Protein Design Automation: Principles and Practice

Thesis by

Bassil I. Dahiyat

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1998

(Submitted August 5, 1997)

Acknowledgements

I would like to thank my advisor, Stephen L. Mayo, for creating a very challenging and rewarding laboratory to work in. I thank Steve for contributing a great deal of enthusiasm for this work, and a lot of the vision, creative input, and faith needed to sustain this project. Also, I appreciate Steve's tremendous efforts devoted to the programming that underlies all the work described in this thesis.

The chance to work with so many talented and dedicated people in the Mayo group has been a very important part of my time at Caltech. I received a great deal of help both directly in my work, through discussions and from the environment created by the group. Thanks to: Jiye Luo, Yinghai Xu, Jane Takenouchi, Alyce Su, Scott Ross, Sandy Malakauskas, Ben Gordon, Fred Lee, Barry Olafson, Chantal Morgan, Monica Smith, Arthur Street, Cathy Sarisky and Marie Ary.

Jane Sanders and Gary Hathaway often went above and beyond the call of duty providing mass spectra and helpful advice, and Dirk Krapf did the same with amino acid analysis and sequencing. Suzanna Horvath and her group, John Racs and Darcy Campbell, have been very generous with their time and reagents. Tom Meade and Faiz Kayyem helped a great deal, especially in the early days, with advice and the use of equipment, as did Michelle Parks. The members of the Rees group, especially Leemor Joshua-Tor, and the Bjorkman group often helped with advice and discussions. Ron Rock, my roommate, was very helpful the many times I bothered him. My thesis committee, Bill Goddard, Sunney Chan, and Doug Rees were very supportive through the whole process of completing the thesis.

My most special thanks are to my family, for their patience and forbearance of the demands of my work. And to Kim, who has supported and encouraged me so much in the last year and a half. I also thank Kim for bringing joy to my life and for showing me so many good things about others and myself.

Abstract

We have conceived and implemented a cyclical protein design strategy that couples theory, computation and experimental testing. Our goal is an objective, quantitative design algorithm that is based on the physical properties that determine protein structure and stability and which is not limited to specific folds or motifs. Such a method should escape the lack of generality that has resulted from design approaches based on system-specific heuristics and/or subjective considerations. A critical component of the development of our methods has been their experimental testing and validation. The use of a design cycle coupling theory, computation, and experiment has improved our understanding of the physical chemistry governing protein design and hence enhanced the performance of the design algorithm.

Our protein design automation algorithm objectively predicts protein sequences likely to achieve a desired fold by using a side-chain selection algorithm that explicitly and quantitatively considers specific side-chain to backbone and side-chain to side-chain interactions. 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. We subdivided the sequence selection problem into regions of proteins expected to be dominated by different factors: the tightly packed buried core, the solvent exposed surface, and the boundary between core and surface. We assessed the accuracy of a scoring function or combination of scoring functions by experimentally testing their sequence predictions. Improvements to the scoring function were derived from the experimental data and incorporated into the design algorithm. In this manner, we developed a scoring function for the core of a protein that considers packing interactions and hydrophobic solvation energy. In order to design boundary residues effectively, the usually neglected effect of exposed hydrophobic surface area was addressed. Scoring functions for the design of surface residues were developed that account for hydrogen bonding interactions and secondary structure propensities of amino acids. These potential functions were used to successfully redesign several proteins. The integration of these scoring functions was tested by designing the sequence for an entire protein and solving the NMR solution structure of the designed protein. This work reports the first successful automated design and experimental validation of a novel sequence for an entire protein.

Table of Contents

List of Figures and Tables

Figures

Tables

Chapter 1

Introduction

Efforts to control protein structure, and hence function, have generated a great deal of interest in recent years (1, 2). The ability to design novel proteins or redesign existing proteins has obvious practical impact for therapeutic and industrial biotechnology, fields based on proteins. Already, modified functions and specificities have been engineered into proteins by both rational and random library design methods (3, 4, 5), conferring new functions such as peptide ligation or novel substrate recognition. The functions of various proteins have been combined by linking different proteins together (6), and efforts to improve the physical properties of proteins, such as stability or solubility, have shown modest success (7, 8). The design of completely novel structures has been elusive (9, 10), but small peptides with novel sequences that bind receptor targets with high affinity have been selected for and demonstrate the potential for creation of new structures with biological activity. Protein design efforts have also shed light on the factors that control protein stability and folding by allowing critical tests of various structure determinants, such as hydrophobicity and electrostatics (11, 12).

Protein design is the problem of finding a sequence of amino acids that will take a desired protein structure and perform a desired function. Given the practical limits on the size of random sequence libraries, designing all but the smallest protein structures will necessarily depend on knowledge of the physical properties that determine protein structure. Rational design efforts typically incorporate effects 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 (13, 14). 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 (9, 15, 16). 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 (11). Among non de novo approaches, domains of naturally occurring proteins have been modified or coupled together to achieve a desired tertiary organization (6, 17). Iterative mutation of proteins and testing of the new sequences has also been used to design proteins (12, 18, 19, 20).

Though 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 systematically applied specific packing interactions (21, 22, 23, 24, 25). 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. Even more limiting is the inability of these design techniques to generalize to other protein motifs.

Recently, several groups have applied and experimentally tested systematic, quantitative methods for protein design (10, 21, 24, 25, 26, 27, 28). Quantitative approaches to protein design should aid the development of improved, second-generation protein designs because changes to the underlying design principles can be objectively incorporated into the method and avoid subjective biasing of the results. Quantitative techniques are necessary to automate the design procedure and to reduce the time needed to try successive generations of designs. Also, though some proposed algorithms rely on statistical distributions of amino acids to design proteins (23), most are based on physical chemical potentials that screen possible sequences for compatibility with the desired protein fold. Such methods should escape the lack of generality that has resulted from design approaches based on system-specific heuristics and/or subjective considerations.

The use of a quantitative, physical chemistry based method for protein design requires an accurate set of potential functions for scoring of sequences. Unfortunately, current molecular mechanics forcefields are more suited to modeling small perturbations in protein systems and for the refinement of experimentally determined structures. The requirements of protein design are more difficult. The selection of a sequence, which is the essence of protein design, is a comparison of different molecules and is fundamentally more challenging than the comparison of different conformations of the same molecule, which is the most common use of forcefields. Also, the tremendous number of possible sequences for a protein of a given length creates an immense optimization problem. In order to search this vast sequence space efficiently, many design approaches reduce the degrees of freedom in the protein structure by using a fixed protein fold template, in addition to fixed bond lengths and angles. Interactions between residues are often precalculated to speed optimization algorithm performance. This reduction of the energy calculations to pairwise terms requires recasting of the energy expressions and novel methods for computing certain components, such as surface area based solvation potentials. To date, these techniques, which screen possible sequences for compatibility with the desired protein fold, have focused mostly on the redesign of protein cores. Also, the redesign efforts have met with mixed success due to the imperfect sequence scoring functions used (29).

We have sought to expand the range of computational protein design by developing quantitative design methods for residues of all parts of a protein: the buried core, the solvent exposed surface, and the boundary between core and surface (28, 30, 31). A critical component of the development of our methods has been their experimental testing and validation. The use of a design cycle coupling theory, computation, and experiment has improved our understanding of the physical chemistry governing protein design and hence enhanced the performance of the design algorithm (28). The following chapters detail the development of the various parts of the design algorithm and their integration and testing for the design of a complete protein.

Since protein design is the problem of finding an amino acid sequence to take a desired fold, a complete quantitative design algorithm requires both a quantitative description of the target backbone structure and a sequence optimization algorithm to find the best amino acid sequence for the fold. Though backbone specification and sequence selection are coupled, we chose to develop the sequence selection algorithm first because it can be readily tested on known backbones from solved protein structures. Conversely, the testing and development of a backbone generation technique requires an accurate sequence selection algorithm, since the function of the designed backbone is to serve as an effective scaffold for sequence optimization. Further, the redesign of known structures to take different, and hopefully improved, properties is an important goal that can be achieved with sequence selection alone. Recent results indicate that the sequence selection algorithm is not sensitive to even fairly large perturbations in backbone geometry and should be robust enough to accommodate computer-generated backbones (32).

We subdivided the sequence selection problem into regions of proteins expected to be dominated by different factors: the tightly packed buried core, the solvent exposed surface, and the boundary between core and surface. In Chapter II, the tremendous combinatorial optimization problem inherent to sequence selection is addressed using the Dead-End Elimination Theorem (DEE) (33). In addition, the development of a scoring function for core residue sequences is presented. In Chapter III, the balance between van der Waals packing forces and hydrophobic solvation was examined to determine bounds on the need for packing specificity in protein core design. Also, analysis and experimental characterization of designed sequences suggested factors important for the design of boundary residues. Chapter IV presents the initial development of scoring functions for the surface of a protein. In Chapter V, the entire sequence selection algorithm was tested by designing a complete protein sequence and solving the NMR solution structure of the designed protein. This test of the algorithm was successful, though further tests to examine the generality of the design algorithm to other protein structures are needed.

References

1. C. O. Pabo, *Nature* **301**, 200 (1983).

2. M. H. J. Cordes, A. R. Davidson, R. T. Sauer, *Curr. Op. Struct. Biol.* **6**, 3-10 (1996).

3. Q. D. Dang, E. R. Guinto, E. Dicera, *Nature Biotechnology* **15**, 146-149 (1997).

4. D. Y. Jackson, et al., *Science* **266**, 243-247 (1994).

5. B. Li, et al., *Science* **270**, 1657-1660 (1995).

6. J. L. Pomerantz, P. A. Sharp, C. O. Pabo, *Science* **267**, 93-96 (1995).

7. J. Davies, L. Riechmann, *Prot. Eng.* **9**, 531-537 (1996).

8. F. M. Zhang, et al., *Nature* **387**, 206-209 (1997).

9. T. M. Handel, S. A. Williams, W. F. DeGrado, *Science* **261**, 879-885 (1993).

10. S. F. Betz, W. F. Degrado, *Biochemistry* **35**, 6955-6962 (1996).

11. S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, M. H. Hecht, *Science* **262**, 1680-1685 (1993).

12. S. Nautiyal, D. N. Woolfson, D. S. King, T. Alber, *Biochemistry* **34**, 11645-11651 (1995).

13. T. P. Quinn, N. B. Tweedy, R. W. Williams, J. S. Richardson, D. C. Richardson, *Proc. Natl. Acad. Sci. USA* **91**, 8747-8751 (1994).

14. M. H. Hecht, J. S. Richardson, D. C. Richardson, R. C. Ogden, *Science* **249**, 884-891 (1990).

15. W. F. DeGrado, Z. R. Wasserman, J. D. Lear, *Science* **243**, 622-628 (1989).

16. L. Regan, W. F. DeGrado, *Science* **241**, 976-978 (1988).

17. A. Pessi, et al., *Nature* **362**, 367-369 (1993).

18. M. D. Struthers, R. P. Cheng, B. Imperiali, *Science* **271**, 342-345 (1996).

19. E. K. O'shea, K. J. Lumb, P. S. Kim, *Current Biology* **3**, 658-667 (1993).

20. P. B. Harbury, T. Zhang, P. S. Kim, T. Alber, *Science* **262**, 1401-1407 (1993).

21. J. H. Hurley, W. A. Baase, B. W. Matthews, *J. Mol. Biol.* **224**, 1143-1154 (1992).

22. H. Kono, J. Doi, *Proteins: Structure, Function and Genetics* **19**, 244-255 (1994).

23. D. T. Jones, *Protein Sci.* **3**, 567-574 (1994).

24. H. W. Hellinga, F. M. Richards, *Proc. Natl. Acad. Sci. USA* **91**, 5803-5807 (1994).

25. J. R. Desjarlais, T. M. Handel, *Protein Sci.* **4**, 2006-2018 (1995).

26. P. B. Harbury, B. Tidor, P. S. Kim, *Proc. Natl. Acad. Sci. USA* **92**, 8408-8412 (1995).

27. M. Klemba, K. H. Gardner, S. Marino, N. D. Clarke, L. Regan, *Nature Struc. Biol.* **2**, 368-373 (1995).

28. B. I. Dahiyat, S. L. Mayo, *Protein Sci.* **5**, 895-903 (1996).

29. G. A. Lazar, J. R. Desjarlais, T. M. Handel, *Protein Sci.* **6**, 1167-1178 (1997).

30. B. I. Dahiyat, S. L. Mayo, *Protein Sci.* **6**, 1333-1337 (1997).

31. B. I. Dahiyat, S. L. Mayo, *Proc. Natl. Acad. Sci. USA* **94**, 10172-10177 (1997).

32. A. Su, S. L. Mayo, *Protein Sci.* **6**, 1701-1707 (1997).

33. J. Desmet, M. De Maeyer, B. Hazes, I. Lasters, *Nature* **356**, 539-542 (1992).

Chapter 2

Automated Core Sequence Selection

The text of this chapter is adapted from a published manuscript that was coauthored with Professor Stephen L. Mayo.

B. I. Dahiyat and S. L. Mayo, *Protein Sci.*, **5**, 895 (1996).

Abstract

We have conceived and implemented a cyclical protein design strategy that couples theory, computation and experimental testing. 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. 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. 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 circular dichroism spectroscopy and size exclusion chromatography. All peptides were dimeric and nearly 100% helical at 1 °C with melting temperatures ranging from 24-57 °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.

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 (Figure 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 (1, 2) 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 (3). Specific tertiary interactions based on the three-dimensional juxtaposition of atoms 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 over 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 (Figure 2). This potential energy surface is not guaranteed to match the actual potential energy surface which 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 (3) 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 (4, 5, 6) 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 forcefield 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 (Dahiyat and Mayo, unpublished 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 (Figure 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 (7). 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 readily synthesized 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**) (8). The **a** and **d** positions are usually hydrophobic and buried at the dimer interface while the other positions are usually polar and solvent exposed (Figure 3). The backbone needed for input to the simulation module was taken from the crystal structure of GCN4-p1 (9). 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 three minutes, 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 one million step Monte Carlo search run at a temperature of 1000 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 three 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).

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. Circular dichroism spectroscopy (CD) 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 (Figure 4A). The ellipticity values at 222 nm indicate that all of the peptides are >85% helical (approximately -28000 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 (Figure 4B), with 6 of the 8 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 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 (10).

The characterization of the PDA peptides demonstrates the successful design of several stable dimeric helical coiled coils. The sequences were automatically generated 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 nuclear magnetic resonance 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 exchange and chemical shift dispersion comparable to GCN4-p1 (Dahiyat, Xu and Mayo, unpublished results) (11, 12).

Data analysis and design feedback

A detailed analysis of the correspondence between the theoretical and experimental potential surfaces (Figure 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 (Figure 5A). The modest correlation, 0.67, in the plot shows that while 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 systematically examined using quantitative structure activity relationships (QSAR), which is a statistical technique commonly used in structure based drug design (13).

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) (14), 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 (14). 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 (15).

To assess the predictive power of these QSAR equations, as well as their robustness, cross validation analysis was carried out. 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 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 (16, 17). Further, simpler structural measures, such as number of buried atoms, that reflect underlying principles such as hydrophobic solvation (18) 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 (19).

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 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 (Figure 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 (17).

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 extensively characterized by Sauer and coworkers (20). Specifically, a cluster of three buried residues, V36, M40 and V47, were randomly mutated to Val, Met, Leu, Ile or Phe. Of the 125 possible combinations, 78 were generated. Also, this dataset has been used to test several computational schemes and can serve as a basis for comparing different forcefields (19, 21, 22). 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 (21, 22), that 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 wildtype has the lowest energy of the 125 possible sequences and the correlation is essentially equivalent to previously published results which demonstrates 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 non-buried positions. However, even with these obstacles, strategies such as PDA that address packing and specific tertiary interactions will be an important part of protein design in the future.

Methods and Materials

Sequence optimization: dead end elimination and Monte Carlo search

Our rotamer library is similar to that used by Desmet and coworkers (4). χ_1 and χ_2 angle values of rotamers for all amino acids except Met, Arg and Lys were expanded plus and minus one standard deviation 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 forcefield (23).

A Lennard-Jones 12-6 potential with radii and well depth parameters from the Dreiding forcefield was used for van der Waals interactions. Non-bonded 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 (4). The template consisted of the protein backbone and the side chains of residue positions not to be optimized. No intra-side-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 non-optimized 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 (24) 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} (25). 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 (26).

DEE was implemented with a novel addition to the improvements suggested by Goldstein (6). 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" (4, 5, 6). However, unification can cause an unmanageable increase in the number of super rotamers per super residue position and can lead to intractably slow performance since 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 1000 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, CA).

Coiled coil sequence prediction

Homodimeric coiled coils were modeled on the backbone coordinates of GCN4-p1, PDB ascension code 2ZTA (9, 27). 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 forcefield. Solvent accessible surface areas were calculated with the Connolly algorithm (25) 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 non-bonded energies. Charge equilibration charges (28) and Gasteiger (29) charges were used to generate electrostatic energies. Charge equilibration charges were manually adjusted 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, CA). 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 10000 generations of random crossover mutations were performed. If a child had a better score than the worst equation in the population, the child 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 (30). 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 to 4 minutes. The combined activity score is that of Hellinga and Richards (22).

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 by 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.

Circular dichroism

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 two degree 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 (31). The T_m's were reproducible to within one degree. Peptide concentrations were determined from the tyrosine absorbance at 275 nm (32).

Size exclusion chromatography

Size exclusion chromatography was performed with a Synchropak GPC 100 column (25 cm by 4.6 mm) at pH 7.0 in 50 mM phosphate and 150 mM NaCl at 0 °C. GCN4-p1 and p-LI (10) were used as size standards. 10 μl 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.

Acknowledgements

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

1. C. O. Pabo, *Nature* **301**, 200 (1983).
2. J. U. Bowie, R. Luthy, D. Eisenberg, *Science* **253**, 164-170 (1991).
3. J. W. Ponder, F. M. Richards, *J. Mol. Biol.* **193**, 775-791 (1987).
4. J. Desmet, M. De Maeyer, B. Hazes, I. Lasters, *Nature* **356**, 539-542 (1992).
5. J. Desmet, M. De Maeyer, I. Lasters, in *The protein folding problem and tertiary structure prediction* K. Merz Jr, S. Le Grand, Eds. (Birkhauser, Boston, 1994) pp. 307-337.
6. R. F. Goldstein, *Biophys. J.* **66**, 1335-1340 (1994).
7. N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, *J. Chem. Phys.* **21**, 1087-1092 (1953).
8. C. Cohen, D. A. D. Parry, *Proteins: Structure, Function and Genetics* **7**, 1-15 (1990).
9. E. K. O'Shea, J. D. Klemm, P. S. Kim, T. Alber, *Science* **254**, 539-544 (1991).
10. P. B. Harbury, T. Zhang, P. S. Kim, T. Alber, *Science* **262**, 1401-1407 (1993).
11. T. G. Oas, L. P. McIntosh, E. K. O'Shea, F. W. Dahlquist, P. S. Kim, *Biochemistry* **29**, 2891-2894 (1990).
12. E. M. Goodman, P. S. Kim, *Biochemistry* **30**, 11615-11620 (1991).
13. A. J. Hopfinger, *Journal of Medicinal Chemistry* **28**, 1133-1139 (1985).
14. D. Rogers, A. J. Hopfinger, *Journal of Chemical Information and Computer Science* **34**, 854-866 (1994).
15. K. A. Dill, *Biochemistry* **29**, 7133-7155 (1990).
16. D. Eisenberg, A. D. McLachlan, *Nature* **319**, 199-203 (1986).
17. L. Wesson, D. Eisenberg, *Protein Sci.* **1**, 227-235 (1992).
18. M. K. Chan, S. Mukund, A. Kletzin, M. W. W. Adams, D. C. Rees, *Science* **267**, 1463-1469 (1995).
19. W. F. van Gunsteren, A. E. Mark, *J. Mol. Biol.* **227**, 389-395 (1992).

20. W. A. Lim, R. T. Sauer, *J. Mol. Biol.* **219**, 359-376 (1991).

21. C. Lee, M. Levitt, *Nature* **352**, 448-451 (1991).

22. H. W. Hellinga, F. M. Richards, *Proc. Natl. Acad. Sci. USA* **91**, 5803-5807 (1994).

23. S. L. Mayo, B. D. Olafson, W. A. Goddard III, *J. Phys. Chem.* **94**, 8897-8909 (1990).

24. B. Lee, F. M. Richards, *J. Mol. Biol.* **55**, 379-400 (1971).

25. M. L. Connolly, *Science* **221**, 709-713 (1983).

26. G. Perrot, et al., *Journal of Computational Chemistry* **13**, 1-11 (1992).

27. F. C. Bernstein, et al., *J. Mol. Biol.* **112**, 535-542 (1977).

28. A. K. Rappe, W. A. Goddard III, *J. Phys. Chem.* **95**, 3358-3363 (1991).

29. J. Gasteiger, M. Marsili, *Tetrahedron* **36**, 3219-3288 (1980).

30. L. J. Beamer, C. J. Pabo, *J. Mol. Biol.* **227**, 177-196 (1992).

31. C. R. Cantor, P. R. Schimmel, *Biophysical Chemistry* (W. H. Freeman and Company, New York, 1980), vol. 3.

32. B. M. P. Huyghues-Despointes, J. M. Scholtz, R. L. Baldwin, *Protein Sci.* **2**, 1604-1611 (1993).

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

Name	Sequence	Rank	Energy
PDA-3H*	R**M**KQ**L**EDK**V**EE**L**LSKNYH**L**ENE**V**AR**L**KKL**V**GER	1	-118.1
PDA-3A	R**M**KQ**L**EDK**V**EE**L**LSKNYH**L**ENE**V**AR**L**KKL**A**GER	2	-115.3
PDA-3G	R**M**KQ**L**EDK**V**EE**L**LSKNYH**L**ENE**M**AR**L**KKL**V**GER	5	-112.8
PDA-3B	R**L**KQ**M**EDK**V**EE**L**LSKNYH**L**ENE**V**AR**L**KKL**V**GER	6	-112.6
PDA-3D	R**L**KQ**M**EDK**V**EE**L**LSKNYH**L**ENE**V**AR**L**KKL**A**GER	13	-109.7
PDA-3C	R**M**KQ**W**EDK**A**EE**L**LSKNYH**L**ENE**V**AR**L**KKL**V**GER	14	-109.6
PDA-3F	R**M**KQ**F**EDK**V**EE**L**LSKNYH**L**ENE**V**AR**L**KKL**V**GER	56	-103.9
PDA-3E	R**M**KQ**L**EDK**V**EE**L**LSKNYH**A**ENE**V**AR**L**KKL**V**GER	70	-103.1

*Matches GCN4-p1 wildtype sequence.

Table 2. CD data and calculated structural properties of the PDA peptides.

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)	Rb	E_{CQ} (kcal/ mol)	E_{CG} (kcal/ mol)	E_{vdW} (kcal/ mol)	Npb	Pb
3H	33000	57	-118.1	2967	2341	1830	28	-234	-308	409	207	128
3A	30300	48	-115.3	2910	2361	1725	26	-232	-312	400	203	128
3B	28200	47	-112.6	2977	2372	1830	28	-242	-306	379	210	127
3G	30700	47	-112.8	3003	2383	1878	32	-240	-309	439	212	128
3F	28800	39	-103.9	3000	2336	1872	28	-188	-302	420	212	128
3D	27800	39	-109.7	2920	2392	1725	26	-240	-310	370	206	127
3C	24100	26	-109.6	2878	2400	1843	26	-149	-304	398	215	129
3E	27500	24	-103.1	2882	2361	1674	24	-179	-309	411	203	127

*E_{MC} is the Monte Carlo energy; ΔA_{np} and ΔA_p are the changes in solvent accessible non-polar 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; Rb is the number of side chain rotatable bonds (excluding methyl rotors); Npb and Pb are the number of buried non-polar and polar atoms, respectively.

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

QSAR equation	LOF	r^2	CV r^2
$-1.44{*}E_{MC} + 0.14{*}\Delta A_{np} - 0.73{*}Npb$	16.23	.98	.78
$-1.78{*}E_{MC} + 0.20{*}\Delta A_{np} - 2.43{*}Rot$	23.13	.97	.75
$-1.59{*}E_{MC} + 0.17{*}\Delta A_{np} - 0.05{*}Vol$	24.57	.97	.36
$-1.54{*}E_{MC} + 0.11{*}\Delta A_{np}$	25.45	.91	.80
$-1.60{*}E_{MC} + 0.09{*}\Delta A_{np} - 0.12{*}\Delta A_{p}$	33.88	.96	.90

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 buried rotatable bonds.

Figure 2-1. Protein design automation cycle.

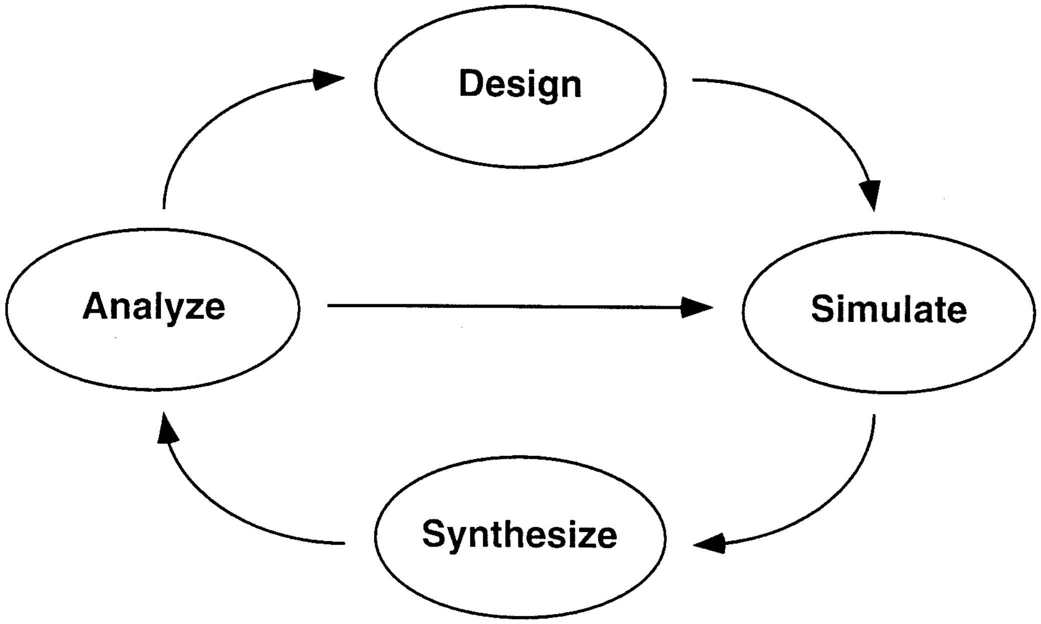

Design
Analyze
Simulate
Synthesize

Figure 2-2. 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.

