routine to find high-scoring sequences and to avoid searching low-scoring regions of sequence space. Hence, the DEE algorithm and the Monte Carlo search are both critical for providing candidate sequences for experimental testing.

Model system and experimental testing

The homodimeric coiled coil of α helices was selected as the initial design target. Coiled coils are synthesized readily by solid-phase techniques and their helical secondary structure and dimeric tertiary organization ease characterization. Their sequences display a seven-residue periodic HP pattern called a heptad repeat, (**a**·**b**·**c**·**d**·**e**·**f**·**g**) (Cohen & Parry, 1990). The **a** and **d** positions are usually hydrophobic and buried at the dimer interface, whereas the other positions are usually polar and solvent exposed (Fig. 3). The backbone needed for input to the simulation module was taken from the crystal structure of GCN4-p1 (O'Shea et al., 1991). The 16 hydrophobic **a** and **d** positions were optimized in the crystallographically determined fixed field of the rest of the protein. Homodimer sequence symmetry was enforced, only rotamers from hydrophobic amino acids (A, V, L, I, M, F, Y, and W) were considered, and the asparagine at an **a** position, Asn 16, was not optimized.

Optimizing the 16 **a** and **d** positions each with 238 possible hydrophobic rotamers results in 238^{16} or 10^{38} rotamer sequences. The DEE algorithm finds the global optimum in 3 min, including rotamer energy calculation time. The DEE solution matches the naturally occurring GCN4-p1 sequence of **a** and **d** residues for all of the 16 positions. A 10^6-step Monte Carlo search run at a temperature of 1,000 K generated the list of sequences rank ordered by their score. To test reproducibility, the search was repeated three times with different random number seeds and all trials provided essentially identical results. The second best sequence is a Val 30 to Ala mutation and lies 3 kcal/mol above the ground state sequence. Within the top 15 sequences, up to six mutations from the ground state sequence are tolerated, indicating that a variety of packing arrangements are available even for a small coiled coil. Eight sequences with a range of stabilities were selected for experimental testing, including six from the top 15 and two more about 15 kcal/mol higher in energy, the 56th and 70th in the list (Table 1).

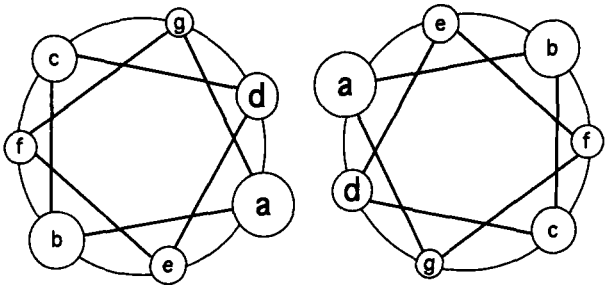

Fig. 3. Helical wheel diagram of a coiled coil. One heptad repeat is shown viewed down the major axes of the helices. The **a** and **d** positions define the solvent-inaccessible core of the molecule (Cohen & Parry, 1990).

The designed **a** and **d** sequences were synthesized using the GCN4-p1 sequence for the **b**·**c** and **e**·**f**·**g** positions. Standard solid-phase techniques were used and, following HPLC purification, the identities of the peptides were confirmed by mass spectrometry. CD spectroscopy was used to assay the secondary structure and thermal stability of the designed peptides. The CD spectra of all the peptides at 1 °C and a concentration of 40 μM exhibit minima at 208 and 222 nm and a maximum at 195 nm, which are diagnostic for α helices (Fig. 4A). The ellipticity values at 222 nm indicate that all of the peptides are >85% helical (approximately −28,000 deg cm^2/dmol), with the exception of PDA-3C, which is 75% helical at 40 μM but increases to 90% helical at 170 μM (Table 2). The melting temperatures (T_m's) show a broad range of values (Fig. 4B), with six of the eight peptides melting at greater than physiological temperature. Also, the T_m's were not correlated to the number of sequence differences from GCN4-p1. Single amino acid changes resulted

Table 1. *Partial Monte Carlo list from coiled coil prediction consisting of the peptides synthesized and characterized*[a]

Name	Sequence	Rank	Energy
PDA-3H[b]	RMKQLEDKVEELLSKNYHLENEVARLKKLVGER	1	−118.1
PDA-3A	RMKQLEDKVEELLSKNYHLENEVARLKKLAGER	2	−115.3
PDA-3G	RMKQLEDKVEELLSKNYHLENEMARLKKLVGER	5	−112.8
PDA-3B	RLKQMEDKVEELLSKNYHLENEVARLKKLVGER	6	−112.6
PDA-3D	RLKQMEDKVEELLSKNYHLENEVARLKKLAGER	13	−109.7
PDA-3C	RMKQWEDKAEELLSKNYHLENEVARLKKLVGER	14	−109.6
PDA-3F	RMKQFEDKVEELLSKNYHLENEVARLKKLVGER	56	−103.9
PDA-3E	RMKQLEDKVEELLSKNYHAENEVARLKKLVGER	70	−103.1

[a] Monte Carlo rank and score are listed and the **a** and **d** positions are indicated by bold type to highlight the optimized positions. The fixed **b**·**c** and **e**·**f**·**g** positions are also included in order to show the complete sequences that were synthesized and tested.

[b] Matches GCN4-p1 wild-type sequence.

