EXPLORATION OF THE DETERMINANTS OF PROTEIN STRUCTURE AND STABILITY BY PROTEIN DESIGN

Thesis by

Catherine Sarisky

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2005

(Defended April 27, 2005)

Acknowledgements

Many thanks to (in no particular order) Ben Gordon, Arthur Street, Bassil Dahiyat, Chantal Morgan, Alyce Su, Niles Pierce, Monica Breckow, Rhonda DiGiusto, Shira Jacobsen, J. J. Plecs, Shannon Marshall, Possu Huang, Premal Shah, Deepshikha Datta, Pavel Strop, Scott Ross, Julie Archer Mayo, and John Love for help, advice, and friendship during my time in the Mayo lab. You folks made the lab a great place to learn and do research. Thanks especially to Shannon Marshall for numerous interesting discussions, and to Chantal Morgan for teaching me how to take apart and reassemble nearly every instrument in the lab.

Thanks to my advisor, Steve Mayo, for providing a great work environment and the freedom to take my projects in interesting directions.

Scott Ross taught me everything I know about protein NMR. Thank you for all your help and for always being willing to "talk shop."

Thanks to Cynthia Carlson for assistance with paperwork (and nagging) above and beyond the call of duty. I couldn't have finished without you.

Thanks to Marie Ary for reading drafts of a number of the manuscripts included within this thesis.

Thanks to Rich Roberts for chairing my committee and for good advice, not all of which I took when it was first offered. Thanks also to the other members of my committee, Doug Rees and Peter Dervan, for their ongoing support and helpful feedback.

Thanks to Pamela Pape-Lindstrom and Rene Kratz for being great role models and for encouragement to finish.

Thanks to my family for believing in me and for always pushing just a little bit harder.

Thanks to my husband, Tim Johann, for agreeing to live through another Ph.D., despite having "done his time" by completing his the year I started mine, for preparing numerous gourmet dinners and countless bean burritos while I worked, and for support and encouragement.

Thanks to my new colleagues at Radford University for providing a desk at which to write. Thanks to my current class of R.U. students who have started calling me "Ms. (soon to be Dr.) Sarisky" in their emails. Yes, the final exam is cumulative. So is real life.

Abstract

Optimization of Rotamers by Iterative Techniques (ORBIT) has been used to calculate novel sequences for several small proteins. A partial sequence design (20 of 28 residues) is described for the zinc finger Zif268 (ββα) motif. The designed peptide folds without a metal cofactor, despite its small size and the avoidance of the disulfides and unnatural amino acids that are often used to stabilize peptide structures. The utility of ORBIT for predicting the relative stabilities of a series of ββα peptides was investigated. A good correlation between theoretical and experimental stabilities was observed except when the turn residues were changed. This observation led to the discovery that some of these peptides had an unexpected turn conformation. This information was used to design a peptide that is more stable than the original peptide sequence produced with ORBIT.

The tolerance of ORBIT for altered backbone coordinates was investigated using the protein domain Gβ1. It was determined that altering the coordinates of the backbone template used in ORBIT altered the sequences selected, but that the fold did not change as a result. The Gβ1 domain was also used to parameterize a methionine inclusion penalty, allowing the inclusion of methionine in ORBIT design calculations while preventing indiscriminate inclusion of methionine at sites where a less flexible side-chain will fit.

Lastly, some preliminary work on using ORBIT to design DNA binding interfaces is discussed.

Table of Contents

1. Introduction

Proteins are polymers composed of twenty different amino acid monomers. The order of these monomers is specified by the gene that codes for each protein, resulting in a unique sequence of amino acid residues for each protein. Proteins range in size from fewer than 20 amino acid residues in small peptides to tens of thousands of amino acid residues in the largest.

Protein chains fold to form common elements of local (secondary) structure such as alpha helices and beta sheets. Both structures include hydrogen bonds between backbone atoms as well as hydrophobic and hydrophilic contacts between sidechains. Secondary structure elements are connected to each other by turns and loops, which have more irregular conformations.

Elements of secondary structure come together to form more complex (tertiary) structures. Nearly all proteins include a solvent-inaccessible core containing primarily hydrophobic side-chains, with polar and charged side-chains on the solvent-exposed outer surface of the protein. Protein structure is further complicated by the formation of quaternary structure between protein chains. Homodimers are common, as are more complex multi-subunit assemblies.

Some protein folds are further stabilized by the formation of disulfide bonds between cysteine residues. Disulfide linkages reduce the conformational freedom of the protein, causing a relative stabilization of the folded state. Metal binding sites can also stabilize proteins. In the case of some small motifs like zinc fingers, the protein may fold only in

the presence of the metal cofactor. Post-translational modifications such as phosphorylation and glycosylation can also alter protein structures.

Many smaller proteins assume their folded state without external help. In most proteins studied, the primary structure (the amino acid sequence) is sufficient to specify the tertiary structure. Protein folding has become a field of great interest in the past decade. The various sequencing efforts have resulted in the availability of numerous sequences for putative proteins. However, having the primary sequence of the protein is often insufficient to deduce the structure or function of the protein, unless a closely related protein has already been studied. Although there has been some progress in predicting secondary structure from primary sequence and in threading new sequences onto the structures of closely related proteins, the goal of calculating the detailed three-dimensional structure based on only the protein sequence remains elusive. Two difficulties with protein folding are the lack of knowledge of the true energy function that governs protein stability and the impossibly large conformational space that must be searched for the optimum configuration.

As a way to further understanding of the protein folding problem, some researchers study inverse folding, or protein design. In inverse folding, the target three dimensional structure is specified and an amino acid sequence that will assume this three dimensional structure is calculated. Inverse folding exchanges the difficulties of an intractable search of conformational space for a difficult search of sequence space. Computational protein design offers the promise of *in silico* consideration of more protein sequences than can possibly be evaluated experimentally, despite great improvements in *in vitro* and *in vivo*

selection techniques.

The goal of protein design is to produce novel sequences for proteins. These novel protein sequences may assume the same fold as the wild-type sequence, but with improved thermodynamic stability, novel or improved ligand binding affinity, or new catalytic function. The present work will focus on modifications to protein stability, as a thorough understanding of stability is a precursor to the ability to successfully design for binding or catalysis.

ORBIT (Optimization of Rotamers By Iterative Techniques) is a software package developed by this research group. ORBIT can, in principle, be used to find the optimal protein sequence for a given fold. Development of ORBIT has focused on several areas: (1) Development of a suitable energy function. The energy function must be fast, decomposable to pair-wise interactions, and accurate. As the exact energy function for calculating the energy of a folded protein is not known, the energy functions used in ORBIT are approximations. The energy function includes physically valid terms, such as van der Waals interactions, and non-physical terms, such as propensity and negative design terms. Many terms, such as the methionine inclusion penalty discussed in Chapter 3, have some physical grounding but have been parameterized based on experimental data. (2) Optimization of the computational methods used to determine the optimal sequence based on the energy function. The improvements made by others to ORBIT over the course of this work greatly improved the speed with which results could be obtained for the later calculations.

ORBIT calculations require a fixed protein backbone, which is "decorated" with the side-chains that give an optimal score for the energy function. Side-chains are selected from a set of commonly observed side-chain orientations, called rotamers. The use of a limited set of rotamers reduces the computational difficulty of the calculation, at the possible cost of missing the optimal conformation if the necessary side-chain conformation is not present in the rotamer set.

Early calculations used backbone coordinates derived from protein structures deposited in the Protein Data Bank (PDB). Missing hydrogen atoms were added and the structures subjected to a brief energy minimization to remove any steric clashes. The resulting backbone coordinates provided the template for side-chain selection. Altered backbones, such as those discussed in Chapter 3, led to the generation of different sequences in ORBIT, although the structures assumed by the altered sequences more closely resembled the unaltered backbones than the altered backbones used to generate them. These results may indicate that positioning of secondary structure elements is specified by more than just the volume of the core sidechains.

ORBIT has been used to produce sequences that assume the target fold, to stabilize proteins, and to remove or introduce binding sites. Improvements are still needed to incorporate additional negative design issues, especially when designing turns, as will be discussed in Chapter 2. In Chapter 3, the impact of altering the wild-type backbone will be explored. In the second part of Chapter 3, use of the design cycle reveals that methionine can be included within the rotamer set allowed for protein cores, provided that a penalty term is incorporated in the energy function. In Chapter 4, some

preliminary work on extending ORBIT's capabilities to designing DNA binding proteins is discussed.

2. Studies with ββα folds.

Chapter Introduction

The difficulty of computational protein design increases exponentially with the number of positions to be designed. These computational restrictions make the use of small protein model systems desirable. Researchers designing small motifs by modeling have also favored small motifs[1] or motifs with regular repeats[2] to minimize complexity. The zinc finger ββα motif is in some ways ideal for design. It contains a small beta sheet (two strands and a turn) and an alpha helix, allowing the researcher to demonstrate that both secondary structure types can be designed. At 28 residues, it is small enough for chemical synthesis and sufficiently small for structure elucidation by ^{1}H NMR, without the need to label a sample.

In this laboratory, the Zif268 backbone[3] has been used as the template for the design of ββα motifs. Wild-type Zif268 contains a zinc ion binding site, with two cysteine residues on the β-hairpin and two histidine residues on the helix composing the binding site. Wild-type Zif268 does not fold in the absence of a divalent metal ion cofactor. The designed ββα motifs reported here and elsewhere[4] fold reversibly without the need for a cofactor, although the unfolding transitions are broad due to low cooperativity.

Small motifs also have their drawbacks. The minimal buried hydrophobic surface area results in low thermal stability and broad transitions during thermal denaturation. Although there are relatively few signals present in the NMR, dispersion is much less than that seen in slightly larger protein models like Gβ1, discussed in the next chapter.

Poor dispersion may result from increased motion in core residues or from the inability of such a small protein to provide much variation in magnetic environment for similar residues. It should be noted that FSD-1 does not exhibit ANS binding, indicating that it is not a molten globule. However, the poor NMR dispersion and low protection factors for amide backbone protons are consistent with the presence of more flexibility in this system than in larger proteins.

[1] Cobos E.S., Pisabarro M.T., Vega M.C., Lacroix E., Serrano L., Ruiz-Sanz J., Martinez J.C. (2004). A miniprotein scaffold used to assemble the polyproline II binding epitope recognized by SH3 domains. *J. Mol. Biol.,* 342(1), 355–65.

[2] Munson M., O'Brien R., Sturtevant J.M., Regan L. (1994), Redesigning the hydrophobic core of a four-helix-bundle protein. *Protein Sci.,* 3, 2015–2022.

[3] Pavletich N.P., Pabo C.O. (1991). Zinc finger DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* 252, 809–817.

[4] Dahiyat B.I. and S. L. Mayo (1997) De Novo Protein Design: Fully Automated Sequence Selection. *Science* 278, 82–87.

De novo protein design: towards fully automated sequence selection

Bassil I. Dahiyat, Catherine A. Sarisky and Stephen L. Mayo

Originally published in *J. Mol. Biol.* 273, 789–796, 1997.

Abstract

Several groups have applied and experimentally tested systematic, quantitative methods to protein design with the goal of developing general design algorithms. 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. Our goal is an objective, quantitative design algorithm that is based on the physical properties that determine protein structure and stability and that is not limited to specific folds or motifs. We chose the ββα motif typified by the zinc finger DNA binding module to test our design methodology. Using previously published sequence scoring functions developed with a combined experimental and computational approach and the Dead-End Elimination theorem to search for the optimal sequence, we designed 20 out of 28 positions in the test motif. The resulting sequence has less than 40% homology to any known sequence and does not contain any metal binding sites or cysteine residues. The resulting peptide, pda8d, is highly soluble and monomeric and circular dichroism measurements showed it to be folded with a weakly cooperative thermal unfolding transition. The NMR solution structure of pda8d was solved and shows that it is well-defined with a backbone ensemble rms deviation of 0.55 Å. Pda8d folds into the desired ββα motif with well-defined

elements of secondary structure and tertiary organization. Superposition of the pda8d backbone to the design target is excellent, with an atomic rms deviation of 1.04 Å.

Introduction

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. Several groups have applied and experimentally tested systematic, quantitative methods to protein design with the goal of developing general design algorithms (Hellinga *et al.* 1991, Hurley *et al.* 1992, Desjarlais and Handel 1995, Harbury *et al.* 1995, Klemba *et al*. 1995, Betz and Degrado 1996, Dahiyat and Mayo 1996). 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. A critical component of the development of these methods has been their experimental testing and validation. Our goal is an objective, quantitative design algorithm that is based on the physical properties that determine protein structure and stability and that is not limited to specific folds or motifs. This work reports the initial computational and experimental results of combining our core, surface, and boundary methodologies for the design of a small protein motif.

In selecting a motif to test the integration of our design methodologies, 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 sites. We chose the ββα motif typified by the zinc finger DNA binding module (Pavletich and Pabo 1991). Though it consists of less than 30 residues, this motif contains sheet, helix, and turn structures. Further, recent work by Imperiali and coworkers, who 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 takes this structure has demonstrated the ability of this fold to form in the absence of metal ions (Struthers *et al*. 1996a).

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 (Dahiyat and Mayo 1996). 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 (Ponder and Richards 1987). The immense search problem presented by rotamer sequence optimization is overcome by application of the Dead-End Elimination (DEE) theorem (Desmet *et al*. 1992, Goldstein 1994, De Maeyer *et al*. 1997). Our implementation of the DEE theorem extends its utility to sequence design and rapidly finds the globally optimal sequence in its optimal conformation.

In previous work we determined the different contributions of core, surface, and

boundary residues to the scoring of a sequence arrangement. The core of a coiled coil and of the streptococcal protein 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 non-polar surface area (Dahiyat and Mayo 1996 and Dahiyat and Mayo 1997b). 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 are designed using a hydrogen-bond potential and secondary structure propensities in addition to a van der Waals potential (Dahiyat and Mayo 1997a). Coiled coils designed with such a scoring function were 10 to 12°C more thermally stable than the naturally occurring analog. 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. We use these scoring functions without modification here in order to provide a rigorous test of the generality of our current algorithm.

Sequence design

The sequence selection algorithm requires structure coordinates that define the target motif's backbone. The Brookhaven Protein Data Bank (PDB) (Bernstein *et al*. 1977) was examined for high resolution structures of the ββα motif, and the second zinc finger module of the DNA binding protein Zif268 (PDB code 1zaa) was selected as our design

template (Pavletich and Pabo 1991). The backbone of the second module aligns very closely with the other two zinc fingers in Zif268 and with zinc fingers in other proteins and is therefore representative of this fold class. Twenty-eight residues were taken from the crystal structure starting at lysine 33 in the numbering of PDB entry 1zaa, which corresponds to our position 1. The first 12 residues comprise the β sheet with a tight turn at the sixth and seventh positions. Two residues connect the sheet to the helix, which extends through position 26 and is capped by the last two residues.

In order to assign the residue positions in the template structure into core, surface or boundary classes, the extent of side-chain burial in Zif268 and the direction of the C^{α}-C^{β} vectors were examined. The small size of this motif limits to one (position 5) the number of residues that can be assigned unambiguously to the core while six residues (positions 3, 12, 18, 21, 22, and 25) were classified as boundary. Three of these residues are from the sheet (positions 3, 5, and 12) and four are from the helix (positions 18, 21, 22, and 25). One of the zinc binding residues of Zif268 is in the core and two are in the boundary, but the fourth, position 8, has a C^{α}-C^{β} vector directed away from the protein's geometric center and is therefore classified as a surface position. The other surface positions considered by the design algorithm are 4, 9, and 11 from the sheet; 15, 16, 17, 19, 20, and 23 from the helix; and 14, 27, and 28, which cap the helix ends. The remaining exposed positions, which either were in turns, had irregular backbone dihedrals, or were partially buried, were not included in the sequence selection for this initial study. 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. The scoring functions used were identical to our previous work (Figure 1 legend).

In total, 20 out of 28 positions of the template were optimized during sequence selection. The algorithm first selects Gly for all positions with ϕ angles greater than 0° in order to minimize backbone strain (residues 9 and 27). The 18 remaining residues were split into two sets and optimized separately to speed the calculation. One set contained the one core, the six boundary positions and position 8, which resulted in 1.2×10^9 possible amino acid sequences corresponding to 4.3×10^{19} rotamer sequences. The other set contained the remaining ten surface residues, which had 10^{10} possible amino acid sequences and 4.1×10^{23} rotamer sequences. The two groups do not interact strongly with each other making their sequence optimizations mutually independent, though there are strong interactions within each group. Each optimization was carried out with the non-optimized positions in the template set to the crystallographic coordinates.

The optimal sequences found from the two calculations were combined and are shown in Figure 1 aligned with the sequence from the second zinc finger of Zif268. Even though all of the hydrophilic amino acids were considered at each of the boundary positions, only non-polar amino acids were selected. The calculated seven core and boundary positions form a well-packed buried cluster. The Phe side-chains selected by the algorithm at the zinc binding His positions, 21 and 25, are 80% buried and the Ala at 5 is 100% buried while the Lys at 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 Zif268 (Figure 2). The calculated optimal configuration buried ~830 $Å^2$ of non-polar surface area, with Phe12 (96% buried) and Leu18 (88% buried) anchoring the cluster. On the helix surface, the algorithm positions Asn14 as a helix N-cap with a hydrogen bond between its side-chain carbonyl oxygen and the backbone amide proton of residue 16. The six 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. Positions 4 and 11 on the exposed sheet surface were selected to be Thr, one of the best β-sheet forming residues (Kim and Berg 1993, Minor and Kim 1994, Smith *et al*. 1994).

Combining the 20 designed positions with the Zif268 amino acids at the remaining eight sites results in a peptide with overall 39% (11/28) homology to Zif268, which reduces to 15% (3/20) homology when only the designed positions are considered. A BLAST (Altschul *et al.* 1990) search of the non-redundant protein sequence database of the National Center for Biotechnology Information finds weak homology, less than 40%, to several zinc finger proteins and fragments of other unrelated proteins. None of the alignments had significance values less than 0.26. By objectively selecting 20 out of 28 residues on the Zif268 template, a peptide with little homology to known proteins and no zinc binding site was designed.

Experimental characterization

The far UV circular dichroism (CD) spectrum of the designed molecule, pda8d, shows a

maximum at 195 nm and minima at 218 nm and 208 nm, which is indicative of a folded structure (Figure 3a). The thermal melt is weakly cooperative, with an inflection point at 39°C, and is completely reversible (Figure 3b). 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 transitions observed for other short peptides (Weiss and Keutmann 1990, Scholtz *et al.* 1991, Struthers *et al*. 1996b).