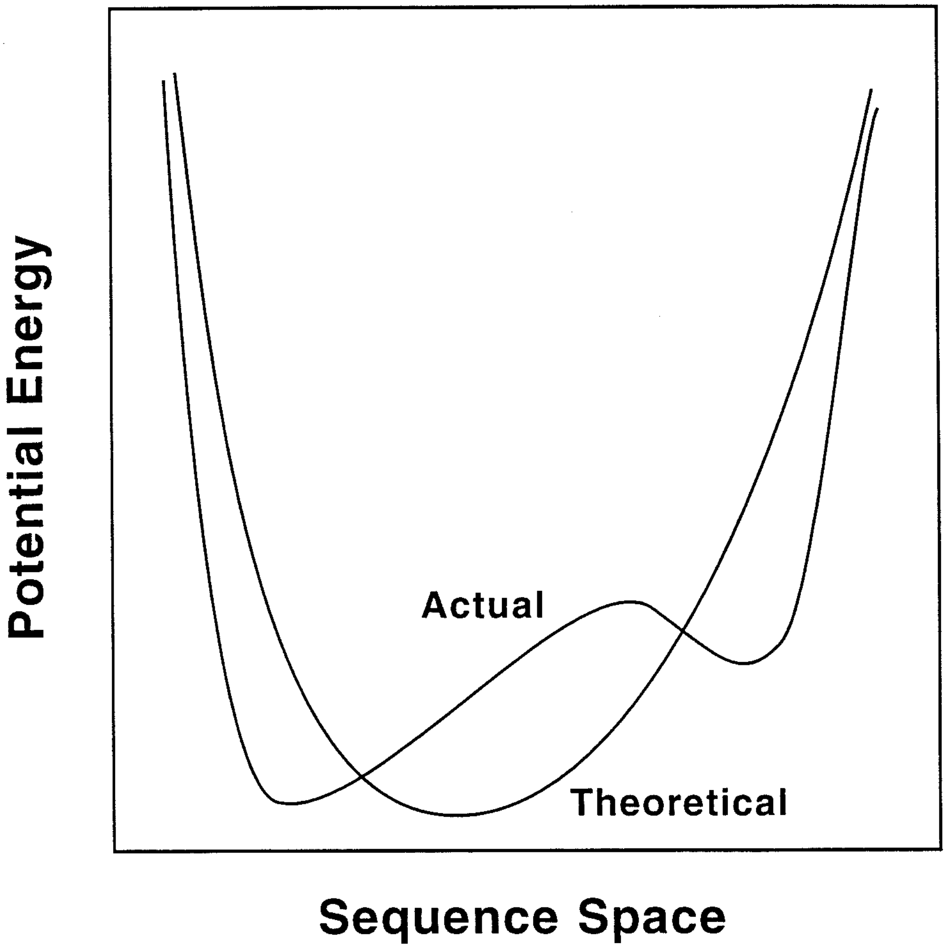

Potential Energy
Actual
Theoretical
Sequence Space

Figure 2-3. 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 (8).

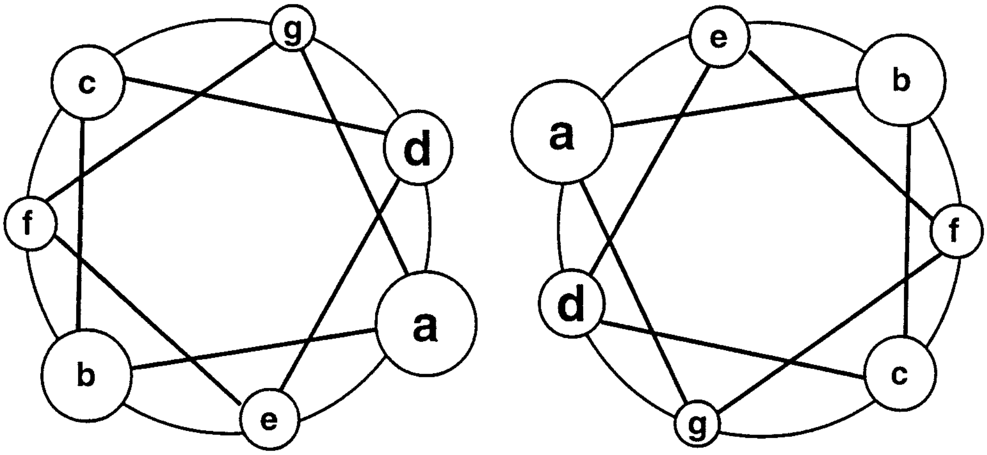

g
c
d
f
a
b
e
e
b
a
f
d
c
g

Figure 2-4. 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[θ]/dT^{-1} versus T (inset).

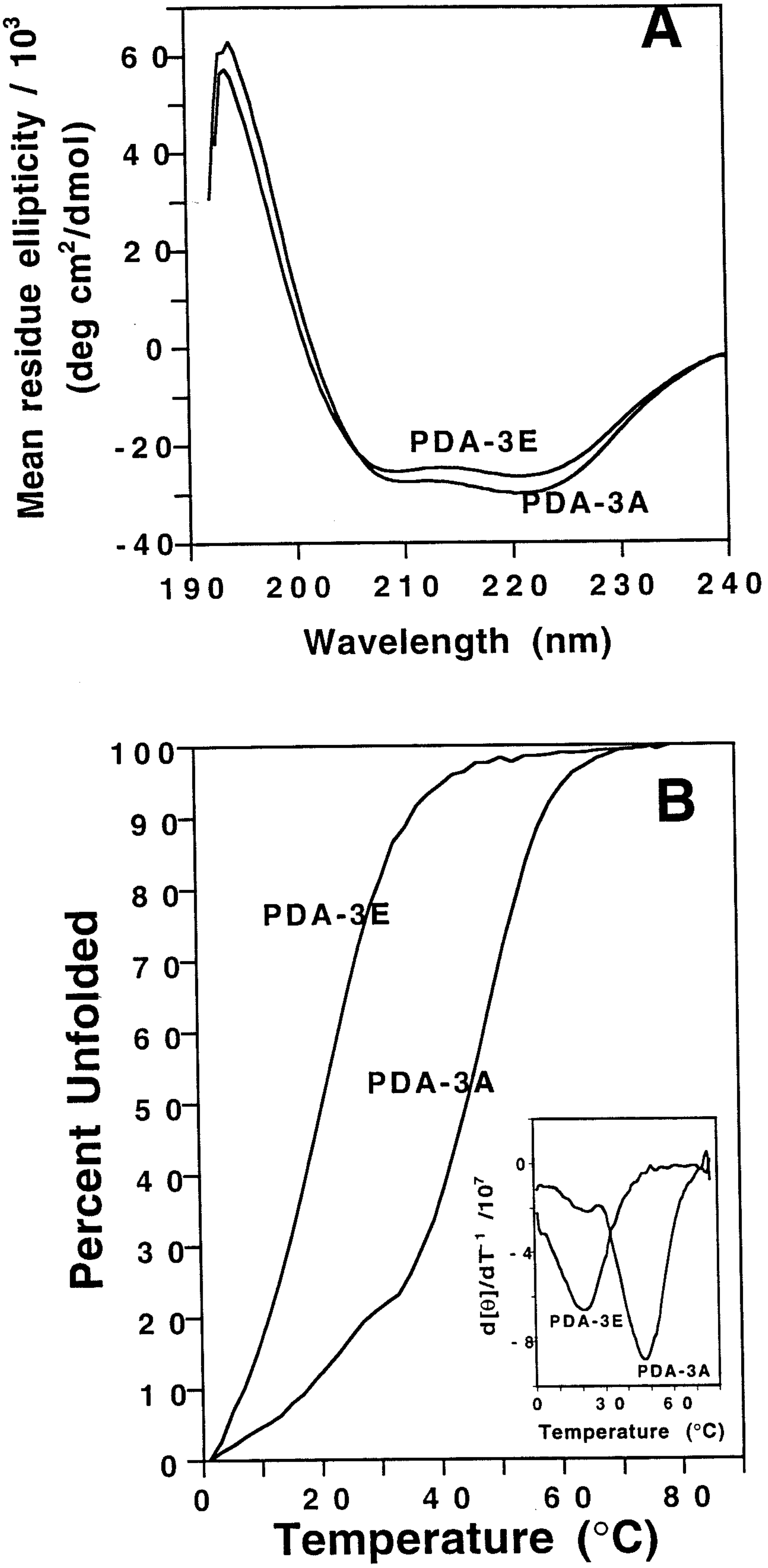

A
Mean residue ellipticity / 10^3 (deg cm^2/dmol)
60
40
20
0
-20
-40
190
200
210
220
230
240
Wavelength (nm)
PDA-3E
PDA-3A
B
Percent Unfolded
100
90
80
70
60
50
40
30
20
10
0
0
20
40
60
80
Temperature (°C)
PDA-3E
PDA-3A
d[θ]/dT^{-1} /10^7
0
-4
-8
0
30
60
Temperature (°C)
PDA-3E
PDA-3A

Figure 2-5. Comparison of simulation cost functions to experimental T_m's. (A) shows the initial cost function which contains only a van der Waals term for the 8 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.

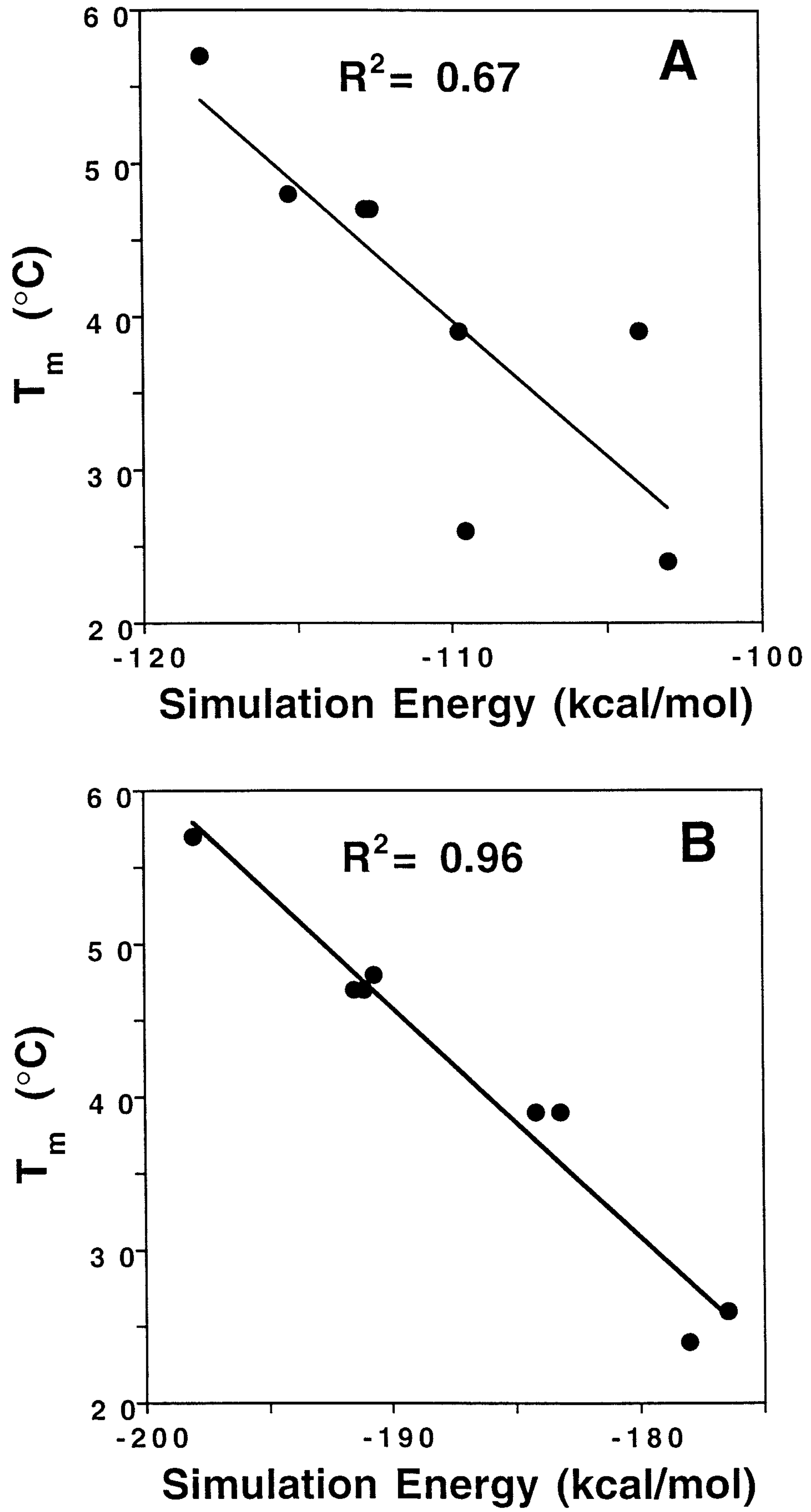
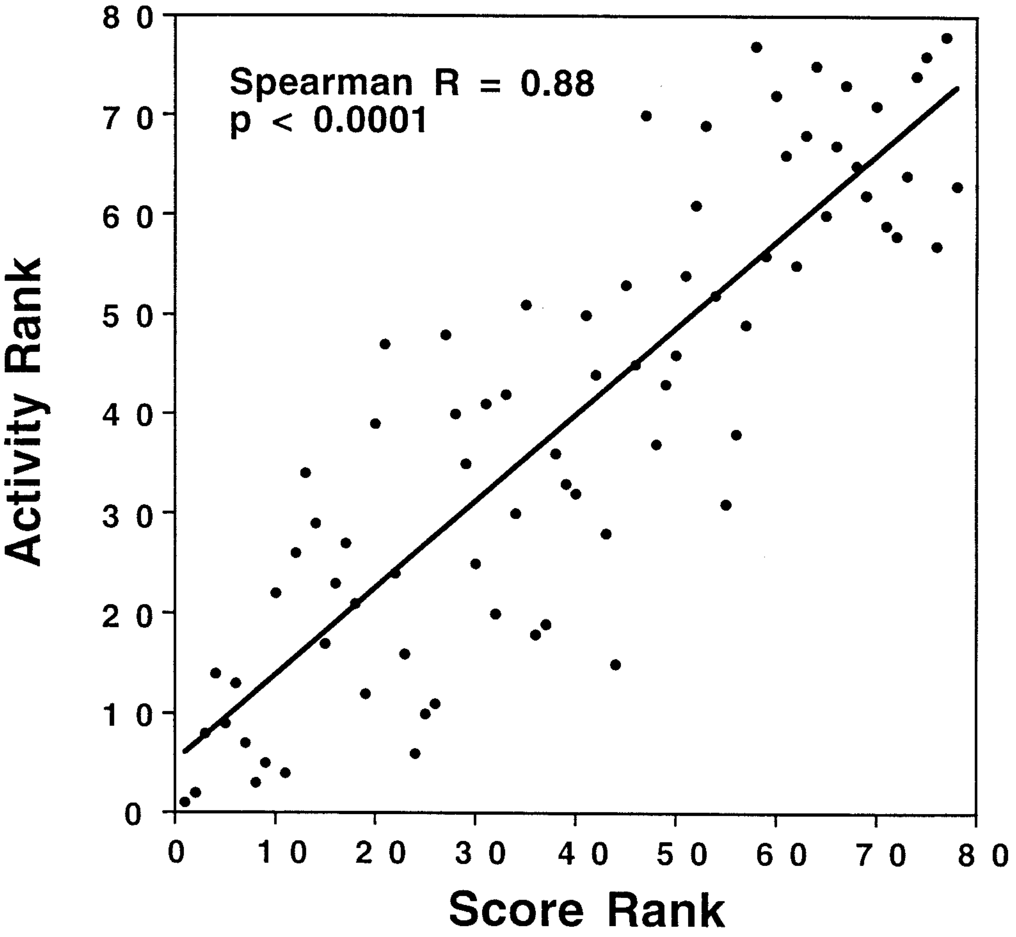

A
R^2= 0.67
T_m (°C)
60
50
40
30
20
-120
-110
-100
Simulation Energy (kcal/mol)
B
R^2= 0.96
T_m (°C)
60
50
40
30
20
-200
-190
-180
Simulation Energy (kcal/mol)

Figure 2-6. Rank correlation of energy predicted by the simulation module versus the combined activity score of λ repressor mutants (20, 22).


Spearman R = 0.88
p < 0.0001
Activity Rank
Score Rank
0
10
20
30
40
50
60
70
80

Chapter 3

Probing The Role Of Packing Specificity In Protein Design

The text of this chapter is adapted from a published manuscript that was coauthored with Professor Stephen L. Mayo.

B. I. Dahiyat and S. L. Mayo, *Proc. Natl. Acad. Sci., USA,* **94**, 10172, (1997).

Summary

Using a protein design algorithm that quantitatively considers side-chain packing, the effect of specific steric constraints on protein design was assessed in the core of the streptococcal protein G β1 domain. The strength of packing constraints used in the design was varied, resulting in core sequences that reflected differing amounts of packing specificity. The structural flexibility and stability of several of the designed proteins were experimentally determined and showed a trend from well-ordered to highly mobile structures as the degree of packing specificity in the design decreased. This trend both demonstrates that the inclusion of specific packing interactions is necessary for the design of native-like proteins and defines a useful range of packing specificity for the design algorithm. In addition, an analysis of the modeled protein structures suggested that penalizing for exposed hydrophobic surface area can improve design performance.

The placement of hydrophobic amino acids into protein cores is critical for maintaining the stable, highly ordered native structures of naturally occurring proteins(1, 2, 3, 4). Many designed proteins have been constructed to form a nonpolar core by simply selecting a suitable pattern of hydrophobic and polar residues (HP pattern). While the correct secondary structure and overall tertiary organization seem to have been attained in several cases, most designs appear to lack the structural ordering of native proteins(5, 6, 7). Several lines of evidence suggest that the omission of specific packing interactions as a design criterion is a cause of disorder in designed proteins. Computational design methods that systematically incorporate side-chain packing have successfully redesigned the hydrophobic cores of proteins and resulted in well-ordered structures(8, 9). Further, a coiled-coil design that selected core residue packing arrangements to disfavor competing structures resulted in a protein with greater native-like character than was achieved with previous efforts(10). Conversely, lattice simulations suggest that selection of an optimal HP pattern can result in uniquely folded sequences(11). In this study, we seek to quantitatively assess both the degree to which specific packing interactions are necessary for the design of well-ordered proteins and the tolerance of native-like structure to variations in core packing patterns.

Previous studies that have examined the role of core packing on protein structure demonstrate that while some variation in the buried positions of a protein is allowed, there are limits on the sequences that result in stable, native-like folds(2, 12, 13, 14, 15). Though providing great insight on the importance of core interactions, the conclusions from these studies are difficult to generalize to other proteins and do not provide a framework to assess designed proteins. We propose the use of an automated side-chain selection algorithm, which explicitly and quantitatively considers specific side-chain packing interactions(9), as the basis of a method to define the need for packing constraints in protein design. Our side-chain selection algorithm screens all possible sequences and finds the optimal amino acid type and side-chain orientation for a given backbone. In order to correctly account for the torsional flexibility of side chains and the geometric specificity of side-chain placement, we consider a discrete set of all allowed conformers of each side chain, called rotamers(16, 17). The immense search problem presented by rotamer sequence optimization is overcome by application of the Dead-End Elimination (DEE) theorem(18, 19, 20). Our implementation of the DEE theorem extends its utility to sequence design and rapidly finds the globally optimal sequence in its optimal conformation. Scoring of sequence arrangements includes an atomic van der Waals potential, which captures the two main features of steric packing interactions: excluded volume and the weakly attractive dispersive force. Protein cores designed with this and with similar(8) algorithms result in stable, well-ordered proteins.

The referenced sequence prediction algorithms do not predict a wide variety of packing arrangements for a given backbone, but rather select a single family of closely related core sequences, indicating that designs produced by these algorithms are highly determined by packing specificity. Two factors are likely to be responsible for this stringency: the use of a fixed backbone and the highly restrictive repulsive (excluded volume) component of the van der Waals potential. The repulsive component, and the likelihood that different side-chain arrangements cause packing clashes, can be modulated, however, by scaling the van der Waals radii of the atoms in the simulation. We implement this modulation in the packing constraints by varying a radius scale factor, α (Equation 1). R_0 and D_0 are the van der Waals radius and well depth, respectively, and E_{vdw} and R are the energy and interatomic distance.

Equation 1

$$E_{vdw} = D_0\left\{\left(\frac{\alpha R_0}{R}\right)^{12} - 2\left(\frac{\alpha R_0}{R}\right)^{6}\right\}$$

By predicting core sequences with various radii scalings and then experimentally characterizing the resulting proteins, a rigorous study of the importance of packing effects on protein design is possible.

By using a protein design algorithm to assess the bounds of effective steric constraints on core packing, these bounds can be readily incorporated back into the algorithm to improve design performance, a critical advantage of the approach, since an underlying goal of understanding packing constraints in protein design is to use this information to develop better designs. Specifically, a reduced van der Waals steric constraint can compensate for the restrictive effect of a fixed backbone and discrete side-chain rotamers in the simulation and could allow a broader sampling of sequences compatible with the desired fold. The use of experimental data to test our designs and subsequently to improve our design algorithm is the central feature of our overall protein design strategy. In previous work, experimental feedback was able to improve core design performance by suggesting the use of a hydrophobic solvation potential to refine the selection of nonpolar residues and by providing effective solvation parameters for use in the simulation(9). Similarly, this study should provide practical improvements to our sequence scoring potential in addition to generally assaying the role of packing specificity in protein structure.

Model system core sequence predictions

An ideal model system to study core packing is the β1 immunoglobulin-binding domain of streptococcal protein G (Gβ1)(21) (Figure 1). Its small size, 56 residues, renders computations more tractable and simplifies production of the protein by either synthetic or recombinant methods. Several crystal structures(22) and a solution structure(21) are available to provide backbone templates for the side-chain selection algorithm. In addition, the energetics and structural dynamics of Gβ1 have been extensively characterized and can serve as the reference point for a well-ordered native protein(23, 24, 25, 26). Gβ1 is highly thermostable with a melting temperature of 87 °C. Perhaps most critical for a core packing study, Gβ1 contains no disulfide bonds and does not require a cofactor or metal ion to fold, but rather relies upon the burial of its hydrophobic core for stability. Further, Gβ1 contains sheet, helix and turn structures and is without the repetitive side-chain packing patterns found in coiled coils or some helical bundles. This lack of periodicity reduces the bias from a particular secondary or tertiary structure and necessitates the use of an objective side-chain selection algorithm to examine packing effects.

Sequence positions that constitute the core were chosen by examining the side-chain solvent accessible surface area of Gβ1. Any side chain exposing less than 10% of its surface was considered buried (Figure 1). Eleven residues meet this criteria, with seven from the β sheet (positions 3, 5, 7, 20, 43, 52 and 54), three from the helix (positions 26, 30, and 34) and one in an irregular secondary structure (position 39). These positions form a contiguous core.

The remainder of the protein structure, including all other side chains and the backbone, was used as the template for sequence selection calculations at the eleven core positions.

All possible core sequences consisting of alanine, valine, leucine, isoleucine, phenylalanine, tyrosine or tryptophan (A, V, L, I, F, Y or W) were considered. Our rotamer library was similar to that used by Desmet and coworkers(18). In order to minimize possible errors that might arise from the discreteness of the library, additional rotamers were created with χ_1 and χ_2 angle values increased or decreased by one standard deviation about the mean value from the Ponder and Richards library. Optimizing the sequence of the core of Gβ1 with 217 possible hydrophobic rotamers considered at all 11 positions results in 217^{11}, or $5x10^{25}$, rotamer sequences. Our scoring function consisted of two components: a van der Waals energy term and an atomic solvation term favoring burial of hydrophobic surface area. The van der Waals radii of all atoms in the simulation were scaled by a factor α (Eqn. 1) to change the importance of packing effects. Radii were not scaled for the buried surface area calculations. Global optimum sequences for various values of the radius scaling factor α were found using the Dead-End Elimination theorem (Table 1). Optimal sequences, and their corresponding proteins, are named by the radius scale factor used in their design. For example, the sequence designed with a radius scale factor of $\alpha = 0.90$ is called α90.

α100 was designed with $\alpha = 1.0$ and hence serves as a baseline for full incorporation of steric effects. The α100 sequence is very similar to the core sequence of Gβ1 (Table 1) even though no information about the naturally occurring sequence was used in the side-chain selection algorithm. Variation of α from 0.90 to 1.05 caused little change in the optimal sequence, demonstrating the algorithm's robustness to minor parameter perturbations. Further, the packing arrangements predicted with $\alpha = 0.90 - 1.05$ closely match Gβ1 with average χ angle differences of only 4° from the crystal structure. The homology and conformational similarity to Gβ1 imply that, when packing constraints are used, backbone conformation strongly determines a single family of well packed core designs. Nevertheless, the constraints on core packing were being modulated by α as demonstrated by Monte Carlo searches for other low energy sequences. Several alternate sequences and packing arrangements are in the 20 best sequences found by the Monte Carlo procedure when $\alpha = 0.90$. These alternate sequences score much worse when $\alpha = 0.95$, and when $\alpha = 1.0$ or 1.05 only strictly conservative packing geometries have low energies, with low energy sequences consisting entirely of point mutations to smaller residues. Therefore, $\alpha = 1.05$ and $\alpha = 0.90$ define the high and low ends, respectively, of a range where packing specificity dominates sequence design.

Position 7 is exceptional because the crystal structure has a leucine at this position with a nearly eclipsed χ_2 of 111°. This strained χ_2 is unlikely to be an artifact of the structure determination since it is present in two crystal forms and a solution structure(21, 22). Our rotamer library does not contain any eclipsed rotamers and no staggered leucine rotamers pack well at this position. Instead, the side-chain selection algorithm chose valine or isoleucine rotamers that conserved the χ_1 dihedral and are still able to pack well. We expect the removal of the strained leucine rotamer to stabilize the protein, a prediction that is tested in the experimental section of this work.

For $\alpha < 0.90$, extensive changes in the optimal sequence occur (Table 1). The role of packing is reduced enough to let the hydrophobic surface potential begin to dominate, thereby increasing the size of the residues selected for the core. A significant change in the optimal sequence appears between $\alpha = 0.90$ and 0.85 with both $\alpha 85$ and $\alpha 80$ containing three additional mutations relative to $\alpha 90$. Also, $\alpha 85$ and $\alpha 80$ have a 15% increase in total side-chain volume relative to Gβ1. As α drops below 0.80 an additional 10% increase in side-chain volume and numerous mutations occur, showing that packing constraints have been overwhelmed by the drive to bury nonpolar surface. Though the jumps in volume and shifts in packing arrangement appear to occur suddenly for the optimal sequences, examination of the suboptimal low energy sequences by Monte Carlo sampling demonstrates that the changes are not abrupt. For example, the $\alpha 85$ optimal sequence is the 11th best sequence when $\alpha = 0.90$, and similarly, the $\alpha 90$ optimal sequence is the 9th best sequence when $\alpha = 0.85$.

For $\alpha > 1.05$ atomic van der Waals repulsions are so severe that most amino acids cannot find any allowed packing arrangements, which results in the selection of alanine for many positions. This stringency is likely an artifact of the large atomic radii and does not reflect increased packing specificity accurately. Rather, $\alpha = 1.05$ is the upper limit for the usable range of van der Waals scales within our modeling framework.

Experimental characterization of core designs

Variation of the van der Waals scale factor α results in four regimes of packing specificity: regime 1 where $0.9 \leq \alpha \leq 1.05$ and packing constraints dominate the sequence selection; regime 2 where $0.8 \leq \alpha < 0.9$ and the hydrophobic solvation potential begins to compete with packing forces; regime 3 where $\alpha < 0.8$ and hydrophobic solvation dominates the design; and, regime 4 where $\alpha > 1.05$ and van der Waals repulsions appear to be too severe to allow meaningful sequence selection. Sequences that are optimal designs were selected from each of the regimes for synthesis and characterization. They are α90 from regime 1, α85 from regime 2, α70 from regime 3 and α107 from regime 4. For each of these sequences, the calculated amino acid identities of the eleven core positions are shown in Table 1; the remainder of the protein sequence matches Gβ1. The stability and structural order of each of the four sequences were assessed. The goal was to study the relation between the degree of packing specificity used in the core design and the extent of native-like character in the resulting proteins.