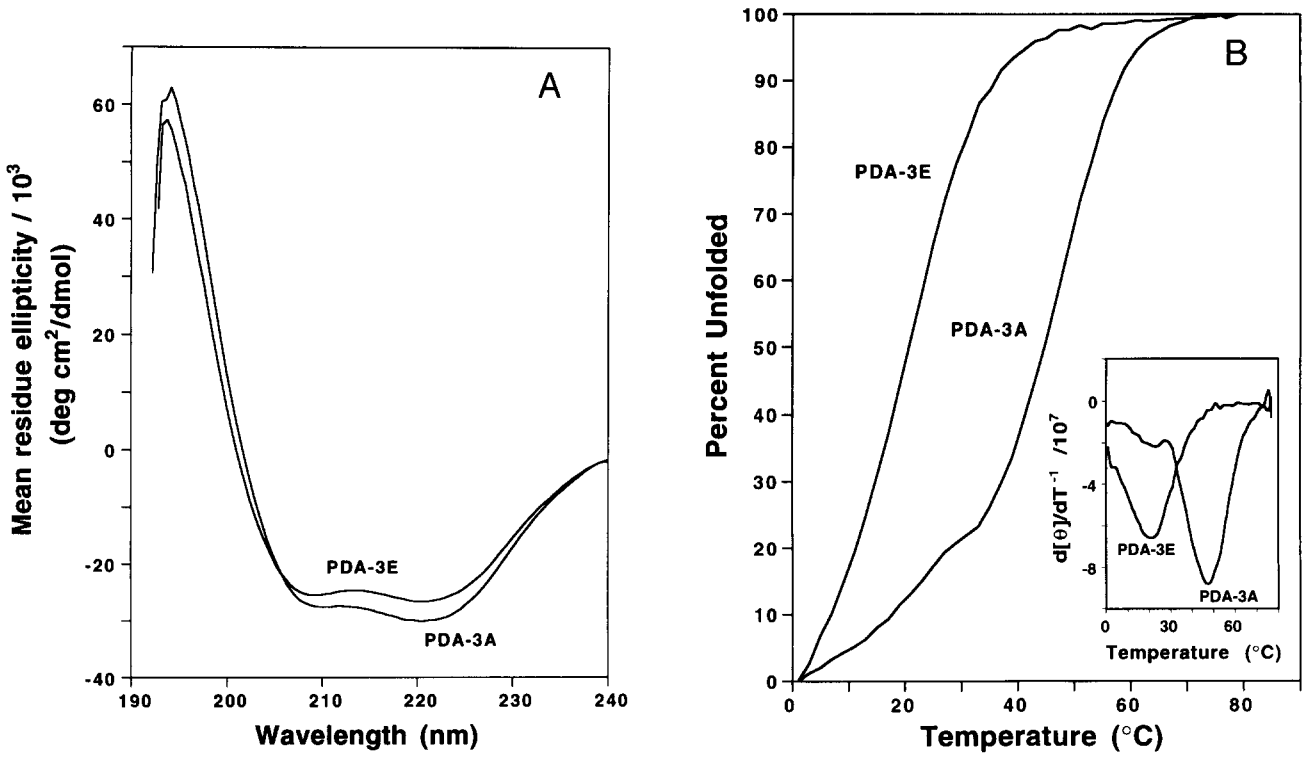

Fig. 4. Typical CD data. **A:** Spectra of PDA-3A and PDA-3E show the minima at 222 and 208 nm and the maximum at 195 nm characteristic of α helices. **B:** Thermal melts of these peptides monitored at 222 nm were used to calculate T_m's from the minima of plots of $d[\theta]/dT^{-1}$ versus T (inset).

in some of the most and least stable peptides, demonstrating the importance of specificity in sequence selection.

Size-exclusion chromatography confirmed the dimeric nature of these designed peptides. Using coiled coil peptides of known oligomerization state as standards, the PDA peptides migrated as dimers. This result is consistent with the appearance of β-branched residues at **a** positions and leucines at **d** positions, which have been shown previously to favor dimerization over other possible oligomerization states (Harbury et al., 1993).

The characterization of the PDA peptides demonstrates the successful design of several stable dimeric helical coiled coils. The sequences were generated automatically in the context of the design paradigm by the simulation module using well-defined inputs that explicitly consider the HP patterning and steric specificity of protein structure. Two-dimensional NMR experiments aimed at probing the specificity of the tertiary packing are the focus of further studies on these peptides. Initial experiments show significant protection of amide protons from chemical ex-

Table 2. *CD data and calculated structural properties of the PDA peptides*[a]

Name	$-[\theta]_{222}$ (deg cm^2/dmol)	T_m (°C)	E_{MC} (kcal/mol)	ΔA_{np} (Å^2)	ΔA_p (Å^2)	Vol (Å^3)	Rot bonds	E_{CQ} (kcal/mol)	E_{CG} (kcal/mol)	E_{vdW} (kcal/mol)	Npb	Pb
PDA-3H	33,000	57	−118.1	2,967	2,341	1,830	28	−234	−308	409	207	128
PDA-3A	30,300	48	−115.3	2,910	2,361	1,725	26	−232	−312	400	203	128
PDA-3B	28,200	47	−112.6	2,977	2,372	1,830	28	−242	−306	379	210	127
PDA-3G	30,700	47	−112.8	3,003	2,383	1,878	32	−240	−309	439	212	128
PDA-3F	28,800	39	−103.9	3,000	2,336	1,872	28	−188	−302	420	212	128
PDA-3D	27,800	39	−109.7	2,920	2,392	1,725	26	−240	−310	370	206	127
PDA-3C	24,100	26	−109.6	2,878	2,400	1,843	26	−149	−304	398	215	129
PDA-3E	27,500	24	−103.1	2,882	2,361	1,674	24	−179	−309	411	203	127