Sedimentation equilibrium studies at 100 μM and both 7°C and 25°C give a molecular mass of 3490, in good agreement with the calculated mass of 3362, indicating the peptide is monomeric. At concentrations greater than 500 μM, however, the data do not fit well to an ideal single species model. When the data were fit to a monomer-dimer-tetramer model, dissociation constants of 0.5 to 1.5 mM for monomer-to-dimer and greater than 4 mM for dimer-to-tetramer were found, though the interaction was too weak to accurately measure these values. Diffusion coefficient measurements using the water-sLED pulse sequence (Altieri *et al.* 1995) agreed with the sedimentation results: at 100 μM pda8d has a diffusion coefficient close to that of a monomeric zinc finger control, while at 1.5 mM the diffusion coefficient is similar to that of protein Gβ1, a 56-residue protein. The CD spectrum of pda8d is concentration independent from 10 μM to 2.6 mM. NMR COSY spectra taken at 2.1 mM and 100 μM were almost identical with five of the H^{α}-HN cross-peaks shifted no more than 0.1 ppm and the rest of the cross-peaks remaining unchanged. These data indicate that pda8d undergoes a weak association at high concentration, but this association has essentially no effect on the peptide's structure.

The NMR chemical shifts of pda8d are well-dispersed, suggesting that the protein is folded and well-ordered. The H^{α}-HN fingerprint region of the TOCSY spectrum is well-resolved with no overlapping resonances (Figure 4a) and all of the H^{α} and HN resonances have been assigned. All unambiguous sequential and medium-range NOEs are shown in Figure 4b. H^{α}-HN and/or HN-HN NOEs were found for all pairs of residues except R6-I7 and K16-E17, both of which have degenerate HN chemical shifts, and P2-Y3 which have degenerate H^{α} chemical shifts. An NOE is present, however, from a P2 H^{δ} to the Y3 HN analogous to sequential HN-HN connections. Also, strong K1 H^{α} to P2 H^{δ} NOEs are present and allowed completion of the resonance assignments.

The structure of pda8d was determined using 354 NOE restraints (12.6 restraints per residue) that were non-redundant with covalent structure. An ensemble of 32 structures (Figure 4c) was obtained using X-PLOR (Brunger 1992) with standard protocols for hybrid distance geometry-simulated annealing. The structures in the ensemble had good covalent geometry and no NOE restraint violations greater than 0.3 Å. As shown in Table 1, the backbone was well-defined with a root-mean-square (rms) deviation from the mean of 0.55 Å when the disordered termini (residues 1, 2, 27, and 28) were excluded. The rms deviation for the backbone (3 to 26) plus the buried side-chains (residues 3, 5, 7, 12, 18, 21, 22, and 25) was 1.05 Å.

The NMR solution structure of pda8d shows that it folds into a ββα motif with well-defined secondary structure elements and tertiary organization, which matches the design target. A direct comparison of the design template, the backbone of the second zinc finger

of Zif268, to the pda8d solution structure highlights their similarity (Figure 4d). Alignment of the pda8d backbone to the design target is excellent, with an atomic rms deviation of 1.04 Å (Table 1). Pda8d and the design target correspond throughout their entire structures, including the turns connecting the secondary structure elements.

In conclusion, the experimental characterization of pda8d shows that it is folded and well-ordered with a weakly cooperative thermal transition, and that its structure is an excellent match to the design target. To our knowledge, pda8d is the shortest sequence of naturally occurring amino acids that folds to a unique structure without metal binding, oligomerization, or disulfide bond formation (McKnight *et al.* 1996). The successful design of pda8d supports the use of objective, quantitative sequence selection algorithms for protein design. Also, this work is an important step towards the goal of the successful automated design of a complete protein sequence. Though our algorithm requires a template backbone as input, recent work indicates that it is not sensitive to even fairly large perturbations in backbone geometry (Su and Mayo 1997). This robustness suggests that the algorithm can be used to design sequences for *de novo* backbones.

Acknowledgements

We thank Scott Ross for assistance with NMR studies, Pak Poon of the UCLA Molecular Biology Institute for sedimentation equilibrium studies, and Gary Hathaway of the Caltech Protein and Peptide Microanalytical Laboratory for mass spectra. We acknowledge financial support from 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

A.S. Altieri, D.P. Hinton and R.A. Byrd, Association of biomolecular systems *via* pulsed field gradient NMR self-diffusion measurements. *J. Am. Chem. Soc.* **117** (1995), pp. 7566–7567.

S.F. Altschul, W. Gish, W. Miller, E.W. Myers and D.J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215** (1990), pp. 403–410.

F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, Jr, M.D. Brice, J.R. Rodgers *et al.*, The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112** (1977), pp. 535–542.

S.F. Betz and W.F. Degrado, Controlling topology and native-like behavior of de novo-designed peptides: design and characterization of antiparallel 4-stranded coiled coils. *Biochemistry* **35** (1996), pp. 6955–6962.

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

M.L. Connolly, Solvent accessible surfaces of proteins and nucleic acids. *Science* **221** (1983), pp. 709–713.

B.I. Dahiyat and S.L. Mayo, Protein design automation, *Protein Sci.* **5** (1996), pp. 895–903.

B.I. Dahiyat and S.L. Mayo, Automated design of the surface positions of protein helices. *Protein Sci.* **6** (1997), pp. 1333–1337.

B.I. Dahiyat and S.L. Mayo, Probing the role of packing specificity in protein design. *Proc. Natl Acad. Sci. USA* (1997), 94, pp. 10172-10177.

M. De Maeyer, J. Desmet and I. Laster, All in one: a highly detailed rotamer library improves both accuracy and speed in the modeling of sidechains by dead-end elimination. *Folding Design* **2** (1997), pp. 53–66.

J.R. Desjarlais and T.M. Handel, De novo design of the hydrophobic cores of proteins. *Protein Sci.* **4** (1995), pp. 2006–2018.

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

R.L. Dunbrack and M. Karplus, Backbone-dependent rotamer library for proteins: an application to side-chain prediction. *J. Mol. Biol.* **230** (1993), pp. 543–574.

R.F. Goldstein, Efficient rotamer elimination applied to protein side-chains and related spin-glasses. *Biophys. J.* **66** (1994), pp. 1335–1340.

P.B. Harbury, B. Tidor and P.S. Kim, Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc. Natl Acad. Sci. USA* **92** (1995), pp. 8408–8412.

H.W. Hellinga, J.P. Caradonna and F.M. Richards, Construction of new ligand-binding

sites in proteins of known structure 2. Grafting of buried transition-metal binding site into *Escherichia coli* thioredoxin. *J. Mol. Biol.* **222** (1991), pp. 787–803.

J.H. Hurley, W.A. Baase and B.W. Matthews, Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J. Mol. Biol.* **224** (1992), pp. 1143–1154.

C.W.A. Kim and J.M. Berg, Thermodynamic β-sheet forming propensities measured using a zinc finger host peptide. *Nature* **362** (1993), pp. 267–270.

M. Klemba, K.H. Gardner, S. Marino, N.D. Clarke and L. Regan, Novel metal-binding proteins by design. *Nature Struct. Biol.* **2** (1995), pp. 368–373.

R. Koradi, M. Billeter and K. Wuthrich, Molmol: a program for the display and analysis of macromolecular structures. *J. Mol. Graph.* **14** (1996), pp. 51–55.

J. Kuszewski, M. Nilges and A.T. Brunger, Sampling and efficiency of metrix matrix distance geometry: a novel "partial" metrization algorithm. *J. Biomol. NMR* **2** (1992), pp. 33–56.

B. Lee and F.M. Richards, The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55** (1971), pp. 379–400.

S.L. Mayo, B.D. Olafson and W.A. Goddard III, Dreiding: a generic forcefield for molecular simulations. *J. Phys. Chem.* **94** (1990), pp. 8897–8909.

C.J. McKnight, D.S. Doering, P.T. Matsudaira and P.S. Kim, A thermostable 35-residue

subdomain within villin headpiece. *J. Mol. Biol.* **260** (1996), pp. 126–134.

D.L. Minor and P.S. Kim, Measurement of the β-sheet-forming propensities of amino acids. *Nature* **367** (1994), pp. 660–663.

V. Munoz and L. Serrano, Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. *Proteins: Struct. Funct. Genet.* **20** (1994), pp. 301–311.

M. Nilges, G.M. Clore and A.M. Gronenborn, Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Letters* **229** (1988), pp. 317–324.

M. Nilges, J. Kuszewski and A.T. Brünger, Sampling properties of simulated annealing and distance geometry. In: J.C Hoch, F.M Poulsen and C. Redfield, Editors, *Computational Aspects of the Study of Biological Macromolecules by NMR*, Plenum Press, New York (1991), pp. 451–457.

N.P. Pavletich and C.O. Pabo, Zinc finger DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* **252** (1991), pp. 809–817.

M. Piotto, V. Saudek and V. Sklenar, Gradient tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. *J. Biomol. NMR* **2** (1992), pp. 661–665.

J.W. Ponder and F.M. Richards, Tertiary templates for proteins-use of packing criteria in the enumeration of allowed sequences for different structural classes. *J.*

Mol. Biol. **193** (1987), pp. 775–791.

J.M. Scholtz, S. Marqusee, R.L. Baldwin, E.J. York, J.M. Stewart, M. Santoro *et al.*, Calorimetric determination of the enthalpy change for the alpha-helix to coil transition of an alanine peptide in water. *Proc. Natl Acad. Sci. USA* **88** (1991), pp. 2854–2858.

C.K. Smith, J.M. Withka and L. Regan, A thermodynamic scale for the β-sheet-forming tendencies of amino acids. *Biochemistry* **33** (1994), pp. 5510–5517.

M.D. Struthers, R.P. Cheng and B. Imperiali, Design of a monomeric 23-residue polypeptide with defined tertiary structure. *Science* **271** (1996), pp. 342–345.

M.D. Struthers, R.P. Cheng and B. Imperiali, Economy in protein design: evolution of a metal-independent ββα motif based on the zinc finger domains. *J. Am. Chem. Soc.* **118** (1996), pp. 3073–3081.

A. Su and S.L. Mayo, Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci.* **6** (1997), pp. 1701–1707.

M.A. Weiss and H.T. Keutmann, Alternating zinc finger motifs in the male-associated protein ZFY: defining architectural rules by mutagenesis and design of an aromatic swap second-site revertant. *Biochemistry* **29** (1990), pp. 9808–9813.

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

Table 1

NMR structure determination: distance restraints, structural statistics, atomic root-mean-square (rms) deviations, and comparison to the template. <SA> is the 32 simulated annealing structures, SA is the unminimized average structure, and SD is the standard deviation.

Distance restraints	
Intraresidue	148
Sequential	94
Short range (|i-j| = 2–5 residues)	78
Long range (|i-j| > 5 residues)	34
Total	354

Structural statistics	<SA> ± SD
Rms deviation from distance restraints (Å)	.049 ± .004
Rms deviation from idealized geometry (Å)	
Bonds (Å)	0.0051 ± 0.0004
Angles (degrees)	0.76 ± 0.04
Impropers (degrees)	0.56 ± 0.04

Atomic rms deviations (Å)*	<SA> vs. SA ± SD
Backbone	0.55 ± 0.03

Backbone + nonpolar side-chains	1.05 ± 0.06
Heavy atoms	1.25 ± 0.04

Rms deviations between the experimental structure to the template (Å)*

	SA vs. template
Backbone	1.04
Heavy atoms	2.15

*Atomic rms deviations are for residues 3 to 26, inclusive. The first two residues were highly disordered and had only sequential and intraresidue contacts. Residue 27 had one |i-j|=3 contact; residue 28 had one |i-j|=2 and one |i-j|=5 contact. <SA> is the 32 simulated annealing structures, SA is the average structure, and SD is the standard deviation. The design target is the backbone of Zif268.

Figures

Figure 1. Sequence of pda8d aligned with the second zinc finger of Zif268. The boxed positions were designed using the sequence selection algorithm. The coordinates of PDB record 1zaa (Pavletich and Pabo 1991) from residues 33 to 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 hydrogen atoms on the structure which was then conjugate gradient minimized for 50 steps using the Dreiding force field (Mayo et al 1990). As in our previous work (Dahiyat and Mayo 1997a), a backbone-dependent rotamer library was used (Dunbrack and Karplus 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. All His rotamers were protonated on both N^δ and N^ε. A Lennard-Jones 12-6 potential with van

der Waals radii scaled by 0.9 (Dahiyat and Mayo 1997b) was used for van der Waals interactions for all residues. An atomic solvation parameter of 23 cal/mol/Å^2 was used to favor hydrophobic burial and to penalize solvent exposure for core and boundary residues (Dahiyat and Mayo 1996, Dahiyat and Mayo 1997b). To calculate side-chain non-polar 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 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. The Richards definition of solvent-accessible surface area (Lee and Richards 1971) was used and areas were calculated with the Connolly algorithm (Connolly 1983). All residues with hydrogen bond donor or acceptors used a hydrogen bond potential based on the potential used in Dreiding but with more restrictive angle-dependent terms to limit the occurrence of unfavorable hydrogen bond geometries (Dahiyat and Mayo 1997a). A secondary structure propensity potential was used for surface β sheet positions (residues 4 and 11) (Dahiyat and Mayo 1997a). Propensity values from Serrano and coworkers were used (Munoz and Serrano 1994). Sequence optimization was performed with a modified version of DEE (Dahiyat and Mayo 1996). The set consisting of positions 3, 5, 8, 12, 18, 21, 22, and 25 contained 1.2×10^9 possible amino acid sequences and 4.3×10^{19} rotamer sequences. The set consisting of positions 4, 11, 14, 15, 16, 17, 19, 20, 23, and 28 contained 10^{10} possible amino acid sequences and 4.1×10^{23} rotamer sequences. The energy calculations and sequence optimizations took a total of 281 CPU minutes. All

calculations were performed on a Silicon Graphics Power Challenge server with ten R10000 processors running in parallel.

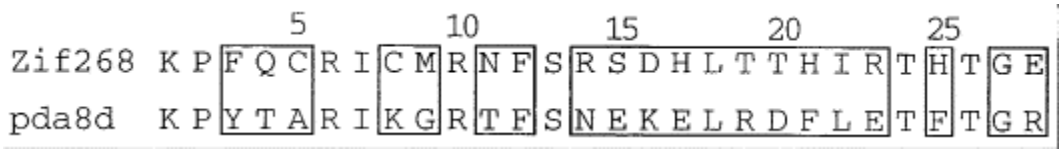

5
10
15
20
25
Zif268 K P F Q C R I C M R N F S R S D H L T T H I R T H T G E
pda8d K P Y T A R I K G R T F S N E K E L R D F L E T F T G R

Figure 2. Comparison of Zif268 and calculated pda8d structures. For clarity, only side-chains from residues 3, 5, 8, 12, 18, 21, 22, and 25 are shown. a, Stereo diagram of Zif268 showing its buried residues and zinc binding site. b, Stereo diagram of the calculated pda8d side-chain orientations showing the same residue positions as in a. Diagrams were made with MOLMOL (Koradi *et al*. 1996).

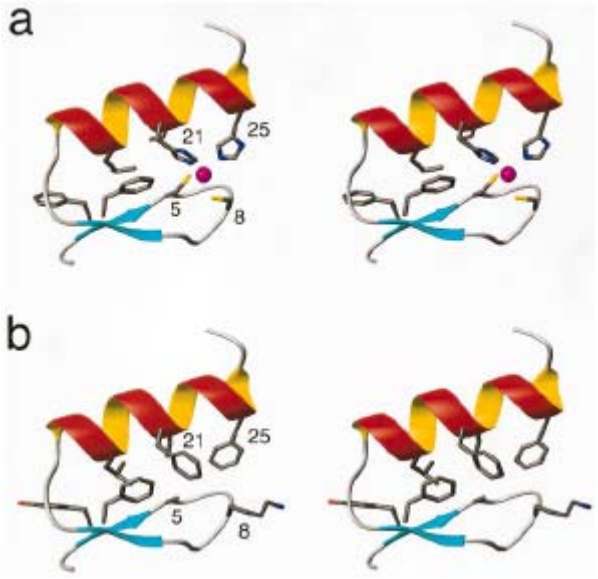

a
21
25
5
8
b
21
25
5
8

Figure 3. CD measurements of pda8d. a, Far UV CD spectrum of pda8d. Protein concentration was 43 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 three scans using a one second integration time and 1 nm increments. All CD data were acquired on an Aviv 62DS spectrometer equipped with a thermoelectric temperature control unit. b, Thermal unfolding of pda8d monitored by CD. Protein concentration was 115 μM in 50 mM sodium phosphate at pH 5.0. Unfolding was monitored at 218 nm in a 1 mm cuvette using 2 degree increments with an averaging time of 40 seconds and an equilibration time of 120 seconds 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. Pda8d 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×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 spectroscopy yielded a molecular weight of 3363 daltons (3362.8 calculated).