Far UV circular dichroism (CD) spectra of the selected proteins are shown in Figure 2A. α90 and α85 have ellipticities and spectra very similar to Gβ1 (not shown), suggesting that their secondary structure content is comparable to that of Gβ1. Conversely, α70 has much weaker ellipticity and a perturbed spectrum, implying a loss of secondary structure relative to Gβ1. α107 has a spectrum characteristic of a random coil. Thermal melts monitored by CD are shown in Figure 2B. α85 and α90 both have cooperative transitions with melting temperatures (T_m's) of 83 °C and at least 92 °C, respectively (Table 2). α107 shows no thermal transition, behavior expected from a fully unfolded polypeptide, and α70 has a broad, shallow transition, centered at ~40 °C, characteristic of partially folded structures. An accurate T_m could not be measured for α90 because the transition was still incomplete at 99 °C. Relative to Gβ1, which has a T_m of 87 °C (23), α85 is slightly less thermostable and α90 is more stable. Chemical denaturation measurements of the free energy of unfolding (ΔG_u) at 25 °C match the trend in T_m's (Table 2). α90 has a larger ΔG_u than that reported for Gβ1(23) while α85 is slightly less stable. It was not possible to measure ΔG_u for α70 or α107 because they lack discernible transitions.

The extent of chemical shift dispersion in the proton NMR spectrum of each protein was assessed to gauge each protein's degree of native-like character (Figure 3). α90 possesses a highly dispersed spectrum, the hallmark of a well-ordered native protein. α85 has diminished chemical shift dispersion and peaks that are somewhat broadened relative to α90, suggesting a moderately mobile structure that nevertheless maintains a distinct fold. α70's low solubility limited NMR sample concentrations to ~100 μM and its NMR spectrum has almost no dispersion. The broad peaks are indicative of a collapsed but disordered and fluctuating structure. α107 has a spectrum with sharp lines and no dispersion, which is indicative of an unfolded protein.

Amide hydrogen exchange kinetics are consistent with the conclusions reached from examination of the proton NMR spectra. Figure 4 shows the average number of unexchanged amide protons as a function of time for each of the designed proteins (see Methods). α90 protects ~13 protons for over 20 hours of exchange at pH 5.5 and 25 °C. The α90 exchange curve is indistinguishable from Gβ1's (not shown). α85 also maintains a well-protected set of amide protons, a distinctive feature of ordered native-like proteins. The number of protected protons, however, is only about half that of α90. The difference is likely due to higher flexibility in some parts of the α85 structure. In contrast, α70 and α107 were fully exchanged within the three minute dead time of the experiment, indicating highly dynamic structures.

In addition to NMR experiments, near UV CD spectra and the extent of 8-anilino-1-naphthalene sulfonic acid (ANS) binding were used to assess the structural ordering of the proteins. Figure 5 shows that the near UV CD spectra of α85 and α90 have the strong peaks expected for proteins with aromatic residues fixed in a unique tertiary structure. α70 and α107 have the featureless spectra indicative of proteins with highly mobile aromatic residues, such as non-native collapsed states or unfolded proteins. α70 also binds ANS well, as indicated by a three-fold intensity increase and blue shift of the ANS emission spectrum (Figure 6). This strong binding suggests that α70 possesses a cluster of hydrophobic residues accessible to ANS, possibly due to loose packing or partial exposure of the core to solvent. ANS binds α85 weakly, with only a 25% increase in emission intensity, similar to the association seen for some native proteins(27). α90 and α107 cause no change in ANS fluorescence. Finally, size exclusion chromatography was used to assure that the proteins were not forming aggregates. All of the designed proteins elute at the expected time for a 6 kDa protein confirming their monomeric states.

In summary, by integrating the data from a variety of different techniques, we assessed the stability and structural ordering of the optimal sequence designs from the four packing specificity regimes. α90 is a well-packed native-like protein by all criteria, and it is more stable than the naturally occurring Gβ1 sequence, possibly because of reduced torsional strain at position 7 and increased hydrophobic surface burial. α85 is also a stable, ordered protein, albeit with greater motional flexibility than α90, as evidenced by its NMR spectrum and hydrogen exchange behavior. α70 has all the features of a disordered collapsed globule: a non-cooperative thermal transition, no NMR spectral dispersion or amide proton protection, reduced secondary structure content and strong ANS binding. α107 is a completely unfolded chain, likely due to its lack of large hydrophobic residues to hold the core together. The clear trend is a loss of protein ordering as α decreases below 0.90.

Design feedback and scoring function improvement

The effectiveness of the different packing regimes for protein design can be evaluated in light of the experimental data. In regime 1, with $0.9 \leq \alpha \leq 1.05$, the design is dominated by packing specificity resulting in well-ordered proteins as evidenced by α90. In regime 2, with $0.8 \leq \alpha < 0.9$, packing forces are weakened enough to let the hydrophobic force drive larger residues into the core which produces a stable well-packed protein with somewhat increased structural motion. In regime 3, $\alpha < 0.8$, packing forces are reduced to such an extent that the hydrophobic force dominates, resulting in a fluctuating, partially folded structure with no stable core packing. In regime 4, $\alpha > 1.05$, the steric forces used to implement packing specificity are scaled too high to allow reasonable sequence selection and hence produce an unfolded protein. These results indicate that effective protein design requires a consideration of packing effects. Within the context of a protein design algorithm, we have quantitatively defined the range of packing forces necessary for successful designs. Also, we have demonstrated that reduced specificity can be used to design protein cores with alternative packings, presumably by relaxing the dependence of sequence selection on the fixed backbone and discrete rotamers used in the simulation.

To take advantage of the benefits of reduced packing constraints, protein cores should be designed with the smallest α that still results in structurally ordered proteins. The optimal protein sequence from regime 2, α85, is stable and well packed, suggesting $0.8 \leq \alpha < 0.9$ as a good range. NMR spectra and hydrogen exchange kinetics, however, clearly show that α85 is not as structurally ordered as α90. The packing arrangements predicted by our algorithm for W43 in α85 and α90 present a possible explanation (Figure 7). For α90, W43 is predicted to pack in the core with the same conformation as in the crystal structure of Gβ1. In α85, the larger side chains at positions 34 and 54, leucine and phenylalanine respectively, compared to alanine and valine in α90, force W43 to expose 91 Å^2 of nonpolar surface compared to 19 Å^2 in α90. The hydrophobic driving force this exposure represents seems likely to stabilize alternate conformations that bury W43 and thereby could contribute to α85's conformational flexibility. In contrast to the other core positions, a residue at position 43 can be mostly exposed or mostly buried depending on its side-chain conformation. We designate positions with this characteristic as boundary positions, which pose a difficult problem for protein design because of their potential to either strongly interact with the protein's core or with solvent.

A scoring function that penalizes the exposure of hydrophobic surface area might assist in the design of boundary residues. Dill and coworkers used an exposure penalty to improve protein designs in a theoretical study(11). A nonpolar exposure penalty would favor packing arrangements that either bury large side chains in the core or replace the exposed amino acid with a smaller or more polar one. We implemented a side-chain nonpolar exposure penalty in our optimization framework and used a penalizing solvation parameter with the same magnitude as the hydrophobic burial parameter.

The results of adding a hydrophobic surface exposure penalty to our scoring function are shown in Tables 3 A-D. When $\alpha = 0.90$, the optimal sequence does not change and the next 14 best sequences, found by Monte Carlo sampling, change very little, except to penalize those that expose W43, namely the 11th, 12th, 14th and 15th ranked sequences (Table 3 A and B). This minor effect is not surprising, since steric forces still dominate for $\alpha = 0.90$ and most of these sequences expose very little surface area. In contrast, when $\alpha = 0.85$ the nonpolar exposure penalty dramatically alters the ordering of low energy sequences (Table 3 C and D). The α85 sequence, the former ground state, drops to 7th and the rest of the 15 best sequences expose far less hydrophobic area because they bury W43 in a conformation similar to α90 (Figure 7). The exceptions are the 8th and 14th best sequences, which reduce the size of the exposed boundary residue by replacing W43 with an isoleucine, and the 13th best sequence which replaces W43 with a valine. The new ground state sequence is very similar to α90, with a single valine to isoleucine mutation, and should share α90's stability and structural order. Burying W43 restricts sequence selection in the core somewhat, but the reduced packing forces for $\alpha = 0.85$ still produce more sequence variety than $\alpha = 0.90$ (Table 3 B and D). The exposure penalty complements the use of reduced packing specificity by limiting the gross overpacking and solvent exposure that occurs when the core's boundary is disrupted. Adding this constraint should allow lower packing forces to be used in protein design, resulting in a broader range of high-scoring sequences and reduced bias from fixed backbone and discrete rotamers. Both of these benefits are critical for the success of *de novo* design algorithms because they ease the requirement for accurate backbone specification and provide a greater variety of candidate sequences.

To examine the effect of substituting a smaller residue at a boundary position, we selected the 13th best sequence of the $\alpha = 0.85$ optimization with exposure penalty (Table 3 D) for synthesis and characterization. This sequence, α85W43V, replaces W43 with a valine but is otherwise identical to α85. Though the 8th and 14th sequences also have a smaller side chain at position 43, additional changes in their sequences relative to α85 would complicate interpretation of the effect of the boundary position change. Also, α85W43V has a significantly different packing arrangement compared to Gβ1, with 7 out of 11 positions altered, but only an 8% increase in side-chain volume. Hence, α85W43V is a test of the tolerance of this fold to a different, but nearly volume conserving, core. The far UV CD spectrum of α85W43V is very similar to that of Gβ1 with an ellipticity at 218 nm of -14000 deg cm^2/dmol. While the secondary structure content of α85W43V is native-like, its T_m is 65 °C, nearly 20 °C lower than α85. In contrast to α85W43V's decreased stability, its NMR spectrum has greater chemical shift dispersion than α85 (Figure 3). The amide hydrogen exchange kinetics show a well protected set of about four protons after 20 hours (Figure 4). This faster exchange relative to α85 is explained by α85W43V's significantly lower stability(28). α85W43V appears to have improved structural specificity at the expense of stability, a phenomenon observed previously in coiled coils(29). By using an exposure penalty, the design algorithm produced a protein with greater native-like character.

Conclusion

We have quantitatively defined the role of packing specificity in protein design and have provided practical bounds for the role of steric forces in our protein design algorithm. This study differs from previous work because of the use of an objective, quantitative algorithm to vary packing forces during design. Further, by using the minimum effective level of steric forces, we were able to design a wider variety of packing arrangements that were compatible with the Gβ1 fold. This broader sampling of sequences reflects a reduction in the bias arising from the fixed backbone and discrete rotamers used in our side-chain selection algorithm. Finally, we have identified a difficulty in the design of side chains that lie at the boundary between the core and the surface of Gβ1, and we have implemented a nonpolar surface exposure penalty in our sequence design scoring function that addresses this problem.

Methods

Sequence optimization: DEE and Monte Carlo. The protein structure was modeled on the backbone coordinates of Gβ1, PDB record 1pga(22, 30). Atoms of all side chains not optimized were left in their crystallographically determined positions. The program BIOGRAF (Molecular Simulations Incorporated, 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(31). The rotamer library, DEE optimization and Monte Carlo search followed the methods of our previous work(9). A Lennard-Jones 12-6 potential was used for van der Waals interactions, with atomic radii scaled for the various cases as discussed in the text. The Richards definition of solvent-accessible surface area(32) was used and areas were calculated with the Connolly algorithm(33). An atomic solvation parameter, derived from our previous work, of 23 cal/mol/Å^2 was used to favor hydrophobic burial and to penalize solvent exposure. To calculate side-chain nonpolar exposure in our optimization framework, we first consider the total hydrophobic area exposed by a rotamer in isolation. This exposure is decreased by the area buried in rotamer/template contacts, and the sum of the areas buried in rotamer/rotamer contacts, quantities that are calculated as pairwise interactions between rotamers as is required for DEE. The remaining exposed area is then converted to a penalty energy using a solvation parameter with the same magnitude as for hydrophobic burial but with opposite sign.

Peptide synthesis and purification. With the exception of the eleven core positions designed by the sequence selection algorithm, the sequences synthesized match Protein Data Bank entry 1pga. Peptides were synthesized on an Applied Biosystems 433A automated peptide synthesizer using Fmoc chemistry, HBTU activation and pre-derivatized HMP resin. Modified 0.25 mmol coupling cycles were used with extended reaction and deprotection times. Peptides were cleaved by treating approximately 200 mg of resin with 2 mL of trifluoroacetic acid (TFA) and 100 µl water, 100 µl thioanisole, 50 µl ethanedithiol and 150 mg phenol as scavengers. The peptides were isolated by precipitation into cold methyl t-butyl ether and purified by reverse-phase HPLC on a Vydac C8 column (25 cm x 22 mm) with a linear acetonitrile-water gradient containing 0.1% TFA. Peptides were then lyophilized and stored at -20 °C. Matrix assisted laser desorption mass spectrometry found all molecular weights to be within one unit of the expected masses.

CD and fluorescence spectroscopy and size exclusion chromatography. The solution conditions for all experiments were 50 mM sodium phosphate buffer at pH 5.5 and 25 °C unless otherwise noted. Circular dichroism spectra were acquired on an Aviv 62DS spectrometer. A 1 mm pathlength cell was used and the temperature was controlled by a thermoelectric unit. Peptide concentration was approximately 20 µM and spectra were baseline corrected with a buffer blank. Thermal melts were monitored at 218 nm using 2° increments with an averaging time of 10 s and an equilibration time of 120 s. T_m's were defined as the maxima of the derivative of the melting curve. Reversibility for each of the proteins was confirmed by comparing room temperature CD spectra from before and after heating. Guanidinium chloride denaturation measurements followed published methods(34). Samples were equilibrated for at least four hours and the signal was monitored at 218 nm with an averaging time of 60 s. Protein concentrations were determined by UV spectrophotometry.

Fluorescence experiments were performed on a Hitachi F-4500 in a 1 cm pathlength cell. Both peptide and ANS concentrations were 50 μM. The excitation wavelength was 370 nm and emission was monitored from 400 to 600 nm at a resolution of 0.2 nm. Spectra were baseline corrected with blank buffer samples.

Size exclusion chromatography was performed with a PolyLC hydroxyethyl A column (20 cm x 9 mm) at pH 5.5 in 50 mM sodium phosphate at 0 °C. Ribonuclease A, carbonic anhydrase and Gβ1 were used as molecular weight standards. 10 μl injections of 1 mM peptide were made, except α70 for which 100 μl injections of 100 μM solution were made. The flowrate was 0.2 ml/min. Peptide concentrations during the separation were approximately 15 μM as estimated from peak heights monitored at 275 nm.

Nuclear magnetic resonance spectroscopy. NMR samples were prepared in 90/10 H_2O/D_2O and 50 mM sodium phosphate buffer at pH 5.5. Spectra were acquired on a Varian Unityplus 600 MHz spectrometer at 25 °C. 32 transients were acquired with 1.5 seconds of solvent presaturation used for water suppression. Samples were approximately 1 mM, except for α70 which had limited solubility of about 100 μM. For hydrogen exchange studies, an NMR sample was prepared, the pH was adjusted to 5.5 and a spectrum was acquired to serve as an unexchanged reference. This sample was lyophilized, reconstituted in D_2O and repetitive acquisition of spectra was begun immediately at a rate of 75 s per spectrum. Data acquisition continued for approximately 20 hours, then the sample was heated to 99 °C for three minutes to fully exchange all protons. After cooling to 25 °C, a final spectrum was acquired to serve as the fully exchanged reference. All spectra were processed with identical phase and baseline slope corrections and integrated. The areas of all exchangeable amide peaks were normalized by a set of non-exchanging aliphatic peaks, and the fraction of total amide exchange was calculated using the unexchanged and fully exchanged reference spectra. pH values, uncorrected for isotope effects, were measured for all the samples after data acquisition and the time axis was normalized to correct for minor differences in pH(35).

Acknowledgements

We wish to thank D. B. Gordon for helpful discussions, S. Ross for assistance with the NMR spectroscopy, and G. Hathaway of the Caltech Protein and Peptide Microanalytical Facility for mass spectra. This work was supported by the Rita Allen 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

1. D. Shortle, W. Stites, A. Meeker, *Biochemistry* **29**, 8033-8041 (1990).

2. W. A. Lim, R. T. Sauer, *J. Mol. Biol.* **219**, 359-376 (1991).

3. F. M. Richards, W. A. Lim, *Quarterly Reviews of Biophysics* **26**, 423-498 (1993).

4. K. A. Dill, et al., *Protein Sci.* **4**, 561-602 (1995).

5. L. Regan, W. F. DeGrado, *Science* **241**, 976-978 (1988).

6. M. H. Hecht, J. S. Richardson, D. C. Richardson, R. C. Ogden, *Science* **249**, 884-891 (1990).

7. S. Kamtekar, J. M. Schiffer, H. Xiong, J. M. Babik, M. H. Hecht, *Science* **262**, 1680-1685 (1993).

8. J. R. Desjarlais, T. M. Handel, *Protein Sci.* **4**, 2006-2018 (1995).

9. B. I. Dahiyat, S. L. Mayo, *Protein Sci.* **5**, 895-903 (1996).

10. S. F. Betz, W. F. Degrado, *Biochemistry* **35**, 6955-6962 (1996).

11. S. Sun, R. Brem, H. S. Chan, K. A. Dill, *Prot. Eng.* **8**, 1205-1213 (1995).

12. W. A. Lim, R. T. Sauer, *Nature* **339**, 31-36 (1989).

13. W. A. Lim, D. C. Farruggio, R. T. Sauer, *Biochemistry* **31**, 4324-4333 (1992).

14. M. Munson, R. O'Brien, J. M. Sturtevant, L. Regan, *Protein Sci.* **3**, 2015-2022 (1994).

15. M. Munson, et al., *Protein Sci.* **5**, 1584-1593 (1996).

16. J. W. Ponder, F. M. Richards, *J. Mol. Biol.* **193**, 775-791 (1987).

17. R. L. Dunbrack, M. Karplus, *J. Mol. Biol.* **230**, 543-574 (1993).

18. J. Desmet, M. De Maeyer, B. Hazes, I. Lasters, *Nature* **356**, 539-542 (1992).

19. J. Desmet, M. De Maeyer, I. Lasters, in *The protein folding problem and tertiary structure prediction* K. Merz Jr, S. Le Grand, Eds. (Birkhauser, Boston, 1994) pp. 307-337.

20. R. F. Goldstein, *Biophys. J.* **66**, 1335-1340 (1994).

21. A. M. Gronenborn, et al., *Science* **253**, 657-661 (1991).

22. T. Gallagher, P. Alexander, P. Bryan, G. L. Gilliland, *Biochemistry* **33**, 4721-4729 (1994).

23. P. Alexander, S. Fahnestock, T. Lee, J. Orban, P. Bryan, *Biochemistry* **31**, 3597-3603 (1992).

24. J. J. Barchi, B. Grasberger, A. M. Gronenborn, G. M. Clore, *Protein Sci.* **3**, 15-21 (1994).

25. J. Kuszewski, G. M. Clore, A. M. Gronenborn, *Protein Sci.* **3**, 1945-1952 (1994).

26. J. Orban, P. Alexander, P. Bryan, D. Khare, *Biochemistry* **34**, 15291-15300 (1995).

27. G. V. Semisotnov, et al., *Biopolymers* **31**, 119-128 (1991).

28. S. L. Mayo, R. L. Baldwin, *Science* **262**, 873-876 (1993).

29. P. B. Harbury, T. Zhang, P. S. Kim, T. Alber, *Science* **262**, 1401-1407 (1993).

30. F. C. Bernstein, et al., *J. Mol. Biol.* **112**, 535-542 (1977).

31. S. L. Mayo, B. D. Olafson, W. A. Goddard III, *J. Phys. Chem.* **94**, 8897-8909 (1990).

32. B. Lee, F. M. Richards, *J. Mol. Biol.* **55**, 379-400 (1971).

33. M. L. Connolly, *Science* **221**, 709-713 (1983).

34. C. N. Pace, *Methods Enzymol.* **131**, 266-280 (1986).

35. C. A. Rohl, J. M. Scholtz, E. J. York, J. M. Stewart, R. L. Baldwin, *Biochemistry* **31**, 1263-1269 (1992).

36. R. Koradi, M. Billeter, K. Wuthrich, *J. Mol. Graph.* **14**, 51-55 (1996).

Table 1

		Gβ1 sequence										
α	vol	TYR	LEU	LEU	ALA	ALA	PHE	ALA	VAL	TRP	PHE	VAL
		3	5	7	20	26	30	34	39	43	52	54
0.60	1.40	TRP	TRP	PHE	ILE	PHE	TRP	ILE	LEU	PHE	TRP	TYR
0.70	1.28	TRP	TYR	ILE	ILE	PHE	TRP	LEU	ILE	PHE	LEU	ILE
0.75	1.23	PHE	ILE	PHE	ILE	VAL	TRP	VAL	LEU	|	|	ILE
0.80	1.13	PHE	|	ILE	|	|	|	ILE	ILE	|	TRP	ILE
0.85	1.15	PHE	|	ILE	|	|	|	LEU	ILE	|	TRP	PHE
0.90	1.01	PHE	|	ILE	|	|	|	|	ILE	|	|	|
0.95	1.01	PHE	|	ILE	|	|	|	|	ILE	|	|	|
1.0	0.99	PHE	|	VAL	|	|	|	|	ILE	|	|	|
1.05	0.93	PHE	|	ALA	|	|	|	|	|	|	|	|
1.075	0.83	ALA	ALA	ILE	|	|	ILE	|	|	|	ILE	ILE
1.10	0.77	ALA	|	ALA	|	|	ALA	|	|	|	ILE	ILE
1.15	0.68	ALA	ALA	ALA	|	|	ALA	|	|	|	LEU	|
1.25	0.64	ALA	ALA	ALA	|	|	ALA	|	|	|	LEU	ALA

DEE was used to determine optimal sequences for the core positions of Gβ1 as a function of van der Waals radius scale factor α. The Gβ1 sequence and position numbers are shown at the top. Vol is the fraction of core side-chain volume relative to the Gβ1 sequence. A vertical bar indicates identity with the Gβ1 sequence. Both a van der Waals potential and a hydrophobic solvation potential (23 cal/mol/Å^2 favoring burial) were used to score sequences(9).

Table 2

	T_m (°C)	ΔG_u (kcal/mol)	m (kcal/mol/M)	C_m (M)	$[\theta]_{218}$ (deg cm2 /dmol)	Protected protons
α90	>92	-7.2	1.7	4.2	-15500	13
α85	83	-6.1	1.6	3.8	-13700	6
α70	40	-	-	-	-6800	0
α107	-	-	-	-	-3000	0

Stability of designed proteins. T_m is the melting temperature measured by circular dichroism. ΔG_u is the unfolding free energy determined using guanidinium chloride as the denaturant, m is the free energy dependence on denaturant concentration and C_m is the denaturation midpoint. $[\theta]_{218}$ is the ellipticity at 218 nm. Protected protons are the number of unexchanged protons after 20 hours of hydrogen deuterium exchange at pH 5.5 and 25 °C.

Table 3 A

$\alpha = 0.90$

Rank	A_{np}	TYR	LEU	LEU	ALA	ALA	PHE	ALA	VAL	TRP	PHE	VAL
		3	5	7	20	26	30	34	39	43	52	54
1	30	PHE	|	ILE	|	|	|	|	ILE	|	|	|
2	30	|	|	ILE	|	|	|	|	ILE	|	|	|
3	30	PHE	|	ILE	|	|	|	|	ILE	|	TYR	|
4	38	PHE	|	ILE	|	|	|	|	ILE	|	TRP	|
5	30	|	|	ILE	|	|	|	|	ILE	|	TYR	|
6	38	|	|	ILE	|	|	|	|	ILE	|	TRP	|
7	30	PHE	|	VAL	|	|	|	|	ILE	|	|	|
8	30	|	|	VAL	|	|	|	|	ILE	|	|	|
9	29	PHE	|	ILE	|	|	|	|	|	|	|	|
10	29	|	|	ILE	|	|	|	|	|	|	|	|
11	109	PHE	|	ILE	|	|	|	LEU	ILE	|	TRP	PHE
12	104	PHE	|	ILE	|	|	|	LEU	ILE	|	|	PHE
13	30	PHE	|	VAL	|	|	|	|	ILE	|	TYR	|
14	109	|	|	ILE	|	|	|	LEU	ILE	|	TRP	PHE
15	104	|	|	ILE	|	|	|	LEU	ILE	|	|	PHE

The 15 best sequences for the core positions of Gβ1 using $\alpha = 0.90$ without an exposure penalty. A_{np} is the amount of exposed hydrophobic surface area of the core residues. The naturally occurring Gβ1 sequence is shown at the top of each table with residue numbers. DEE was used to find the optimum and a Monte Carlo search to find the other sequences.

Table 3 B

α = 0.90 exposure penalty												
Rank	A_{np}	TYR	LEU	LEU	ALA	ALA	PHE	ALA	VAL	TRP	PHE	VAL
		3	5	7	20	26	30	34	39	43	52	54
1	30	PHE	|	ILE	|	|	|	|	ILE	|	|	|
2	30	|	|	ILE	|	|	|	|	ILE	|	|	|
3	30	PHE	|	ILE	|	|	|	|	ILE	|	TYR	|
4	30	|	|	ILE	|	|	|	|	ILE	|	TYR	|
5	30	PHE	|	VAL	|	|	|	|	ILE	|	|	|
6	38	PHE	|	ILE	|	|	|	|	ILE	|	TRP	|
7	30	|	|	VAL	|	|	|	|	ILE	|	|	|
8	29	PHE	|	ILE	|	|	|	|	|	|	|	|
9	38	|	|	ILE	|	|	|	|	ILE	|	TRP	|
10	29	|	|	ILE	|	|	|	|	|	|	|	|
11	29	|	|	ILE	|	|	|	|	|	|	|	|
12	30	PHE	|	VAL	|	|	|	|	ILE	|	TYR	|
13	30	|	|	VAL	|	|	|	|	ILE	|	TYR	|
14	29	PHE	|	ILE	|	|	|	|	|	|	TYR	|
15	104	PHE	|	ILE	|	|	|	LEU	ILE			PHE

The 15 best sequences for the core positions of Gβ1 using α = 0.90 with an exposure penalty.

Table 3 C

α = 0.85												
Rank	A_{np}	TYR	LEU	LEU	ALA	ALA	PHE	ALA	VAL	TRP	PHE	VAL
		3	5	7	20	26	30	34	39	43	52	54
1	109	PHE	|	ILE	|	|	|	LEU	ILE	|	TRP	PHE
2	109	|	|	ILE	|	|	|	LEU	ILE	|	TRP	PHE
3	104	PHE	|	ILE	|	|	|	LEU	ILE	|	|	PHE
4	104	|	|	ILE	|	|	|	LEU	ILE	|	|	PHE
5	108	PHE	|	ILE	|	|	|	LEU	|	|	TRP	PHE
6	62	PHE	|	ILE	|	|	|	LEU	ILE	VAL	TRP	PHE
7	103	PHE	|	ILE	|	|	|	LEU	ILE	|	TYR	PHE
8	109	PHE	|	VAL	|	|	|	LEU	ILE	|	TRP	PHE
9	30	PHE	|	ILE	|	|	|	|	ILE	|	|	|
10	38	PHE	|	ILE	|	|	|	|	ILE	|	TRP	|
11	108	|	|	ILE	|	|	|	LEU	|	|	TRP	PHE
12	62	|	|	ILE	|	|	|	LEU	ILE	VAL	TRP	PHE
13	109	PHE	|	ILE	|	|	TYR	LEU	ILE	|	TRP	PHE
14	103	|	|	ILE	|	|	|	LEU	ILE	|	TYR	PHE
15	109	|	|	VAL	|	|	|	LEU	ILE	|	TRP	PHE

The 15 best sequences for the core positions of Gβ1 using α = 0.85 without an exposure penalty.