[a] E_{MC} is the Monte Carlo energy; ΔA_{np} and ΔA_p are the changes in solvent-accessible nonpolar and polar surface areas upon folding, respectively; E_{CQ} is the electrostatic energy using equilibrated charges; E_{CG} is the electrostatic energy using Gasteiger charges; E_{vdW} is the van der Waals energy; Vol is the side-chain van der Waals volume; Rot bonds is the number of side-chain rotatable bonds (excluding methyl rotors); Npb and Pb are the number of buried nonpolar and polar atoms, respectively.

change and chemical shift dispersion comparable to GCN4-p1 (B.I. Dahiyat, Y. Xu, & S.L. Mayo, unpubl. results) (Oas et al., 1990; Goodman & Kim, 1991).

Data analysis and design feedback

A detailed analysis of the correspondence between the theoretical and experimental potential surfaces (Fig. 2), and hence an estimate of the accuracy of the simulation cost function, was enabled by the collection of experimental data. Using thermal stability as a measure of design performance, melting temperatures of the PDA peptides were plotted against the sequence scores found in the Monte Carlo search (Fig. 5A). The modest correlation, 0.67, in the plot shows that, although an exclusively van der Waals scoring function can screen for stable sequences, it does not accurately predict relative stabilities. In order to address this issue, correlations between calculated structural properties and T_m's were examined systematically using quantitative structure–activity relationships (QSAR), which is a statistical technique used commonly in structure-based drug design (Hopfinger, 1985).

Table 2 lists various molecular properties of the PDA peptides in addition to the van der Waals-based Monte Carlo scores and the experimentally determined T_m's. A wide range of properties was examined, including molecular mechanics components, such as electrostatic energies, and geometric measures, such as volume. The goal of QSAR is the generation of equations that closely approximate the experimental quantity, in this case T_m, as a function of the calculated properties. Such equations suggest which properties can be used in an improved cost function. The PDA analysis module employs genetic function approximation (GFA) (Rogers & Hopfinger, 1994), a novel method to optimize QSAR equations that selects which properties are to be included and the relative weightings of the properties using a genetic algorithm. GFA accomplishes an efficient search of the space of possible equations and robustly generates a list of equations ranked by their correlation to the data.

Equations are scored by lack of fit (LOF), a weighted least-square error measure that resists overfitting by penalizing equations with more terms (Rogers & Hopfinger, 1994). GFA optimizes both the length and the composition of the equations and, by generating a set of QSAR equations, clarifies combinations of properties that fit well and properties that recur in many equations. All of the top five equations that correct the simulation energy (E_{MC}) contain burial of nonpolar surface area, ΔA_{np} (Table 3). The presence of ΔA_{np} in all of the top equations, in addition to the low LOF of the QSAR containing only E_{MC} and ΔA_{np}, strongly implicates nonpolar surface burial as a critical property for predicting peptide stability. This conclusion is not surprising given the role of the hydrophobic effect in protein energetics (Dill, 1990).

To assess the predictive power of these QSAR equations, as well as their robustness, cross validation analysis was conducted. Each peptide was sequentially removed from the data set and the coefficients of the equation in question were refit. This new equation was then used to predict the withheld data point. When all of the data points had been predicted in this manner, their correlation to the measured T_m's was computed (Table 3). Only the $E_{MC}/\Delta A_{np}$ QSAR and the $E_{MC}/\Delta A_{np}/\Delta A_p$ QSAR performed well in cross validation. The $E_{MC}/\Delta A_{np}$ equation could not be expected to fit the data as smoothly as QSAR's with three