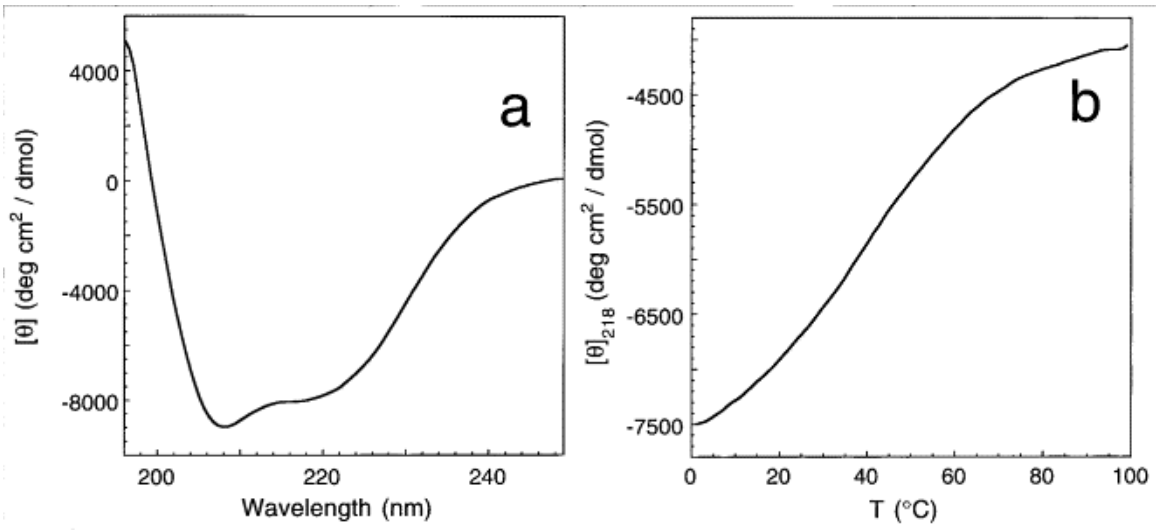
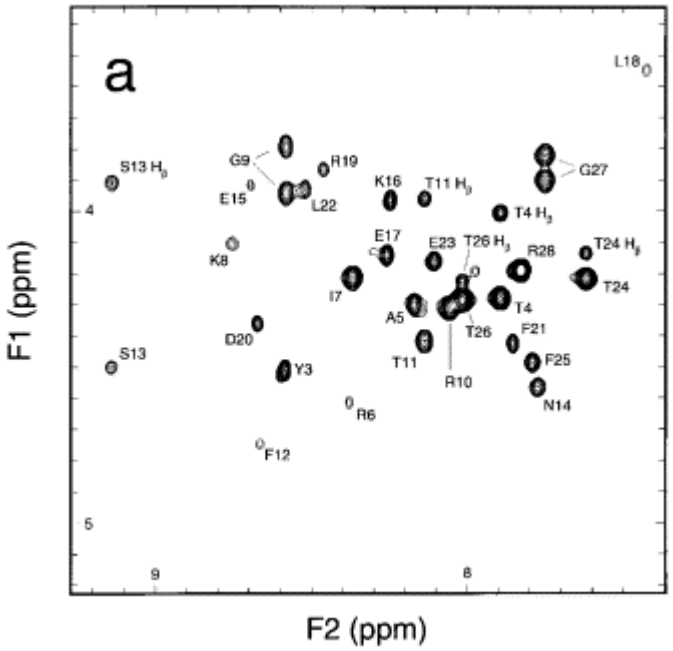
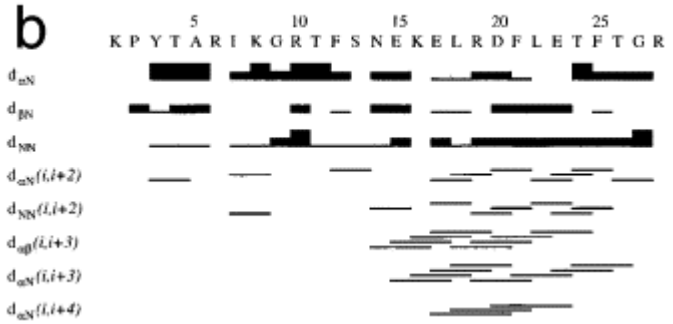

a
4000
0
-4000
-8000
[θ] (deg cm² / dmol)
200
220
240
Wavelength (nm)
b
-4500
-5500
-6500
-7500
[θ]218 (deg cm² / dmol)
0
20
40
60
80
100
T (°C)

Figure 4. NMR spectra and solution structure of pda8d. a, TOCSY H^{α}-HN fingerprint region of pda8d. 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 were prepared in 90/10 $H_2O/^2H_2O$ or 99.9% ^2H_2O with 50 mM sodium phosphate at pH 5.0. Sample pH was adjusted using a glass electrode with no correction for the effect of ^2H_2O on measured pH. All spectra for assignments were collected at 7°C. Sample concentration was approximately 2 mM. NMR assignments were based on standard homonuclear methods using DQF-COSY, NOESY, and TOCSY spectra (Wuthrich 1986). NOESY and TOCSY spectra were acquired with 2048 points in F_2 and 512 increments in F_1 and DQF-COSY spectra were acquired with 4096 points in F_2 and 1024 increments in F_1. All spectra were acquired with a spectral width of 7500 Hz and 32 transients. NOESY spectra were recorded with mixing times of 100 and 200 ms and TOCSY spectra were recorded with an isotropic mixing time of 80 ms. In TOCSY and DQF-COSY spectra water suppression was achieved by presaturation during a relaxation delay of 1.5 and 2.0 seconds, respectively. Water suppression in the NOESY spectra was accomplished with the WATERGATE pulse sequence (Piotto *et al*. 1992). Chemical shifts were referenced to the HO^2H resonance. Spectra were zero-filled in both F_2 and F_1 and apodized with a shifted Gaussian in F_2 and a cosine bell in F_1 (NOESY and TOCSY) or a 30° shifted sine bell in F_2 and a shifted Gaussian in F_1 (DQF-COSY). Water-sLED experiments (Altieri *et al.* 1995) were run at 25°C at 1.5 mM, 400 μM and 100 μM in 99.9% ^2H_2O with 50 mM sodium phosphate at pH 5.0. Axial gradient field strength was

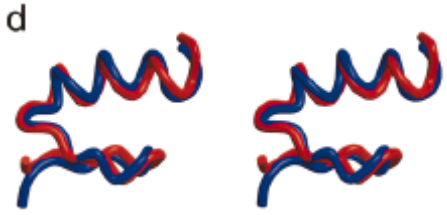
varied from 3.26 to 53.1 G/cm and a diffusion time of 50 ms was used. Spectra were processed with 2 Hz line broadening and integrals of the aromatic and high field aliphatic protons were calculated and fit to an equation relating resonance amplitude to gradient strength in order to extract diffusion coefficients (Altieri *et al.* 1995). Diffusion coefficients were 1.48×10^{-7}, 1.62×10^{-7}, and 1.73×10^{-7}, cm^2/s at 1.5 mM, 400 μM, and 100 μM, respectively. The diffusion coefficient for the zinc finger monomer control was 1.72×10^{-7} cm^2/s and for protein Gβ1 was 1.49×10^{-7} cm^2/s. b, NMR assignments summary and NOE connectivities of pda8d. Bars represent unambiguous connectivities and the bar thickness of the sequential connections is indexed to the intensity of the resonance. c, Solution structure of pda8d. Stereoview showing the best fit superposition of the 32 converged simulated annealing structures from X-PLOR (Brunger 1992). The backbone C^{α} trace is shown in blue. 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 antiparallel strands from positions 3 to 6 (back strand) and 9 to 12 (front strand), with a hairpin turn at residues 7 and 8, followed by a helix from positions 15 to 26. The termini, residues 1, 2, 27, and 28, have very few NOE restraints and are disordered. NOEs were classified into three distance-bound ranges based on cross-peak intensity: 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. Standard hybrid distance geometry-simulated annealing protocols were followed (Nilges *et al*. 1988, Nilges *et al*. 1991, Kuszewski *et al*. 1992). Ninty-eight distance geometry

structures were generated and, following regularization and refinement, resulted in an ensemble of 32 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°. Coordinates will be deposited with the Brookhaven Protein Data Bank and are available from the authors on request until processed and released. d, Comparison of pda8d solution structure and the design target. Stereoview of the best fit superposition of the average NMR structure of pda8d (blue) and the backbone of Zif268 (red). Residues 3 to 26 were used in the fit.


a
F1 (ppm)
F2 (ppm)
L18
S13 Hβ
G9
R19
E15
L22
K16
T11 Hβ
G27
T4 Hγ
E17
E23
T26 Hγ
R28
T24 Hβ
K8
I7
T24
A5
T4
T26
F21
D20
S13
T11
F25
Y3
R10
N14
R6
F12
4
5
9
8


b
5
10
15
20
25
K P Y T A R I K G R T F S N E K E L R D F L E T F T G R
dαN
dβN
dNN
dαN(i,i+2)
dNN(i,i+2)
dαβ(i,i+3)
dαN(i,i+3)
dαN(i,i+4)


c


d

Structure determination details

The structure of pda8d was determined using 354 NOE restraints (12.6 restraints per residue) that were non-redundant with covalent structure. An ensemble of 32 structures was obtained using X-PLOR with standard protocols for hybrid distance geometry-simulated annealing. The structures in the ensemble had good covalent geometry and no NOE restraint violations greater than 0.3 Å. As shown in Table 1, the backbone was well defined with a root-mean-square (rms) deviation from the mean of 0.55 Å when the disordered termini (residues 1, 2, 27, and 28) were excluded. The rms deviation for the backbone (3−26) plus the buried side-chains (residues 3, 5, 7, 12, 18, 21, 22, 25) was 1.05 Å.

The ββα fold: Explorations in sequence space

Catherine A. Sarisky and Stephen L. Mayo

Originally published in *J. Mol. Biol*. 307, 1411−1418, 2000.

Abstract

The computational redesign of the second zinc finger of Zif268 to produce a 28-residue peptide (FSD-1) that assumes a ββα fold without metal binding was recently reported.[1] In order to explore the tolerance of this metal-free fold towards sequence variability, six additional peptides resulting from the ORBIT computational protein design process were synthesized and characterized. The experimental stabilities of five of these peptides are strongly correlated with the energies calculated by ORBIT. However, when a peptide with a mutation in the β-turn is examined, the calculated stability does not accurately predict the experimentally determined stability. The NMR solution structure of a peptide incorporating this mutation (FSD-EY) reveals that the register between the β-strands is different from the model structure used to select and score the sequences. FSD-EY has a type I' turn instead of the target EbaaagbE turn (rubredoxin knuckle). Two additional peptides that have improved side-chain to backbone hydrogen bonding and turn propensity for the target turn were also characterized. Both are of comparable stability to FSD-1. These results demonstrate the robustness of the ORBIT protein design methods and underscore the need for continued improvements in negative design.

Introduction

The first complete computational design of a novel sequence for an entire protein fold was accomplished with the ORBIT (Optimization of Rotamers by Iterative Techniques) protein design algorithm.[1] The calculation was based on the backbone fold of the second zinc finger of Zif268 and resulted in the protein FSD-1. Unlike naturally occurring zinc fingers, which require metal binding for fold stability, the sequence of FSD-1 does not contain a metal binding site but contains a completely hydrophobic core. The stability of FSD-1 is modest, with a melting temperature (T_m) of about 40°C. FSD-1 has been shown by 2D NMR analysis to assume a ββα fold similar to the target backbone. ORBIT selects the optimal amino acid sequence for a target backbone by solving the combinatorial problem of placing amino acid side-chain rotamers on a fixed protein backbone in an arrangement that optimizes the system's total energy. A force field that includes terms for van der Waals, solvation, electrostatics, and hydrogen bonding is used to capture the essential energetic features thought to be responsible for the thermodynamic stability of proteins.[2] The rotamer-space combinatorial search problem is solved using the Dead-End Elimination theorem.[3,4,5,6]

Prior to sequence selection, residues are classified into three groups: core, boundary, and surface, based on solvent exposure.[1,7] Table 1 shows the residue classification for the second zinc finger of Zif268. Residues classified as core are restricted to A, V, L, I, F, Y, and W. The surface group is restricted to A, S, T, D, N, E, Q, H, K, and R. Residues in the boundary are selected from the combined core and surface lists. Because the current version of the ORBIT force field does not consider side-chain entropy loss upon folding,

methionine is not included in any of the calculations. In addition proline and cysteine are also excluded, and for the calculations presented here glycine is required at all positions with positive ϕ angles (positions 9 and 27). Rotamers were generated using the backbone dependent library of Dunbrack and Karplus,[8] as described previously.[9]

Naturally occurring zinc finger proteins show modest sequence variability; for example the three fingers of Zif268 share 45−65% identity.[10] Some of this sequence variability is related to the need to recognize different DNA target sites. Much of the variability, however, appears unrelated to function. Zinc fingers may be relatively insensitive to sequence changes due to the stability imparted by metal binding. Alignments of zinc finger sequences show that only the zinc binding residues are entirely conserved, while other positions can accommodate at least a few amino acids.[11]

In an attempt explore the sequence tolerance of the zinc-free zinc finger $\beta\beta\alpha$ fold and the robustness of the force field used to compute the FSD-1 sequence,[12] sub-optimal sequences were generated with the use of a Monte Carlo simulated annealing protocol that used the FSD-1 ground state sequence as the starting point of the simulation.[13] Several of the sequences resulting from this simulation were synthesized and their properties analyzed with the goal of assessing the relationship between the computed and experimental stabilities. Analysis of this type can potentially yield insight into the necessary and sufficient components of an effective force field for protein design.

A rank-ordered list by energy of the top 1000 sequences was maintained during the Monte Carlo simulation. The sequence list was subsequently sorted by the number of mutations from the FSD-1 sequence. The top sequence in each mutation category was

then selected for experimental analysis (Table 1). These top scoring sequences are named MC1 through MC6, which correspond to a single- through a six-fold mutation from the FSD-1 sequence, respectively. The first five variants, MC1 through MC5, contain changes only at surface positions, but the six-fold mutant, MC6, includes a change at boundary position 7 (I7Y) in addition to five surface mutations (Figure 2a).

Results

Comparison of the experimental stabilities of the variants is complicated by several factors. The low stabilities of all the peptides prohibit the use of chemical denaturation as a stability probe because the pretransition baselines cannot be accurately determined. The weakly cooperative thermal unfolding transitions prevent accurate determination of melting temperatures. Circular dichroism (CD) spectra, however, can be used to generate a precise although indirect measure of relative stability. In addition nuclear magnetic resonance (NMR) spectra can be used to gain qualitative insight into stability.

CD data analysis included two quantities: the position of the minimum in the far UV region (195 nm to 250 nm), λ_{min}, and the ratio between the intensity at the minimum and the intensity of the shoulder at 218 nm, θ_r (Figure 1). Previous work in this laboratory has shown for the $\beta\beta\alpha$ fold that a red-shifted minimum and high ratio between the shoulder and the minimum are consistent with enhanced stability.[14] An increase in random coil character (that is, a loss in stability) is expected to cause blue-shifting of the minimum, due to contributions from the negative random coil signal at 200 nm.[15]

Peptides MC1 through MC5 show a strong correlation between the energies calculated

using ORBIT and the experimental stability measurements, λ_{min} (Figure 1c) and θ_r (Figure 1d). The λ_{min} and θ_r values for FSD-1, MC1, and MC2 are very similar, as are their computed stabilities. MC3, MC4, and MC5 exhibit increasingly blue-shifted values of λ_{min} and progressively decreasing values of θ_r. MC6 appears to be unexpectedly well-folded by both CD measures despite its low predicted stability. Plots of λ_{min} and θ_r versus computed stability yield correlation coefficients R^2 of 0.96, when the MC6 data are excluded. The unexpected apparent stability of MC6 is also seen in a qualitative analysis of the peptides' 1D ^{1}H NMR spectra. MC6 exhibits better chemical shift dispersion than would be expected from its calculated energy (data not shown). Although some improvement in dispersion may result from the presence of an additional aromatic amino acid at position 7, the improvement is also seen in regions distant from the β-hairpin containing position 7.

Because much of the aberrant character of MC6 was thought to be related to the incorporation of Tyr at boundary position 7 and not to the surface mutations, the I7Y mutation was studied in isolation. A Q1E mutation was also introduced into the FSD-1 background in order to prevent N-terminal cyclization.[16] The resulting peptide, FSD-EY, is significantly more stable than any of the previous variants including FSD-1, as measured by λ_{min} and θ_r (Figure 1b).

In order to determine the source of the unexpected fold stabilization, the structure of FSD-EY was solved using standard ^{1}H homonuclear NMR methods. Structural statistics are shown in Table 2. FSD-EY assumes a ββα fold as expected (Figure 2b). The α-helix

is clearly defined from residues 15 to 24, and of the 34 structures in the structure ensemble, 26 contain a two-stranded β-sheet. The helix and the second β-strand are quite similar to the target backbone (Figure 2a). There are substantial differences, however, between the experimentally determined backbone and the target backbone used to select and score the sequences. The position of the turn has changed from residues 6 through 9 in the target fold to residues 8 and 9 in FSD-EY resulting in a β-sheet register shift of two residues (Figure 2d). The FSD-EY β-sheet is formed by hydrogen bonding between residues 5 and 12 and residues 7 and 10 compared to residues 3 and 12 and 5 and 10 for the target fold. Half of the members of the structure ensemble of FSD-EY contain a two-residue type I' turn with Lys 8 at the first position and Gly 9 at the second compared to the four-residue rubredoxin knuckle[17] found in the target fold.[10]

Discussion

The appearance of a type I' turn in the FSD-EY structure motivated an analysis of sequence preferences for both type I' turns and rubredoxin knuckles. Rubredoxin knuckles are described as EbaaagbE turns in the SLoop database.[18,19] The EbaaagbE turn is defined as a six-residue loop that connects two β-strands (the initial and final "E"). The conformations of the six loop residues are indicated by the letters "baaagb," where "b" corresponds to a β-space backbone conformation, "a" corresponds to an α-space backbone conformation, and "g" corresponds to glycine. For the numbering scheme used here, the "aaag" residues correspond to positions 6 through 9 (Figure 2d). This turn is often observed in proteins with metal binding via two cysteines (as in zinc fingers and

rubredoxin). The rubredoxin knuckle turn geometry buries three backbone amides (at positions 7, 8, and 10) that are stabilized by side-chain to backbone hydrogen bonds involving the side-chains of residues 5 and 8. In rubredoxin, for example, the buried amides form novel hydrogen bonds to the sulfur atoms in the cysteine side-chains.[17] In the ββα fold of interest here, hydrogen bonds are formed between the backbone amides of residues 7 and 8 and the side-chain of cysteine residue 5, and the backbone amide of residue 10 and side-chain of cysteine residue 8. In the absence of metal binding, the turn can be stabilized by more traditional side-chain to backbone hydrogen bonding. The SLoop database indicates that 31 of the 48 members of the EbaaagbE turn have putative hydrogen bond acceptors at the positions corresponding to both residues 5 and 8. An additional 14 turn members have an acceptor at one of these positions. Despite the observed preference for satisfying the hydrogen bonding potential of the buried amides in the EbaaagbE turn, the computed sequences (including FSD-1) fail to provide side-chains at positions 5 and 8 that can accept hydrogen bonds from the backbone amides of residues 7, 8, and 10.

The absence of a hydrogen bond acceptor at position 5 in the computed sequences is related to the definition of this position as a "core" position where only hydrophobic amino acids are allowed in the sequence selection calculations. In all cases, the computed amino acid identity at position 5 is Ala. Position 8, on the other hand, is classified as a surface position where several potential hydrogen bond acceptors are allowed. The failure of ORBIT to select an amino acid at position 8 capable of forming the indicated hydrogen bonds may be the result of a failure in the computational model

used to score sequence arrangements. The absence of hydrogen bond acceptors at positions 5 and 8 in the computed sequences ultimately requires the uncompensated desolvation of three backbone amides, which could significantly destabilize the target fold (and allow the population of alternative turn geometries).