Table 3 D

α = 0.85 exposure penalty

Rank	A_{np}	TYR	LEU	LEU	ALA	ALA	PHE	ALA	VAL	TRP	PHE	VAL
		3	5	7	20	26	30	34	39	43	52	54
1	30	PHE	|	ILE	|	|	|	|	ILE	|	|	ILE
2	29	PHE	|	ILE	|	|	|	ILE	ILE	|	|	|
3	29	PHE	ILE	PHE	|	|	|	|	ILE	|	|	|
4	30	|	|	ILE	|	|	|	|	ILE	|	|	ILE
5	29	|	|	ILE	|	|	|	ILE	ILE	|	|	|
6	29	|	ILE	PHE	|	|	|	|	ILE	|	|	|
7	109	PHE	|	ILE	|	|	|	LEU	ILE	|	TRP	PHE
8	52	PHE	|	ILE	|	|	|	LEU	ILE	ILE	|	PHE
9	29	|	|	ILE	|	|	|	|	ILE	|	|	|
10	29	PHE	|	ILE	|	|	|	|	ILE	|	|	|
11	109	|	|	ILE	|	|	|	LEU	ILE	|	TRP	PHE
12	38	PHE	|	ILE	|	|	|	|	ILE	|	TRP	ILE
13	62	PHE	|	ILE	|	|	|	LEU	ILE	VAL	TRP	PHE
14	52	|	|	ILE	|	|	|	LEU	ILE	ILE	|	PHE
15	30	PHE	|	ILE	|	|	|	|	ILE	|	TYR	ILE

The 15 best sequences for the core positions of Gβ1 using α = 0.85 with an exposure penalty.

Figure 3-1. Ribbon diagram of Gβ1 (PDB code 1pga) showing the side chains of the 11 core positions used in this study. Carbon atoms are gray, oxygen atoms are red and nitrogen atoms are blue. The core positions (residues 3, 5, 7, 20, 26, 30, 34, 39, 43, 52 and 54) bury greater than 90% of their side-chain surface area in the Gβ1 structure. Y3, A26 and F30 are obscured by the helix. Figures were produced with MOLMOL(36).

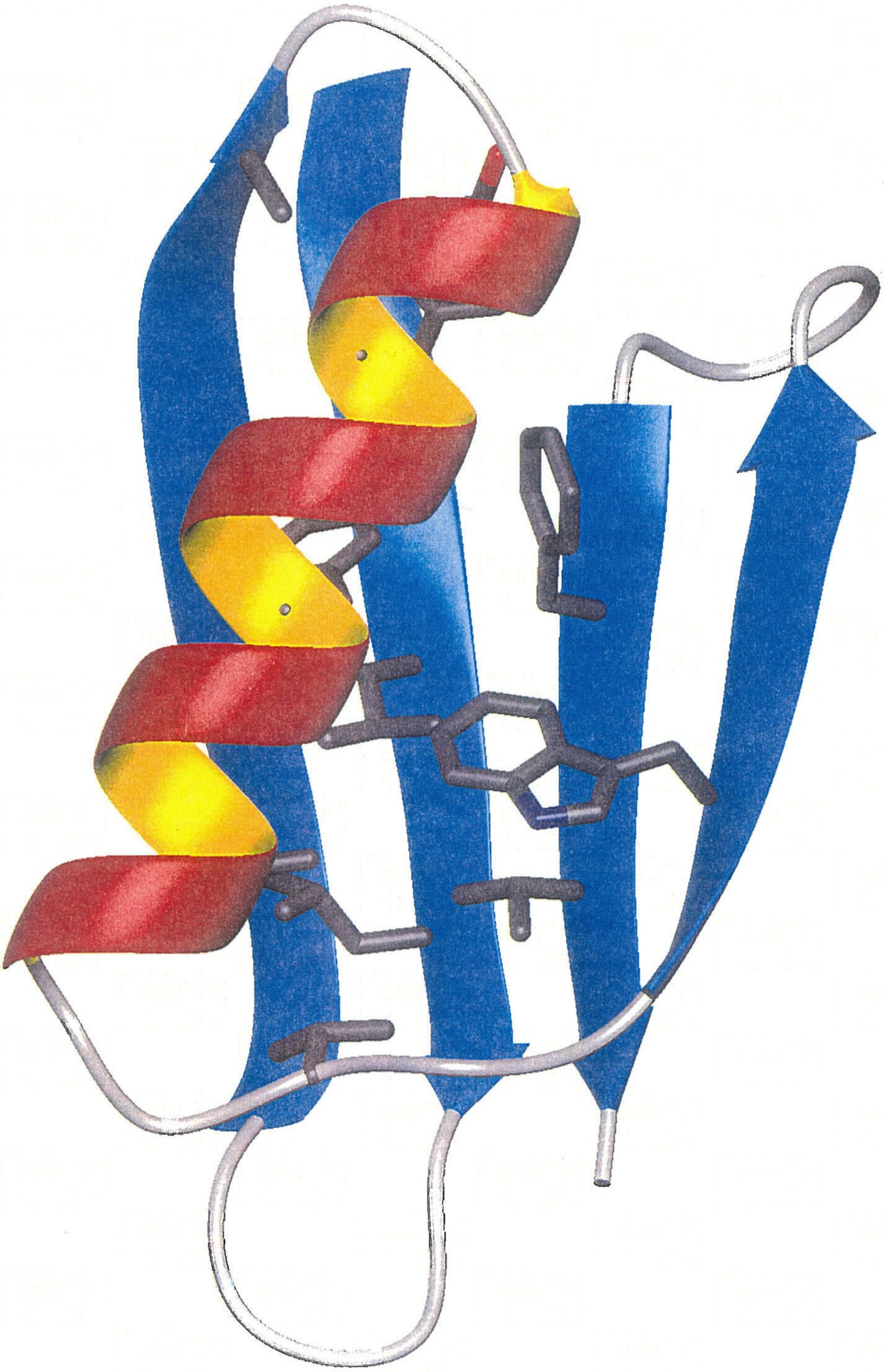


Figure 3-2. Secondary structure and thermal stability of α90, α85, α70 and α107. A) Far UV CD spectra. α90 and α85 have large ellipticities indicative of high secondary structure content. α70 displays much less secondary structure and α107 has the spectrum of a random coil. B) Thermal denaturation monitored by CD. α90 and α85 have sharp transitions at > 92° and 83° C, respectively, indicative of cooperative unfolding, while α70 had a broad transition centered ~40° C and α107 had no detectable unfolding transition.

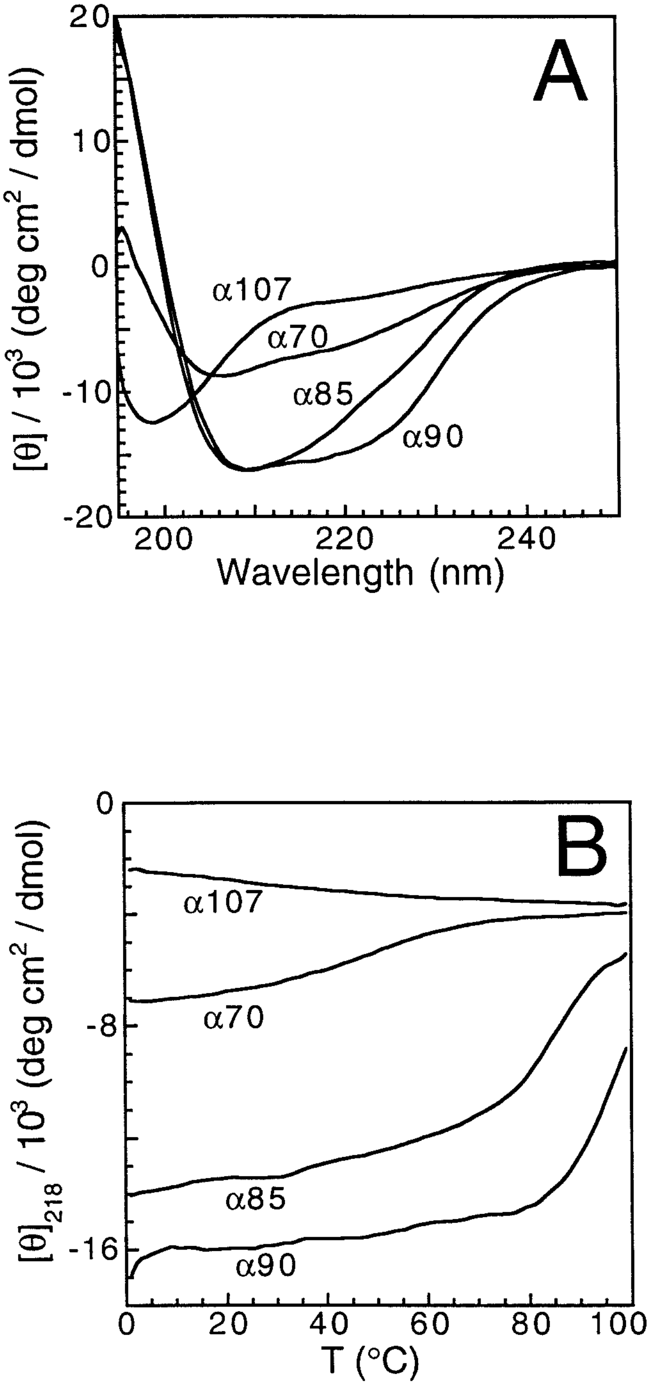

A
20
10
0
-10
-20
[θ] / 10^3 (deg cm^2 / dmol)
200
220
240
Wavelength (nm)
α107
α70
α85
α90
B
0
-8
-16
$[\theta]_{218}$ / 10^3 (deg cm^2 / dmol)
0
20
40
60
80
100
T (°C)
α107
α70
α85
α90

Figure 3-3. Proton NMR spectra of α90, α85, α70, α107 and α85W43V. The decrease in dispersion from α90 to α85 to α70 reflects a graded decrease in protein structural order, from the highly ordered α90 structure to the fluctuating, mobile α70 structure. α107 appears unfolded. α85W43V has narrower lines and greater dispersion than α85, indicating that the single W to V mutation reduced conformational flexibility relative to α85. The sharp peaks at 8.45 and 0.15 ppm in the α70 spectrum are impurities.

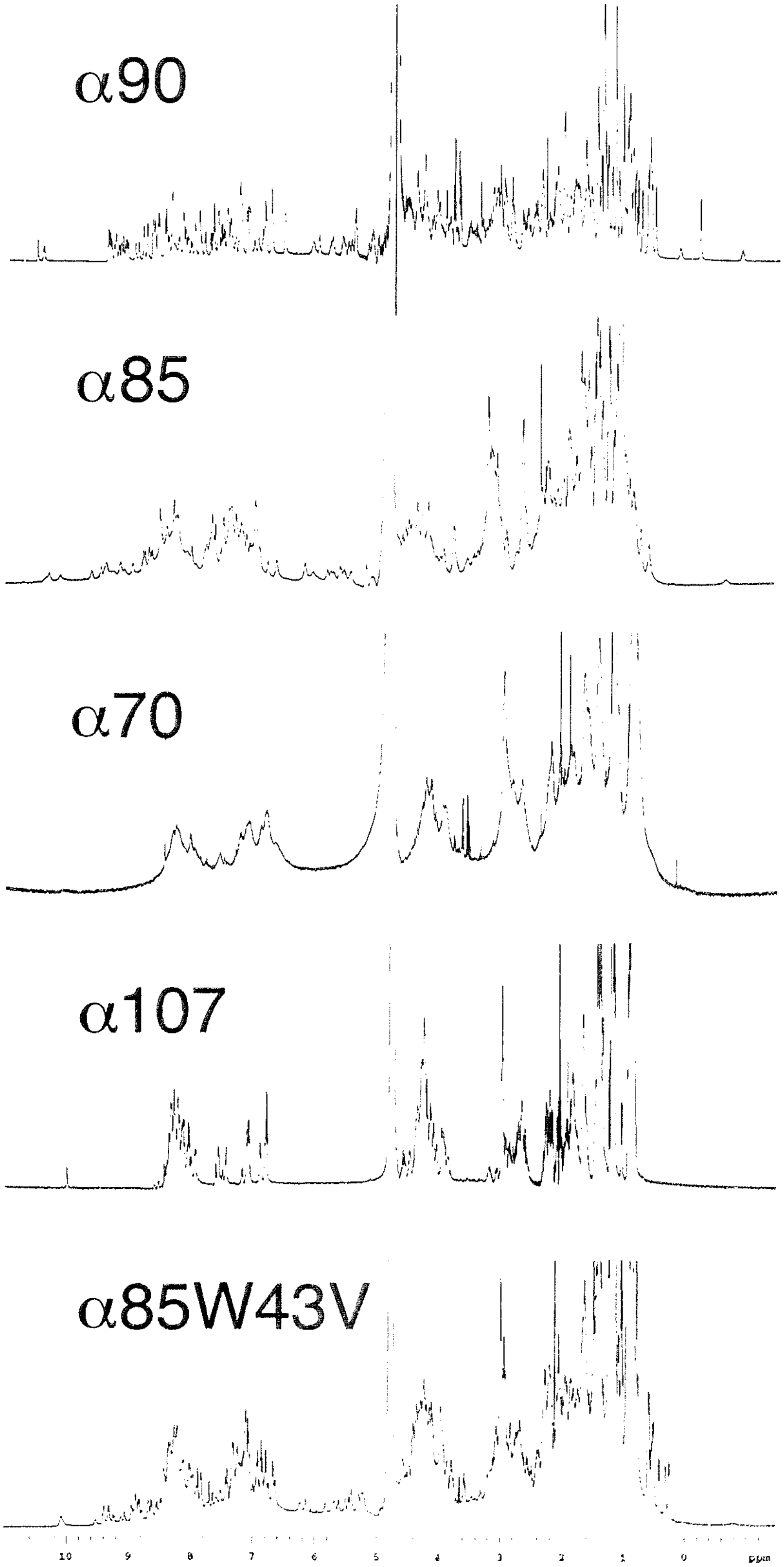

α90
α85
α70
α107
α85W43V
10
9
8
7
6
5
4
3
2
1
0
ppm

Figure 3-4. Amide hydrogen-deuterium exchange kinetics of α90, α85, α70, α107 and α85W43V. Total area of exchangeable peaks, expressed as number of protons, as a function of exchange time at 25 °C and pH 5.5. Exchange times were corrected for slight pH differences between the samples. Areas were normalized to non-exchanging peak areas and fully exchanged and non-exchanged spectra were used as references for 0 and 55 protons, respectively. α90, α85 and α85W43V protect ~13, ~6 and ~5 protons, respectively, for at least 20 hours. α70 and α107 are fully exchanged in the 3 minute deadtime of the experiment.

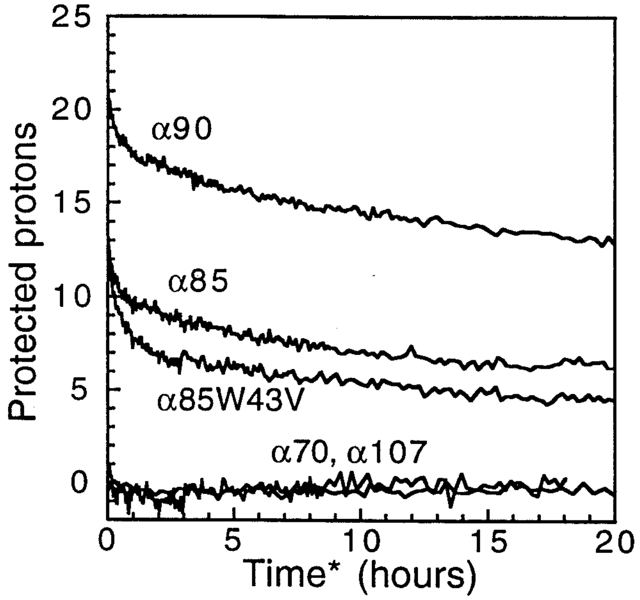

25
20
15
10
5
0
Protected protons
α90
α85
α85W43V
α70, α107
0
5
10
15
20
Time* (hours)

Figure 3-5. Near UV CD spectra of α90, α85, α70 and α107. Strong peaks in the spectra of α90 and α85 indicate distinct tertiary environments for their aromatic residues, while the featureless spectra of α70 and α107 are consistent with highly mobile aromatic residues in fluctuating structures.

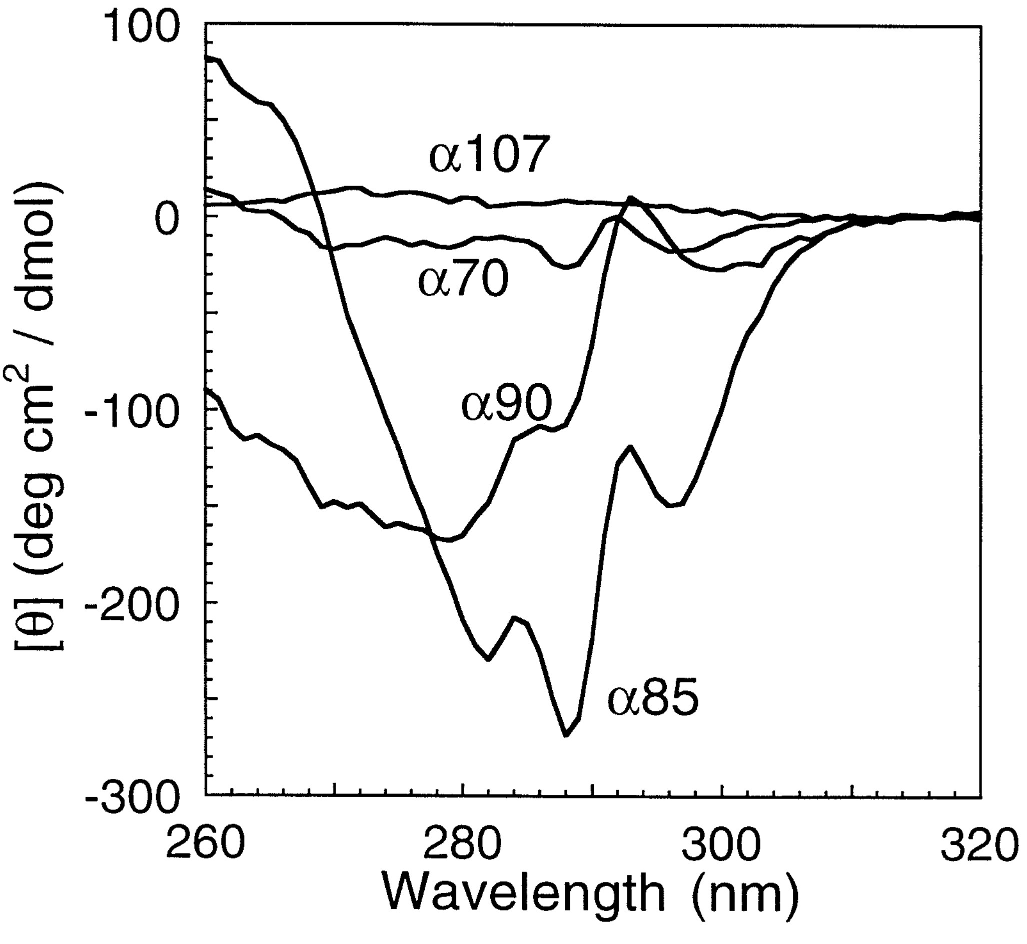

100
0
-100
-200
-300
[θ] (deg cm^2 / dmol)
260
280
300
320
Wavelength (nm)
α107
α70
α90
α85

Figure 3-6. ANS flourescence alone and in the presence of α90, α85, α70 and α107. The lack of fluorescence enhancement for α90 and α107 indicate no ANS binding, and the moderate increase in fluorescence for α85 is consistent with weak interaction. The three-fold enhancement for α70, however, indicates strong binding and is consistent with a hydrophobic cluster accessible to ANS, such as in a collapsed globule.

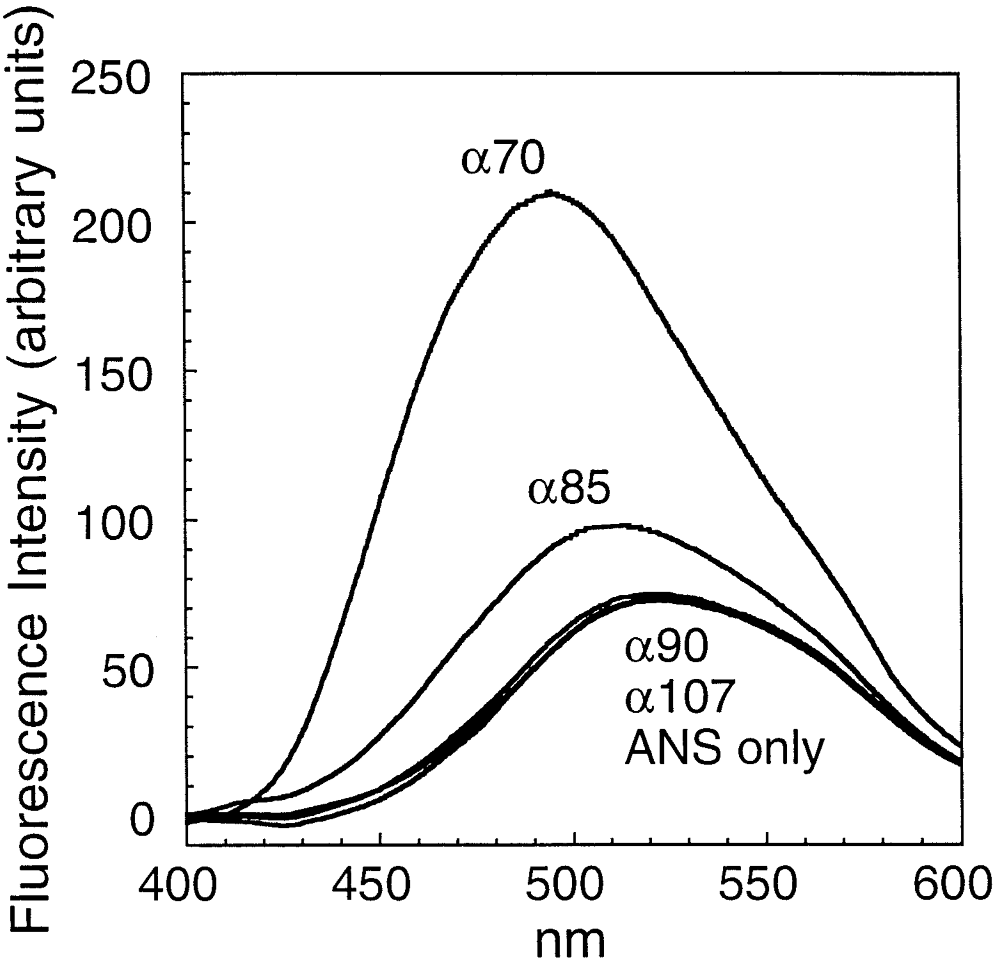
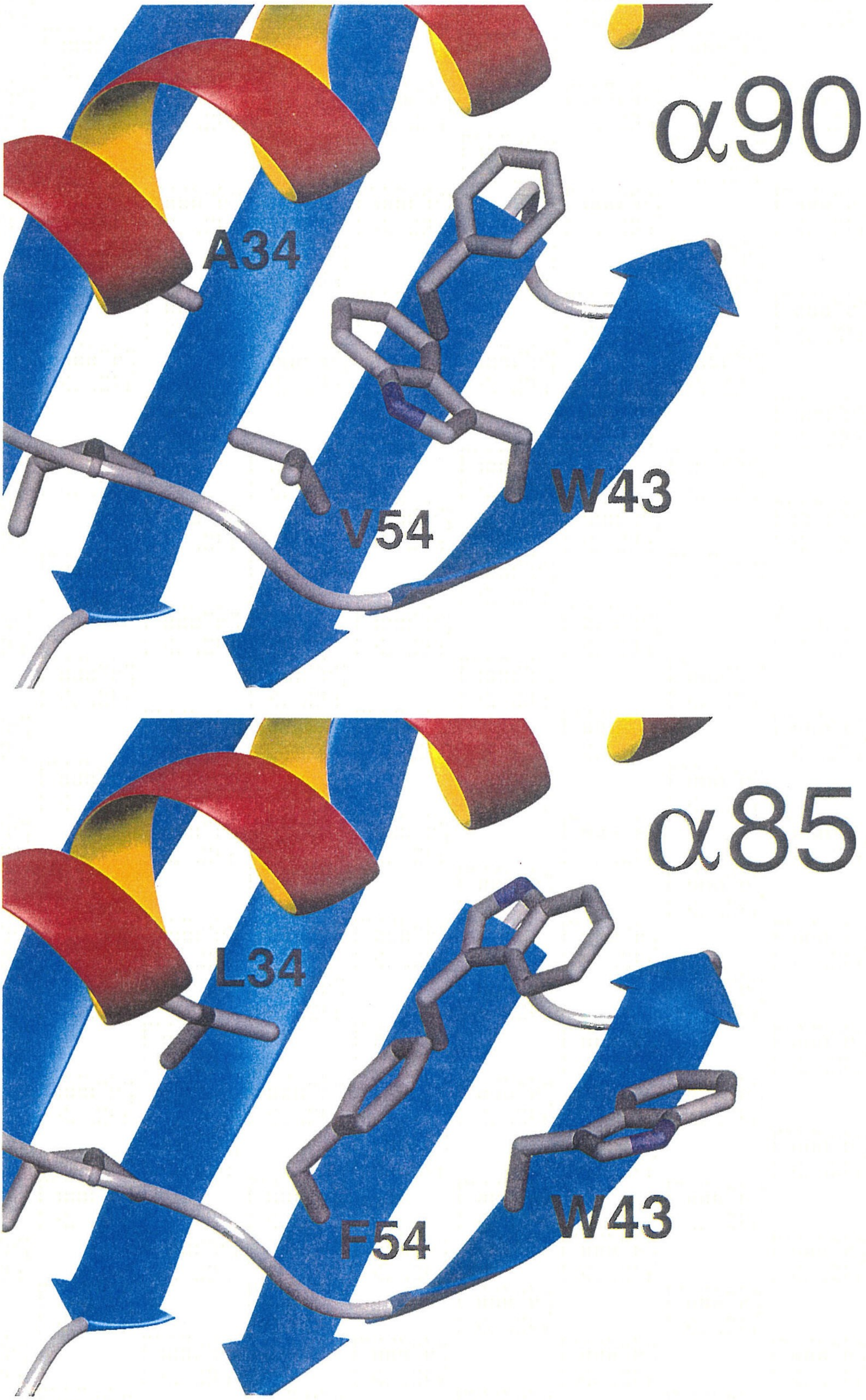

Fluorescence Intensity (arbitrary units)
250
200
150
100
50
0
400
450
500
550
600
nm
α70
α85
α90
α107
ANS only

Figure 3-7. Core packing arrangements predicted by DEE for α90 (top) and α85 (bottom). For clarity only side chains for residues 34, 39, 43, 52 and 54 are shown. In α90, W43 has a conformation that results in over 90% burial of its surface area. In α85, W43 is only 46% buried and is rotated into solvent to avoid steric clashes with L34 and F52, which occupy a much larger volume than A34 and V52 in α90. The reduced packing forces used to design α85 allow larger residues into the core which disrupt the packing at the core's boundary.


α90
A34
V54
W43
α85
L34
F54
W43

Chapter 4

Automated Design of the Surface Positions of Protein Helices

The text of this chapter is adapted from a published manuscript that was coauthored with Professor Stephen L. Mayo and D. Benjamin Gordon.

B. I. Dahiyat, D. B. Gordon, and S. L. Mayo, *Protein Sci.*, **6**, 1333 (1997).

Abstract

Using a protein design algorithm that quantitatively considers side-chain interactions, the design of surface residues of α helices was examined. Three scoring functions were tested: a hydrogen-bond potential, a hydrogen-bond potential in conjunction with a penalty for uncompensated burial of polar hydrogens, and a hydrogen-bond potential in combination with helix propensity. The solvent exposed residues of a homodimeric coiled coil based on GCN4-p1 were designed by using the Dead-End Elimination Theorem to find the optimal amino acid sequence for each scoring function. The corresponding peptides were synthesized and characterized by circular dichroism spectroscopy and size exclusion chromatography. The designed peptides were dimeric and nearly 100% helical at 1 °C, with melting temperatures from 69-72 °C, over 12 °C higher than GCN4-p1, while a random hydrophilic sequence at the surface positions produced a peptide that melted at 15 °C. Analysis of the designed sequences suggests that helix propensity is the key factor in sequence design for surface helical positions.