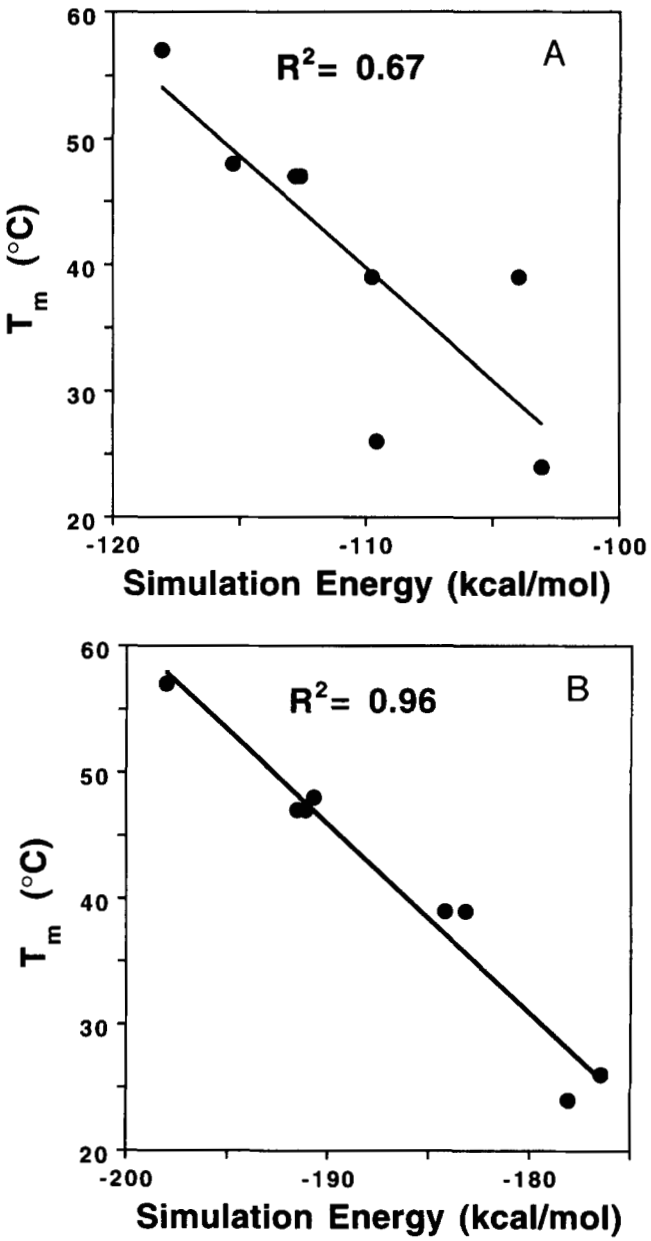

Fig. 5. Comparison of simulation cost functions to experimental T_m's. **A:** Initial cost function, which contains only a van der Waals term for the eight PDA peptides. **B:** Improved cost function containing polar and nonpolar surface area terms weighted by atomic solvation parameters derived from QSAR analysis; 16 cal/mol/Å^2 favors nonpolar surface burial and 86 cal/mol/Å^2 opposes polar surface burial.

terms and hence had a lower cross validated r^2. However, all other two-term QSAR's had LOF scores greater than 48 and cross validation correlations less than 0.55 (data not shown). The QSAR analysis independently predicted with no subjective bias that consideration of nonpolar and polar surface area burial is necessary to improve the simulation. This result is consistent with previous studies on atomic solvation potentials (Eisenberg & McLachlan, 1986; Wesson & Eisenberg, 1992). Further, simpler structural measures, such as number of buried atoms, which reflect underlying principles such as hydrophobic solvation

Table 3. *Top five QSAR equations generated by GFA with LOF, correlation coefficient, and cross validation scores*[a]

QSAR equation	LOF	r^2	CV r^2
$-1.44 * E_{MC} + 0.14 * \Delta A_{np} - 0.73 *$ Npb	16.23	0.98	0.78
$-1.78 * E_{MC} + 0.20 * \Delta A_{np} - 2.43 *$ Rot	23.13	0.97	0.75
$-1.59 * E_{MC} + 0.17 * \Delta A_{np} - 0.05 *$ Vol	24.57	0.97	0.36
$-1.54 * E_{MC} + 0.11 * \Delta A_{np}$	25.45	0.91	0.80
$-1.60 * E_{MC} + 0.09 * \Delta A_{np} - 0.12 * \Delta A_p$	33.88	0.96	0.90

[a] ΔA_{np} and ΔA_p are nonpolar and polar surface buried upon folding, respectively. Vol is side-chain volume, Npb is the number of buried nonpolar atoms, and Rot is the number of rotatable bonds.

(Chan et al., 1995), were not deemed as significant by the QSAR analysis. These results justify the cost of calculating actual surface areas, though in some studies simpler potentials have been shown to perform well (van Gunsteren & Mark, 1992).

ΔA_{np} and ΔA_p were introduced into the simulation module to correct the cost function. Contributions to surface burial from rotamer/template and rotamer/rotamer contacts were calculated and used in the interaction potential. Independently counting buried surface from different rotamer pairs, which is necessary in DEE, leads to overestimation of burial because the radii used in the determination of solvent-accessible surfaces are much larger than the van der Waals contact radii and hence can overlap greatly in a close-packed protein core. To account for this discrepancy, the areas used in the QSAR were recalculated using the pairwise area method and a new $E_{MC}/\Delta A_{np}/\Delta A_p$ QSAR equation was generated. The ratios of the E_{MC} coefficient to the ΔA_{np} and ΔA_p coefficients are scale factors that are used in the simulation module to convert buried surface area into energy, i.e., atomic solvation parameters. Thermal stabilities are predicted well by this cost function (Fig. 5B). In addition, the improved cost function still predicts the naturally occurring GCN4-p1 sequence as the ground state. The surface area to energy scale factors, 16 cal/mol/Å^2 favoring nonpolar area burial and 86 cal/mol/Å^2 opposing polar area burial, are similar in sign, scale, and relative magnitude to solvation potential parameters derived from small molecule transfer data (Wesson & Eisenberg, 1992).

λ *repressor mutants*

To demonstrate the generality of the cost function, other proteins were examined using the simulation module. A library of core mutants of the DNA-binding protein λ repressor has been characterized extensively by Sauer and coworkers (Lim & Sauer, 1991). Specifically, a cluster of three buried residues, V36, M40, and V47, were randomly mutated to Val, Met, Leu, Ile, or Phe. Seventy-eight of the 125 possible combinations were generated. Also, this data set has been used to test several computational schemes and can serve as a basis for comparing different force fields (Lee & Levitt, 1991; van Gunsteren & Mark, 1992; Hellinga & Richards, 1994). The simulation module, using the cost function found by QSAR, was used to find the optimal conformation and energy for each mutant sequence. All hydrophobic residues within 5 Å of the three mutation sites were also left free to be relaxed by the algorithm. This 5-Å sphere contained 12 residues, a significantly larger problem than previous efforts (Lee & Levitt, 1991; Hellinga & Richards, 1994), which were rapidly optimized by the DEE component of the simulation module. The rank correlation of the predicted energy to the combined activity score proposed by Hellinga and Richards is shown in Figure 6. The wild type has the lowest energy of the 125 possible sequences, and the correlation is essentially equivalent to previously published results, which demonstrate that the QSAR-corrected cost function is not specific for coiled coils and can model other proteins adequately.