Comparison of the propensity of the amino acids in the turn region for the two turn types indicates that the FSD-EY sequence has high propensity for the common type I' turn and low propensity for the unusual EbaaagbE turn. The amino acids selected by ORBIT for residues 5 through 10 are under-represented in the EbaaagbE turn. In the SLoop database of 48 examples of this turn, alanine is observed just once at position 5. Lysine is observed four times at position 6. Isoleucine (present in Zif268, FSD-1, and MC1 through MC5) occurs once and tyrosine (present in MC6 and FSD-EY) occurs twice at position 7. Lysine is never observed at position 8. ORBIT performs appreciably better at position 9, where glycine occurs 31 times in the database.

Compounding the problem of residues with poor propensity for the EbaaagbE turn, some of the amino acids have good propensities for the type I' turn. In their paper on β-turn potentials, Hutchinson and Thornton[20] report that tyrosine is significantly favored at the *i* position (position 7) in a type I' turn. Isoleucine is also regularly observed at this position, although it has no statistically significant preference. In the *i+2* position (position 9), glycine is highly favored, and the remaining residues have neutral propensities. Thus it is not surprising that the type I' turn is predominant in the FSD-EY mutant, as the amino acids present in the turn region have relatively good propensities for

the type I' turn and relatively poor propensities for the EbaaagbE turn.

In order to examine the possibility of stabilizing the zinc-free ββα fold by specifically stabilizing the EbaaagbE turn structure (and to provide data that could be useful in improving the ORBIT computational model) two additional variants were synthesized and analyzed. The first variant, FSD-ED, contains Asp at position 5. The second variant, FSD-EDS, contains Asp at position 5 plus a Ser at position 8. These substitutions were made to examine the role of hydrogen bond acceptors in the turn. In the SLoop database of sequences with the EbaaagbE turn, Asp and Ser appear 15 times at positions 5 and 8, respectively. Both FSD-ED and FSD-EDS show similar behavior to FSD-1, as measured by CD and 1D ^{1}H NMR (data not shown). It appears that the putative formation of hydrogen bonds between the side-chains and the buried backbone amides compensates for burial of polar groups within the core, but does not provide additional stability.

Conclusions

Five of six FSD-1 sequence variants (MC1 through MC5) show good agreement between CD-based measures of stability and the stabilities computed by the ORBIT design algorithm. The unexpected high stability of MC6 (and FSD-EY) appears to result from the incorporation of a tyrosine in the turn between the two β-strands of the zinc finger ββα fold and the subsequent switch in turn structure from the uncommon rubredoxin knuckle to the more common type I' turn. The ability of the FSD-EY sequence to achieve an alternative turn geometry underscores the need for both better force field descriptions of side-chain rotamers interacting with the target structure and

negative design approaches aimed at selecting sequence solutions that have poor stability on alternative structures. Target state force field optimization is particularly desirable for the type of polar interactions often seen between amino acid side-chains and protein backbones. Although comprehensive negative design approaches that consider all (or many) alternative structures are not currently available, the use of amino acid turn propensities, and potentials derived from them,[20] could allow the direct incorporation of negative design for turns by scoring turn sequences by their predicted difference in energy on the target turn versus their energies on known alternative turn geometries.

Acknowledgements

We thank Scott Ross for assistance with NMR data collection and helpful discussions regarding the FSD-EY solution structure. Shannon Marshall, Chantal Morgan, Bassil Dahiyat, and Alyce Su provided insights into peptide synthesis, purification, and characterization by CD. Supported by the Howard Hughes Medical Institute (S.L.M.) and the National Science Foundation (C.A.S.). Coordinates for FSD-EY have been deposited in the Protein Data Bank with accession code 1FME.

References

1. Dahiyat, B. I. & Mayo, S. L. (1997). De Novo Protein Design: Fully Automated Sequence Selection. *Science* **278**, 82-87.

2. Gordon, D. B., Marshall, S. A. & Mayo, S. L. (1999). Energy functions for protein design. *Current Opinion in Structural Biology* **9**, 509-513.

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

4. Goldstein, R. F. (1994). Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.* **66**, 1335-1340.

5. De Maeyer, M., Desmet, J. & Laster, I. (1997). All in one: a highly detailed rotamer library improves both accuracy and speed in modelling of sidechains by dead-end elimination. *Folding Design* **2**(1), 53-66.

6. Pierce, N. A., Spriet, J. A., Desmet, J. & Mayo, S. L. (2000). Conformational Splitting: A More Powerful Criterion for Dead-End Elimination. *J. Comp. Chem* **21**(11), 999-1009.

7. Street, A. G. & Mayo, S. L. (1999). Computational protein design. *Structure* **7**, R105-R109.

8. Dunbrack, R. L. & Karplus, M. (1993). Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J. Mol. Biol.* **230**, 543-574.

9. Dahiyat, B. I., Sarisky, C. A. & Mayo, S. L. (1997). *De Novo* Protein Design: Towards Fully Automated Sequence Selection. *Journal of Molecular Biology* **273**, 789-796.

10. Elrod-Erickson, M., Rould, M. A., Nekludova, L. & Pabo, C. O. (1996). Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA complexes. *Structure* **4**, 1171-1180.

11. Zang, W.-Q., Veldhoen, N. & Romaniuk, P. J. (1995). Effects of Zinc Finger Mutations on the Nucleic Acid Binding Activities of Xenopus Transcription Factor IIIA. *Biochemistry* **34**, 15545-15552.

12. Regan, L. (1998). Proteins to Order? *Structure* **6**, 1-4.

13. Dahiyat, B. I. & Mayo, S. L. (1996). Protein Design Automation. *Protein Science* **5**, 895-903.

14. Su, A. (1998). Backbone Flexibility in Protein Design: Theory and Experiment. Ph.D, California Institute of Technology.

15. Fasman, G. D., Ed. (1996). Circular dichroism and the conformational analysis of biomolecules. New York: Plenum Press.

16. Heinemann, F. S. & Ozols, J. (1998). Isolation and structural analysis of microsomal membrane proteins. *Frontiers in Bioscience* **3**, 483-493.

17. Blake, P. R. & Summers, M. F. (1994). Probing the unusually similar metal

coordination sites of retroviral zinc fingers and iron-sulfur proteins by nuclear magnetic resonance. *Adv. Biophys. Chem.* **4**, 1-30.

18. Donate, L. E., Rufino, S. D., Canard, L. H. J. & Blundell, T. L. (1996). Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: A database for modeling and prediction. *Protein Science* **5**, 2600-2616.

19. Burke, D. F., Deane, C. M. & Blundell, T. L. (1999). A browsable and searchable web interface to the database of structurally based classification of loops - SLoop. *Bioinformatics*.

20. Hutchinson, E. G. & Thornton, J. M. (1994). A revised set of potentials for β-turn formation in proteins. *Protein Science* **3**, 2207-2216.

21. Shakhnovich, E. I. & Gutin, A. M. (1993). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. USA* **90**, 7195-7199.

22. Shakhnovich, E. I. & Gutin, A. M. (1993). A new approach to the design of stable proteins. *Protein Engineering* **6**(8), 793-800.

23. Laskowski, R. A., Rullmann, J. A., MacAarthur, M. W., Kaptein, R. & Thornton, J. M. (1996). AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J. Biomol NMR* **8**, 477-486.

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

New York.

25. Koradi, R., Billeter, M. & Wuthrich, K. (1996). Molmol: a program for the display and analysis of macromolecular structures. *J. Mol. Graph* **14**, 51-55.

Tables

Table 1. List of peptides.

		Sequence Number																													
Peptide	Score	1		3		5		7		9		11		13		15		17		19		21		23		25		27	
FSD-1	-315.7	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
MC1	-315.6	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	R	|	|	|	|
MC2	-315.4	E	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	R	|	|	|	|
MC3	-315.0	E	|	|	|	|	|	|	|	K	|	|	|	|	|	|	|	|	|	|	|	|	|	R	|	|	|	|
MC4	-313.9	E	|	|	|	|	|	|	|	|	|	|	|	|	K	R	|	|	|	|	|	|	|	R	|	|	|	|
MC5	-313.4	E	|	|	|	|	|	|	|	|	K	|	|	|	|	K	R	|	|	|	|	|	|	|	R	|	|	|	|
MC6	-312.5	E	|	|	|	|	|	Y	E	|	K	|	|	|	|	K	R	|	|	|	|	|	|	|	R	|	|	|	|
FSD-EY	-314.2	E	|	|	|	|	|	Y	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
FSD-ED	-313.8	E	|	|	|	D	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
FSD-EDS	-310.5	E	|	|	|	D	|	|	S	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|	|
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
Class		s	s	b	s	c	s	b	s	s	s	s	b	s	s	s	s	s	b	s	s	b	b	s	s	b	s	s	s

The FSD-1 sequence was generated by ORBIT as previously described.[1] The MC sequences were generated using a Monte Carlo simulated annealing protocol, similar to that described previously;[13] 1000 annealing cycles with 10^6 steps per cycle were used. The high and low temperatures for the annealing cycles were 10,000 K and 100 K, respectively. The energies are calculated with the assumption that the unfolded energies

for all the sequences are the same. Although this is not generally true, in the case of a fixed binary pattern this assumption is consistent with the random energy model, which states that sequences with the same composition have isoenergetic unfolded states.[21,22] The selection of the remaining sequences is described in the main text. These energies were calculated for comparison, as these sequences do not appear on the Monte Carlo list. Residues that are identical to the FSD-1 sequence are indicated by a vertical bar (|) in the table. "Class" is the residue classification into core (c), boundary (b), and surface (s) groups. Peptides were synthesized with an Applied Biosystems 433A peptide synthesizer using FMOC chemistry. The peptides were cleaved from resin using TFA and purified by reverse phase HPLC on a C8 column with a water-acetonitrile gradient containing 0.1% TFA. Peptide masses were confirmed by matrix assisted laser desorption mass spectroscopy.

Table 2. Experimental restraints and structure statistics for FSD-EY

NOE distance restraints	
Intraresidue	122
Sequential	91
Medium range ($2 \leq \|i-j\| \leq 4$)	50
Long range ($\|i-j\| > 4$)	53
RMSDs from data	
Distance restraints (Å)	0.048 ± 0.002
RMSDs from ideal geometry	
Bonds (Å)	0.0035 ± 0.0002
Angles (°)	0.59 ± 0.04
Impropers (°)	0.42 ± 0.04
Ensemble atomic RMSDs (Å)	
Backbone (residues 3–26)	0.40
Heavy atoms	1.17
Ensemble Ramachandran statistics	
Residues in most favored regions (%)	63.5
Residues in additionally allowed regions (%)	34.4
Residues in generously allowed regions (%)	1.7
Residues in disallowed regions (%)	0.4

Ramachandran plot statistics were generated with PROCHECK-NMR.[23]

Figures

Figure 1. Experimental peptide data. a, CD wavelength scans of peptides FSD-1, MC3, MC5, and MC6. b, CD wavelength scans of FSD-1 and FSD-EY. All CD spectra were acquired on an Aviv 62DS spectrometer with thermoelectric temperature control at 1 °C in 50 mM sodium phosphate at pH 5.0. Peptide concentrations were 50 μM. c, Correlation between the ORBIT energy score and λ_{min}, and d, Correlation between the ORBIT energy score and θ_r, for FSD-1 and MC1 to MC5 (filled circles). MC6 (open circle) is omitted from the linear fit.

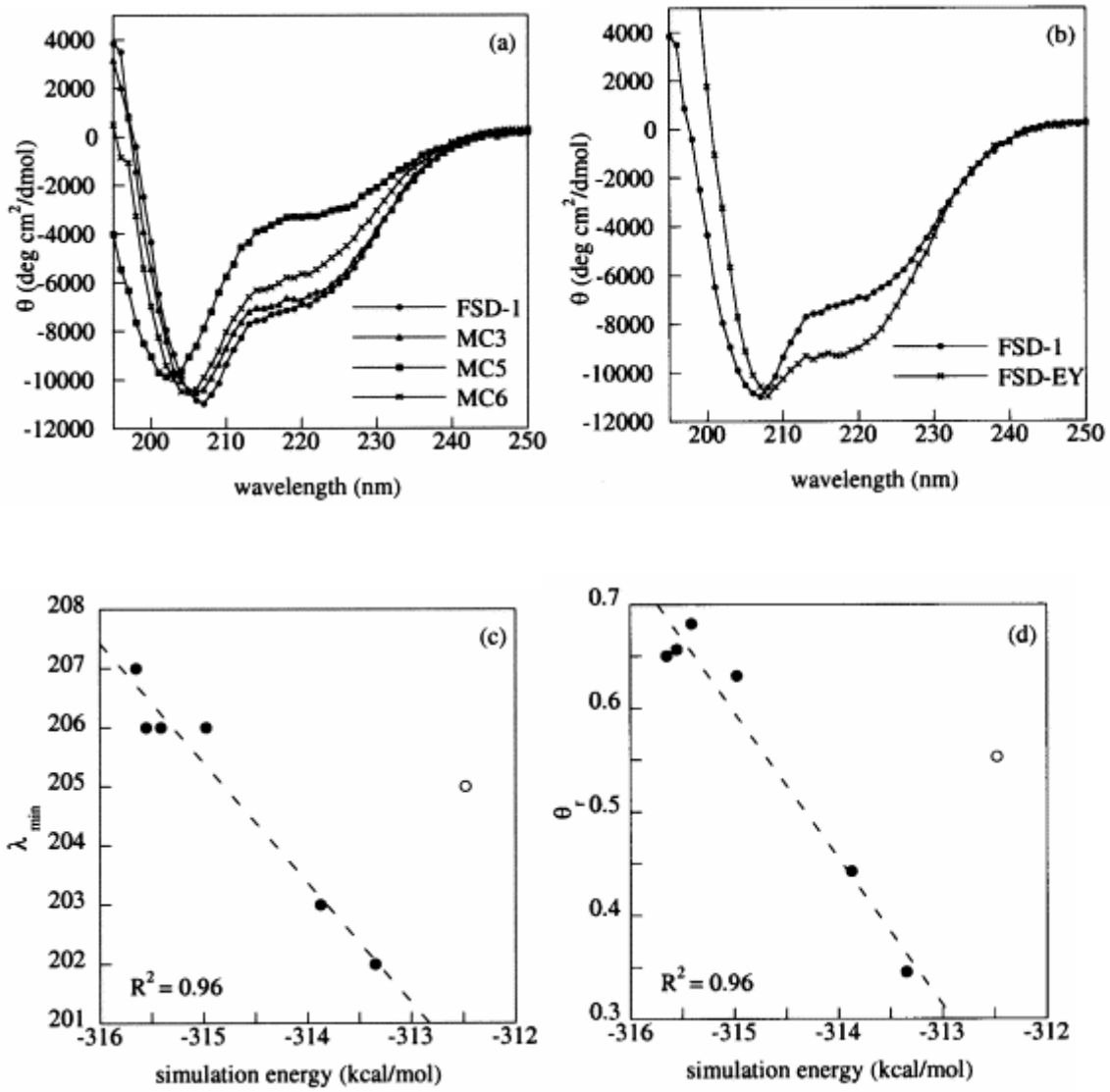
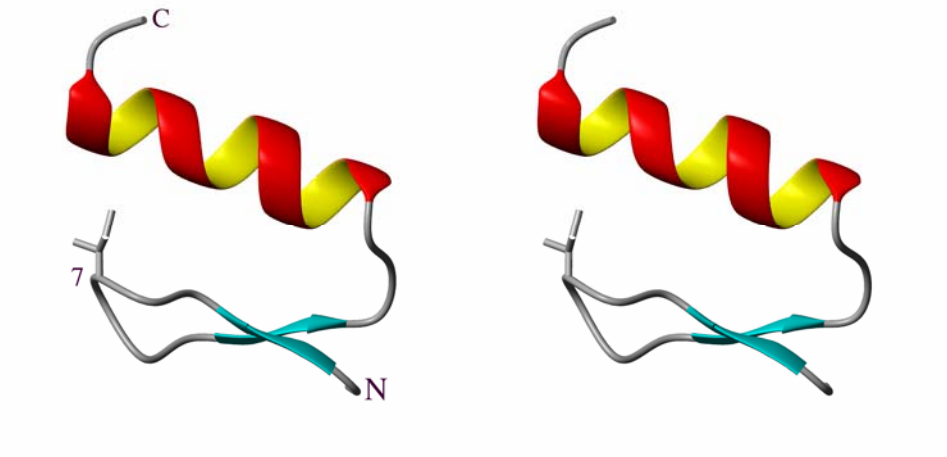
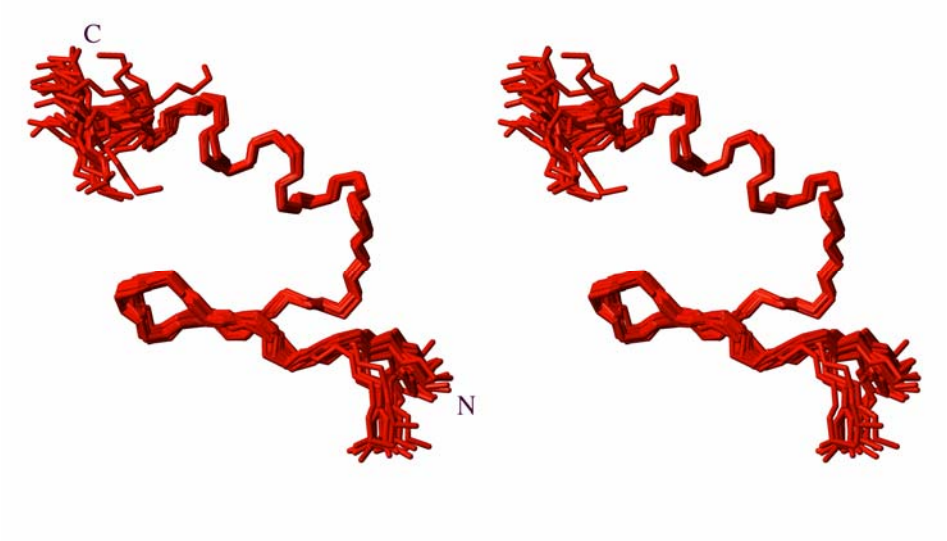
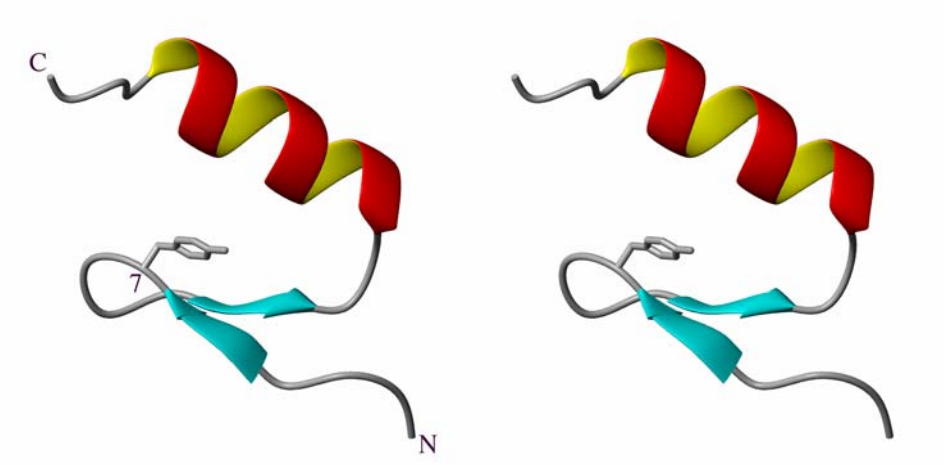