Several groups have proposed and tested systematic, quantitative methods for protein design that screen possible sequences for compatibility with the desired protein fold (1, 2, 3, 4, 5, 6, 7, 8). These algorithms consider the spatial positioning and steric complementarity of side chains by explicitly modeling the atoms of sequences under consideration. To date, such techniques have typically focused on designing the cores of proteins and have scored sequences with van der Waals and sometimes hydrophobic solvation potentials. We seek to extend this sequence selection approach to the design of the solvent exposed residues of proteins as part of an effort to develop a complete de novo design algorithm. In this study, we consider the design of the surface positions of α helices.

Although mutagenesis studies suggest that surface positions are more tolerant of substitutions than core positions (9, 10), surface residues can still have a significant effect on protein structure and stability. To assess the importance of surface residue selection for protein design, we used several scoring functions to compute sequences for the surface positions of our model helical protein, the coiled coil GCN4-p1 (11). By experimentally characterizing the resulting proteins, the performance of each scoring function was assessed and the effect of surface sequence changes on the protein's stability was determined.

GCN4-pl, a homodimeric coiled coil, was selected as the model system because it can be readily synthesized by solid phase techniques and its helical secondary structure and dimeric tertiary organization ease characterization. The sequences of homodimeric coiled coils display a seven residue periodic hydrophobic and polar pattern called a heptad repeat (**a**·**b**·**c**·**d**·**e**·**f**·**g**) (12). The **a** and **d** positions are buried at the dimer interface and are usually hydrophobic, whereas the **b**, **c**, **e**, **f**, and **g** positions are solvent exposed and usually polar

(Figure 1). Examination of the crystal structure of GCN4-p1 (11) shows that the **b**, **c**, and **f** side chains extend into solvent and expose at least 55% of their surface area. In contrast, the **e** and **g** residues bury from 50 to 90% of their surface area by packing against the **a** and **d** residues of the opposing helix. We selected the 12 **b**, **c**, and **f** residue positions for surface sequence design: positions 3, 4, 7, 10, 11, 14, 17, 18, 21, 24, 25, and 28 using the numbering from PDB entry 2zta (13). The remainder of the protein structure, including all other side chains and the backbone, was used as the template for sequence selection calculations. The symmetry of the dimer and lack of interactions of surface residues between the subunits allowed independent design of each subunit, thereby significantly reducing the size of the sequence optimization problem.

All possible sequences of hydrophilic amino acids (D, E, N, Q, K, R, S, T, A, and H) for the 12 surface positions were screened by our design algorithm. The torsional flexibility of the amino acid side chains was accounted for by considering a discrete set of all allowed conformers of each side chain, called rotamers (14, 15). Optimizing the 12 **b**, **c**, and **f** positions each with 10 possible amino acids results in 10^{12} possible sequences which corresponds to ~10^{28} rotamer sequences when using the Dunbrack and Karplus backbone-dependent rotamer library. The immense search problem presented by rotamer sequence optimization is overcome by application of the Dead-End Elimination (DEE) theorem (16, 17, 18). Our implementation of the DEE theorem extends its utility to sequence design and rapidly finds the globally optimal sequence in its optimal conformation.

We examined three potential-energy functions for their effectiveness in scoring surface sequences. Each candidate scoring function was used to design the **b**, **c**, and **f** positions of the model coiled coil and the resulting peptide was synthesized and characterized to assess design performance. A hydrogen-bond potential was used to check if predicted hydrogen bonds can contribute to designed protein stability, as expected from studies of hydrogen bonding in proteins and peptides (19, 20). Optimizing sequences for hydrogen bonding, however, often buries polar protons that are not involved in hydrogen bonds. This uncompensated loss of potential hydrogen-bond donors to water prompted examination of a second scoring scheme consisting of a hydrogen-bond potential in conjunction with a penalty for burial of polar protons (21). We tested a third scoring scheme which augments the hydrogen bond potential with the empirically derived helix propensities of Baldwin and coworkers (22). Although the physical basis of helix propensities is unclear, they can have a significant effect on protein stability and can potentially be used to improve protein designs (23, 24, 25, 26, 27). A van der Waals potential was used in all cases to account for packing interactions and excluded volume.

Several other sequences for the **b**, **c** and **f** positions were also synthesized and characterized to help discern the relative importance of the hydrogen-bonding and helix-propensity potentials. The sequence designed with the hydrogen-bond potential was randomly scrambled, thereby disrupting the designed interactions but not changing the helix propensity of the sequence. Also, the sequence with the maximum possible helix propensity, all positions set to alanine, was made. Finally, to serve as undesigned controls, the naturally occurring GCN4-p1 sequence and a sequence randomly selected from the hydrophilic amino acid set were synthesized and studied.

Results

The surface sequences of all of the peptides examined in this study are shown in Table 1. Sequence 6A, designed with a hydrogen-bond potential, has a preponderance of Arg and Glu residues that are predicted to form numerous hydrogen bonds to each other. These long chain amino acids are favored because they can extend across turns of the helix to interact with each other and with the backbone. When the optimal geometry of the scrambled 6A sequence, 6D, was found with DEE, far fewer hydrogen bonding interactions were present and its score was much worse than 6A's. 6B, designed with a polar hydrogen burial penalty in addition to a hydrogen-bond potential, is still dominated by long residues such as Lys, Glu and Gln but has fewer Arg. Because Arg has more polar hydrogens than the other amino acids, it more often buries nonhydrogen-bonded protons and therefore is disfavored when using this potential function. 6C was designed with a hydrogen-bond potential and helix propensity in the scoring function and consists entirely of Ala and Arg residues, the amino acids with the highest helix propensities (22). The Arg residues form hydrogen bonds with Glu residues at nearby **e** and g positions. The random hydrophilic sequence, 6E, possesses no hydrogen bonds and scores very poorly with all of the potential functions used. The scrambled GCN4-p1 sequence, 6G, exposes a tyrosine residue.

The secondary structures and thermal stabilities of the peptides were assessed by circular dichroism (CD) spectroscopy. The CD spectra of the peptides at 1 °C and 40 μM are characteristic of α helices, with minima at 208 and 222 nm, except for the random surface sequence peptides 6E and 6G (Fig. 2A). 6E and 6G have spectra suggestive of a mixture of α helix and random coil with $[\theta]_{222}$ of -12000 deg cm^2/dmol and -14000 deg cm^2/dmol respectively, while all the other peptides are greater than 90% helical with $[\theta]_{222}$ of less than -30000 deg cm^2/dmol. The melting temperatures (T_m's) of the designed peptides are 12-16 °C higher than the T_m of GCN4-p1, except for 6E which has a T_m of 15 °C and 6G which has a T_m of 22 °C (Fig. 2B). CD spectra taken before and after melts were identical indicating reversible thermal denaturation. The redesign of surface positions of this coiled coil produces structures that are much more stable than wildtype GCN4-p1, while a random hydrophilic sequence largely disrupts the peptide's stability. A random sequence with wildtype composition is also greatly destabilized, suggesting specific interactions are present in the naturally occurring sequence.

Size exclusion chromatography (SEC) showed that all the peptides were dimers except for 6F, the all Ala surface sequence, which migrated as a tetramer. These data show that surface redesign did not change the tertiary structure of these peptides, in contrast to some core redesigns (28). In addition, nuclear magnetic resonance (NMR) spectra of the peptides at ~1 mM showed chemical shift dispersion similar to GCN4-p1 (data not shown).

Discussion

Peptide 6A, designed with a hydrogen-bond potential, melts at 71 °C versus 57 °C for GCN4-p1, demonstrating that rational design of surface residues can produce structures that are markedly more stable than naturally occurring coiled coils. This gain in stability is probably not due to improved hydrogen bonding since 6D, which has the same surface amino acid composition as 6A but a scrambled sequence and no predicted hydrogen bonds, also melts at 71 °C. Further, 6B was designed with a different scoring function and has a different sequence and set of predicted hydrogen bonds but a very similar T_m of 72 °C.

An alternative explanation for the increased stability of these sequences relative to GCN4-p1 is their higher helix propensity. The long polar residues selected by the hydrogen bond potential, Lys, Glu, Arg and Gln, are also among the best helix formers (22). Since the effect of helix propensity is not as dependent on sequence position as that of hydrogen bonding, especially far from the helix ends, little effect would be expected from scrambling the sequence of 6A. A rough measure of the helix propensity of the surface sequences, the sum of the standard free energies of helix propagation ($\Sigma\Delta G°$) (22), corresponds to the peptides' thermal stabilities (Table 1). Though $\Sigma\Delta G°$ matches the trend in peptide stability, it is not quantitatively correlated to the increased stability of these coiled coils.

Peptide 6C was designed with helix propensity as part of the scoring function and it has a $\Sigma\Delta G°$ of -2.041 kcal/mol. Though 6C is more stable than GCN4-p1, its T_m of 69 °C is slightly lower than 6A and 6B, in spite of 6C's higher helix propensity. Similarly, 6F has the highest helix propensity possible with an all Ala sequence and a $\Sigma\Delta G°$ of -3.096 kcal/mol, but its T_m of 73 °C is only marginally higher than that of 6A or 6B. 6F also migrates as a tetramer during SEC, not a dimer, likely because its poly(Ala) surface exposes a large hydrophobic patch that could mediate association. Though the results for 6C and 6F support the conclusion that helix propensity is important for surface design, they point out possible limitations in using propensity exclusively. Increasing propensity does not necessarily confer the greatest stability on a structure, perhaps because other factors are being effected unfavorably. Also, as is evident from 6F, changes in the tertiary structure of the protein can occur. The destabilization of peptide 6G relative to GCN4-p1 suggests that specific interactions, such as the hydrogen bond formed by Asp 7, play an important role in the stabilization of GCN4-p1. The similar stabilities for 6D and 6A imply that no specific interactions are present in 6A which suggests that multiple mechanisms can greatly effect helical stability.

The characterization of these peptides clearly shows that surface residues have a dramatic impact on the stability of α-helical coiled coils. The wide range of stabilities displayed by the different surface designs is notable, with greater than a 50 °C spread between the random hydrophilic sequence (T_m 15 °C) and the designed sequences (T_m 69 - 72 °C). This result is consistent with studies on other proteins that demonstrated the importance of solvent exposed residues (23, 24, 29, 30). Further, these designs have significantly higher T_m's than the wildtype GCN4-p1 sequence, demonstrating that surface residues can be used to improve stability in protein design (26). Though helix propensity appears to be more important than hydrogen bonding in stabilizing the designed coiled coils, hydrogen bonding could be important in the design and stabilization of other types of secondary structure.

Methods and materials

Sequence design: Scoring functions and DEE

The protein structure was modeled on the backbone coordinates of GCN4-p1, PDB record 2zta (11, 13). Atoms of all side chains not optimized were left in their crystallographically determined positions. The program BIOGRAF (Molecular Simulations Incorporated, 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 (31). The symmetry of the dimer and lack of interactions of surface residues between the subunits allowed independent design of each subunit. All computations were done using the first monomer to appear in 2zta (chain A). A backbone-dependent rotamer library was used (15). χ_3 angles that were undetermined from the database statistics were assigned the following values: Arg, -60°, 60°, and 180°; Gln, -120°, -60°, 0°, 60°, 120°, and 180°; Glu, 0°, 60°, and 120°; Lys, -60°, 60°, and

180°. χ_4 angles that were undetermined from the database statistics were assigned the following values: Arg, -120°, -60°, 60°, 120°, and 180°; Lys, -60°, 60°, and 180°. Rotamers with combinations of χ_3 and χ_4 that resulted in sequential g^+/g^- or g^-/g^+ angles were eliminated. Uncharged His rotamers were used. A Lennard-Jones 12-6 potential with van der Waals radii scaled by 0.9 (Dahiyat & Mayo, submitted) was used for van der Waals interactions. The hydrogen bond potential consisted of a distance-dependent term and an angle-dependent term:

$$E_{HB} = D_0 \left\{ 5\left(\frac{R_0}{R}\right)^{12} - 6\left(\frac{R_0}{R}\right)^{10} \right\} F(\theta, \phi, \varphi),$$

where R_0 (2.8 Å) and D_0 (8 kcal/mol) are the hydrogen-bond equilibrium distance and well-depth, respectively, and R is the donor to acceptor distance. This hydrogen bond potential is based on the potential used in DREIDING, with more restrictive angle-dependent terms to limit the occurrence of unfavorable hydrogen bond geometries. The angle term varies depending on the hybridization state of the donor and acceptor:

sp^3 donor - sp^3 acceptor $\quad F = \cos^2\theta \cos^2(\phi - 109.5),$

sp^3 donor - sp^2 acceptor $\quad F = \cos^2\theta \cos^2\phi,$

sp^2 donor - sp^3 acceptor $\quad F = \cos^4\theta,$

sp^2 donor - sp^2 acceptor $\quad F = \cos^2\theta \cos^2(\max[\phi, \varphi]),$

where θ is the donor-hydrogen-acceptor angle, ϕ is the hydrogen-acceptor-base angle (the base is the atom attached to the acceptor, for example the carbonyl carbon is the base for a carbonyl oxygen acceptor), and φ is the angle between the normals of the planes defined by the six atoms attached to the sp^2 centers (the supplement of φ is used when φ is less than 90°). The hydrogen-bond function is only evaluated when 2.6 Å $\leq R \leq$ 3.2 Å, $\theta > 90°$, $\phi - 109.5° < 90°$ for the sp^3 donor - sp^3 acceptor case, and $\phi > 90°$ for the sp^3 donor - sp^2 acceptor case; no switching functions were used. Template donors and acceptors that were involved in template-template hydrogen bonds were not included in the donor and acceptor lists. For the purpose of exclusion, a template-template hydrogen bond was considered to exist when 2.5 Å $\leq R \leq$ 3.3 Å and $\theta \geq 135°$. A penalty of 2 kcal/mol for polar hydrogen burial, when used, was only applied to buried polar hydrogens not involved in hydrogen bonds, where a hydrogen bond was considered to exist when E_{HB} was less than -2 kcal/mol. This penalty was not applied to template hydrogens. The hydrogen-bond potential was also supplemented with a weak coulombic term that included a distance-dependent dielectric constant of *40R*, where *R* is the interatomic distance. Partial atomic charages were only applied to polar functional groups. A net formal charge of +1 was used for Arg and Lys and a net formal charge of –1 was used for Asp and Glu. Energies associated with α-helical propensities were calculated using the following equation:

$$E_{\alpha} = 10^{N_{ss}(\Delta G^{\circ}_{aa} - \Delta G^{\circ}_{ala})} - 1,$$

where ΔG° is the standard free energy of helix propagation (22), and N_{ss} is the propensity scale factor which was set to 3.0. This potential was selected in order to scale the propensity energies to a similar range as the other terms in the scoring function. The DEE optimization followed the methods of our previous work (1). Calculations were performed on either a 12 processor, R10000-based Silicon Graphics Power Challenge or a 512 node Intel Delta.

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 by 22 mm) with a linear acetonitrile-water gradient containing 0.1% TFA. Peptides were then lyophilized and stored at -20 °C until use. Matrix assisted laser desorption mass spectrometry found all molecular weights to be within one unit of the expected masses.

CD and NMR

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 two degree 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 (32). The T_m's were reproducible to within one degree. Peptide concentrations were determined by quantitative amino acid analysis. NMR samples were prepared in 90/10 H_2O/D_2O and 50 mM sodium phosphate buffer at pH 7.0. Spectra were acquired on a Varian Unityplus 600 MHz spectrometer at 25 °C. 32 transients were acquired with 1.5 seconds of solvent presaturation used for water suppression. Samples were ~1 mM.

Size exclusion chromatography

Size exclusion chromatography was performed with a PolyLC hydroxyethyl A column (20 cm x 9 mm) at pH 7.0 in 50 mM phosphate and 150 mM NaCl at 0 °C. GCN4-p1 and p-LI (28) were used as size standards for dimer and tetramer, respectively. 5 µl injections of ~1 mM peptide solution were chromatographed at 0.50 ml/min and monitored at 214 nm. Samples were run in triplicate.

Acknowledgements

We wish to thank G. Hathaway and D. Krapf of the Caltech Protein and Peptide Microanalytical Laboratory for mass spectra and amino acid analysis and the Caltech Center for Advanced Computing Research for use of the Intel Delta. This work was supported by the Rita Allen 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. D.B.G. is partially supported by NIH Training Grant GM 07616C-19.

References

1. B. I. Dahiyat, S. L. Mayo, *Protein Sci.* **5**, 895-903 (1996).

2. J. R. Desjarlais, T. M. Handel, *Protein Sci.* **4**, 2006-2018 (1995).

3. H. W. Hellinga, J. P. Caradonna, F. M. Richards, *J. Mol. Biol.* **222**, 787-803 (1991).

4. J. H. Hurley, W. A. Baase, B. W. Matthews, *J. Mol. Biol.* **224**, 1143-1154 (1992).

5. M. Klemba, K. H. Gardner, S. Marino, N. D. Clarke, L. Regan, *Nature Struc. Biol.* **2**, 368-373 (1995).

6. P. B. Harbury, B. Tidor, P. S. Kim, *Proc. Natl. Acad. Sci. USA* **92**, 8408-8412 (1995).

7. S. F. Betz, W. F. Degrado, *Biochemistry* **35**, 6955-6962 (1996).

8. S. Nautiyal, D. N. Woolfson, D. S. King, T. Alber, *Biochemistry* **34**, 11645-11651 (1995).

9. J. F. Reidhaar-Olson, R. T. Sauer, *Science* **241**, 53-57 (1988).

10. J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, R. T. Sauer, *Science* **247**, 1306-1310 (1990).

11. E. K. O'Shea, J. D. Klemm, P. S. Kim, T. Alber, *Science* **254**, 539-544 (1991).

12. C. Cohen, D. A. D. Parry, *Proteins: Structure, Function and Genetics* **7**, 1-15 (1990).

13. F. C. Bernstein, et al., *J. Mol. Biol.* **112**, 535-542 (1977).

14. J. W. Ponder, F. M. Richards, *J. Mol. Biol.* **193**, 775-791 (1987).

15. R. L. Dunbrack, M. Karplus, *J. Mol. Biol.* **230**, 543-574 (1993).

16. J. Desmet, M. De Maeyer, B. Hazes, I. Lasters, *Nature* **356**, 539-542 (1992).

17. J. Desmet, M. De Maeyer, I. Lasters, in *The protein folding problem and tertiary structure prediction* K. Merz Jr, S. Le Grand, Eds. (Birkhauser, Boston, 1994) pp. 307-337.

18. R. F. Goldstein, *Biophys. J.* **66**, 1335-1340 (1994).

19. D. F. Stickle, L. G. Presta, K. A. Dill, G. D. Rose, *J. Mol. Biol.* **226**, 1143-1159 (1992).

20. B. M. P. Huyghues-Despointes, T. M. Klingler, R. L. Baldwin, *Biochemistry* **34**, 13267-13271 (1995).

21. D. Eisenberg, A. D. McLachlan, *Nature* **319**, 199-203 (1986).

22. A. Chakrabartty, T. Kortemme, R. L. Baldwin, *Protein Sci.* **3**, 843-852 (1994).

23. K. T. O'Neil, W. F. DeGrado, *Science* **250**, 646-651 (1990).

24. X. J. Zhang, W. A. Baase, B. W. Matthews, *Biochemistry* **30**, 2012-2017 (1991).

25. M. Blaber, X. J. Zhang, B. W. Matthews, *Science* **260**, 1637-1640 (1993).

26. E. K. O'shea, K. J. Lumb, P. S. Kim, *Current Biology* **3**, 658-667 (1993).

27. V. Villegas, A. R. Viguera, F. X. Aviles, L. Serrano, *Folding and Design* **1**, 29-34 (1996).

28. P. B. Harbury, T. Zhang, P. S. Kim, T. Alber, *Science* **262**, 1401-1407 (1993).

29. D. L. Minor, P. S. Kim, *Nature* **367**, 660-663 (1994).

30. C. K. Smith, J. M. Withka, L. Regan, *Biochemistry* **33**, 5510-5517 (1994).

31. S. L. Mayo, B. D. Olafson, W. A. Goddard III, *J. Phys. Chem.* **94**, 8897-8909 (1990).

32. C. R. Cantor, P. R. Schimmel, *Biophysical Chemistry* (W. H. Freeman and Company, New York, 1980), vol. 3.

Table 1. Sequences and properties of synthesized peptides

Peptide	Design method	Surface Sequence	T_m	$\Sigma\Delta G°$	N
		bcf bcf bcf bcf	(°C)	(kcal/mol)	
GCN4-p1	none	KQD EES YHN ARK	57	3.831	2
6A	HB	EKD RER RRE RRE	71	2.193	2
6B	HB + PB	EKQ KER ERE ERQ	72	2.868	2
6C	HB + HP	ARA AAA RRR ARA	69	-2.041	2
6D	scrambled HB	REE RRR EDR KRE	71	2.193	2
6E	random polar	NTR AKS ANH NTQ	15	4.954	2
6F	poly(Ala)	AAA AAA AAA AAA	73	-3.096	4
6G	random GCN4	ENQ AKE RSH KDY	22	3.831	-

For clarity only the designed surface residues are shown and they are grouped by position (**b**, **c**, and **f**). The sequence numbers of the designed positions are: 3, 4, 7, 10, 11, 14, 17, 18, 21, 24, 25, and 28. Melting temperatures (T_m's) were determined by circular dichroism and oligomerization states (N) were determined by size exclusion chromatography. $\Sigma\Delta G°$ is the sum of the standard free energy of helix propagation of the 12 **b**, **c**, and **f** positions (22). Abbreviations for design methods are: hydrogen bonds (HB), polar hydrogen burial penalty (PB), and helix propensity (HP).

Figure 4-1. Helical wheel diagram of a dimeric coiled coil (12). One heptad repeat is shown viewed down the major axes of the helices. The **b**, **c**, and **f** positions define the solvent-exposed surface of the molecule, as determined by accessible surface area calculations.

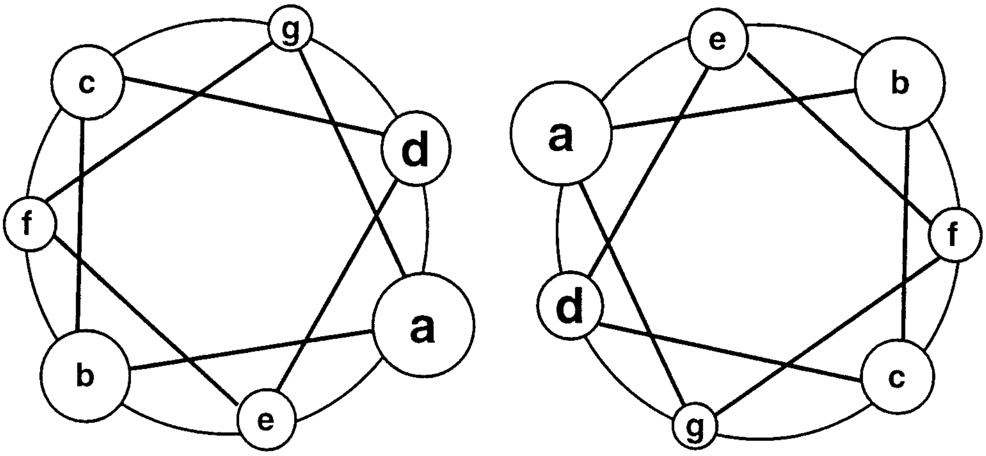
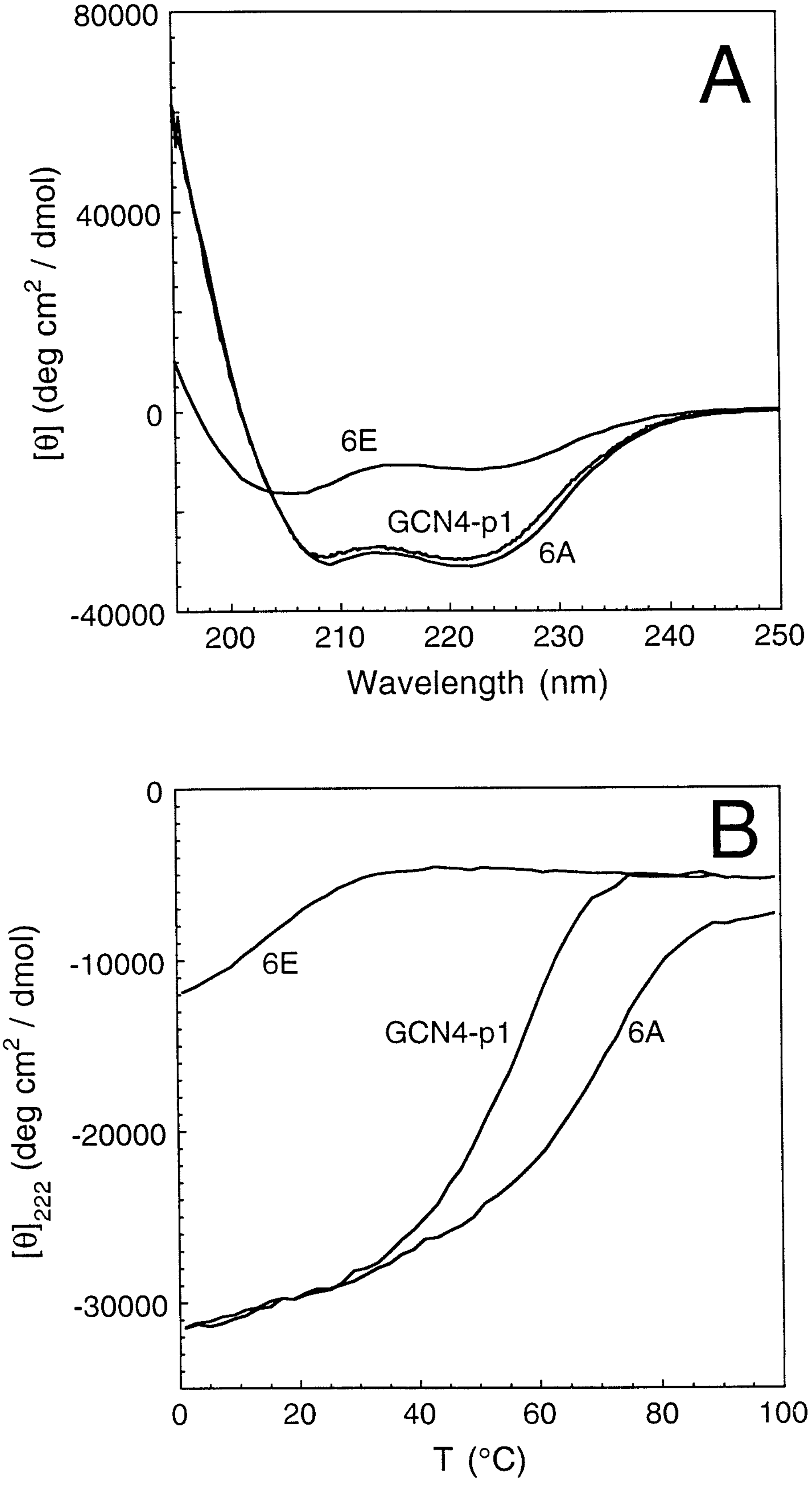

g
c
d
f
a
b
e
e
b
a
f
d
c
g

Figure 4-2. Typical CD data. **A**: Spectra of 6A, 6E and wildtype GCN4-p1. The spectra of 6A and GCN4-p1 are nearly identical and are consistent with 100% α helix while the spectrum of 6E is a mixture of α helix and random coil. **B**: Thermal melts of peptides 6A, 6E and GCN4-p1 monitored at 222 nm were used to calculate T_m's from the minima of plots of $d[\theta]/dT^{-1}$ versus T (Table 1).