Concluding remarks

A full circuit of the PDA cycle has been completed. The cores of stable peptides that achieve the target fold have been designed by a largely automated computational design procedure that includes specific tertiary interactions and systematically incorporates experimental feedback. By using DEE, the simulation module can very rapidly find the optimal sequence from the vast number of possibilities. Further, generating a list of candidate sequences and synthesizing and experimentally characterizing them allowed a quantitative analysis of properties important to successful design. A critical feature that had been missing from the simulation, the effect of solvation, was derived from the data and incorporated into the cost function. This feedback improved design performance and, importantly, was not based on subjective interpretation of the data.

The PDA design cycle and its elements can be used in the future as part of de novo protein design, protein redesign, and mutation strategies. Significant challenges that lie ahead include the generation of de novo backbone structures for use in the simulation module, improvement of polar residue rotamer libraries, and the treatment of partially buried and nonburied positions. However, even with these obstacles, strategies such as PDA,

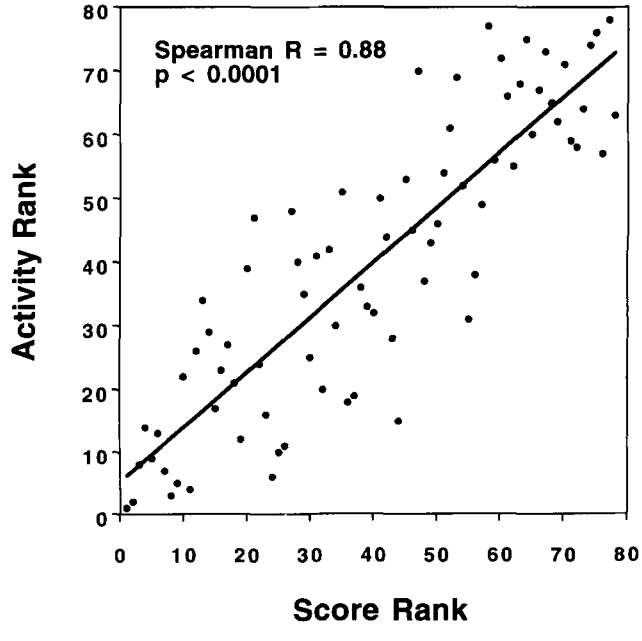

Fig. 6. Rank correlation of energy predicted by the simulation module versus the combined activity score of λ repressor mutants (Lim & Sauer, 1991; Hellinga & Richards, 1994).

which address packing and specific tertiary interactions, will be an important part of protein design in the future.

Methods and materials

Sequence optimization: DEE and Monte Carlo search

Our rotamer library is similar to that used by Desmet and coworkers (Desmet et al., 1992). χ_1 and χ_2 angle values of rotamers for all amino acids except Met, Arg, and Lys were expanded ± 1 SD about the mean value from the Ponder and Richards library in order to minimize possible errors that might arise from the discreteness of the library. χ_3 and χ_4 angles that were undetermined from the database statistics were assigned values of 0° and 180° for Gln and 60°, −60°, and 180° for Met, Lys, and Arg. The number of rotamers per amino acid is: Gly, 1; Ala, 1; Val, 9; Ser, 9; Cys, 9; Thr, 9; Leu, 36; Ile, 45; Phe, 36; Tyr, 36; Trp, 54; His, 54; Asp, 27; Asn, 54; Glu, 69; Gln, 90; Met, 21; Lys, 57; Arg, 55. The cyclic amino acid Pro was not included in the library. Further, all rotamers in the library contained explicit hydrogen atoms. Rotamers were built with bond lengths and angles from the Dreiding force field (Mayo et al., 1990).

A Lennard-Jones 12-6 potential with radii and well depth parameters from the Dreiding force field was used for van der Waals interactions. Nonbonded interactions for atoms connected by one or two bonds were not considered. van der Waals radii for atoms connected by three bonds were scaled by 0.5. Rotamer/rotamer pair energies and rotamer/template energies were calculated in a manner consistent with the published DEE algorithm (Desmet et al., 1992). The template consisted of the protein backbone and the side chains of residue positions not to be optimized. No intraside-chain potentials were calculated. This scheme scored the packing geometry and eliminated bias from rotamer internal energies. Prior to DEE, all rotamers with template interaction energies greater than 25 kcal/mol were eliminated. Also, any rotamer whose interaction was greater than 25 kcal/mol with all other rotamers at another residue position was eliminated. A program called PDA_SETUP was written that takes as input backbone coordinates, including side chains for positions not optimized, a rotamer library, a list of positions to be optimized, and a list of the amino acids to be considered at each position. PDA_SETUP outputs a list of rotamer/template and rotamer/rotamer energies.