(a)
θ (deg cm²/dmol)
wavelength (nm)
FSD-1
MC3
MC5
MC6
(b)
θ (deg cm²/dmol)
wavelength (nm)
FSD-1
FSD-EY
(c)
λ min
R² = 0.96
simulation energy (kcal/mol)
(d)
θ r
R² = 0.96
simulation energy (kcal/mol)

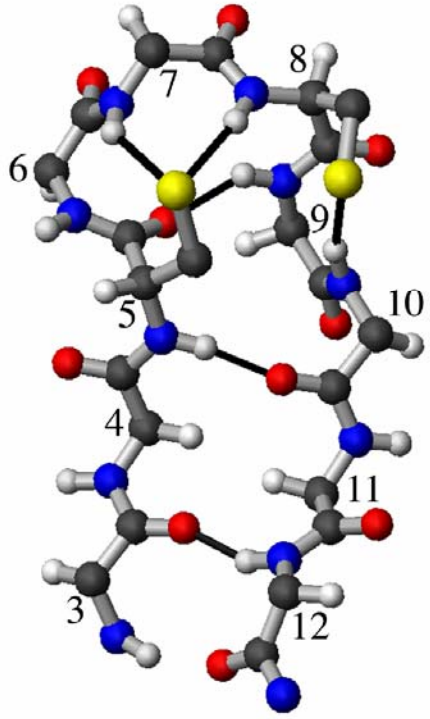
Figure 2. Model and solution structures. a, Stereoview of the second finger of Zif268 showing Ile 7, which is present in Zif268, FSD-1, and MC1 through MC5. b, Stereoview showing the 42 members of the ensemble of structures for FSD-EY. The RMSD between members of the ensemble is 0.40 Å when the backbone atoms of residues 3–26 are considered. NMR data were collected on a Varian UnityPLUS 600 MHz spectrometer. NMR samples were prepared in 50 mM sodium phosphate at pH* 5.0. Solvent was either 90% H_2O/10% D_2O or 99.9% D_2O. Peptide concentration was 2 mM. TOCSY, DQF-COSY, and WATERGATE NOESY spectra were collected using the 90% H_2O sample. An additional NOESY spectrum was collected using the D_2O sample. Assignments were made using standard techniques.[24] The structure ensemble was generated as previously described.[9] c, Ribbon diagram of the average FSD-EY structure, showing Tyr 7. d, Comparison of the model turn, type EbaaagbE (left), and the FSD-EY turn, type I' (right). Hydrogen bonds are indicated with black bars. Figures were created with MOLMOL.[25]


C
7
N

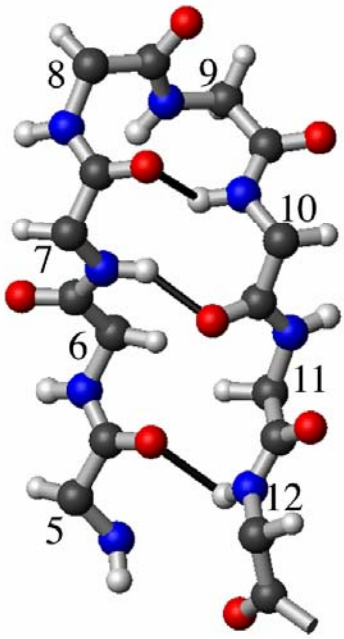

C
N



EbaaagbE turn



Type I' turn



Other projects with ββα folds

Dipole restrictions in the ββα helix

In the form used for generation of the FSD and MC sequences, the energy function used for ORBIT did not explicitly consider the macrodipole present in alpha helices. This macrodipole has been shown to impact sequence preferences for the first and last four residues in an alpha helix.[1] Morgan and Mayo[2] showed that sequences selected with ORBIT for engrailed homeodomain could be significantly improved by limiting polar N- and C-terminal residues to a smaller subset. We recalculated the optimal sequence for the Zif268 backbone, using Morgan and Mayo's results. The resulting peptide, nc2 (K16E/R21Q/D20K), exhibited reduced CD signal at 220 nm when compared to FSD-1. The T_m transition was comparable to FSD-1. An NMR data set was collected for this peptide. Based on reduced dispersion relative to FSD-1, the decision was made not to solve the structure.

Re-evaluation of FSD-1

With the results on FSD-EY clearly indicating a change in beta sheet register, we re-evaluated the NMR data used to determine the structure of FSD-1. There were two restraints that held the beta strands in roughly the Zif268 register. The first was an NOE crosspeak assigned to 3δ#-12HN (a putative interaction between the degenerate δ protons on Tyr-3 and the amide proton of Phe 12). The assignment of this cross-peak was complicated by several other NH residues at nearly the same chemical shift, and the

presence of some cyclization and/or deamidation at Gln-1, resulting in two sets of peaks for nearby residues. Further consideration of the assignment in this region and comparison with the spectra of MC2 reveals that this cross-peak is more appropriately assigned to 3δ#-4HN'. Not only is the alignment better, but this is the appropriate assignment to make in the case of ambiguity, as assignments close in primary structure should be chosen over assignments distant in primary structure in the case of ambiguity.

The second restraint that held FSD-1 in the Zif268 register was a pair of 3-12 and 12-3 hydrogen bond restraints, based on deuterium exchange. The overlap in this region is poor, making determination of protection difficult. Further, addition of hydrogen bond restraints is generally considered acceptable only in the presence of a set of cross-strand NOEs, which are mostly missing or ambiguous in FSD-1. Lacking these NOEs, the evidence for a possible 3-12/12-3 pair of hydrogen bonds is ambiguous at best.

The FSD-1 sequence does assume a ββα fold. However, there is no clear evidence allowing discrimination between an EbaaagbE and a type I' turn.

[1] Huyghues-Despointes, B., Scholtz, J., Baldwin, R. (1993). Effect of a single aspartate on helix stability at different positions in a neutral alanine-based peptide. *Protein Science* 2, 1604–1611.

[2] Marshall S.A., Morgan C.S. and Mayo S.L. (2002). Electrostatics significantly affect the stability of designed homeodomain variants. *J. Mol. Biol.* 316, 189–199.

3. Studies with Gβ1.

Chapter Introduction

The β1 domain of streptococcal protein G (Gβ1 or GB1) has been frequently used to study protein structure, thermodynamics, and folding. This small 56-residue protein is readily over-expressed in *E. coli* (as described below) or chemically synthesized.[1] Although small, the protein assumes a compact globular fold, with both an α helix and a β sheet included in its structure. This fold has been used for protein folding studies,[2,3] as a host for propensity studies,[4] and as a scaffold for decoration by researchers working in protein design.[1,5,6] From these studies, we know that even suboptimal sequences can assume the correct fold.

This chapter contains two studies in which Gβ1 is used for protein design. In the first study, I solved the NMR structure of Δ0, a Gβ1 core variant with a backbone based on the wild-type backbone. This structure was compared to the NMR structure of a Gβ1 core variant, Δ1.5, generated by core sequence selection on a backbone that was raised by 1.5Å above the sheet. Although the volume of the core residues was significantly higher for Δ1.5 compared to Δ0, the two sequences assume nearly identical folds, without the increase in the helix to sheet distance present in the template.

In the second study, core mutations in Gβ1 were used to study the impact of the inclusion of methionine in protein design calculations. Prior to this study, ORBIT design calculations were generally performed with a rotamer set that did not include methionine,

as early attempts that included methionine tended to produce sequences with numerous methionine side-chains in the core. In this study, a Monte Carlo search was used to identify sequences with energy scores similar to Δ0, the most stable known core sequence. Sixteen sequences with favorable energy scores and zero to two methionines included in the core were over-expressed and characterized by circular dichroism. This data set was used to calibrate a penalty for inclusion of methionine in designed protein cores. The methionine inclusion penalty resulting from this study has since been used in ORBIT to improve computational designs of lysozyme.[7]

[1] Dahiyat, B.I. and S.L. Mayo (1997). Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci.,* 94, 10172–10177.

[2] Ding K, Louis J.M., Gronenborn A.M. (2004). Insights into conformation and dynamics of protein GB1 during folding and unfolding by NMR. *J. Mol. Biol.*, 335(5), 1299–307.

[3] Nauli S., Kuhlman B., Le Trong I., Stenkamp R.E., Teller D., Baker D. (2002). Crystal structures and increased stabilization of the protein G variants with switched folding pathways NuG1 and NuG2. *Protein Sci.*, 11(12), 2924–31.

[4] Minor, D.L. Jr, Kim P.S. (1994). Context is a major determinant of beta-sheet propensity. *Nature*, 371, 264–7.

[5] Farinas E., Regan L. (1998) The de novo design of a rubredoxin-like Fe site. *Protein Sci.*, 7(9), 1939–46.

[6] Malakauskas S.M. and Mayo S.L. (1998) Design, Structure, and Stability of a Hyperthermophilic Protein Variant. *Nature Struct. Biol .*, 5, 470.

[7] Mooers B.H., Datta D., Baase W.A., Zollars E.S., Mayo S.L., Matthews B.W. (2003). Repacking the Core of T4 lysozyme by automated design. *J. Mol Biol.* 332(3), 741–56.

Designed protein G core variants fold to native-like structures: Sequence selection by ORBIT tolerates variation in backbone specification

Scott A. Ross, Catherine A. Sarisky, Alyce Su and Stephen L. Mayo

Originally published in *Protein Science* 10(2), 450–4, 2000.

Abstract

The solution structures of two computationally designed core variants of the ß1 domain of streptococcal protein G (Gß1) were solved by ^{1}H NMR methods to assess the robustness of amino acid sequence selection by the ORBIT protein design package under changes in protein backbone specification. One variant has mutations at three of 10 core positions and corresponds to minimal perturbations of the native Gß1 backbone. The other, with mutations at six of 10 positions, was calculated for a backbone in which the separation between Gß1's α-helix and β-sheet was increased by 15% relative to native Gß1. Exchange broadening of some resonances and the complete absence of others in spectra of the sixfold mutant bespeak conformational heterogeneity in this protein. The NMR data were sufficiently abundant, however, to generate structures of similar, moderately high quality for both variants. Both proteins adopt backbone structures similar to their target folds. Moreover, the sequence selection algorithm successfully predicted all core χ_1 angles in both variants, five of six χ_2 angles in the threefold mutant and four of seven χ_2 angles in the sixfold mutant. We conclude that ORBIT calculates sequences that fold specifically to a geometry close to the template, even when the

template is moderately perturbed relative to a naturally occurring structure. There are apparently limits to the size of acceptable perturbations: In this study, the larger perturbation led to undesired dynamic behavior.

Introduction

It is now well known that protein backbones undergo small but global rearrangements to accommodate changes in hydrophobic core packing when core amino acid residues are mutated (Baldwin *et al.* 1993; Lim *et al.* 1994). Understanding this interplay between sequence and structure is particularly important for protein design. Most computational design methods presented to date presuppose a rigid backbone structure (for review, see Street and Mayo 1999), though several groups have reported efforts to treat both backbone structural variability and side-chain selection (Su and Mayo 1997; Harbury et al. 1998; Desjarlais and Handel 1999). In our approach, the global fold of a protein is decomposed via supersecondary structure parameterization. Variation of supersecondary structure parameter values then provides new fixed-backbone templates for input to a sequence selection algorithm.

In particular, we studied the immunoglobulin binding ß1 domain of streptococcal protein G (Gß1), a 56-residue domain comprising a four-stranded β-sheet and an α-helix. Four parameters were derived that fix the position and orientation of the helix with respect to the sheet: the distance between the helix center and the sheet plane, two angles defining the orientation of the helix axis with respect to the sheet plane, and an angle defining rotation about the helix axis. Each of these parameters was varied incrementally (up to

±1.5 Å for the helix-sheet distance and up to ±10° for the angles) to generate novel backbones. The backbones were then used as templates for core residue sequence selection calculations with the ORBIT (Optimization of Rotamers by Iterative Techniques) protein design programs, which utilize the dead-end elimination theorem to solve the rotamer space combinatorial optimization problem (Desmet *et al*. 1992; Pierce *et al.* 2000). The ten most buried residues in the crystal structure of the wild-type protein (excluding glycines) were included in the calculation: backbone variation and subsequent sequence selection resulted in mutations at three to six of these positions (Su and Mayo 1997).

Gß1 variants containing the optimal sequences calculated in this fashion were expressed and purified for analysis. Thermal stabilities were assessed by circular dichroism (CD) spectroscopy; fold specificities were evaluated by a qualitative consideration of chemical shift dispersion in 1D ^{1}H nuclear magnetic resonance (NMR) spectra. It was found that small perturbations of the backbone yielded small changes in core sequence (three of 10 positions) and that the proteins containing those sequences were similar to Gß1 in thermal stability and chemical shift dispersion. Many of the sequences calculated for more extensively displaced backbones also yielded well-folded proteins, judged by chemical shift dispersion. Several of these latter variants, however, are destabilized relative to the wild-type protein.

Analysis at this level establishes that the sequence selection algorithm is tolerant of small variations in backbone specification: when a nonnative but native-like backbone is used as a template, a sequence is calculated that yields a well-folded, thermostable protein. It

is of considerable interest to know, further, how closely the folded protein matches the target structure and, particularly, how accurately the algorithm predicts core side-chain packing under backbone perturbations.

We report here the solution structures of two Gß1 variants determined by ^{1}H NMR: one minimally perturbed (a threefold mutant) and one extensively perturbed (a sixfold mutant). When the native Gß1 backbone is used as a template, the lowest-energy calculated sequence has three conservative mutations relative to the wild-type sequence: Y3F, L7I, and V39I (Dahiyat and Mayo 1997). These mutations have been rationalized in terms of the details of the calculation (Su and Mayo 1997). Experimentally, the protein containing this sequence (designated $\Delta h_{0.9}$[+0.00 Å] in the previous study, referred to hereafter as Δ0) was found to be slightly more stable than wild-type, with a melting temperature (T_m) of 91°C (T_m of Gß1 is 89°C). The Δ0 sequence was also obtained by sequence selection with several different backbones in which the orientation of the helix with respect to the sheet was varied by small amounts. Thus Δ0 represents the optimal sequence for backbones close to the native fold. Displacement of the template helix from the sheet plane by +1.50 Å yields the sixfold mutant, which contains the three core substitutions of Δ0 plus F30L, A34I, and F52W. Among the extensively perturbed variants of the earlier study, this protein (previously designated $\Delta h_{1.0}$[+1.50 Å], referred to hereafter as Δ1.5) was the best behaved, with chemical shift dispersion comparable to wild-type and a T_m of 73°C.

Results and discussion

Standard sets of 2D ^{1}H NMR data were collected for Δ0 and Δ1.5. Spin systems were assigned for all residues of Δ0. Core residue side-chains were completely assigned; other side-chain assignments are >95% complete. Good dispersion of chemical shifts and narrow linewidths in the Δ0 spectra indicate that this protein favors a single conformation under the experimental conditions. The Δ1.5 data, by contrast, contain evidence of conformational dynamics. While resonance assignments for this protein are also ~95% complete, no spin system was found for E27, and cross peaks to the backbone amide protons of T25, T51, and T53 are broadened and of low intensity. The chemical shifts of the ring protons of W52 are similar to random coil values, and the indole imino proton signal from this residue is absent, suggesting that its side-chain is conformationally labile and accessible to solvent. Also, the Hε and Hζ ring protons of F3 could not be assigned definitively.

Families of structures consistent with the data were generated by standard distance geometry/simulated annealing methods (Nilges *et al.* 1988, 1991). The structures of both molecules are well defined, and their stereochemical quality is good (Table 1). Both proteins have the characteristic protein G fold. The Δ0 sequence adopts a fold quite similar to its template, that is, the native Gß1 backbone (Fig. 1a). The rms deviation (rmsd) between atoms in the minimized mean experimental backbone and atoms in the crystallographic backbone is 0.92 Å (excluding two residues at the N-terminus, for which

few experimental restraints exist). Δ1.5 also closely matches the native Gß1 structure, with a backbone atomic rmsd of 1.03 Å. With a backbone atomic rmsd of 1.26 Å (Fig. 1b), Δ1.5 is somewhat less similar to its own target backbone.

Prediction by ORBIT of core side-chain packing was found to be excellent (Fig. 2a,b). All of the nontrivial core residue χ_1 angles were predicted correctly: the largest deviations between target and experimental structures were 22° (F30) in Δ0 and 35° (L5) in Δ1.5. Somewhat less robust was the χ_2 angle prediction: five of six nontrivial χ_2s were correctly predicted in Δ0, four of seven in Δ1.5. Closer examination of the Δ1.5 core reveals that the residues for which χ_2 is mispredicted (F3, L5, L30) interact with side-chains that are dynamically disordered (E27 and W52, as described above). Misprediction of χ_2 in these residues might be a further indication of conformational heterogeneity in this portion of the protein.

A previous study found that Gß1 variants with multiple core mutations form stable well-folded proteins (Gronenborn et al. 1996). We have extended this result herein, showing that a native-like fold is retained with changes at as many as six of ten core positions. The Δ0 and Δ1.5 structures demonstrate, furthermore, that the sequences generated by ORBIT from perturbed backbone templates lead to correctly folded proteins and that ORBIT predicts core side-chain conformations in such proteins reasonably well. Similar success in predicting fold specificity and core packing has been demonstrated for the ROC

algorithm in a study of a designed core variant of ubiquitin (Johnson *et al.* 1999). In that study, a detailed analysis of backbone and core side-chain dynamics showed small but significant differences between wild-type and variant proteins. Our sixfold mutant, Δ1.5, the sequence obtained from the largest backbone perturbation we attempted, also shows unintended dynamic behavior. Much of this behavior may be caused by two aspects of the F52W mutation. First, the experimental Δ1.5 backbone more closely resembles the wild-type than the calculated backbone, so the core is overpacked. The bulk of the W52 side-chain must be compensated in ways (such as local structural fluctuations) other than global displacement of the helix from the sheet plane. Second, burial of the W52 imino proton in the hydrophobic core without a hydrogen-bonding partner may also contribute to the conformational exchange.