A
80000
40000
0
-40000
[θ] (deg cm^2 / dmol)
6E
GCN4-p1
6A
200
210
220
230
240
250
Wavelength (nm)
B
0
-10000
-20000
-30000
[θ]$_{222}$ (deg cm^2 / dmol)
6E
GCN4-p1
6A
0
20
40
60
80
100
T (°C)

Chapter 5

De novo Protein Design: Fully Automated Sequence Selection

The text of this chapter is adapted from a published manuscript that was coauthored with Professor Stephen L. Mayo.

B. I. Dahiyat and S. L. Mayo, *Science*, **278**, 82 (1997).

Abstract

We report the first fully automated design and experimental validation of a novel sequence for an entire protein. Using a computational design algorithm based on physical chemical potential functions, we screened a virtual combinatorial library of 1.9 x 10^{27} possible amino acid sequences for compatibility with the design target, a ββα protein motif. A BLAST search shows that the designed sequence, Full Sequence Design 1 (FSD-1), has very low identity to any known protein sequence. The solution structure of FSD-1 was solved using nuclear magnetic resonance spectroscopy and indicates that FSD-1 forms a compact, well ordered structure that is in excellent agreement with the design target structure.

De novo protein design has received considerable attention recently, and significant advances have been made toward the goal of producing stable, well-folded proteins with novel sequences (1). These efforts have generated insight into the factors that control protein folding. In addition, the design of proteins with novel structures and functions heralds new approaches to biotechnology (2). In order to broaden the scope and power of protein design techniques, several groups have developed and experimentally tested systematic, quantitative methods for protein design with the goal of developing general design algorithms (3, 4). To date, these techniques, which screen possible sequences for compatibility with the desired protein fold, have focused mostly on the redesign of protein cores.

We have sought to expand the range of computational protein design by developing quantitative design methods for residues of all parts of a protein: the buried core, the solvent exposed surface, and the boundary between core and surface (4, 5, 6). Our goal is an objective, quantitative design algorithm that is based on the physical properties that determine protein structure and stability and which is not limited to specific folds or motifs. Such a method should escape the lack of generality that has resulted from design approaches based on system-specific heuristics and/or subjective considerations. A critical component of the development of our methods has been their experimental testing and validation. The use of a design cycle coupling theory, computation, and experiment has improved our understanding of the physical chemistry governing protein design and hence enhanced the performance of the design algorithm (4). This work reports the first successful automated design and experimental validation of a novel sequence for an entire protein.

Sequence selection. Our design methodology consists of an automated side-chain selection algorithm that explicitly and quantitatively considers specific side-chain to backbone and side-chain to side-chain interactions (4). The side-chain selection algorithm screens all possible sequences and finds the optimal sequence of amino acid types and side-chain orientations for a given backbone. In order to correctly account for the torsional flexibility of side chains and the geometric specificity of side-chain placement, we consider a discrete set of all allowed conformers of each side chain, called rotamers (7). The immense search problem presented by rotamer sequence optimization is overcome by application of the Dead-End Elimination (DEE) theorem (8). Our implementation of the DEE theorem extends its utility to sequence design and rapidly finds the globally optimal sequence in its optimal conformation (4).

In previous work we determined the different contributions of core, surface, and boundary residues to the scoring of a sequence arrangement. We assessed the accuracy of a scoring function or combination of scoring functions by experimentally testing their sequence predictions. Improvements to the scoring function were derived from the experimental data and incorporated into the design algorithm. The core of a coiled coil and of the streptococcal protein G β1 (Gβ1) domain were successfully redesigned using a van der Waals potential to account for steric constraints and an atomic solvation potential favoring the burial and penalizing the exposure of nonpolar surface area (4, 6). Effective solvation parameters and the appropriate balance between packing and solvation terms were found by systematic analysis of experimental data and feedback into the simulation. Solvent exposed residues on the surface of a protein were designed using a hydrogen-bond potential and secondary structure propensities in addition to a van der Waals potential. Coiled coils designed with such a scoring function were 10-12 °C more thermally stable than the naturally occurring analogue (5). Residues that form the boundary between the core and surface require a combination of the core and the surface scoring functions. The algorithm considers both hydrophobic and hydrophilic amino acids at boundary positions, while core positions are restricted to hydrophobic amino acids and surface positions are restricted to hydrophilic amino acids.

In order to assess the capability of our design algorithm, we have computed the entire amino acid sequence for a small protein motif. We sought a protein fold that would be small enough to be both computationally and experimentally tractable, yet large enough to form an independently folded structure in the absence of disulfide bonds or metal binding. We chose the ββα motif typified by the zinc finger DNA binding module (9). Though it consists of less than 30 residues, this motif contains sheet, helix, and turn structures. Recent work has demonstrated the ability of this fold to form in the absence of metal ions or disulfide bonds. Imperiali and coworkers designed a 23 residue peptide, containing an unusual amino acid (D-proline) and a non-natural amino acid (3-[1,10-phenanthrol-2-yl]-L-alanine), that achieved this fold (10); our initial characterization of a partially computed sequence indicated that it also forms this fold (11). In computing the full sequence for this target fold, we use the scoring functions from our previous work without modification in order to provide a test of the algorithm's generality (12). The ββα motif was not used in any of our prior work to develop the design methodology.

The sequence selection algorithm requires structure coordinates that define the target motif's backbone (N, Cα, C and O atoms and Cα-Cβ vectors). The Brookhaven Protein Data Bank (PDB) (13) was examined for high resolution structures of the ββα motif, and the second zinc finger module of the DNA binding protein Zif268 was selected as our design template (9, 14). In order to assign the residue positions in the template structure into core, surface or boundary classes, the orientation of the Cα-Cβ vectors was assessed relative to a solvent accessible surface computed using only the template Cα atoms (15). The small size of this motif limits to one (position 5) the number of residues that can be assigned unambiguously to the core while seven residues (positions 3, 7, 12, 18, 21, 22, and 25) were classified as boundary and the remaining 20 residues were assigned to the surface. Interestingly, while three of the zinc binding positions of Zif268 are in the boundary or core, one residue, position 8, has a Cα-Cβ vector directed away from the protein's geometric center and is classified as a surface position. As in our previous studies, the amino acids considered at the core positions during sequence selection were A, V, L, I, F, Y, and W; the amino acids considered at the surface positions were A, S, T, H, D, N, E, Q, K, and R; and the combined core and surface amino acid sets (16 amino acids) were considered at the boundary positions. Two of the residue positions (9 and 27) have ϕ angles greater than 0° and are set to Gly by the sequence selection algorithm to minimize backbone strain.

The total number of amino acid sequences that must be considered by the design algorithm is the product of the number of possible amino acid types at each residue position. The ββα motif residue classification described above results in a virtual combinatorial library of 1.9×10^{27} possible amino acid sequences (16). A corresponding peptide library consisting of only a single molecule for each 28 residue sequence would have a mass of 11.6 metric tons (17). In order to accurately model the geometric specificity of side-chain placement, we explicitly consider the torsional flexibility of amino acid side chains in our sequence scoring by representing each amino acid with a discrete set of allowed conformations, called rotamers (18). As a result, the design algorithm must consider all rotamers for each possible amino acid at each residue position. The total size of the search space for the ββα motif is therefore 1.1 x 10^{62} possible rotamer sequences. We use a search algorithm based on an extension of the DEE theorem to solve the rotamer sequence optimization problem (4, 8). Efficient implementation of the DEE theorem has made complete protein sequence design tractable for about 50 residues on current parallel computers in a single calculation. The rotamer optimization problem for the ββα motif required 90 CPU hours to find the optimal sequence (19, 20).

The optimal sequence, shown in Figure 1, is called Full Sequence Design-1 (FSD-1). Even though all of the hydrophilic amino acids were considered at each of the boundary positions, the algorithm selected only nonpolar amino acids. The eight core and boundary positions are predicted to form a well-packed buried cluster. The Phe side chains selected by the algorithm at the zinc binding His positions of Zif268, positions 21 and 25, are over 80% buried and the Ala at position 5 is 100% buried while the Lys at position 8 is greater than 60% exposed to solvent (Figure 2). The other boundary positions demonstrate the strong steric constraints on buried residues by packing similar side chains in an arrangement similar to that of Zif268 (Figure 2). The calculated optimal configuration for core and boundary residues buries ~1150 Å^2 of nonpolar surface area. On the helix surface, the algorithm positions Asn 14 as a helix N-cap with a hydrogen bond between its side-chain carbonyl oxygen and the backbone amide proton of residue 16. The eight charged residues on the helix form three pairs of hydrogen bonds, though in our coiled coil designs helical surface hydrogen bonds appeared to be less important than the overall helix propensity of the sequence (5).

Positions 4 and 11 on the exposed sheet surface were selected to be Thr, one of the best β-sheet forming residues (21).

Figure 1 shows the alignment of the sequences for FSD-1 and Zif268. Only 6 of the 28 residues (21%) are identical and only 11 (39%) are similar. Four of the identities are in the buried cluster, which is consistent with the expectation that buried residues are more conserved than solvent exposed residues for a given motif (22). A BLAST (23) search of the FSD-1 sequence against the non-redundant protein sequence database of the National Center for Biotechnology Information did not find any zinc finger protein sequences. Further, the BLAST search found only low identity matches of weak statistical significance to fragments of various unrelated proteins. The highest identity matches were 10 residues (36%) with p values ranging from 0.63 - 1.0. Random 28 residue sequences that consist of amino acids allowed in the ββα position classification described above produced similar BLAST search results, with 10 or 11 residue identities (36 - 39%) and p values ranging from 0.35 - 1.0, further suggesting that the matches found for FSD-1 are statistically insignificant. The very low identity to any known protein sequence demonstrates the novelty of the FSD-1 sequence and underscores that no sequence information from any protein motif was used in our sequence scoring function.

In order to examine the robustness of the computed sequence, the sequence of FSD-1 was used as the starting point of a Monte Carlo simulated annealing run. The Monte Carlo search finds high scoring, suboptimal sequences in the neighborhood of the optimal solution (4). The energy spread from the ground-state solution to the 1000th most stable sequence is about 5 kcal/mol indicating that the density of states is high. The amino acids comprising the core of the molecule, with the exception of position 7, are essentially invariant (Figure 1). Almost all of the sequence variation occurs at surface positions, and typically involves conservative changes. Asn 14, which is predicted to form a helix N-cap, is among the most conserved surface positions. The strong sequence conservation observed for critical areas of the molecule suggests that if a representative sequence folds into the design target structure, then perhaps thousands of sequences whose variations do not disrupt the critical interactions may be equally competent. Even if billions of sequences would successfully achieve the target fold, they would represent only a vanishingly small proportion of the 10^{27} possible sequences.

Experimental validation. FSD-1 was synthesized in order to characterize its structure and assess the performance of the design algorithm (24). The far UV circular dichroism (CD) spectrum of FSD-1 shows minima at 220 nm and 207 nm, which is indicative of a folded structure (Figure 3a) (25). The thermal melt is weakly cooperative, with an inflection point at 39 °C (Figure 3b), and is completely reversible (data not shown). The broad melt is consistent with a low enthalpy of folding which is expected for a motif with a small hydrophobic core. This behavior contrasts the uncooperative thermal unfolding transitions observed for other folded short peptides (26). FSD-1 is highly soluble (greater than 3 mM) and equilibrium sedimentation studies at 100 μM, 500 μM and 1 mM show the protein to be monomeric (27). The sedimentation data fit well to a single species, monomer model with a molecular mass of 3630 at 1 mM, in good agreement with the calculated monomer mass of 3488. Also, far UV CD spectra showed no concentration dependence from 50 μM to 2 mM, and nuclear magnetic resonance (NMR) COSY spectra taken at 100 μM and 2 mM were essentially identical.

The solution structure of FSD-1 was solved using homonuclear 2D ^{1}H NMR spectroscopy (28). NMR spectra were well dispersed indicating an ordered protein structure and easing resonance assignments. Proton chemical shift assignments were determined with standard homonuclear methods (29). Unambiguous sequential and short-range NOEs are shown in Figure 4 and indicate helical secondary structure from residues 15 to 26 in agreement with the design target.

The structure of FSD-1 was determined using 284 experimental restraints (10.1 restraints per residue) that were non-redundant with covalent structure including 274 NOE distance restraints and 10 hydrogen bond restraints involving slowly exchanging amide protons (30). Structure calculations were performed using X-PLOR (31) with standard protocols for hybrid distance geometry-simulated annealing (32). An ensemble of 41 structures converged with good covalent geometry and no distance restraint violations greater than 0.3 Å (Figure 5 and Table 1). The backbone of FSD-1 is well defined with a root-mean-square (rms) deviation from the mean of 0.54 Å (residues 3-26). Considering the buried side chains (residues 3, 5, 7, 12, 18, 21, 22, and 25) in addition to the backbone gives an rms deviation of 0.99 Å, indicating that the core of the molecule is well ordered. The stereochemical quality of the ensemble of structures was examined using PROCHECK (33). Not including the disordered termini and the glycine residues, 87% of the residues fall in the most favored region and the remainder in the allowed region of ϕ,ψ space. Modest heterogeneity is present in the first strand (residues 3-6) which has an average backbone angular order parameter (34) of $<S> = 0.96 \pm 0.04$ compared to the second strand (residues 9-12) with an $<S> = 0.98 \pm 0.02$ and the helix (residues 15-26) with an $<S> = 0.99 \pm 0.01$. Overall, FSD-1 is notably well ordered and, to our knowledge, is the shortest sequence consisting entirely of naturally occurring amino acids that folds to a unique structure without metal binding, oligomerization or disulfide bond formation (35).

The packing pattern of the hydrophobic core of the NMR structure ensemble of FSD-1 (Tyr 3, Ile 7, Phe 12, Leu 18, Phe 21, Ile 22, and Phe 25) is similar to the computed packing arrangement. Five of the seven residues have χ_1 angles in the same gauche$^+$, gauche$^-$ or trans category as the design target, and three residues match both χ_1 and χ_2 angles. The two residues that do not match their computed χ_1 angles are Ile 7 and Phe 25, which is consistent with their location at the less constrained, open end of the molecule. Ala 5 is not involved in its expected extensive packing interactions and instead exposes about 45% of its surface area because of the displacement of the strand 1 backbone relative to the design template. Conversely, Lys 8 behaves as predicted by the algorithm with its solvent exposure (60%) and χ_1 and χ_2 angles matching the computed structure. Most of the solvent exposed residues are disordered which precludes examination of the predicted surface residue hydrogen bonds. Asn 14, however, forms a helix N-cap from its sidechain carbonyl oxygen as predicted, but to the amide of Glu 17, not Lys 16 as expected from the design. This hydrogen bond is present in 95% of the structure ensemble and has a donor-acceptor distance of 2.6 ± 0.06 Å. In general, the side chains of FSD-1 correspond well with the design algorithm predictions, but further refinement of the scoring function and rotamer library should improve side-chain placement.

In Figure 6, we compare the average restrained minimized structure of FSD-1 and the design target. The overall backbone rms deviation of FSD-1 from the design target is 1.98 Å for residues 3-26 and only 0.98 Å for residues 8-26 (Table 2). The largest difference between FSD-1 and the target structure occurs from residues 4-7, with a displacement of 3.0-3.5 Å of the backbone atom positions of strand 1. The agreement for strand 2, the strand to helix turn, and the helix is remarkable, with the differences nearly within the accuracy of the structure determination. For this region of the structure, the rms difference of ϕ,ψ angles between FSD-1 and the design target is only 14 ± 9°. In order to quantitatively assess the similarity of FSD-1 to the global fold of the target, we calculated their supersecondary structure parameters (Table 1) (36, 37), which describe the relative orientations of secondary structure units in proteins. The values of θ, the inclination of the helix relative to the sheet, and Ω, the dihedral angle between the helix axis and the strand axes, are nearly identical. The height of the helix above the sheet, h, is only 1 Å greater in FSD-1. A study of protein core design as a function of helix height for Gβ1 variants demonstrated that up to 1.5 Å variation in helix height has little effect on sequence selection (37). The comparison of secondary structure parameter values and backbone coordinates highlights the excellent agreement between the experimentally determined structure of FSD-1 and the design target, and demonstrates the success of our algorithm at computing a sequence for this ββα motif.

The quality of the match between FSD-1 and the design target demonstrates the ability of our algorithm to design a sequence for a fold that contains the three major secondary structure elements of proteins: sheet, helix, and turn. Since the ββα fold is different from those used to develop the sequence selection methodology, the design of FSD-1 represents a successful transfer of our algorithm to a new motif. We are currently testing the performance of the algorithm on several different motifs, with the hope that its basis in physical chemistry and the absence of heuristics and subjective considerations will allow further generalization. Also, we are exploring the generation of novel backbone templates for use as input to our fully automated sequence selection algorithm to enable the design of new protein folds. Recent results indicate that the sequence selection algorithm is not sensitive to even fairly large perturbations in backbone geometry and should be robust enough to accommodate computer-generated backbones (37).

By using an optimization technique based on the DEE theorem, we were able to circumvent the combinatorial size and complexity of protein design and apply objective criteria for amino acid sequence selection. The key to using a computational optimization technique for the FSD-1 design, and for the continued development of the methodology in the future, is the tight coupling of theory, computation, and experiment used to develop a better understanding of the factors controlling protein structure and stability. By using a quantitative design method and improving the accuracy of the physicochemical description of protein systems, we are able to apply the power of computational approaches to protein design. Given that the reported sequence was computed with only a four GigaFLOPS computer (19), and that TeraFLOPS computers are now available with PetaFLOPS computers on the drawing board, the prospect for pursuing larger and more complex designs is excellent.

References

1. M. H. J. Cordes, A. R. Davidson, R. T. Sauer, *Curr. Op. Struct. Biol.* **6**, 3 (1996).
2. D. Y. Jackson *et al., Science* **266**, 243 (1994); B. Li *et al., ibid* **270**, 1657 (1995); J. S. Marvin *et al., Proc. Natl. Acad. Sci. USA* **94**, 4366 (1997).
3. H. W. Hellinga, J. P. Caradonna, F. M. Richards, *J. Mol. Biol.* **222**, 787 (1991); J. H. Hurley, W. A. Baase, B. W. Matthews, *ibid* **224**, 1143 (1992); J. R. Desjarlais and T. M. Handel, *Protein Sci.* **4**, 2006 (1995); P. B. Harbury, B. Tidor, P. S. Kim, *Proc. Natl. Acad. Sci. USA* **92**, 8408 (1995); M. Klemba, K. H. Gardner, S. Marino, N. D. Clarke, L. Regan, *Nature Struc. Biol.* **2**, 368 (1995); S. F. Betz and W. F. Degrado, *Biochemistry* **35**, 6955 (1996).
4. B. I. Dahiyat and S. L. Mayo, *Protein Sci.* **5**, 895 (1996).
5. B. I. Dahiyat and S. L. Mayo, *ibid*, in press.
6. B. I. Dahiyat and S. L. Mayo, *Proc. Natl. Acad. Sci. USA*, in press.
7. J. W. Ponder and F. M. Richards, *J. Mol. Biol.* **193**, 775 (1987).
8. [J. Desmet, M. De Maeyer, B. Hazes, I. Lasters, *Nature* **356**, 539 (1992); R. F. Goldstein, *Biophys. J.* **66**, 1335 (1994); M. De Maeyer, J. Desmet, I. Laster, *Folding and Design* **2**, 53 (1997)] DEE finds and eliminates rotamers that are mathematically provable to be inconsistent (or dead ending) with the global minimum energy solution of the system. A rotamer, *r*, at some residue position, *i*, will be dead ending if when compared with some other rotamer, *t*, at the same residue position the following inequality is satisfied:

$$E(i_r) - E(i_t) + \sum_j \min_s \{E(i_r j_s) - E(i_t j_s)\} > 0$$

where $E(i_r)$ and $E(i_t)$ are rotamer/template energies, $E(i_r j_s)$ and $E(i_t j_s)$ are rotamer/rotamer energies for rotamers on residues *i* and *j*, and the function min_s selects the rotamer *s* on residue *j* that minimizes the function's argument. Iterative application of the above elimination criterion results in a rapid and substantial reduction in the combinatorial size of the problem; however, this rarely results in finding the ground state solution and application of similar but higher order elimination criteria are required. Application of the higher order elimination criteria are computationally intensive and dominate the total computation time because they involve interactions between three or more residue positions.

9. N. P. Pavletich and C. O. Pabo, *Science* **252**, 809 (1991).

10. M. D. Struthers, R. P. Cheng, B. Imperiali, *ibid* **271**, 342 (1996).

11. B. I. Dahiyat and S. L. Mayo, unpublished results.

12. A Lennard-Jones 12-6 potential with van der Waals radii scaled by 0.9 (6) was used for van der Waals interactions for all residues. An atomic solvation parameter of 23.2 cal/mol/Å^2 was used to favor hydrophobic burial and to penalize solvent exposure for core and boundary residues (4, 6). The Richards definition of solvent-accessible surface area [B. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971)] was used and areas were calculated with the Connolly algorithm (40). All residues with hydrogen bond donor and/or acceptors used a hydrogen bond potential based on the potential used in Dreiding (41) but with more restrictive angle-dependent terms to limit the occurrence of unfavorable hydrogen bond geometries (5). A secondary structure propensity

potential was used for surface β sheet positions where the i-1 and i+1 residues were also in β sheet conformations (5). Propensity values from Serrano and coworkers were used [V. Munoz and L. Serrano, *Proteins: Struct. Funct. Genet.* **20**, 301 (1994)].

13. F. C. Bernstein *et al., J. Mol. Biol.* **112**, 535 (1977).

14. The coordinates of PDB record 1zaa (9, 13) from residues 33-60 were used as the structure template. In our numbering, position 1 corresponds to 1zaa position 33. The program BIOGRAF (Molecular Simulations, Incorporated, San Diego, CA) was used to generate explicit hydrogens on the structure which was then conjugate gradient minimized for 50 steps using the Dreiding force field (41).

15. A solvent accessible surface for only the Cα atoms of the target fold was generated using the Connolly algorithm (40) with a probe radius of 8.0 Å, a dot density of 10 $Å^{-2}$ and a Cα radius of 1.95 Å. A residue was classified as a core position if the distance from its Cα, along its Cα-Cβ vector, to the solvent accessible surface was greater than 5.0 Å, and if the distance from its Cβ to the nearest surface point was greater than 2.0 Å. The remaining residues were classified as surface positions if the sum of the distances from their Cα, along their Cα-Cβ vector, to the solvent accessible surface plus the distance from their Cβ to the closest surface point was less than 2.7 Å. All remaining residues were classified as boundary positions. The classifications for Zif268 were used as computed except that positions 1, 17, and 23 were converted from the boundary to the surface class to account for end effects from the proximity of chain termini to these residues in the tertiary structure and inaccuracies in the assignment.

16. One core position (7 possible amino acids), 7 boundary positions (16 possible amino acids), 18 surface positions (10 possible amino acids) and 2 positions with ϕ greater than 0° (1 possible amino acid) result in 7 * 16^7 * 10^{18} * 1^2= 1.88 x 10^{27} possible amino acid sequences.

17. 1.88 x 10^{27} peptide molecules, with an average mass of 3712 daltons for the possible compositions allowed by the residue position classification, would weigh (1.88 x 10^{27} * 3712 daltons) / 6.02 x 10^{23} daltons/gram = 1.159 x 10^7 grams = 11.6 metric tons.

18. As in our previous work (5), a backbone-dependent rotamer library was used [R. L. Dunbrack and M. Karplus, *J. Mol. Biol.* **230**, 543 (1993)]. χ_1 and χ_2 angle values of rotamers for all aromatic amino acids, and χ_1 angle values for all other hydrophobic amino acids were expanded ±1 standard deviation about the mean value reported in the Dunbrack and Karplus library. χ_3 angles that were undetermined from the database statistics were assigned the following values: Arg, -60°, 60°, and 180°; Gln, -120°, -60°, 0°, 60°, 120°, and 180°; Glu, 0°, 60°, and 120°; Lys, -60°, 60°, and 180°. χ_4 angles that were undetermined from the database statistics were assigned the following values: Arg, -120°, -60°, 60°, 120°, and 180°; Lys, -60°, 60°, and 180°. Rotamers with combinations of χ_3 and χ_4 that resulted in sequential g^+/g^- or g^-/g^+ angles were eliminated. All rotamers contained explicit hydrogen atoms and were built with bond lengths and angles from the Dreiding force field (41). All His rotamers were protonated on both Nδ and Nε.

19. All calculations were performed on a Silicon Graphics Power Challenge server with 10 R10000 processors running in parallel. Peak performance is 3.9 GigaFLOPS (FLOPS = floating point operations per second).

20. The sequence optimization consists of two phases: pairwise rotamer energy calculations and DEE searching. The DEE optimization was initially run with control parameters set for optimal speed followed by a DEE-based, residue pairwise, round robin optimization. The energy calculations took 53 CPU hours and sequence optimizations took 37 CPU hours.

21. C. W. A. Kim and J. M. Berg, *Nature* **362**, 267 (1993); D. L. Minor and P. S. Kim, *ibid* **367**, 660 (1994); C. K. Smith, J. M. Withka, L. Regan, *Biochemistry* **33**, 5510 (1994).

22. J. U. Bowie, J. F. Reidhaar-Olson, W. A. Lim, R. T. Sauer, *Science* **247**, 1306 (1990).

23. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).

24. FSD-1 was synthesized using standard solid phase Fmoc chemistry on an Applied Biosystems 433A automated peptide synthesizer. The peptide was cleaved from the resin with TFA and purified by reversed phase high performance liquid chromatography on a Vydac C8 column (25 cm x 10 mm) with a linear acetonitrile-water gradient containing 0.1% TFA. Peptide was lyophilized and stored at -20 °C. Matrix assisted laser desorption mass spectrometry yielded a molecular weight of 3489.7 daltons (3489.0 calculated).

25. Protein concentration was 50 μM in 50 mM sodium phosphate at pH 5.0. The spectrum was acquired at 1 °C in a 1 mm cuvette and was baseline corrected with a buffer blank. The spectrum is the average of 3 scans using a 1 s integration time and 1 nm increments. All CD data were acquired on an Aviv 62DS spectrometer equipped with a thermoelectric temperature control unit. Thermal unfolding was

monitored at 218 nm in a 1 mm cuvette using 2° increments with an averaging time of 40 s and an equilibration time of 120 s per increment. Reversibility was confirmed by comparing 1 °C CD spectra from before and after heating to 99 °C. Peptide concentrations were determined by UV spectrophotometry.

26. J. M. Scholtz *et al.*, *Proc. Natl. Acad. Sci. USA* **88**, 2854 (1991); M. A. Weiss and H. T. Keutmann, *Biochemistry* **29**, 9808 (1990); M. D. Struthers, R. P. Cheng, B. Imperiali, *J. Am. Chem. Soc.* **118**, 3073 (1996).

27. Sedimentation equilibrium studies were carried out on a Beckman XL-A ultracentrifuge equipped with an An-60 Ti analytical rotor at a rotor speed of 40,000 rpm. Protein concentration was 100 µM, 500 µM or 1 mM in 50 mM sodium phosphate at pH 5.0 and 7 °C. Absorption was monitored at 286 nm (500 µM and 1 mM) or 234 nm (100 µM). Concentration profiles were fit to an ideal single species model which resulted in randomly distributed residuals.