The pairwise solvation potential was implemented in two components to remain consistent with the DEE methodology: rotamer/template and rotamer/rotamer burial. For the rotamer/template buried area, the reference state was defined as the rotamer in question at residue i with the backbone atoms only of residues $i-1$, i, and $i+1$. The area of the side chain was calculated with the backbone atoms excluding solvent but not counted in the area. The folded state was defined as the area of the rotamer in question at residue i, but now in the context of the entire template structure, including nonoptimized side chains. The rotamer/template buried area is the difference between the reference and the folded states. The rotamer/rotamer reference area is simply the sum of the areas of the isolated rotamers. The folded state is the area of the two rotamers placed in their relative positions on the protein scaffold, but with no template atoms present. The Richards definition of solvent-accessible surface area (Lee & Richards, 1971) was used, with a probe radius of 1.4 Å and Drieding van der Waals radii. Carbon and sulfur, and all attached hydrogens, were considered nonpolar. Nitrogen and oxygen, and all attached hydrogens, were considered polar. Surface areas were calculated with the Connolly algorithm using a dot density of 10 $Å^{-2}$ (Connolly, 1983). In more recent implementations of PDA_SETUP, the MSEED algorithm of Scheraga has been used in conjunction with the Connolly algorithm to speed up the calculation (Perrot et al., 1992).

DEE was implemented with a novel addition to the improvements suggested by Goldstein (1994). As has been noted, exhaustive application of the $R = 1$ rotamer elimination and $R = 0$ rotamer-pair flagging equations and limited application of the $R = 1$ rotamer-pair flagging equation routinely fails to find the global solution. This problem can be overcome by unifying residues into "super residues" (Desmet et al., 1992, 1994; Goldstein, 1994). However, unification can cause an unmanageable increase in the number of super rotamers per super residue position and can lead to intractably slow performance because the computation time for applying the $R = 1$ rotamer-pair flagging equation increases as the fourth power of the number of rotamers. These problems are of particular importance for protein design applications given the requirement for large numbers of rotamers per residue position. In order to limit memory size and to increase performance, we developed a heuristic that governs which residues (or super residues) get unified and the number of rotamer (or super rotamer) pairs that are included in the $R = 1$ rotamer-pair flagging equation. A manuscript detailing this implementation of DEE is in preparation. A program called PDA_DEE was written that takes a list of rotamer energies from PDA_SETUP and outputs the global minimum sequence in its optimal conformation with its energy.

The Monte Carlo search starts at the global minimum sequence found by DEE. A residue was picked randomly and changed to a random rotamer selected from those allowed at that site. A new sequence energy was calculated and, if it met the Boltzman criteria for acceptance, the new sequence was used as the starting point for another jump. If the Boltzman test failed, then another random jump was attempted from the previous sequence. A list of the best sequences found and their energies was maintained throughout the search. Typically 10^6 jumps were made, 100 sequences saved, and the temperature was set to 1,000 K. After the search was over, all of the saved sequences were quenched by changing the temperature to 0 K, fixing the amino acid identity, and trying every possible rotamer jump at every position. The search was implemented in a program called PDA_MONTE, whose input was a global optimum solution from PDA_DEE and a list of rotamer energies from PDA_SETUP. The output was a list of the best sequences rank ordered by their score.

PDA_SETUP, PDA_DEE, and PDA_MONTE were implemented in the CERIUS2 software development environment (Biosym/Molecular Simulations, San Diego, California).

Coiled coil sequence prediction

Homodimeric coiled coils were modeled on the backbone coordinates of GCN4-p1, PDB ascension code 2ZTA (Bernstein et al., 1977; O'Shea et al., 1991). Atoms of all side chains not optimized were left in their crystallographically determined positions. The program BIOGRAF (Biosym/Molecular Simulations, San Diego, CA) was used to generate explicit hydrogens on the structure which was then conjugate gradient minimized for 50 steps

using the Dreiding forcefield. The HP pattern was enforced by only allowing hydrophobic amino acids into the rotamer groups for the optimized **a** and **d** positions. The hydrophobic group consisted of Ala, Val, Leu, Ile, Met, Phe, Tyr, and Trp for a total of 238 rotamers per position. Homodimer symmetry was enforced by penalizing by 100 kcal/mol rotamer pairs that violate sequence symmetry. Different rotamers of the same amino acid were allowed at symmetry related positions. The asparagine that occupies the **a** position at residue 16 was left in the template and not optimized. A 10^6 step Monte Carlo search run at a temperature of 1000 K generated the list of candidate sequences rank ordered by their score. To test reproducibility, the search was repeated three times with different random number seeds and all trials provided essentially identical results. The Monte Carlo searches took about 90 minutes. All calculations in this work were performed on a Silicon Graphics 200 MHz R4400 processor.

Data analysis and design feedback

Properties were calculated using BIOGRAF and the Dreiding force field. Solvent-accessible surface areas were calculated with the Connolly algorithm (Connolly, 1983) using a probe radius of 1.4 Å and a dot density of 10 $Å^{-2}$. Volumes were calculated as the sum of the van der Waals volumes of the side chains that were optimized. The number of buried polar and nonpolar heavy atoms were defined as atoms, with their attached hydrogens, that expose less than 5 $Å^2$ in the surface area calculation. Electrostatic energies were calculated using a dielectric of one and no cutoff was set for calculation of nonbonded energies. Charge equilibration charges (Rappe & Goddard, 1991) and Gasteiger charges (Gasteiger & Marsili, 1980) were used to generate electrostatic energies. Charge equilibration charges were adjusted manually to provide neutral backbones and neutral side chains in order to prevent spurious monopole effects. The selection of properties was limited by the requirement that properties could not be highly correlated. Correlated properties cannot be differentiated by QSAR techniques and only create redundancy in the derived relations.