These results suggest several avenues for improvement of the design protocol. The method used to generate the Δ1.5 template neglected the loops connecting helix and sheet. Experimentally, we found that the Δ1.5 sequence does not achieve the helix-sheet separation specified in the Δ1.5 template; explicit consideration of loop length during backbone specification might enable us to achieve better agreement between target and experimental structures. In addition, further terms in the ORBIT scoring function, such as a penalty for burial of uncompensated polar hydrogens (implemented subsequent to this study), may lead to more favorable sequence selection and, hence, improved fold specificity.

Materials and methods

Designed proteins were expressed and purified as previously described (Su and Mayo 1997). For NMR experiments, 5–15 mg of lyophilized protein was dissolved in 700 μL buffer (50 mM sodium phosphate in either 90% H_2O/10% D_2O at pH 6.0 or 99.9% D_2O, pD 6.0), yielding 1–3 mM protein concentration. NMR experiments were performed on a Varian UnityPlus 600-MHz spectrometer equipped with a Nalorac *Z*-axis gradient probe. DQF-COSY, TOCSY, and NOESY spectra were acquired at 25°C for the structure determinations. Additional data sets were acquired at 35°C to facilitate resonance assignments. TOCSY spectra were acquired with mixing times of 25 and 80 msec, NOESY spectra with mixing times of 75, 100, and 150 msec. The spectral width in all experiments was 7500 Hz. The TOCSY and NOESY spectra were recorded with $256t_1$ * $1024t_2$ complex points, the DQF-COSY spectra with $512t_1$ * $2048t_2$ complex points. Amide hydrogen exchange rates were measured by following the time course of the disappearance of amide-α proton cross-peaks in magnitude-mode COSY spectra ($256t_1$ * $2048t_2$ points) for protonated, lyophilized protein resuspended in 99.9% D_2O. E.COSY spectra were also acquired, with $625t_1$ * $2048t_2$ complex points. All spectra were processed with VNMR (Varian).

Resonance assignment was performed using ANSIG (Kraulis 1989) for the Δ0 data and NMRCOMPASS (MSI) for the Δ1.5 data. Cross-peaks in the 75 msec mixing time NOESY spectra were assigned for use as distance restraints. Poorer dispersion in the Δ1.5 spectra than in the Δ0 spectra necessitated additional steps in assigning NOESY

cross peaks, as follows. A table of putative NOESY cross-peak assignments was generated automatically in NMRCOMPASS. Proton pairs separated by >10 Å in the Δ1.5 template were discarded as possible assignments, yielding a partially assigned restraint set (Nilges *et al.* 1997). The subset of unambiguously assigned restraints taken from this set was used to calculate an initial ensemble of structures. The minimized mean of this ensemble was then used to calculate a new set of interproton distances, which were again used to filter the NOESY cross-peak assignments, this time with a 5-Å distance cutoff. After the second cycle of distance filtering, remaining ambiguous restraints were discarded. This approach resulted in a comparable number of distance restraints for the two proteins (Table 1). The χ_1 restraints were obtained from coupling constant measurements in E.COSY spectra combined with patterns of intraresidue NOEs (Wagner *et al.* 1987). These angular restraints were found to improve the quality and precision of the ensemble of Δ1.5 structures but not that of the Δ0 structures. Hence, χ_1 restraints were not used in refinement of the Δ0 ensemble. Handling of experimental restraints was otherwise as previously described (Malakauskas and Mayo 1998).

Standard hybrid distance geometry/simulated annealing protocols were used to find structures consistent with experimental restraints (Nilges et al. 1988, 1991). Distance geometry structures (100) were generated, regularized, and refined, resulting in ensembles of structures (68 for Δ0, 81 for Δ1.5) with no restraint violations >0.3 Å, rmsds from idealized bond lengths <0.01Å, and rmsds from idealized bond angles <1°. Statistics for the 40 lowest-energy structures of each of these ensembles are compiled in

Table 1.

Assignment details for Δ0

Spectra were collected under two sets of conditions: 35ºC/pH 5 and 25ºC/pH 6, to allow comparison with structures being determined for other mutants. TOCSY, COSY, and NOESY were collected in 50 mM phosphate buffer (90% H_2O/10% D_2O) under both sets of conditions. NOESY spectra were also collected in 10% D_2O. An ECOSY was collected at 35ºC/pH 5.

Initial proton chemical shift assignments were made for the 35ºC/pH 5 data set, with occasional use of the 25ºC/pH 6 data set to resolve ambiguities. Unambiguous NOESY cross-peaks were assigned by hand to confirm a protein G type fold. Ambiguous peaks were assigned using the interproton distances from the preliminary structures and careful consideration of alignment. All NOESY peaks could be assigned, except in the methyl-methyl and aliphatic-aliphatic region, where extensive overlap and instrumental artifacts precluded full assignment. Cross-peaks in the HCα-HCα were assigned on the D_2O NOESY, due to artifacts from water in the 90/10 spectrum. The ECOSY spectrum was used to make stereospecific assignments for six pairs of β methylene protons. Two of the three non-degenerate pairs of glycine alpha protons, four pairs of terninal amide protons on Asn and Gln, and the γ methyl groups of V30 and V55 were stereospecifically assigned near the end of the structure determination, on the basis of the NOESY data.

Acknowledgements

We thank Monica Breckow for assistance with molecular biology protocols. This work was supported by the Howard Hughes Medical Institute (S.L.M.). C.A.S. is partially supported by an NSF graduate research fellowship. Coordinates and NMR restraints have been deposited in the Protein Data Bank. Accession numbers for the coordinates are 1fd6 and 1fcl for Δ0 and Δ1.5, respectively.

References

Baldwin, E.P., Hajiseyedjavadi, O., Baase, W.A., and Matthews, B.W. 1993. The role of backbone flexibility in the accomodation of variants that repack the core of T4 lysozyme. *Science* **262:** 1715–1718.

Dahiyat, B.I. and Mayo, S.L. 1997. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci.* 94: 10172–10177.

Desjarlais, J.R. and Handel, T.M. 1999. Side-chain and backbone flexibility in protein core design. *J. Mol. Biol.* 289: 305–318.

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

Gallagher, T., Alexander, P., Bryan, P., and Gilliland, G.L. 1994. Two crystal structures of the ß1 immunoglobulin-binding domain of streptococcal protein G and comparison

with NMR. *Biochemistry* 33: 4721–4729.

Gronenborn, A.M., Frank, M.K., and Clore, G.M. 1996. Core mutants of the immunoglobulin binding domain of streptococcal protein G: Stability and structural integrity. *FEBS Lett.* 398: 312–316.

Harbury, P.B., Plecs, J.J., Tidor, B., Alber, T., and Kim, P.S. 1998. High-resolution protein design with backbone freedom. *Science* 282: 1462–1467.

Johnson, E.C., Lazar, G.A., Desjarlais, J.R., and Handel, T.M. 1999. Solution structure and dynamics of a designed hydrophobic core variant of ubiquitin. *Structure* 7: 967–976.

Koradi, R., Billeter, M., and Wüthrich, K. 1996. MOLMOL: A program for display and analysis of macromolecular structures. *J. Mol. Graph.* 14: 51–55.

Kraulis, P.J. 1989. ANSIG: A program for the assignment of protein ^{1}H NMR spectra by interactive computer graphics. *J. Magn. Reson.* 84: 627–633.

Laskowski, R.A., Rullmann, J.A., MacArthur, M.W., Kaptein, R., and Thornton, J.M. 1996. AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR. *J. Biomol. NMR* 8: 477–486

Lim, W.A., Hodel, A., Sauer, R.T., and Richards, F.M. 1994. The crystal structure of a mutant protein with altered but improved hydrophobic core packing. *Proc. Natl. Acad. Sci.* 91: 423–427

Malakauskas, S.M. and Mayo, S.L. 1998. Design, structure, and stability of a

hyperthermophilic protein variant. *Nat. Struct. Biol.* 5: 470–475

Nilges, M., Clore, G.M., and Gronenborn, A.M. 1988. Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Lett.* 229: 317–324.

Nilges, M., Kuszewski, J., and Brünger, A.T. 1991. Sampling and efficiency of metric matrix distance geometry. In *Computational aspects of the study of biological macromolecules by NMR* (ed. J.C. Hoch, et al.), pp. 451–457. Plenum, New York.

Nilges, M., Macias, M.J., O'Donoghue, S.I., and Oschkinat, H. 1997. Automated NOESY interpretation with ambiguous distance restraints: The refined NMR solution structure of the pleckstrin homology domain from ß-spectrin. *J. Mol. Biol.* 269: 408–422.

Pierce, N.A., Spriet, J.A., Desmet, J., and Mayo, S.L. 2000. Conformational splitting: A more powerful criterion for dead-end elimination. *J. Comp. Chem.* 21: 999–1009.

Street, A.G. and Mayo, S.L. 1999. Computational protein design. *Structure* 7: R105–R109.

Su, A. and Mayo, S.L. 1997. Coupling backbone flexibility and amino acid sequence selection in protein design. *Prot. Sci.* 6: 1701–1707

Wagner, G., Braun, W., Havel, T.F., Schaumann, T., Go, N., and Wüthrich, K. 1987. Protein structures in solution by nuclear magnetic resonance and distance geometry—The polypeptide fold of the basic pancreatic trypsin-inhibitor determined using 2 different

algorithms, DISGEO and DISMAN. *J. Mol. Biol.* 196: 611–639.

Table 1. Experimental restraints and structure statistics

	Δ0	Δ1.5
NOE distance restraints		
Intraresidue	208	317
Sequential	145	146
Medium range (2|i-j|4)	67	73
Long range (|i-j|5)	176	161
Hydrogen bond restraints	28	36
χ_1 restraints	0	10
rmsds from data		
Distance restraints (Å)	0.028 ± 0.001	0.029 ± 0.003
χ_1 restraints (°)	n/a	0.57 ± 0.50
RMSDs from ideal geometry		
Bonds (Å)	0.0031 ± 0.0001	0.0033 ± 0.0001
Angles (°)	0.55 ± 0.01	0.58 ± 0.01
Impropers (°)	0.41 ± 0.01	0.42 ± 0.01
Ensemble atomic RMSDs (Å)[a]		
Backbone	0.23	0.23
Heavy atoms	0.74	0.60
Ensemble Ramachandran statistics[b]		
Residues in most favored regions (%)	77.7	80.4
Residues in additionally allowed regions (%)	20.7	19.3
Residues in generously allowed regions (%)	1.4	0.2
Residues in disallowed regions (%)	0.1	0.1

[a] Ensemble RMSDs were calculated for residues 2–56 of both proteins.

[b] Ramachandran analysis was performed with PROCHECK-NMR (Laskowski *et al.* 1996).

Figures

Figure 1. Stereoviews of experimental versus target structures of Gß1 variants. (*a*) Superposition of the minimized mean experimental structure of Δ0 (green) and the crystal structure of Gß1 (red), accession code 1pga (Gallagher *et al.* 1994). (*b*) Superposition of the minimized mean experimental (yellow) and calculated (blue) structures of Δ1.5. Incomplete N-terminal methionine processing results in mixtures of 56 and 57 amino acid proteins, with the 57-mer predominating for more stable variants. The structures presented are the 57-mer of Δ0 and the 56-mer of Δ1.5 (sequence numbering for the 56-mer is used throughout the text). Figures were generated using MOLMOL (Koradi *et al.* 1996).

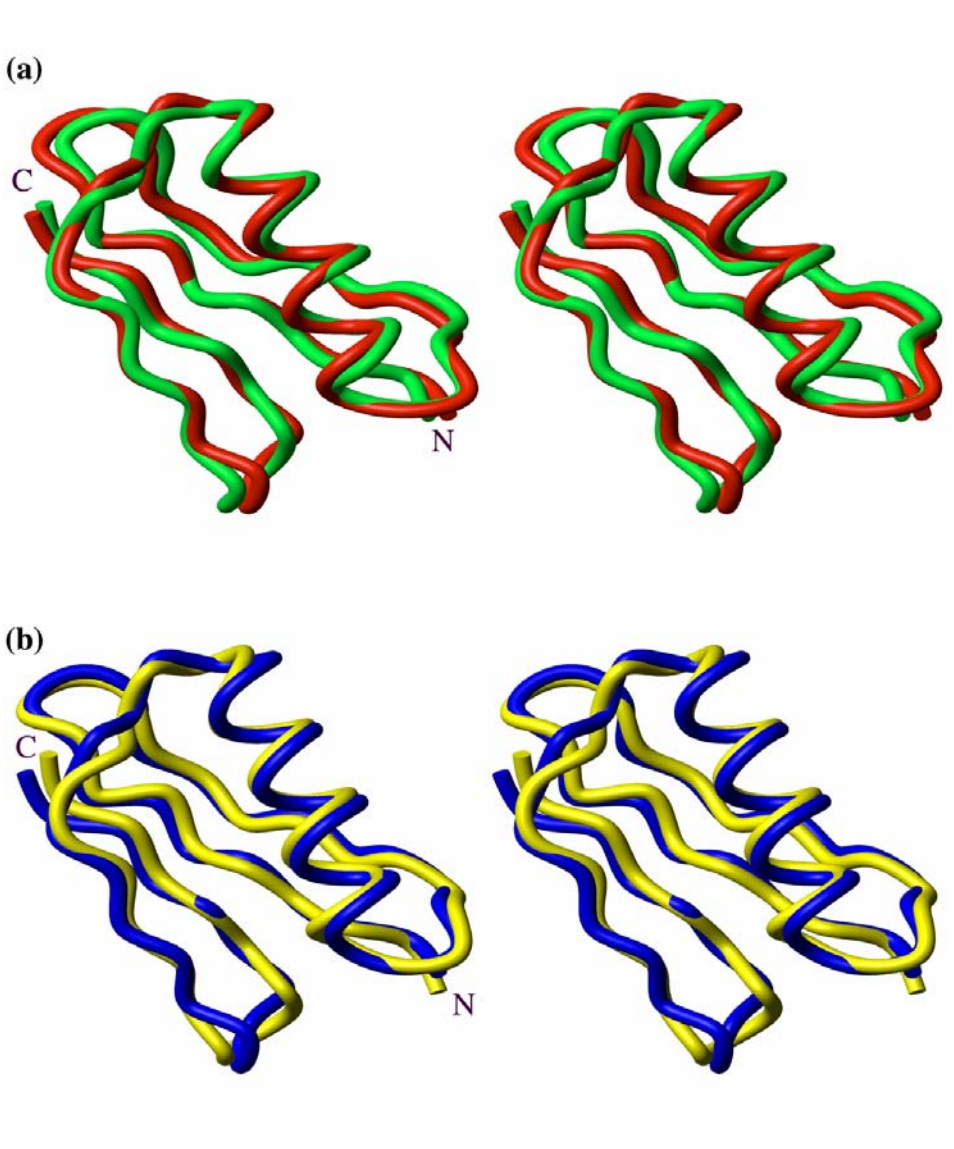
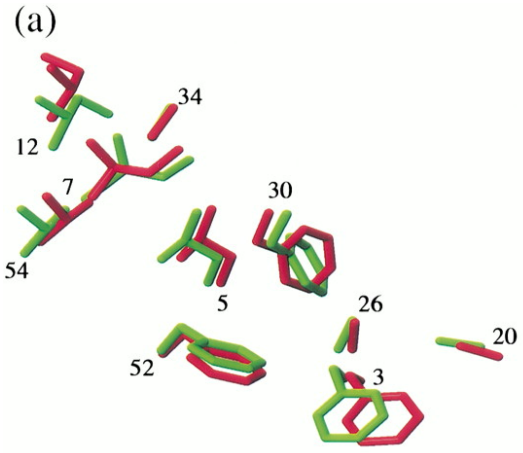
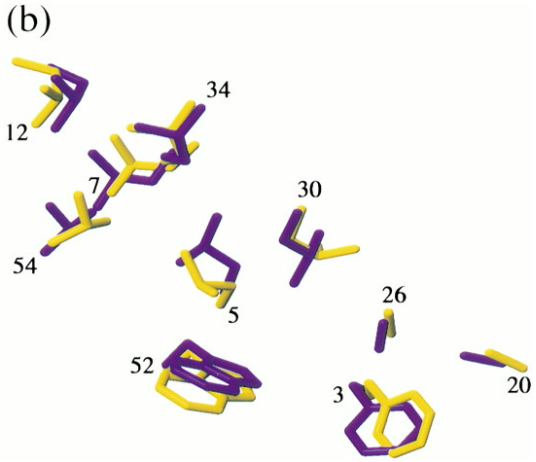

(a)
C
N
(b)
C
N

Figure 2. Side-chain packing in Gß1 variants. (*a*) Core residue heavy atoms of the minimized mean experimental (green) and calculated (red) structures of Δ0. (*b*) Core residue heavy atoms of the minimized mean experimental (yellow) and calculated (blue) structures of Δ1.5. χ_1 and χ_2 angles in the ensemble of NMR structures were found in all cases to be well represented by the values in the minimized mean structures. Residue numbers are located near each residue's Cα atom.


(a)
34
12
7
30
54
5
26
20
52
3


(b)
34
12
7
30
54
5
26
52
3
20

Inclusion of an entropic penalty for methionine in protein design calculations

Abstract

A series of sixteen core mutants of the β1 domain of the streptococcal IgG binding protein (protein Gβ1) were expressed and characterized. The correlation between the calculated energies from the ORBIT protein design process and the experimentally determined melting temperatures is excellent if one just considers the eight mutants that do not contain methionine. When methionine-containing mutants are included, false positives decrease the correlation, as four relatively unstable methionine-containing mutants receive the most favorable energy scores. Addition of a penalty term for the inclusion of methionine eliminates the false positives, restoring the correlation between calculation and experiment for this Gβ1 data set. In studies with lysozyme, the methionine inclusion penalty allows selection of methionine at one key position where it is most crucial to stability, while preventing its indiscriminate selection, which is seen in the absence of a penalty term. The penalty effectively disallows methionine at most core positions, while allowing its selection at positions where it is vital to core packing.