28. NMR data were collected on a Varian Unityplus 600 MHz spectrometer equipped with a Nalorac inverse probe with a self-shielded z-gradient. NMR samples (~2mM) were prepared in 90/10 H_2O/D_2O or 99.9% D_2O with 50 mM sodium phosphate at pH 5.0 (uncorrected glass electrode). All spectra were collected at 7 °C. DQF-COSY [U. Piantini, O. W. Sorensen, R. R. Ernst, *J. Am. Chem. Soc.* **104**, 6800 (1982)], TOCSY [A. Bax and D. G. Davis, *J. Magn. Reson.* **65**, 355 (1985)] and NOESY [J. Jeener, B. H. Meier, P. Bachmann, R. R. Ernst, *J. Chem. Phys.* **71**, 4546 (1979)] spectra were acquired to accomplish resonance assignments and structure determination. NOESY spectra were recorded with mixing times of 200 ms for use during resonance assignments and 100 ms to derive distance restraints. Water suppression was accomplished either

with presaturation during the relaxation delay or pulsed field gradients [M. Piotto, V. Saudek, V. Sklenar, *J. Biomol. NMR* **2**, 661 (1992)]. Spectra were processed using VNMR (Varian Associates, Palo Alto California) and spectra were assigned using ANSIG [P. J. Kraulis, *J. Magn. Reson.* **24**, 627 (1989)].

29. K. Wuthrich, *NMR of Proteins and Nucleic Acids* (John Wiley and Sons, New York, 1986).

30. NOEs were classified into three distance-bound ranges based on crosspeak intensity calibrated to the Tyr 3 Hδ-Hε crosspeak: strong (1.8 to 2.7 Å), medium (1.8 to 3.3 Å) and weak (1.8 to 5.0 Å). Upper bounds for restraints involving methyl protons were increased by 0.5 Å to account for the increased intensity of methyl resonances. All partially overlapped NOEs were set to weak restraints. Hydrogen bond restraints were derived from hydrogen deuterium exchange kinetics measurements followed by 1D ^{1}H spectroscopy. Unambiguously assigned amide peaks for Tyr 3, Phe 12, Leu 18, Phe 21 and Phe 25 were protected from exchange at 7 °C, pH* 5.0. Hydrogen bond restraints (two per hydrogen bond) were only included at the late stages of structure refinement when initial calculations indicated the donor-acceptor pairings.

31. A. T. Brunger, *X-PLOR Version 3.1 A system for X-ray Crystallography and NMR* (Yale University Press, New Haven, 1992).

32. Standard hybrid distance geometry-simulated annealing protocols were followed [M. Nilges, G. M. Clore, A. M. Gronenborn, *FEBS Lett.* **229**, 317 (1988); M. Nilges, J. Kuszewski, A. T. Brunger, in *Computational Aspects of the Study of Biological Macromolecules by NMR* J. C. Hoch, Ed. (Plenum Press, New York, 1991); J. Kuszewski, M. Nilges, A. T.

Brunger, *J. Biomol. NMR* **2**, 33 (1992)]. Following substructure embedding, 18 ps of high temperature (2000 K) dynamics followed by 75 ps of cooling to 100 K produced an ensemble of regularized structures. This ensemble was then refined for six cycles of 50 ps cooling to 10 K, starting at 1000 K (first four cycles) or 500 K (last two cycles), followed by 500 steps of conjugate gradient minimization. A quartic repulsive potential was used for nonbonded contacts and the final REPEL radius scale was 0.8. The force constant for NOE-derived distance restraints was 50 kcal mol^{-1} $Å^{-2}$. 100 distance geometry structures were generated and, following regularization and refinement, resulted in an ensemble, <SA>, of 41 structures with no restraint violations greater than 0.3 Å, rms deviations from idealized bond lengths less than 0.01 Å and rms deviations from idealized bond angles and impropers less than 1°. An average structure, SA, was generated by superimposing and then averaging the coordinates of <SA>. The restrained minimized structure was generated by regularizing the structure of SA as described above followed by one cycle of 500 K refinement and 500 steps of conjugate gradient minimization.

33. R. A. Laskowski, M. W. Macarthur, D. S. Moss, J. M. Thornton, *J. Appl. Crystallogr.* **26**, 283 (1993).

34. S. G. Hyberts, M. S. Goldberg, T. F. Havel, G. Wagner, *Protein Sci.* **1**, 736 (1992).

35. C. J. McKnight, P. T. Matsudaira, P. S. Kim, *Nature Struc. Biol.* **4**, 180 (1997).

36. J. Janin and C. Chothia, *J. Mol. Biol.* **143**, 95 (1980); F. E. Cohen, M. J. E. Sternberg, W. R. Taylor, *ibid* **156**, 821 (1982).

37. A. Su and S. L. Mayo, *Protein Sci.* , in press.

38. R. Koradi, M. Billeter, K. Wuthrich, *J. Mol. Graph.* **14**, 51 (1996).

39. W. J. Becktel and J. A. Schellman, *Biopolymers* **26**, 1859 (1987).

40. M. L. Connolly, *Science* **221**, 709 (1983).

41. S. L. Mayo, B. D. Olafson, W. A. Goddard III, *J. Phys. Chem.* **94**, 8897 (1990).

42. We thank P. Poon and T. Laue for sedimentation equilibrium measurements and helpful discussions; A. Su for assistance calculating super-secondary structure parameters; S. Ross for assistance with NMR measurements; G. Hathaway for mass spectrometry; J. Abelson and P. Bjorkman for critical reading of the manuscript; and R. A. Olofson for helpful discussions. Supported by the Howard Hughes Medical Institute (S.L.M.), the Rita Allen Foundation, the Chandler Family Trust, the Booth Ferris Foundation, the David and Lucile Packard Foundation, the Searle Scholars Program/The Chicago Community Trust, and grant GM08346 from the National Institutes of Health (B.I.D.). Coordinates and NMR restraints have been deposited in the Brookhaven Protein Data Bank with accession numbers 1FSD and R1FSDMR, respectively.

Table 1. NMR structure determination: distance restraints, structural statistics and atomic root-mean-square (rms) deviations. <SA> are the 41 simulated annealing structures, SA is the average structure before energy minimization, $(SA)_r$ is the restrained energy minimized average structure, and SD is the standard deviation.

Distance restraints		
Intraresidue	97	
Sequential	83	
Short range ($\|i-j\|$ = 2-5 residues)	59	
Long range ($\|i-j\|$ > 5 residues)	35	
Hydrogen bond	10	
Total	284	
Structural statistics		
	<SA> ± SD	$(SA)_r$
Rms deviation from distance restraints (Å)	0.043 ± 0.003	0.038
Rms deviation from idealized geometry		
Bonds (Å)	0.0041 ± 0.0002	0.0037
Angles (degrees)	0.67 ± 0.02	0.65
Impropers (degrees)	0.53 ± 0.05	0.51
*Atomic rms deviations (Å)**		
	<SA> vs. SA ± SD	<SA> vs. $(SA)_r$ ± SD
Backbone	0.54 ± 0.15	0.69 ± 0.16
Backbone + nonpolar side chains†	0.99 ± 0.17	1.16 ± 0.18
Heavy atoms	1.43 ± 0.20	1.90 ± 0.29

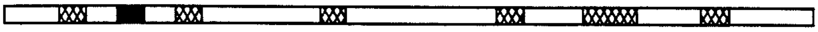
*Atomic rms deviations are for residues 3 to 26, inclusive. Residues 1, 2, 27 and 28 were disordered (ϕ,ψ angular order parameters (34) < 0.78) and had only sequential and $|i-j|=2$ NOEs. †Nonpolar side chains are from residues 3, 5, 7, 12, 18, 21, 22, and 25 which constitute the core of the protein.

Table 2. Comparison of the FSD-1 experimentally determined structure and the design target structure. The FSD-1 structure is the restrained energy minimized average from the NMR structure determination. The design target structure is the second DNA binding module of the zinc finger Zif268 (9).

Atomic rms deviations (Å)		
Backbone, residues 3-26	1.98	
Backbone, residues 8-26	0.98	
Super-secondary structure parameters*		
	FSD-1	Design target
h (Å)	9.9	8.9
θ (degrees)	14.2	16.5
Ω (degrees)	13.1	13.5

$^*h$, θ, Ω are calculated as previously described (36, 37). h is the distance between the centroid of the helix Cα coordinates (residues 15-26) and the least-squares plane fit to the Cα coordinates of the sheet (residues 3-12). θ is the angle of inclination of the principal moment of the helix Cα atoms with the plane of the sheet. Ω is the angle between the projection of the principal moment of the helix onto the sheet and the projection of the average least-squares fit line to the strand Cα coordinates (residues 3-6 and 9-12) onto the sheet.

Figure 5-1. Sequence of FSD-1 aligned with the second zinc finger of Zif268. The bar at the top of the figure shows the residue position classifications: solid bars indicate core positions, hatched bars indicate boundary positions and open bars indicate surface positions. The alignment matches positions of FSD-1 to the corresponding backbone template positions of Zif268. Of the six identical positions (21%) between FSD-1 and Zif268, four are buried (Ile 7, Phe 12, Leu 18, and Ile 22). The zinc binding residues of Zif268 are boxed. Representative non-optimal sequence solutions determined using a Monte Carlo simulated annealing protocol are shown with their rank. Vertical lines indicate identity with FSD-1. The symbols at the bottom of the figure show the degree of sequence conservation for each residue position computed across the top 1000 sequences: filled circles indicate greater than 99% conservation, half-filled circles indicate conservation between 90 and 99%, open circles indicate conservation between 50 and 90%, and the absence of a symbol indicates less than 50% conservation. The consensus sequence determined by choosing the amino acid with the highest occurrence at each position is identical to the sequence of FSD-1.



					5					10					15					20					25			
FSD-1	Q	Q	Y	T	A	K	I	K	G	R	T	F	R	N	E	K	E	L	R	D	F	I	E	K	F	K	G	R
Zif268	K	P	F	Q	C	R	I	C	M	R	N	F	S	R	S	D	H	L	T	T	H	I	R	T	H	T	G	E
Rank																												
10	E	|	|	|	|	|	|	|	|	H	|	|	|	|	|	|	|	|	|	|	|	|	|	R	|	|	|	|
20	|	|	|	|	|	|	|	R	|	T	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
30	|	|	|	|	|	|	F	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	E	|	|
40	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	R	|	E	|	|
50	E	|	|	|	|	|	|	|	|	K	|	|	|	|	K	R	|	|	|	|	|	|	|	|	|	|	|	|
60	E	|	|	|	|	|	Y	|	|	|	|	|	|	|	K	R	|	|	|	|	|	|	|	|	|	|	|	|
70	E	|	|	|	|	|	|	|	|	Q	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
80	|	R	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	R	|	|	|	|
90	|	E	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	R	|	|	|	|
100	|	|	|	|	|	|	|	|	|	|	|	|	|	|	K	R	|	|	|	|	|	|	|	H	|	|	|	|
200	T	|	|	|	|	|	|	|	|	|	|	|	|	|	K	E	|	|	K	K	|	|	|	|	|	|	|	|
300	|	E	|	|	|	|	|	|	|	|	|	|	|	|	K	R	|	|	|	|	|	|	|	|	|	|	|	|
400	E	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	T	|	|
500	E	|	|	|	|	|	|	|	|	K	|	|	|	|	|	R	|	|	|	|	|	|	|	|	|	|	|	|
600	|	|	|	|	|	|	|	|	|	K	|	|	|	|	K	D	|	|	K	K	|	|	|	|	|	|	|	|
700	|	|	|	|	|	|	|	|	|	|	|	|	|	|	K	R	|	|	Q	|	|	|	|	|	|	|	|	|
800	E	|	|	|	|	|	V	|	|	E	|	|	E	|	|	|	R	|	|	|	|	|	|	|	|	|	|	|
900	E	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	K	R	|	|	|	|	|	|	|	|
1000	E	|	|	|	|	|	F	|	|	|	|	|	|	|	K	E	|	|	K	K	|	|	|	|	|	|	|	|
		○	◐	●	●	◐	○	◐	●	○	●	●	○	◐			○	●	○	○	●	●	●		●	◐	●	◐

Figure 5-2. Comparison of Zif268 (9) and computed FSD-1 structures. (**A**) Stereoview of the second zinc finger module of Zif268 showing its buried residues and zinc binding site. (**B**) Stereoview of the computed orientations of buried side chains in FSD-1. For clarity, only side chains from residues 3, 5, 8, 12, 18, 21, 22, and 25 are shown. Color figures were created with MOLMOL (38).

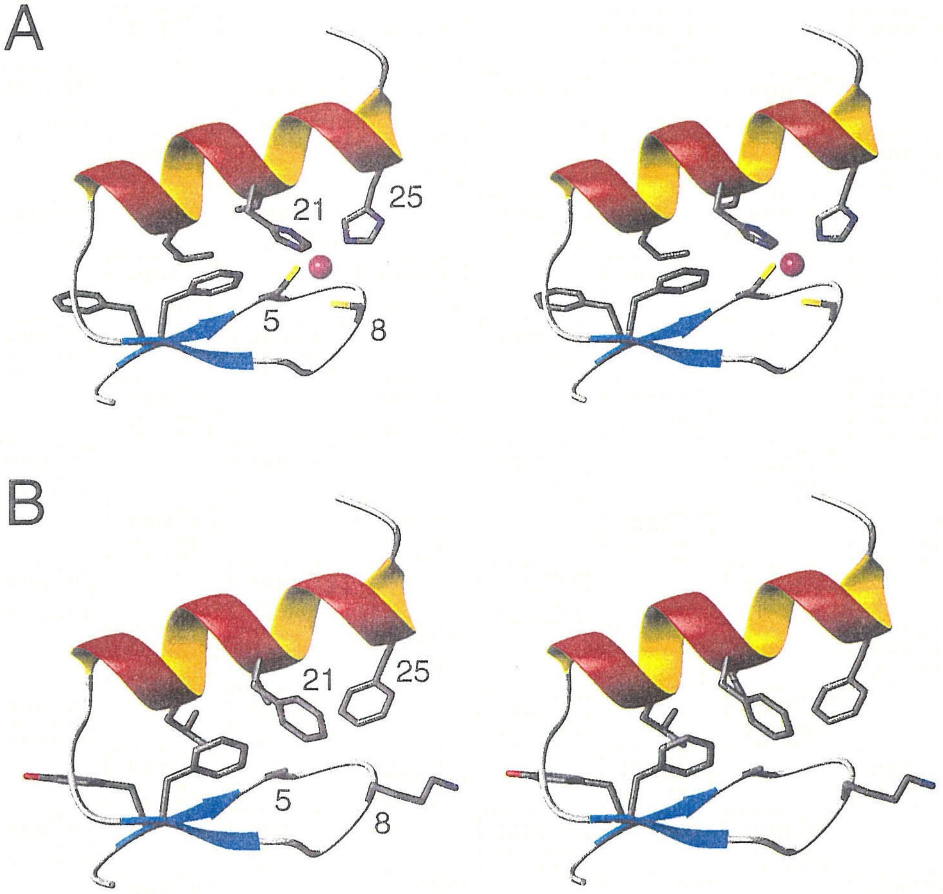

A
21
25
5
8
B
21
25
5
8

Figure 5-3. Circular dichroism (CD) measurements of FSD-1. (**A**) Far UV CD spectrum of FSD-1 at 1 °C. The minima at 220 nm and 207 nm indicate a folded structure. (**B**) Thermal unfolding of FSD-1 monitored by CD. The melting curve has an inflection point at 39 °C. To illustrate the cooperativity of the thermal transition, the melting curve was fit to a two state model (39) and the derivative of the fit is shown (inset). The melting temperature determined from this fit is 42 °C.

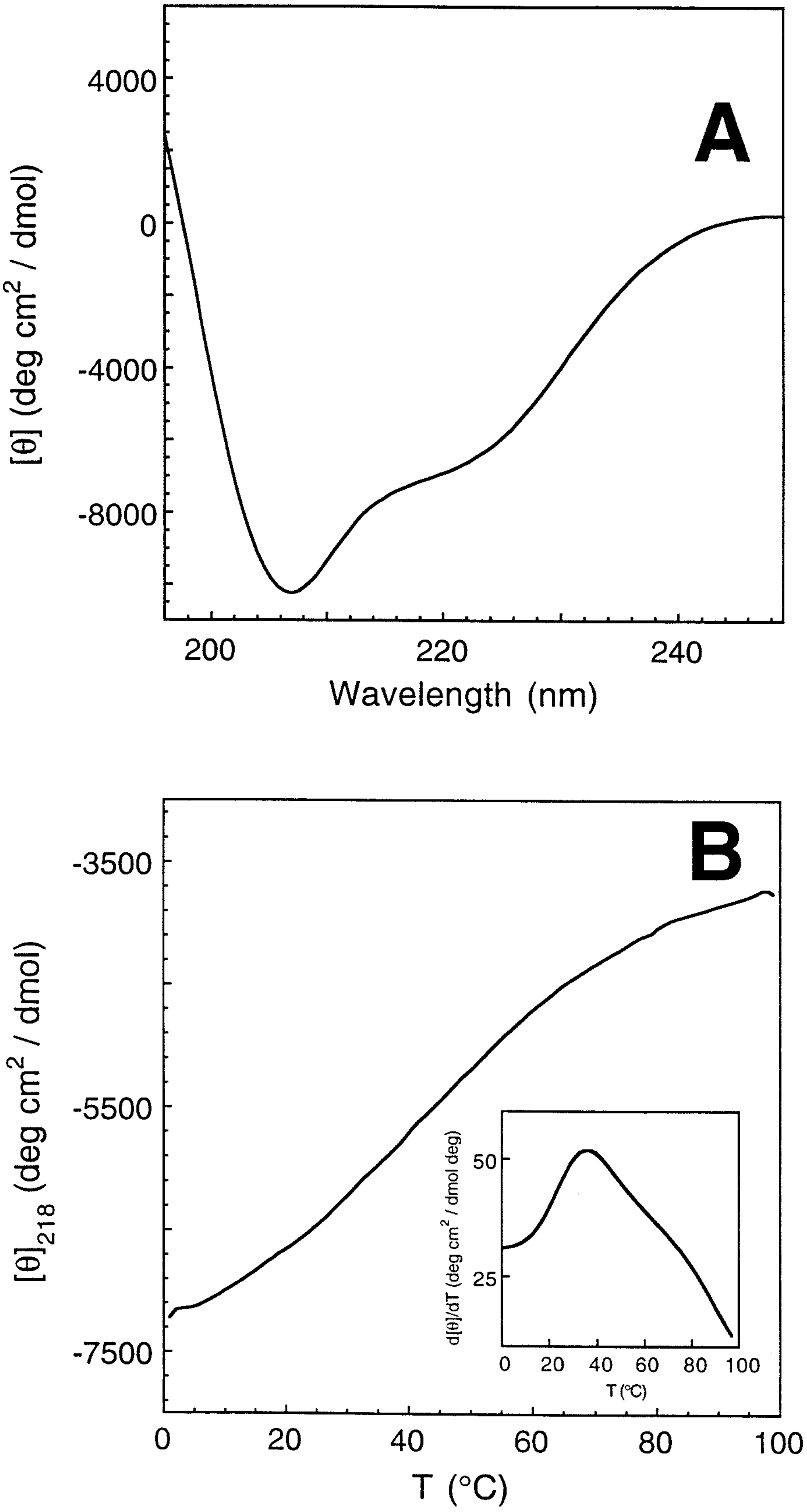

A
4000
0
-4000
-8000
[θ] (deg cm^2 / dmol)
200
220
240
Wavelength (nm)
B
-3500
-5500
-7500
$[\theta]_{218}$ (deg cm^2 / dmol)
0
20
40
60
80
100
T (°C)
50
25
d[θ]/dT (deg cm^2 / dmol deg)
0
20
40
60
80
100
T (°C)

Figure 5-4. Sequential and short-range NOE connectivities of FSD-1. All adjacent residues are connected by Hα-HN, HN-HN and/or Hβ-HN NOE crosspeaks. The helix (residues 15-26) is well defined by short-range connections, as is the hairpin turn at residues 7 and 8.

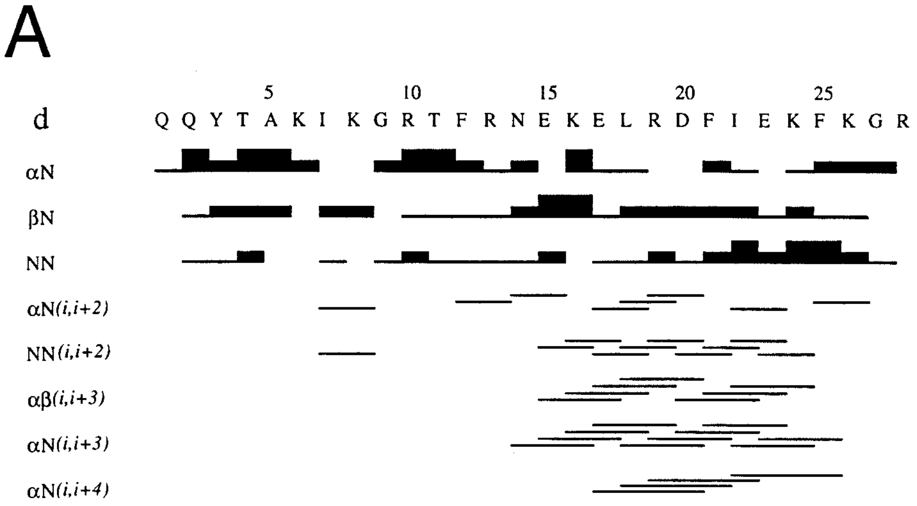

A
5
10
15
20
25
d
Q Q Y T A K I K G R T F R N E K E L R D F I E K F K G R
αN
βN
NN
αN(i,i+2)
NN(i,i+2)
αβ(i,i+3)
αN(i,i+3)
αN(i,i+4)

Figure 5-5. Solution structure of FSD-1. Stereoview showing the best fit superposition of the 41 converged simulated annealing structures from X-PLOR (31). The backbone Cα trace is shown in blue and the sidechain heavy atoms of the hydrophobic residues (Tyr 3, Ala 5, Ile 7, Phe 12, Leu 18, Phe 21, Ile 22, and Phe 25) are shown in magenta. The amino terminus is at the lower left of the figure and the carboxy terminus is at the upper right of the figure. The structure consists of two anti-parallel strands from positions 3-6 (back strand) and 9-12 (front strand), with a hairpin turn at residues 7 and 8, followed by a helix from positions 15-26. The termini, residues 1, 2, 27, and 28, have very few NOE restraints and are disordered.

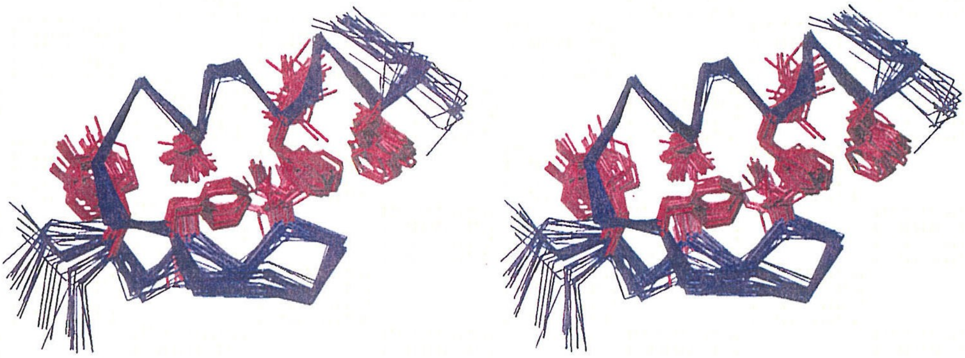
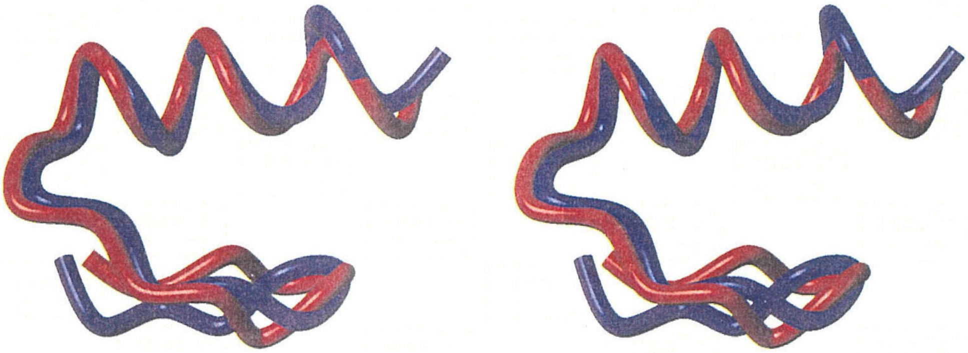


Figure 5-6. Comparison of the FSD-1 structure (blue) and the design target (red). Stereoview of the best fit superposition of the restrained energy minimized average NMR structure of FSD-1 and the backbone of Zif268. Residues 3 to 26 are shown.



Appendix A

Supplemental Experimental Techniques

Appendix A-1. Guanidinium Chloride Protein Denaturation

Denaturation measurements were performed on an Aviv 62A DS spectrometer. High concentration, > 1 mM, protein stock solutions were titrated to the desired pH. Approximately 8 M guanidinium chloride (GdmCl) solutions were titrated to the desired pH and filtered to 0.45 μm. Concentrated GdmCl solutions were not stored cold to prevent precipitation. Typical denaturation experiments required 18 samples to adequately define pre- and post-transition baselines and the transition region. Measurements were done in a 0.1 cm pathlength cuvette to minimize signal loss from GdmCl absorption.

1. Prepared a stock peptide solution in desired buffer, usually 50 mM sodium phosphate pH 5.0, at about 50 μM peptide concentration. Prepared enough peptide stock to make all samples.
2. Prepared a stock peptide solution in the concentrated GdmCl solution containing the same buffer as step 1. The peptide concentrations in the buffer and GdmCl stocks were exactly the same.
3. Mixed the buffer and GdmCl peptide stocks in the appropriate ratios to make samples at each GdmCl concentration. This procedure reduced error from peptide concentration in the denaturation measurement. Effects from non-ideal partial volumes of mixing are less than 1 percent and can be ignored.
4. Equilibrated samples at the desired temperature for at least 4 hours.

5. Measured sample ellipticity at the desired wavelength with at least 2 minutes of signal averaging and 30 seconds of equilibration prior to measurement.

6. GdmCl stock solution concentration was determined by measurement of refractive index in triplicate. Refractive index was converted to concentration following the method of Pace {Pace, 1986 #75}.

Appendix A-2. Protein Oligomerization State By Pulsed-Field Gradient Diffusion Coefficient Measurements

Diffusion coefficient (D) measurements were done using Water-sLED experiments according to the method of Altieri. Experiments were run at 25 °C in 99.9% D_2O with 50 mM sodium phosphate at pH 5.0. Diffusion coefficients could not be accurately measured at low temperature (7 °C) as confirmed by tests on molecular species of various sizes that all gave very similar diffusion coefficients. Axial gradient field strength was varied from 3.26 to 53.1 G/cm and a diffusion time of 50 ms was used. Spectra were processed with 2 Hz line broadening and integrals of the aromatic and high field aliphatic protons were calculated and fit to an equation relating resonance amplitude to the square of gradient strength in order to extract diffusion coefficients {Altieri, 1995 #99}. VNMR pulse sequence watersled was run without RAW water suppression. The diffusion coefficient for a zinc finger monomer control (28 residues) was 1.72×10^{-7} cm^2/s and for protein Gβ1 (56 residues) was 1.49×10^{-7} cm^2/s. Diffusion coefficients showed modest concentration dependence (about 10 percent) from 100 μM to 2 mM probably due to viscosity effects. Reproducibility of measurements was typically better than 10 percent.

Appendix A-3. Peptide Synthesis Variations For Long Sequences

Modifications were made to the standard synthesis protocol provided for the Applied Biosystems 433A, 0.25 ΩMonPrevPeak. The 40 ml reaction vessel was used and only 0.18 mmol of resin was added to increase the reagent excess favoring coupling. Both increased deprotection times and increased coupling times were used to improve the yield of difficult coupling cycles. Coupling cycles were increased to a minimum time of 30 minutes (60 coupling loops in module F) with an increase of 5 minutes every 10 cycles (add time of 10 in module F). Double coupling, with these long coupling times, was used for particularly difficult cycles. At least 4 deprotection loops were done for 4 minutes each. All extended cycles are in SynthAssist chemistry file LongFastMoc2 0.25 MonPrevPeak.