Genetic function approximation (GFA) was performed in the CERIUS2 simulation package version 1.6 (Biosym/Molecular Simulations, San Diego, California). An initial population of 300 equations was generated consisting of random combinations of three properties. Only linear terms were used and initial coefficients were determined by least-squares regression for each set of properties. Redundant equations were eliminated and 10,000 generations of random crossover mutations were performed. If a mutant had a better score than the worst equation in the population, the mutant replaced the worst equation. Also, mutation operators that add or remove terms had a 50% probability of being applied each generation, but these mutations were only accepted if the score was improved. No equation with greater than three terms was allowed. Equations were scored during evolution using the lack of fit (LOF) parameter, a scaled least-square error (LSE) measure that penalizes equations with more terms and hence resists overfitting. LOF is defined as:

$$LOF = \frac{LSE}{\left(1 - \frac{2c}{M}\right)^2},$$

where c is the number of terms in the equation and M is the number of data points. Five different randomized runs were done and the final equation populations were pooled. Only equations containing the simulation energy, E_{MC}, were considered, which resulted in 108 equations ranked by their LOF. General cross validation was performed by removing each data point in turn and then fitting the properties of the equation to the remaining data using least-squares regression. The excluded data point was then predicted by the new equation. When all of the data points had been predicted in this way, a correlation coefficient was calculated for the predicted versus the actual data.

λ Repressor simulation

Template coordinates were taken from PDB file 1LMB (Beamer & Pabo, 1992). The subunit designated chain 4 in the PDB file was removed from the context of the rest of the structure (accompanying subunit and DNA) and, using BIOGRAF, explicit hydrogens were added. The hydrophobic residues with side chains within 5 Å of the three mutation sites (V36, M40, V47) are Y22, L31, A37, M42, L50, F51, L64, L65, I68, and L69. All of these residues are greater than 80% buried except for M42, which is 65% buried, and L64, which is 45% buried. A37 only has one possible rotamer and hence was not optimized. The other nine residues in the 5 Å sphere were allowed to take any rotamer conformation of their amino acid. The mutation sites were allowed any rotamer of the amino acid sequence in question. Depending on the mutant sequence, 5×10^{16} to 7×10^{18} conformations were possible. Rotamer energy and DEE calculation times were 2–4 min. The combined activity score is that of Hellinga and Richards (1994).

Peptide synthesis and purification

Thirty-three residue peptides were synthesized on an Applied Biosystems Model 433A peptide synthesizer using Fmoc chemistry, HBTU activation, and a modified Rink amide resin from Novabiochem. Standard 0.1 mmol coupling cycles were used and amino termini were acetylated. Peptides were cleaved from the resin by treating approximately 200 mg of resin with 2 mL trifluoroacetic acid (TFA) and 100 μL water, 100 μL thioanisole, 50 μL ethanedithiol, and 150 mg phenol as scavengers. The peptides were isolated and purified by precipitation and repeated washing with cold methyl tert-butyl ether followed by reverse-phase HPLC on a Vydac C8 column (25 cm × 22 mm) with a linear acetonitrile-water gradient containing 0.1% TFA. Peptides were then lyophilized and stored at −20 °C until use. Plasma desorption mass spectrometry found all molecular weights to be within one unit of the expected masses.

CD

CD spectra were measured on an Aviv 62DS spectrometer at pH 7.0 in 50 mM phosphate, 150 mM NaCl, and 40 μM peptide. A 1-mm-pathlength cell was used and the temperature was controlled by a thermoelectric unit. Thermal melts were performed in the same buffer using 2° temperature increments with an averaging time of 10 s and an equilibration time of 90 s. T_m values were derived from the ellipticity at 222 nm ($[\theta]_{222}$) by evaluating the minimum of the $d[\theta]_{222}/dT^{-1}$ versus T plot (Cantor & Schimmel, 1980). The T_m's were reproducible to within 1°. Peptide concentrations were determined from the tyrosine absorbance at 275 nm (Huyghues-Despointes et al., 1993).

Size-exclusion chromatography

Size-exclusion chromatography was performed with a Synchropak GPC 100 column (25 cm × 4.6 mm) at pH 7.0 in 50 mM phosphate and 150 mM NaCl at 0 °C. GCN4-p1 and p-LI (Harbury et al., 1993) were used as size standards. Ten-microliter injections of 1 mM peptide solution were chromatographed at 0.20 mL/min and monitored at 275 nm. Peptide concentrations were approximately 60 μM as estimated from peak heights. Samples were run in triplicate.

Acknowledgments

We thank J. Desmet and M. De Maeyer for providing us with their rotamer library and a manuscript before publication; P.J. Bjorkman, J. Holton, E.M. Marzluff, D.B. Gordon, and D.C. Rees for critical comments on the manuscript; and B.D. Olafson for help with partial atomic charges and critical comments on the manuscript. This work was supported by the Rita Allen Foundation, the Chandler Family Trust, the Booth Ferris Foundation, the David and Lucile Packard Foundation, and the Searle Scholars Program/The Chicago Community Trust. B.I.D. is partially supported by NIH training grant GM 08346.

References

Beamer LJ, Pabo CJ. 1992. Refined 1.8 angstrom crystal structure of the lambda repressor operator complex. *J Mol Biol 227*:177–196.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Bowie JU, Luthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*:164–170.

Cantor CR, Schimmel PR. 1980. *Biophysical chemistry*. New York: W. H. Freeman and Company.

Chan MK, Mukund S, Kletzin A, Adams MWW, Rees DC. 1995. Structure of a hyperthermophilic tungstopterin enzyme, aldehyde ferredoxin oxidoreductase. *Science 267*:1463–1469.