Introduction

Cores are a common point of entry for *de novo* protein design, as they may be designed with minimalist energy expressions. Acceptable core designs can be accomplished just by considering packing and restricting core positions to a subset of hydrophobic residues [1-4]. Based on their mutagenesis studies of λ repressor, Lim and Sauer [3] showed that hydrophobicity is the most important predictor of whether a protein with a mutated core will fold. They also found that steric considerations were important in producing variants with wild-type activities. Handel and coworkers successfully redesigned several cores, including 434 cro [1] and ubiquitin [2], using only a Lennard-Jones van der Waals potential. Dahiyat and Mayo found that the addition of terms for solvation can improve the predictive power of the energy expression [4]. Others [5] have used more complex energy expressions for core designs, but the limited available experimental data do not allow evaluation of their assertion that a more complex energy expression actually yields better predictive ability or sequences with enhanced stability. In contrast to the simple energy functions sufficient for core design, designs of surface and partially buried positions require consideration and balancing of many terms, such as electrostatics, hydrogen bonding, structural propensities, and solvation [6].

Upon folding, proteins experience a decrease in entropy. Part of this entropy loss results from loss of backbone flexibility due to formation of secondary and tertiary structure. Another part of this entropy loss results from “freezing” of side-chains into a fixed conformation [7]. These entropy losses are partially offset by changes in

solvation upon folding, but these solvent contributions may be accounted for by consideration of exposed surface areas [8]. While the loss of backbone entropy should be relatively constant for closely related sequences folding to the same structure (excluding the effects of proline and disulfide bonds), the loss of side-chain entropy will depend on the identity and placement of the side-chains. Creamer and coworkers [9-11] have calculated the unfolded entropies of side-chains on the basis of Monte Carlo simulations with small peptide models. With the assumption that a buried side-chain loses all conformational entropy upon folding, these unfolded entropies may be used to calculate the loss of side-chain entropy upon folding.

The energy expression is an integral part of any computational protein design process. The energy expressions used in protein design have been recently reviewed [12, 13]. Design processes such as ORBIT (Optimization of Rotamers by Iterative Techniques) determine the single sequence with the most favorable value for the energy expression, given a fixed backbone. To evaluate the usefulness of the energy expression for protein design, the search technique should be deterministic, so that the energy expression, and not the search method, determines the protein sequence [14].

Energy expressions for protein design can be developed and improved by use of a design cycle [4]. An energy expression is postulated on the basis of existing experimental or theoretical results. The energy expression is then used to generate novel protein sequences, which are characterized. The correlation between experimental and calculated stabilities is examined, and the energy expression is modified to improve the correlation. This cycle is repeated, resulting in experimentally validated improvements in

the energy expression. By beginning with a minimal energy expression and adding additional terms only when they improve the correlation between calculation and experiment, the energy expression is limited to only those terms necessary to ensure a good correlation and good predictive behavior. It is expedient to discretize the problem by the use of rotamer representations of the side-chains and a fixed backbone. Discretization may cause some sequences to receive less favorable energy scores due to van der Waals clashes or failure to correctly score hydrogen bonds, which could be avoided by allowing backbone flexibility or continuous side-chains, at the cost of increased computational complexity.

False negatives (stable sequences receiving poor scores) generally reflect deficiencies in the model, caused in part by discretization of rotamers and the use of a rigid backbone, while false positives (low stability sequences with favorable scores) are indicative of a problem with the energy expression itself, which improperly grants favorable energy scores to some sequences. The necessity of incorporating terms for negative design [15] results in an energy expression that can be used to find a sequence that assumes the desired fold with good stability, but these negative design terms make the energy expression inherently non-physical. Negative design terms reduce false positives, but they may increase the number of false negatives. Using a lattice model, Chiu and Goldstein [16] have shown that the best energy expression for sequence selection is not necessarily one that is physically accurate. Although design can be used as a means to study the physical basis of protein stability, the energy expression most suitable for generating sequences that adopt the target fold with good stabilities may not be the best

one for prediction of protein stability for an existing set of protein sequences, due to negative design issues and computational requirements.

A number of protein design groups have evaluated and parameterized energy expressions based on the correlation between an experimental measure (T_m, ΔG, or a functional assay) and the calculated energy [4, 17]. To test the accuracy of energy expressions intended for core design, several researchers have looked at their abilities to accurately predict the stabilities of a series of core mutants. Lee and Levitt [18] compared the predicted and experimental stabilities and activities of a series of λ repressor core mutants using a van der Waals term and a torsional potential. Kono *et al.* [19] used terms for hydration, side-chain entropy, bond energy, and non-bonded energy to predict the relative stabilities of four conservative malate dehydrogenase core mutants and the wild-type sequence. Parameterization of energy expressions based on an experimental series is hampered by both false positives and false negatives, as they reduce the correlation between the experimental and computational measures.

Of interest to the field of protein design is whether these parameterized energy expressions can be used to generate novel protein sequences with enhanced stability or other desired properties. The ability of an energy expression to predict relative experimental stabilities for a series of mutants selected by some other technique is a less satisfactory measure of success than the use of the energy expression to generate novel sequences. It is sometimes the case that an energy expression that provides a satisfactory correlation between experiment and calculation for a small set of existing sequences will generate novel sequences that do not have good experimental stabilities (N.A. Pierce,

personal communication).

If the energy expression will be used for the selection of a small set of protein sequences with good stabilities, it is useful to the extent that it gives the most favorable scores to stable proteins with the desired solution behavior. Because there are presumably a number of sequences that will have the desired behavior, a few false negatives do not cause significant difficulties. However, false positives are a concern because in general only a limited number of sequences will be experimentally characterized. Thus, a false negative is a missed opportunity to stabilize the protein but does not preclude the evaluation of some of the other acceptable sequences, while a false positive can represent a significant waste of resources to characterize a new protein sequence with unacceptable properties [34].

In protein core design efforts with ORBIT, methionine residues have customarily been disallowed [20-22]. Met is rare in naturally-occurring proteins. The relatively large loss of conformational entropy upon burial of methionine may destabilize designed proteins with high Met content. It has been shown that substituting methionine for many of the core residues of T4 lysozyme is destabilizing [23]. The entropy difference between Met and Leu in an unfolded protein is 3.3 cal/(mol K), and the entropy difference between Met and Ile is 2.7 cal/(mol K) [11]. At relevant temperatures, this corresponds to a 1 kcal/mole entropic penalty per methionine incorporation, assuming that all side-chain entropy is lost upon formation of the hydrophobic core. (Any residual disorder within the core would decrease this penalty.) Thus, burial of Met in the core of a

well-folded protein involves a greater loss of entropy than similarly sized Leu and Ile. Thus, methionine must not be indiscriminately selected during protein design.

Although efforts to design protein cores have been successful without the use of methionine, other lysozyme experiments by Matthews and coworkers [24] show that replacement of certain wild-type methionine residues with other hydrophobic residues destabilizes T4 lysozyme. These data suggest that there may be cases where the optimal sequence for a protein will include one or more methionine residues, despite unfavorable entropy considerations. Although Leu and Ile occupy approximately the same volume as Met, their steric requirements are sufficiently different that they may not be acceptable alternatives at some positions despite their reduced entropy loss upon folding. It is desirable to modify ORBIT to allow incorporation of methionine residues, but adjustment of the energy expression is required to prevent excessive selection of methionines and destabilization of the resulting proteins. Methionine should be chosen only when the stabilization resulting from better packing compensates for the destabilization caused by increased loss of entropy upon folding.

Results and discussion

The β1 domain of the streptococcal IgG binding protein (Gβ1) was used for evaluation of the effects of the inclusion of methionine on designed protein cores. The Gβ1 domain was selected for this study because the ten core residues (as classified by ORBIT) are located in one central cavity, with residues from both α and β

conformational space represented. It has been shown previously that large perturbations to the core sequence of Gβ1 do not cause significant changes to the protein structure [25].

When methionine residues are disallowed, the previously described triple mutant [25–27], Y3F/L7I/V39I (IIV), receives the most favorable energy score. IIV is the most stable known core mutant of Gβ1. Sixteen core mutants of Gβ1 were expressed and characterized (Table 1). The mutants are closely related, with the same amino acid identities at seven of ten core positions, and aliphatic residues I, L, V, or M at the remaining three positions; the variation in core volume between the most underpacked and the most overpacked sequence is roughly three methylene groups. A ribbon diagram indicating the core residues of this protein is shown in Figure 1. The three variable positions in this study are indicated in red. The correlation between the experimental and calculated stabilities is reasonable for the sequences that do not contain methionine ($R=0.76$ for eight mutants, $r_s=0.72$, $p=0.04$, Figure 2a). However, the correlation for the full data set (16 sequences) is poor ($R=0.35$, $r_s=0.13$, $p=0.58$). When methionine is allowed at all core positions, Y3F/L7M/V39L/V54I (MLI) is predicted to be the most stable sequence, as shown in Table 1. However, the MLI mutant is destabilized compared to wild-type, exhibiting a 13ºC decrease in the experimentally determined melting temperature. In addition, mutants MIV, MLV, and MMV, predicted to be more stable than IIV, are also less stable than IIV. MIV and MMV are destabilized relative to the wild-type, and MLV is comparable to the wild-type. The energy function was also evaluated in terms of its ability to rank the relative stabilities of pairs of sequences, after the method of Mendes *et al.* [28]. Performance of the original energy function was poor, with only 51% of pairs ranked correctly. The prediction that these methionine-

containing mutants would be the most stable core sequences indicates that some property of methionine, possibly the higher entropic penalty for folding, is not accurately modeled by the energy expression.

The correlation between experimental and computational stabilities can be improved by the addition of a penalty term for each methionine incorporated in the designed sequence. Use of a 9 kcal/mole Met penalty improved the Spearman rank correlation from r_s=0.13 (p=0.58) to r_s=0.84 (p=0.0004) for the 16 mutants studied. Pair prediction improved to more than 75% success with the inclusion of a penalty term, with minimal gain beyond 8 to 10 kcal/mol, as shown in Figure 3. With the 9 kcal/mole penalty, IIV is correctly predicted to have the highest melting temperature. There are significant false negatives present in the original energy scoring, which are not corrected by the addition of the penalty; however, false negatives far from the global minimum energy sequence are not great concerns if the energy expression will be used to generate novel stable sequences. In the case of LMI and LMV, the poor score likely results in part from an unrepresented side-chain conformation at L7, which is also present in the wild-type. Of greater importance, the false positives (MIV, MVV, MLI, and MMV) are sufficiently penalized by the methionine inclusion penalty to prevent their selection.

Factors in addition to entropy support the use of the methionine inclusion penalty. The relatively large number of Met rotamers in the ORBIT rotamer library improves the likelihood that a Met rotamer will exist that fits into the core without clashes with other side-chains relative to Leu and Ile, which have similar volumes. Thus, the larger penalty term compensates for a bias towards Met, caused by better packing of Met residues in a discrete rotamer and fixed backbone context.

To validate the size of the methionine penalty, we examined data from Matthews and coworkers [24] for a series of mutations in lysozyme. Lysozyme contains four methionines, each of which was mutated to leucine to generate four point mutants. Replacement of two of these methionines is destabilizing as measured by thermal denaturation, while replacement of either of the other two is stabilizing. Energies were calculated for these mutations using the wild-type (WT*) backbone (PDB code 1L63). For each mutated position, the WT* and leucine point mutant sequences were scored with ORBIT, allowing repacking (but no change in side-chain identity) of the mutated residue and any other residues within 5 Å. The differences between the wild-type methionine and the leucine mutant energy scores were compared, as shown in Table 2. The methionine at position 6, which is the most important for retention of stability, is selected despite an 8 to 10 kcal/mol penalty term. The discrimination between methionine and leucine at the other positions is less important, as these positions have only a small effect on stability. This result shows that a methionine penalty of this size does not completely exclude methionine from the sequence selection, but can restrict its occurrence to positions where it is critical to core packing. The core of lysozyme has recently been redesigned using ORBIT. The sequence produced using an 8 kcal/mol Met inclusion penalty is significantly more stable than the sequence produced without the Met penalty, although both are destabilized relative to the cysteine-free wild-type [29]. This result shows that the methionine penalty can improve the sequence selected by ORBIT in a real design case, in addition to improving correlation in an existing data set.

Conclusions

Successful core redesign has often required exclusion of methionine residues for optimal results. Inclusion of a simple energy function term that penalizes methionine inclusion allows methionine to be considered at core positions, while preventing indiscriminate selection. The exact size of this penalty will depend on the energy function used. It is hoped that this set of 16 core mutants of Gβ1 will prove useful to other investigators for optimization of energy functions for core design.

Computational methods

The template structure for the Gβ1 calculations was the file 1pga from the Protein Data Bank. Water atoms were removed, hydrogen atoms were added, and the resulting structure was subjected to 50 steps of steepest descent conjugate gradient minimization using the program BIOGRAF (Molecular Simulations). Ten non-glycine positions are characterized as "core," as previously described [27]. The optimal sequence at these ten positions was calculated with ORBIT, allowing only hydrophobic residues (A, V, L, I, M, F, Y, W) and using type II solvation [30]. A Monte Carlo simulated annealing procedure was used to generate additional sequences that were slightly destabilized relative to the ground state. Energies for these sequences were calculated after repacking to generate the optimal rotamer conformations and lowest possible energies for each sequence. Although all hydrophobic side-chains were allowed at all ten positions in the initial design and in the Monte Carlo procedure, only sequences that varied from each other at three positions (7, 39, and 54) were characterized, as these positions exhibited the most

sequence variability in the Monte Carlo list.

Protein expression and purification

Mutants were generated by sequential rounds of inverse PCR [31] starting from pET-11a plasmids (Novagen) containing the IIV and VIV sequences [27]. Primers were 40 to 45 base pairs long. Template plasmids were digested using DpnI. The resulting plasmids were transformed into *E. coli* XL1 Blue cells. Mutant sequences were verified by sequencing before transformation into *E. coli* BL21(DE3) cells for expression. Proteins were extracted from the cells using a freeze-thaw protocol [32]. After suspension of the protein in PBS buffer and removal of the cells by centrifugation, one volume of acetonitrile was added to precipitate contaminants from the samples. The remaining soluble protein was purified by reverse phase high pressure liquid chromatography on a C8 column using a water/acetonitrile gradient with 0.1% by volume trifluoroacetic acid. All proteins were obtained as mixtures of 56- and 57-mer, due to incomplete N-terminal processing. The two species were readily separated by HPLC. The 57-mer proteins were characterized, as 57-mer was the major species in all cases. Each protein mass was verified by matrix assisted laser desorption mass spectroscopy.

Protein characterization

Mutant proteins were characterized by circular dichroism. The protein concentrations were approximately 50 μM in 50 mM sodium phosphate buffer at pH 5.0.

Wavelength scans from 190 to 250 nm confirmed that the secondary structures of the mutant proteins closely resemble the wild-type protein (data not shown). Thermal denaturation data was collected from 1°C to 99°C in 2°C steps, using a 2-minute equilibration time and a 40-second averaging time for each temperature. Reversibility of the transition was verified by comparison of initial and final wavelength scans at 1°C.

References

1. Desjarlais, J.R. and T.M. Handel, *De novo design of the hydrophobic cores of proteins.* Protein Science, 1995. **4**: p. 2006–2018.
2. Lazar, G.A., J.R. Desjarlais, and T.M. Handel, *De novo design of the hydrophobic core of ubiquitin.* Protein Science, 1997. **6**: p. 1167–1178.
3. Lim, W.A. and R.T. Sauer, *Alternative packing arrangements in the hydrophobic core of λ repressor.* Nature, 1989. **339**: p. 31–36.
4. Dahiyat, B.I. and S.L. Mayo, *Protein design automation.* Protein Science, 1996. **5**: p. 895–903.
5. Jiang, X., et al., *A new approach to the design of uniquely folded thermally stable proteins.* Protein Science, 2000. **9**: p. 403–416.
6. Street, A.G., et al., *Designing protein β-sheet surfaces by Z-score optimization.* Physical Review Letters, 2000. **84**(21): p. 5010–5013.
7. Bromberg, S. and K.A. Dill, *Side-Chain Entropy and Packing in Proteins.* Prot. Sci., 1994. **3**: p. 997–1009.

8. Lee, K.H., et al., *Estimation of changes in side chain configurational entropy in binding and folding: General methods and application to helix formation.* Proteins, 1994. **20**: p. 68–84.

9. Creamer, T.P., *Side-Chain Conformational Entropy in Protein Unfolded States.* Proteins, 2000. **40**: p. 443–450.

10. Creamer, T.P. and G.D. Rose, *Side-chain entropy opposes α-helix formation but rationalizes experimentally determined helix-forming propensities.* PNAS, 1992. **89**: p. 5937–5941.

11. Creamer, T.P. and G.D. Rose, *α-helix-forming propensities in peptides and proteins.* Proteins: structure function and genetics, 1994. **19**: p. 85–97.

12. Mendes, J., R. Guerois, and L. Serrano, *Energy estimation in protein design.* Current Opinion in Structural Biology, 2002. **12**: p. 441–446.

13. Gordon, D.B., S.A. Marshall, and S.L. Mayo, *Energy Functions for Protein Design.* Current Opinion in Structural Biology, 1999. **9**(4): p. 509–513.

14. Voigt, C.A., D.B. Gordon, and S.L. Mayo, *Trading accuracy for speed: a qualitative comparison of search algorithms in protein sequence design.* J. Mol. Biol., 2000. **299**: p. 789–803.

15. Hecht, M.H., et al., Science, 1990. **249**: p. 884–891.

16. Chiu, T.-L. and R.F. Goldstein, *Optimizing potentials for the inverse protein folding problem.* Protein Engineering, 1998. **11**(9): p. 749–752.

17. Desjarlais, J.R. and T.M. Handel, *Side-chain and backbone flexibility in protein core design.* Journal of Molecular Biology, 1999. **289**: p. 305–318.

18. Lee, C. and M. Levitt, *Accurate prediction of the stability and activity*

effects of site-directed mutagenesis on a protein core. Nature, 1991. **352**: p. 448–451.