The extended reaction times tremendously increase cycle times and slow the synthesis. Observation of several synthesis suggests that the extended deprotection times did not have any impact on synthesis quality, and that extended coupling times and double coupling only assisted particularly difficult couplings. A strategy in the future might limit extended reaction times to just hard coupling cycles and use rapid coupling cycles for other steps.

Appendix B

NMR Data Analysis Scripts And Programs

Appendix B-1. Restraint Processing Macro Ovpeakfix.Tcl

```
#! /usr/local/bin/tclsh
#
# This routine has four functions:
# 1)
# Set all specified NOE distance restraints to weak class for a
# given X-PLOR restraints file.  The NOE's are specified in a
# file that contains the nuclei assignments.  The format of the
# file must be
# res# nuc res# nuc   with the smaller res# of a pair coming first.
# Weak distance restraints are 4.0 2.2 1.0 (1.8 - 5.0 Ang).
# The X-PLOR restraints file is argument 0.
# The input file is expected as argument 1.
#
# 2)
# Set all specified NOE distraints to specified class for
# the given X-PLOR restraints file.  Serves to correct for
# bad integral values from the splitting of peaks in the
# first 8 residues in the structure.
# The input file is expected as argument 2.
#
# 3)
# Add specified hbond restraints to the end of the restraints
# file.  Simply cats hbond.restraints to end.  Note, hbond.restraints
# also contains extra restraints from the split peaks that aren't
# labeled in the main NOE spectrum due to overlap but are
# expected to be there.  This was simpler than duplicating peaks
# in ANSIG.
#
# 4)
# Add 0.5 A to all restraints involving methyl protons
# Reads file methlist.dat for all methyl groups.
#
# Bassil Dahiyat
# 1 May 1997
#
# ovpeakfix.tcl
#
if { $argc > 3 } {
  puts stderr "ovpeakfix.tcl:  Too many arguments"
  exit
} elseif { $argc < 3} {
  puts stderr "ovpeakfix.tcl:  Too few arguments.  Specify restraints
file, ovpeaks file and split peak file."
  exit
}
set ovfile  [open [lindex $argv 1] r]
set spltfile      [open [lindex $argv 2] r]
set restrfile     [open [lindex $argv 0] r]
set newfile       [open tmprestraints w]
set methfile      [open methlist.dat r]
#
# Load ovdat arrays, ignoring comments
#
set i 1
gets $ovfile ovdat($i)
while {![eof $ovfile] } {
  if { [string index  [lindex $ovdat($i) 0] 0 ] != "!" && $ovdat($i) !=
"" } {
    incr i
  }
  gets $ovfile ovdat($i)

}
# note:  i is the number of ovpeaks plus 1!!
puts stderr ""
puts stderr "[expr $i - 1] restraints read"
puts stderr ""
puts stderr "Generating modified restraints file."
puts stderr ""
for {set j 1} {$j < $i} {incr j} {
  set ovindex1($j) [lindex $ovdat($j) 0]
  set ovindex2($j) [lindex $ovdat($j) 2]
  set ovnuc1($j)   [lindex $ovdat($j) 1]
  set ovnuc2($j)   [lindex $ovdat($j) 3]
}

#
# Scan restraints file and echo lines until a peak needs
# to be fixed.  Fix the peak, output the fixed line
# and continue.  The scan is independent of the order of the
# ovdat array which slows it down somewhat.
#
# note for format statement usage.  To get left justified
# fields use the - specifier as shown below.
#
gets $restrfile restrline
set ov 0
while {![eof $restrfile] } {
  regsub -all {\)} $restrline " & " restrline
  for {set j 1} {$j < $i} {incr j} {
    if { [lindex $restrline 2] == $ovindex1($j)} {
      if { [lindex $restrline 8] == $ovindex2($j)} {
         if { [lindex $restrline 5] == $ovnuc1($j)} {
           if {[lindex $restrline 11] == $ovnuc2($j)} {
            puts $newfile "[format "assi (resi%3d   and name %-4s)
(resi%3d   and name %-4s)  4.00  2.20  1.00" $ovindex1($j) $ovnuc1($j)
$ovindex2($j) $ovnuc2($j)]"
             if {[lindex $restrline 13] == "4.00" } {
                puts stderr "Already weak restraint $ovindex1($j)
$ovnuc1($j) to $ovindex2($j) $ovnuc2($j) "
             } else  {
             puts stderr "Adjusted restraint $ovindex1($j) $ovnuc1($j)
to $ovindex2($j) $ovnuc2($j) to weak class, 4.0 2.2 1.0."
            }
             set ov 1
           }
         }
      }
    }
  }
  if { !($ov) } {
    puts $newfile $restrline
  }
  set ov 0
  gets $restrfile restrline
}
close $restrfile
close $newfile
set restrfile     [open tmprestraints r]
set newfile [open tmprestraints2 w]
```

```
#
# Load spltdat arrays, ignoring comments
#
set i 1
gets $spltfile spltdat($i)
while {![eof $spltfile] } {
  if { [string index  [lindex $spltdat($i) 0] 0 ] != "!" && $spltdat($i)
!= "" } {
    incr i
  }
  gets $spltfile spltdat($i)

}
# note:  i is the number of spltpeaks plus 1!!
puts stderr ""
puts stderr "[expr $i - 1] restraints read"
puts stderr ""
puts stderr "Generating modified restraints file."
puts stderr ""
for {set j 1} {$j < $i} {incr j} {
  set spltindex1($j) [lindex $spltdat($j) 0]
  set spltindex2($j) [lindex $spltdat($j) 2]
  set spltnuc1($j)   [lindex $spltdat($j) 1]
  set spltnuc2($j)   [lindex $spltdat($j) 3]
  set spltdist($j)   [lindex $spltdat($j) 4]
  set spltmin($j)    [lindex $spltdat($j) 5]
  set spltplus($j)   [lindex $spltdat($j) 6]
}

gets $restrfile restrline
set splt 0
while {![eof $restrfile] } {
  for {set j 1} {$j < $i} {incr j} {
    if { [lindex $restrline 2] == $spltindex1($j)} {
      if { [lindex $restrline 8] == $spltindex2($j)} {
        if { [lindex $restrline 5] == $spltnuc1($j)} {
          if {[lindex $restrline 11] == $spltnuc2($j)} {
            puts $newfile "[format "assi (resi%3d   and name %-4s)
(resi%3d   and name %-4s)  %-4s  %-4s  %-4s" $spltindex1($j)
$spltnuc1($j) $spltindex2($j) $spltnuc2($j) $spltdist($j) $spltmin($j)
$spltplus($j)]"
            if {[lindex $restrline 13] == $spltdist($j) } {
               puts stderr "Already correct restraint $spltindex1($j)
$spltnuc1($j) to $spltindex2($j) $spltnuc2($j) "
            } else  {
            puts stderr "Adjusted restraint $spltindex1($j)
$spltnuc1($j) to $spltindex2($j) $spltnuc2($j) "
            }
            set splt 1
          }
        }
      }
    }
  }
  if { !($splt) } {
    puts $newfile $restrline
  }
  set splt 0
  gets $restrfile restrline
}
close $newfile
close $restrfile
#
# Reopen restraints modified restraints file for read and
# open new tmprestraints file for write
#
set restrfile     [open tmprestraints2 r]
set newfile [open tmprestraints4 w]
#
# Load methdat arrays, ignoring comments
#
set i 1
gets $methfile methdat($i)
while {![eof $methfile] } {
  if { [string index  [lindex $methdat($i) 0] 0 ] != "!" && $methdat($i)
!= "" } {
    incr i
  }
  gets $methfile methdat($i)
}
# note:  i is the number of methpeaks plus 1!!
puts stderr ""
puts stderr "[expr $i - 1] methyls read"
puts stderr ""
puts stderr "Generating modified restraints file."
puts stderr ""
for {set j 1} {$j < $i} {incr j} {
  set methindex($j) [lindex $methdat($j) 0]
  set methnuc($j)   [lindex $methdat($j) 1]
}
gets $restrfile restrline
set meth 0
while {![eof $restrfile] } {
  for {set j 1} {$j < $i} {incr j} {
    if { ([lindex $restrline 2] == $methindex($j) && [lindex $restrline
5] == $methnuc($j)) || ( [lindex $restrline 8] == $methindex($j) &&
[lindex $restrline 11] == $methnuc($j))} {
        set upper [expr [lindex $restrline end] + 0.5]
         puts stderr "Incremented upper bound 0.5 A, [lindex $restrline
2] [lindex $restrline 5] [lindex $restrline 8] [lindex $restrline 11]"
         puts $newfile "[lreplace $restrline end end $upper]"
        set meth 1
    }
  }
  if {!($meth)} {
    puts $newfile $restrline
  }
  set meth 0
  gets $restrfile restrline
}
close $newfile
close $restrfile

exec cat tmprestraints4 hbond.restraints > tmprestraints3
puts stderr " "
puts stderr "New restraints file name [lindex $argv 0].fix"
eval exec cp tmprestraints3 [lindex $argv 0].fix
exec rm tmprestraints tmprestraints2 tmprestraints3 tmprestraints4
```

Appendix B-2. Restraint Processing Macro Noecount.Tcl

```
#! /usr/local/bin/tclsh
#
# Count NOE restraints in an X-PLOR restraints file
# by category including Hbond restraints
# must use standard distance ranges for accurate counting!!!
#
set noefile [open [lindex $argv 0] r]


set totnoe 0
set strongnoe 0
set mednoe 0
set weaknoe 0
set hbond 0
set intra 0
set iplus1 0
set iplus2 0
set iplus3 0
set iplus4 0
set iplus5 0
set long 0
set bad 0
set hb 0
set hb0 0
set hb1 0
set hb2 0
set hb3 0
set hb4 0
set hb5 0
set hbl 0
gets $noefile noeline
while {![eof $noefile] } {

   if { [string index  [lindex $noeline 0] 0 ] != "!" && $noeline != ""
} {
      incr totnoe
      set ecount 0
```

```
     foreach elem $noeline {
       if {([regexp {[0-9]+} $elem ] || [regexp {[0-9]+.[0-9]+} $elem])
&& ![regexp {[A-Z]} $elem ]} {
        incr ecount
        set nums($ecount) $elem
       }
     }
     switch -regexp $nums(3) {
         4.0        {incr weaknoe}
         2.5       {incr strongnoe}
         3.0        {if {[regexp {0.5} $nums(4)]} {incr hbond;set hb 1}
else {incr mednoe}}
        2.0       {incr hbond; set hb 1}
         default {incr bad}
     }
     switch [expr abs($nums(2) - $nums(1))] {
         0          {incr intra;if {$hb} {incr hb0;set hb 0}}
         1   {incr iplus1;if {$hb} {incr hb1;set hb 0}}
         2          {incr iplus2;if {$hb} {incr hb2;set hb 0}}
         3          {incr iplus3;if {$hb} {incr hb3;set hb 0}}
         4          {incr iplus4;if {$hb} {incr hb4;set hb 0}}
         5          {incr iplus5;if {$hb} {incr hb5;set hb 0}}
         default {incr long;if {$hb} {incr hbl;set hb 0}}
     }
   }
  gets $noefile noeline
}
puts stdout ""
puts stdout "Restraints file name:  [lindex $argv 0]"
puts stdout "Total :      $totnoe"
puts stdout "|i-j| = 0 :  $intra\t\t$hb0"
puts stdout "|i-j| = 1 :  $iplus1\t\t$hb1"
puts stdout "|i-j| = 2 :  $iplus2\t\t$hb2"
puts stdout "|i-j| = 3 :  $iplus3\t\t$hb3"
puts stdout "|i-j| = 4 :  $iplus4\t\t$hb4"
puts stdout "|i-j| = 5 :  $iplus5\t\t$hb5"
puts stdout "|i-j| > 5 :  $long\t\t$hbl"
puts stdout ""
puts stdout "Restraint class count"
puts stdout "$strongnoe strong "
puts stdout "$mednoe medium "
puts stdout "$weaknoe weak"
puts stdout "$hbond hbond"
if {$bad > 0} { puts stdout "unknown?? $bad"}
puts stdout ""
```

Appendix B-3. X-PLOR Output Processing Macro Noeviol2.Tcl

```
#! /usr/local/bin/tclsh
#
# Scan X-PLOR output file from dgsa.inp, refine.inp
# and/or accept.inp looking
# for NOE restraint violation entries.  Grabs violations
# and outputs them to a file.
#
# Arguments:
# Requires name of xplor output file as first argument.
# Optional second argument: if set to "filename", names of
# pdb files that have each violation are echoed also.
# Outputs to standard out.
#
# Bassil Dahiyat
# 29 May 1997
```

```
#
# noeviol2.tcl
#
if { $argc > 2 } {
# remember, argc is zero if only argument is the tcl routine
  puts stderr "noeviol.tcl:  Too many arguments"
  exit
} elseif { $argc < 1 } {
  puts stderr "noeviol.tcl:  Too few arguments.  Specify input file and
# strucs"
  exit
} elseif {$argc == 2} {
  puts stderr " "
  puts stderr "noeviol2.tcl: Echoing pdb filenames with violations."
} else {
  puts stderr " "
  puts stderr "noeviol2.tcl: Not echoing pdb filenames with violations."
  puts stderr "noeviol2.tcl: To echo filenames add 'filename' as second
arg."
}
#
# open input and output files
#
set xplorfile [open [lindex $argv 0] r]
#
# Scan input file until NOE viol trigger reached
#

gets $xplorfile chkline
set noeviolcount 0
while { ![eof $xplorfile] } {
  if {[lindex $chkline 0] == "ASSFIL:"  } {
    set fname [lindex $chkline 2]
  }


  if { [lindex $chkline 0] == "set-i-atoms" } {
    incr noeviolcount
    if {[lindex $argv 1] == "filename"} {set fnam($noeviolcount) $fname}
    set noefrom($noeviolcount) {from}
    set noeto($noeviolcount) {to  }
    gets $xplorfile chkline
    while {![eof $xplorfile] && ([lindex $chkline 0] != "set-j-atoms")}
{
      set noefrom($noeviolcount) [lappend noefrom($noeviolcount)
$chkline]
      gets $xplorfile chkline
    }
  }
  if { [lindex $chkline 0] == "set-j-atoms" } {
    gets $xplorfile chkline

    while {![eof $xplorfile] && ([lindex $chkline 0] != "R<average>=")}
{


      set noeto($noeviolcount) [lappend noeto($noeviolcount) $chkline]
      gets $xplorfile chkline
    }
  }
  if { [lindex $chkline 0] == "R<average>=" } {
    set ravg($noeviolcount) $chkline
  }
  gets $xplorfile chkline
#
```

```
}
 puts stdout ""
 puts "Total violators in file: $noeviolcount"
 puts stdout ""
for {set i 1} { $i <= $noeviolcount } {incr i} {
    set violnum($i) 1
    set avgravg [lindex $ravg($i) 1]
    set avgravgd [lindex $ravg($i) 8]
    set avgravgE [lindex $ravg($i) 10]
    for {set j [expr $i + 1]} { $j <= $noeviolcount} {incr j} {
       if { ($noeto($j) == $noeto($i)) && ($noefrom($j) ==
$noefrom($i))} {
          incr violnum($i)
          set noeto($j) "x"
          set noefrom($j) "x"
           set avgravg [expr $avgravg + [lindex $ravg($j) 1] ]
           set avgravgd [expr $avgravgd + [lindex $ravg($j) 8] ]
           set avgravgE [expr $avgravgE + [lindex $ravg($j) 10] ]
           if {[lindex $argv 1] == "filename"} {set fnam($i) [lappend
fnam($i) $fnam($j)]}
       }
    }
    set avgravg  [expr $avgravg  / $violnum($i) ]
    set avgravgd [expr $avgravgd / $violnum($i) ]
    set avgravgE [expr $avgravgE / $violnum($i) ]
    if {  $noeto($i) != "x" } {
    if {$violnum($i) > 20 } {
      puts stdout "!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!"
      puts stdout "!!!!!!OCCURS $violnum($i) TIMES!!!!!!"
      puts stdout "!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!"
    } else {puts stdout "Occurs $violnum($i) times"}
    puts stdout $noefrom($i)
    puts stdout " $noeto($i)"
    puts stdout "Avg R<average> $avgravg [lrange $ravg($i) 2 6]
$avgravgd Avg E(NOE)= $avgravgE"
    puts stdout " "
      if {[lindex $argv 1] == "filename"} {
        set cnt 0
        foreach name $fnam($i) {
           puts -nonewline stdout "$name\t"
           incr cnt
           if { [expr $cnt % 4 ] == 0 } {puts stdout " "}
        }
      }
    puts stdout " "
    puts stdout " "
    }
}
```

Appendix B-4.   X-PLOR Output Processing Macro Sort_pdb.Tcl

```
#! /usr/local/bin/tclsh
#
# sort_pdb.tcl
#
# Bassil Dahiyat
# 12 May 1997
#
# gets the total energy, noe energy and number of violations
# from all the files in the directory called refine*.pdb.
# Expects X-PLOR format files generated by refine.inp or
# similar.  Outputs sorted (by noe energy) list of file
# names in a .nam file and a list of file names with
# the energies in a .E file.  Give name of output file.
#
# generate the egrep output to do sorting on
#
set type [lindex $argv 1]
eval exec egrep violation [glob ${type}*.pd[lindex $argv 2]] > tmplist
eval exec egrep energies: [glob ${type}*.pd[lindex $argv 2]] > tmplist2
set outfile [open tmpfinal w]
#
#
#
set noelist [open tmplist r]
gets $noelist line
while { ![eof $noelist]} {
  set filename [lindex $line 0]
  set noeviol [string trimright [lindex $line 2] {,}]
  set endname [expr [string first . $filename ] - 1]
  set filename  [string range $filename 0 $endname]
  set master1 [lappend master1 [list $noeviol $filename]]
  set master1index [lappend master1index $filename]
  gets $noelist line

}
close $noelist
set noelist [open tmplist2 r]
gets $noelist line
while { ![eof $noelist]} {
  set filename [lindex $line 0]
  set noeenergy [string trimright [lindex $line 7] {,}]
  set totalenergy [string trimright [lindex $line 2] {,}]
  set endname [expr [string first . $filename ] - 1]
  set filename  [string range $filename 0 $endname]
  set master2 [lappend master2 [list $noeenergy $filename $totalenergy]]
  set master2index [lappend master2index $noeenergy]
  gets $noelist line

}
close $noelist
foreach file $master2 {
  set index [lsearch -exact $master1index [ lindex $file 1]]
  set noeviol [lindex [lindex $master1 $index] 0]

  puts $outfile "[format "%9.3f %4d    %-12s %9.3f" [lindex $file 0]
$noeviol [lindex $file 1] [lindex $file 2] ]"
}
close $outfile
exec sort -k1 -n tmpfinal > [lindex $argv 0].E
exec sort -k4 -n tmpfinal > [lindex $argv 0].Etot
eval exec rm -fr tmpfinal tmplist tmplist2
set extractfile [open [lindex $argv 0].Etot r]
set namfile [open [lindex $argv 0].nam w]
gets $extractfile line
while {![eof $extractfile]} {
  puts $namfile "[lindex $line 2].pdb"
gets $extractfile line
}
close $namfile
close $extractfile
```

Appendix B-5. X-PLOR Output Processing Macro Chkrms.Tcl

```
#!/usr/local/bin/tclsh
exec cat [lindex $argv 0].nam | xargs egrep rms-d > junk.rms
```

```
set rmsfile [open junk.rms r]
set outfile [open [lindex $argv 0].rmsE w]
gets $rmsfile rmsline
puts $outfile "\t\tbond\t\tangle\t\timproper\tnoe"
set sumbond 0
set sumang 0
set sumimp 0
set sumnoe 0
set sumbondsq 0
set sumangsq 0
set sumimpsq 0
set sumnoesq 0
set count 0
while {![eof $rmsfile]} {
  incr count
  set rmslist [split [lindex $rmsline 2] {,}]
  set namend [string first ":" [lindex $rmsline 0] ]
  set filename [string range [lindex $rmsline 0] 0  $namend]
  puts -nonewline $outfile "$filename\t"
  set sumbond [expr $sumbond + [lindex $rmslist 0]]
  set sumang [expr $sumang + [lindex $rmslist 1]]
  set sumimp [expr $sumimp + [lindex $rmslist 2]]
  set sumnoe [expr $sumnoe + [lindex $rmslist 3]]
  set sumbondsq [expr $sumbondsq + [lindex $rmslist 0]*[lindex $rmslist
0]]
  set sumangsq [expr $sumangsq + [lindex $rmslist 1]*[lindex $rmslist
1]]
  set sumimpsq [expr $sumimpsq + [lindex $rmslist 2]*[lindex $rmslist
2]]
  set sumnoesq [expr $sumnoesq + [lindex $rmslist 3]*[lindex $rmslist
3]]
  puts -nonewline $outfile "[format "%-8.4g" [lindex $rmslist 0]]\t"
  puts -nonewline $outfile "[format "%-8.4g" [lindex $rmslist 1]]\t"
  puts -nonewline $outfile "[format "%-8.4g" [lindex $rmslist 2]]\t"
  puts            $outfile "[format "%-8.4g" [lindex $rmslist 3]]\t"
  gets $rmsfile rmsline
}
puts $outfile ""
puts -nonewline $outfile "Average rms-d\t"
puts -nonewline $outfile "[format "%-8.4g" [expr $sumbond / $count ]
]\t"
puts -nonewline $outfile "[format "%-8.4g" [expr $sumang / $count ] ]\t"
puts -nonewline $outfile "[format "%-8.4g" [expr $sumimp / $count ] ]\t"
puts            $outfile "[format "%-8.4g" [expr $sumnoe / $count ] ]\t"
set sumbondsq [expr sqrt([expr ($sumbondsq / $count) - ($sumbond /
$count)*($sumbond / $count)])]
set sumangsq [expr sqrt([expr ($sumangsq / $count) - ($sumang /
$count)*($sumang / $count)])]
set sumimpsq [expr sqrt([expr ($sumimpsq / $count) - ($sumimp /
$count)*($sumimp / $count)])]
set sumnoesq [expr sqrt([expr ($sumnoesq / $count) - ($sumnoe /
$count)*($sumnoe / $count)])]
puts -nonewline $outfile "Stddev rms-d\t"
puts -nonewline $outfile "[format "%-8.4g"  $sumbondsq]\t"
puts -nonewline $outfile "[format "%-8.4g"  $sumangsq]\t"
puts -nonewline $outfile "[format "%-8.4g"  $sumimpsq]\t"
puts            $outfile "[format "%-8.4g"  $sumnoesq]\t"
puts $outfile ""
close $rmsfile
exec rm -fr junk.rms

exec cat [lindex $argv 0].nam | xargs egrep energies > junk.E
set rmsfile [open junk.E r]
gets $rmsfile rmsline
```

```
puts $outfile "\t\ttot\tbond\tangle\timprop\tvdw\tnoe"
set sumtot 0
set sumbond 0
set sumang 0
set sumimp 0
set sumvdw 0
set sumnoe 0
set count 0
while {![eof $rmsfile]} {
  incr count
  set namend [string first ":" [lindex $rmsline 0] ]
  set filename [string range [lindex $rmsline 0] 0  $namend]
  puts -nonewline $outfile "$filename\t"
  set sumtot [expr $sumtot + [string trim [lindex $rmsline 2] ,] ]
  set sumbond [expr $sumbond + [string trim [lindex $rmsline 3] ,] ]
  set sumang [expr $sumang + [string trim [lindex $rmsline 4] ,]]
  set sumimp [expr $sumimp + [string trim [lindex $rmsline 5] ,]]
  set sumvdw [expr $sumvdw + [string trim [lindex $rmsline 6] ,]]
  set sumnoe [expr $sumnoe + [string trim [lindex $rmsline 7] ,]]
  puts -nonewline $outfile  "[string trim [lindex $rmsline 2] ,]\t"
  puts -nonewline $outfile  "[string trim [lindex $rmsline 3] ,]\t"
  puts -nonewline $outfile  "[string trim [lindex $rmsline 4] ,]\t"
  puts -nonewline $outfile  "[string trim [lindex $rmsline 5] ,]\t"
  puts -nonewline $outfile  "[string trim [lindex $rmsline 6] ,]\t"
  puts            $outfile  "[string trim [lindex $rmsline 7] ,]\t"
  gets $rmsfile rmsline
}
puts $outfile ""
puts -nonewline $outfile "Average E\t"
puts -nonewline $outfile "[format "%-7.2f" [expr $sumtot / $count ]] "
puts -nonewline $outfile "[format "%-7.2f" [expr $sumbond / $count ]] "
puts -nonewline $outfile "[format "%-7.2f" [expr $sumang / $count ]] "
puts -nonewline $outfile "[format "%-7.2f" [expr $sumimp / $count ]] "
puts -nonewline $outfile "[format "%-7.2f" [expr $sumvdw / $count ]] "
puts $outfile "[format "%-7.2f" [expr $sumnoe / $count ]]"
close $rmsfile
exec rm -fr junk.E
```

Appendix B-6. Operation Of Automatic Spectrum Assignment Program Asgnmr.Exe

Asgnmr.exe operates on an ANSIG crosspeaks export file and outputs an X-PLOR distance restraints file and a modified ANSIG crosspeaks export file. Input to the program consists of an input parameter file containing:

Crosspeak filename
Noesy spectrum name (must match ANSIG name exactly)
NOE distance limit (use decimal):"
PDB coordinate file name (nuclei names must match ANSIG names)"
NOE class minimum intensities, one per line
Minimum distance assignment only? (T or F)

and an ANSIG crosspeaks export file and a PDB coordinate file where the atom names have been converted to match the entries in the ANSIG residue dictionary. Asgnmr.exe extracts nuclei chemical shifts from all spectra in the crosspeaks file and averages them over all instances of that nucleus, and generates distance restraints for the crosspeaks from the spectrum designated by the second input parameter. The program does not consider symmetry in the spectrum. Crosspeak assignments are made with the following priority:

intraresidue, sequential, then all others that meet the distance cutoff. $<R^{-6}>$ distance averaging is done for all nuclei groups. The program will either output all possible assignments that meet the distance limit for a given crosspeak (last input argument F) or only use the assignment with the shortest distance (T). This switch only acts for non-intraresidue and non-sequential crosspeaks; if a crosspeak has any possible intraresidue or sequential assignments, it will take only that one in all cases. If multiple intraresidue or multiple sequential assignments exist for a peak, an error message is generated. The threshold for peak matching is 0.015 ppm. Output consists of numerous diagnostic messages sent to standard output, a file called xplor.noe containing the distance restraints in X-PLOR readable format, and a file called ansig.cpeak containing the new assignments for the noesy spectrum and all old assignments for the other spectra in ANSIG import format.

This automated approach to NOE assignments requires either a preliminary structure or the structure of a close structural homologue to work. The approach is fairly robust at generating structures with the correct global fold. In order to complete a high resolution structure, however, large numbers of inconsistent restraints are introduced that must be removed manually by inspection of the spectra. Given the relative simplicity of manually assigning NOE spectra of small proteins where spectral overlap is not a major problem, it is not clear that automatically assigning the spectra eases or speeds the process of structure solution. Typically, far fewer bad assignments are made during the manual procedure so refinement proceeds rapidly. Automatically assigned spectra require far more effort to deconvolute the inconsistent restraints, and often numerous peaks need to be manually reassigned. Perhaps improvements in the method that accounts for peak overlap, or that use the information about linewidth, peak shape, peak fine structure and peak position obvious to a human observer, will improve the technique in the future.