Cohen C, Parry DAD. 1990. α-Helical coiled coils and bundles: How to design an α-helical protein. *Proteins Struct Funct Genet 7*:1–15.

Connolly ML. 1983. Solvent accessible surfaces of proteins and nucleic acids. *Science 221*:709–713.

DeGrado WF, Wasserman ZR, Lear JD. 1989. Protein design, a minimalist approach. *Science 243*:622–628.

Desjarlais JR, Handel TM. 1995. De novo design of the hydrophobic cores of proteins. *Protein Sci 4*:2006–2018.

Desmet J, De Maeyer M, Hazes B, Lasters I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature 356*:539–542.

Desmet J, De Maeyer M, Lasters I. 1994. The dead-end elimination theorem: A new approach to the side-chain packing problem. In: Merz K Jr, Le Grand S, eds. *The protein folding problem and tertiary structure prediction*. Boston: Birkhauser. pp 307–337.

Dill KA. 1990. Dominant forces in protein folding. *Biochemistry 29*:7133–7155.

Eisenberg D, McLachlan AD. 1986. Solvation energy in protein folding and binding. *Nature 319*:199–203.

Gasteiger J, Marsili M. 1980. Iterative partial equalization of orbital electronegativity—A rapid access to atomic charges. *Tetrahedron 36*:3219–3288.

Goldstein RF. 1994. Efficient rotamer elimination applied to protein side-chains and related spin-glasses. *Biophys J 66*:1335–1340.

Goodman EM, Kim PS. 1991. Periodicity of amide proton exchange rates in a coiled-coil leucine zipper peptide. *Biochemistry 30*:11615–11620.

Handel TM, Williams SA, DeGrado WF. 1993. Metal ion-dependent modulation of the dynamics of a designed protein. *Science 261*:879–885.

Harbury PB, Zhang T, Kim PS, Alber T. 1993. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science 262*:1401–1407.

Hecht MH, Richardson JS, Richardson DC, Ogden RC. 1990. De novo design, expression and characterization of Felix—A four helix bundle protein of native like sequence. *Science 249*:884–891.

Hellinga HW, Richards FM. 1994. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc Natl Acad Sci USA 91*:5803–5807.

Hopfinger AJ. 1985. Computer assisted drug design. *J Med Chem 28*:1133–1139.

Hurley JH, Baase WA, Matthews BW. 1992. Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J Mol Biol 224*:1143–1154.

Huyghues-Despointes BMP, Scholtz JM, Baldwin RL. 1993. Effect of a single aspartate on helix stability at different positions in a neutral alanine based peptide. *Protein Sci 2*:1604–1611.

Jones DT. 1994. De novo protein design using pairwise potentials and a genetic algorithm. *Protein Sci 3*:567–574.

Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH. 1993. Protein design by binary patterning of polar and nonpolar amino acids. *Science 262*:1680–1685.

Kono H, Doi J. 1994. Energy minimization method using automata network for sequence and side-chain conformation prediction from given backbone geometry. *Proteins Struct Funct Genet 19*:244–255.

Lee B, Richards FM. 1971. The interpretation of protein structures: Estimation of static accessibility. *J Mol Biol 55*:379–400.

Lee C, Levitt M. 1991. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature 352*:448–451.

Lim WA, Sauer RT. 1991. The role of internal packing interactions in determining the structure and stability of a protein. *J Mol Biol 219*:359–376.

Mayo SL, Olafson BD, Goddard WA III. 1990. Dreiding—A generic forcefield for molecular simulations. *J Phys Chem 94*:8897–8890.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *J Chem Phys 21*:1087–1092.

Oas TG, McIntosh LP, O'Shea EK, Dahlquist FW, Kim PS. 1990. Secondary structure of a leucine zipper determined by nuclear magnetic resonance spectroscopy. *Biochemistry 29*:2891–2894.

O'Shea EK, Klemm JD, Kim PS, Alber T. 1991. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science 254*:539–544.

Pabo CO. 1983. Designing proteins and peptides. *Nature 301*:200.

Perrot G, Cheng B, Gibson KD, Vila J, Palmer KA, Nayeem A, Maigret B, Scheraga HA. 1992. MSEED: A program for the rapid analytical determination of accessible surface areas and their derivatives. *J Comput Chem 13*:1–11.

Pessi A, Bianchi E, Crameri A, Venturini S, Tramontano A, Sollazzo M. 1993. A designed metal-binding protein with a novel fold. *Nature 362*:367–369.

Pomerantz JL, Sharp PA, Pabo CO. 1995. Structure-based design of transcription factors. *Science 267*:93–96.

Ponder JW, Richards FM. 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.

Quinn TP, Tweedy NB, Williams RW, Richardson JS, Richardson DC. 1994. Betadoublet—De novo design, synthesis and characterization of a beta-sandwich protein. *Proc Natl Acad Sci USA 91*:8747–8751.

Rappe AK, Goddard WA III. 1991. Charge equilibration for molecular dynamics simulations. *J Phys Chem 95*:3358–3363.

Regan L, DeGrado WF. 1988. Characterization of a helical protein designed from first principles. *Science 241*:976–978.

Rogers D, Hopfinger AJ. 1994. Application of genetic function approximation to quantitative structure–property relationships. *J Chem Inf Comput Sci 34*:854–866.

van Gunsteren WF, Mark AE. 1992. Prediction of the activity and stability effects of site-directed mutagenesis on a protein core. *J Mol Biol 227*:389–395.

Wesson L, Eisenberg D. 1992. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci 1*:227–235.