19. Kono, H., et al., *Designing the hydrophobic core of Thermus flavus malate dehydrogenase based on side-chain packing.* Protein Engineering, 1998. **11**(2): p. 47–52.
20. Dahiyat, B.I. and S.L. Mayo, *De Novo Protein Design: Fully Automated Sequence Selection.* Science, 1997. **278**: p. 82–87.
21. Morgan, C.S., *Full sequence design of an alpha-helical protein and investigation of the importance of helix dipole and capping effects in helical protein design*, in *Biochemistry*. 2000, California Institute of Technology: Pasadena.
22. Dahiyat, B.I., C.A. Sarisky, and S.L. Mayo, *De Novo Protein Design: Towards Fully Automated Sequence Selection.* Journal of Molecular Biology, 1997. **273**: p. 789–796.
23. Gassner, N.C., W.A. Baase, and B.W. Matthews, *A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme.* Proceedings of the National Academy of Sciences USA, 1996. **93**: p. 12155–12158.
24. Lipscomb, L.A., et al., *Context-dependent protein stabilization by methionine-to-leucine substitution shown in T4 lysozyme.* Protein Science, 1998. **7**: p. 765–773.
25. Ross, S.A., et al., *Designed protein G core variants fold to native-like structures; sequence selection by ORBIT tolerates variation in backbone specification.* Prot. Sci., 2000. **10**: p. 450–454.
26. Dahiyat, B.I. and S.L. Mayo, *Probing the role of packing specificity in protein*

design. Proc. Natl. Acad. Sci. USA, 1997. **94**: p. 10172–10177.

27. Su, A. and S.L. Mayo, *Coupling backbone flexibility and amino acid selection in protein design.* Protein Science, 1997. **6**: p. 1701–1707.

28. Mendes, J., et al., *Implicit solvation in the self-consistent mean field theory method: sidechain modelling and prediction of folding free energies of protein mutants.* Journal of Computer-Aided Molecular Design, 2001. **15**: p. 721–740.

29. Mooers, B.H.M., et al., *Repacking the Core of T4 Lysozyme by Automated Design.* Journal of Molecular Biology, 2003. **332**: p. 741–756.

30. Street, A.G. and S.L. Mayo, *Pairwise Calculation of Protein Solvent-Accessible Surface Areas.* Folding & Design, 1998. **3**: p. 253–258.

31. Hemsley, A., et al., *A simple method for site-directed mutagenesis using the polymerase chain reaction.* Nucleic Acids Res, 1989. **17**: p. 6545–6551.

32. Johnson, B.H. and M.H. Hecht, *Recombinant proteins can be isolated from E. coli cells by repeated cycles of freezing and thawing.* Biotechnology, 1994. **12**: p. 1357–1360.

33. Koradi, R., M. Billeter, and K. Wuthrich, *Molmol: a program for the display and analysis of macromolecular structures.* J. Mol. Graph, 1996. **14**: p. 51–55.

34. Lazaridis, T. and M. Karplus, *Effective energy functions for protein structure prediction.* Current Opinion in Structural Biology, 2000. **10:** p. 139–145.

Tables

Table 1. Gβ1 mutants and stabilities

Name[a]	Tm (°C)[e]	Energy score[f]	Modified energy score[g]
MLI	78	-159.11	-149.11
MIV	84	-156.41	-146.41
MLV	86	-156.12	146.12
MMV	77	-155.35	-135.35
IIV[b]	91	-155.08	-155.08
MVV	82	-154.82	-144.82
IVV	88	-153.46	-153.46
VIV[c]	89	-153.26	-153.26
LIV[d]	86	-151.72	-151.72
VVV	84	-151.64	-151.64
LVV	84	-150.19	-150.19
MVI	78	-150.12	-140.12
III	88	-150.16	-150.16
LII	84	-146.90	-146.90
LMV	78	-136.31	-126.31
LMI	78	-134.69	-124.69
Wild-type	86	N/A	N/A

[a] Proteins are named with the identities of the residues at positions 7, 39, and 54. All sequences contain the Y3F mutation. Core positions L5, A20, A26, F30, A34, and F52 are unchanged in this series. The protein surface and boundary positions are wild-type

throughout the series.

[b] Previously described by Dahiyat & Mayo (1997) as α90, by Su & Mayo as $\Delta h_{0.9}$[0.00Å], and by Ross *et al.* (2000) as Δ0.

[c] Previously described by Su & Mayo as $\Delta h_{0.9}$[-1.00Å]

[d] Previously predicted by Jiang *et al.* [5] to melt at 4°C higher than the wild-type.

[e] for 57-mer proteins including an N-terminal methionine that is not removed during expression.

[f] The calculated energy without a methionine inclusion penalty. More negative numbers are favorable.

[g] The calculated energy with a 10 kcal/mole methionine inclusion penalty.

Table 2. Effects of the methionine penalty on core calculations in T4 lysozyme

Position	ΔT_m	ΔE_{calc}	Amino acids chosen by ORBIT		Best choice for stability
			No penalty	With penalty	
6	-10.6	-10.04	M	M	M
102	-2.4	-1.89	M	L	M
106	1.7	-1.45	M	L	L
120	1.7	1.28	L	L	L

ΔT_m is the change in melting temperature when methionine is replaced with leucine at the indicated position [24]. ΔE_{calc} is the change in the ORBIT energy score when methionine is replaced with leucine, absent a penalty term. A negative number indicates a predicted destabilization.

Figures

Figure 1. Ribbon diagram of Gβ1 [33]. Core residues are represented by spheres at the beta carbon position. Yellow spheres indicate positions that were constant in this study. Red spheres indicate positions 7, 39, and 54, which were varied during the study.

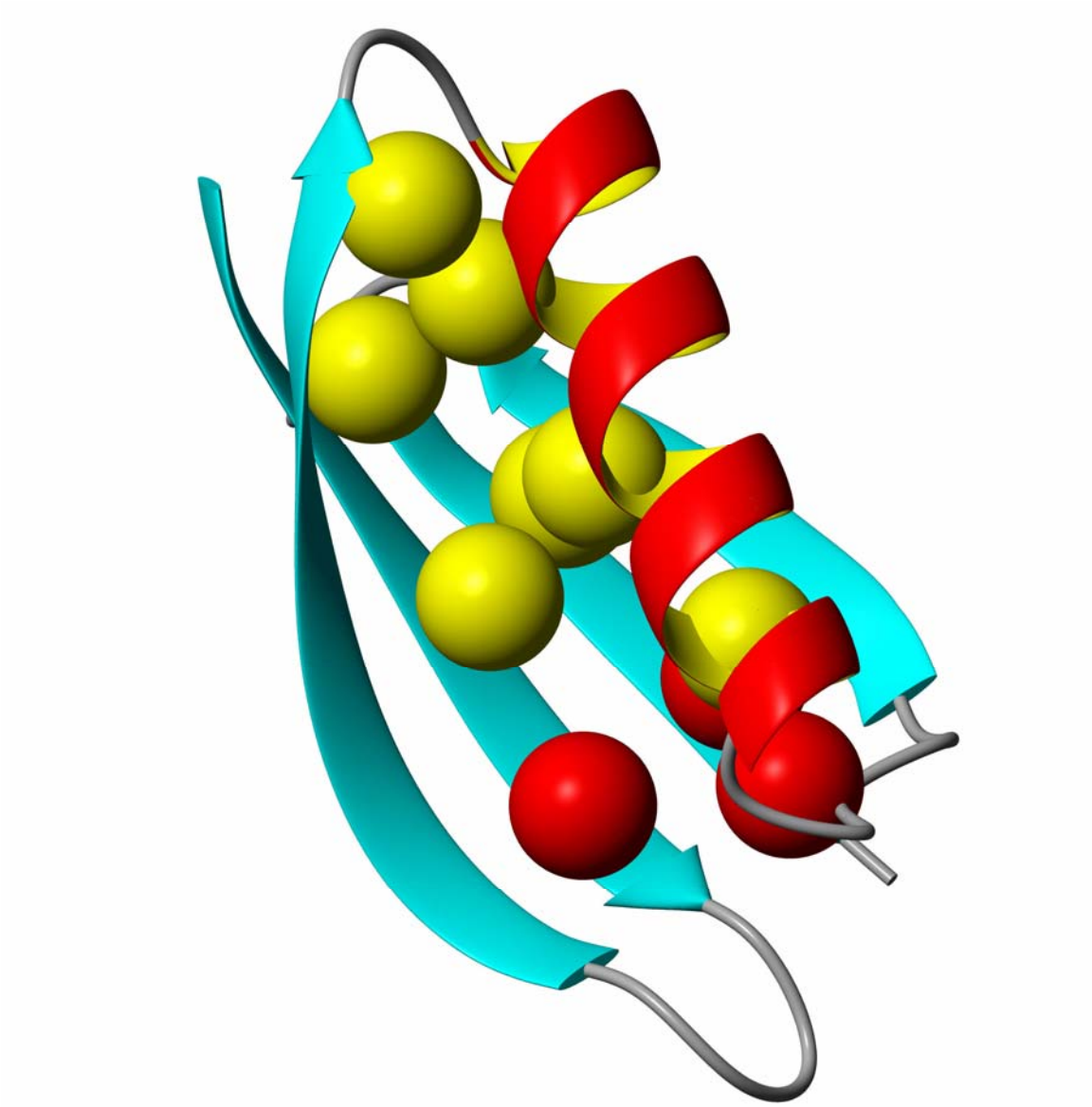
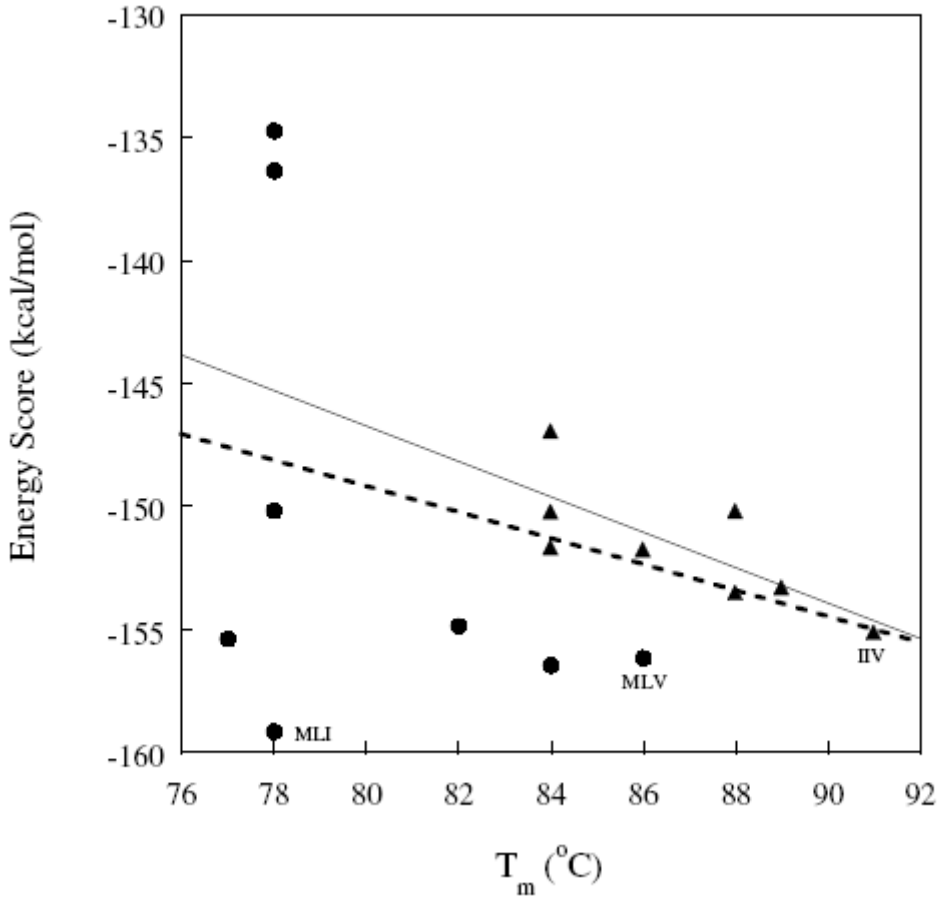


Figure 2. Correlation between experimental and computational results. Proteins containing one or more methionine residues in core positions are represented by filled circles. Proteins not containing a core methionine are indicated with triangles. (a) Correlation between the T_m and the calculated energy without adjustment for methionine content. The solid line indicates the correlation for only the eight sequences that do not contain methionine. The dashed line indicates the correlation when all 16 data points are included in the fit. (b) Correlation between the T_m and the calculated energy after application of a 10 kcal/mole methionine inclusion penalty. The dashed line represents the best linear fit when all 16 data points are included. Sequences discussed in the text are labeled on the graph.


-130
-135
-140
-145
-150
-155
-160
Energy Score (kcal/mol)
76
78
80
82
84
86
88
90
92
T m (°C)
MLI
MLV
IIV

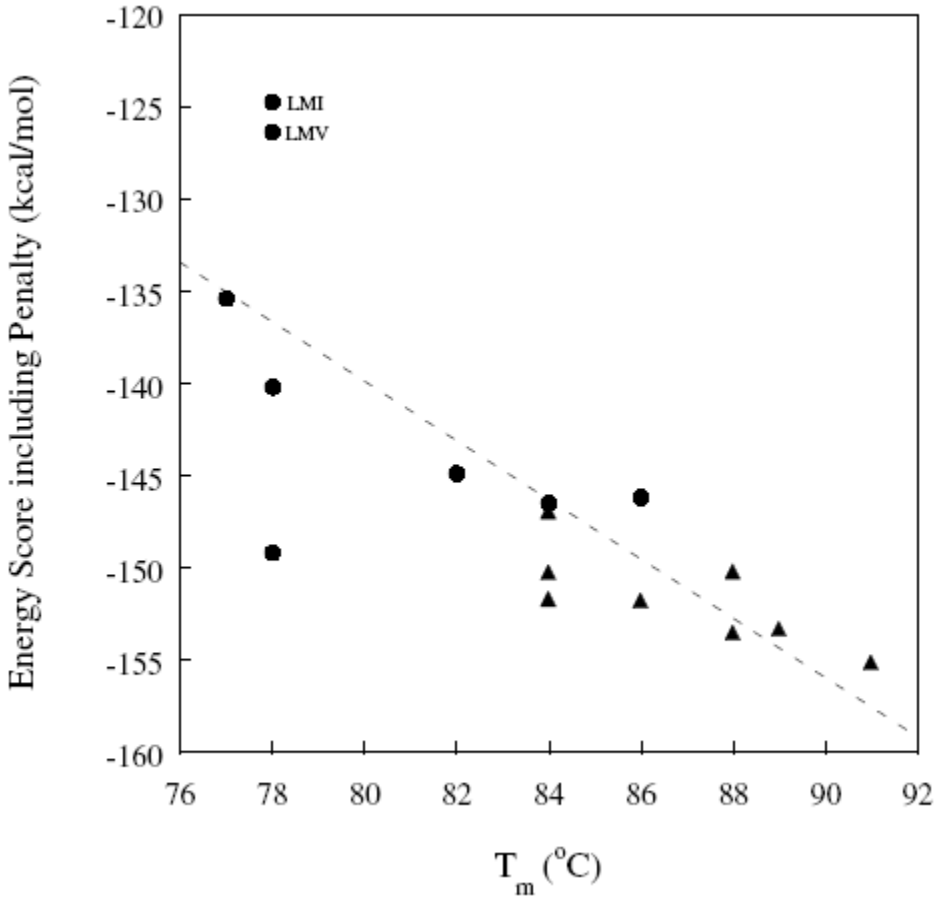
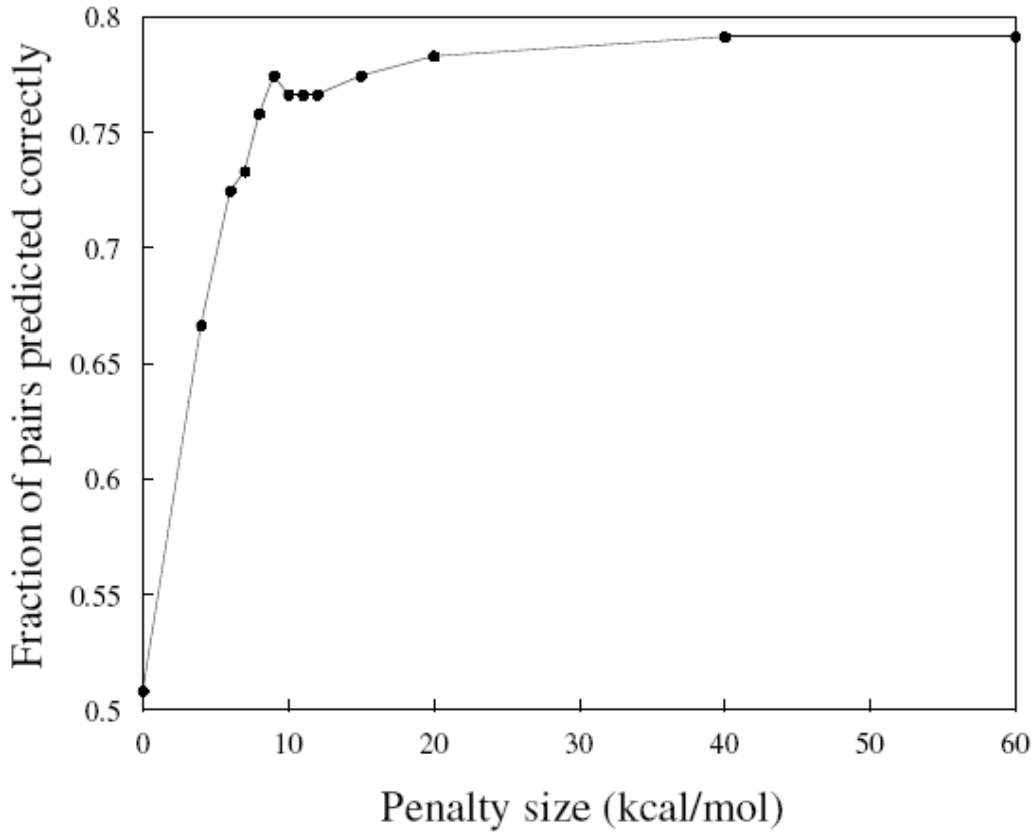

-120
-125
-130
-135
-140
-145
-150
-155
-160
LMI
LMV
Energy Score including Penalty (kcal/mol)
76
78
80
82
84
86
88
90
92
T m (°C)

Figure 3. Pair prediction improvements with the methionine inclusion penalty. The ability of the energy function to qualitatively predict the protein with the higher T_m was considered for each of the 120 possible sequence pairs. The 14 pairs with identical T_ms were counted as incorrect predictions.


Fraction of pairs predicted correctly
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0
10
20
30
40
50
60
Penalty size (kcal/mol)

4. DNA binding.

Abstract

The protein design process ORBIT (Optimization of Rotamers by Iterative Techniques) has been previously used to stabilize a number of proteins including several DNA binding proteins. However, these calculations have always been performed in the absence of DNA, and DNA binding has presumably been destroyed in the process. We now propose to use ORBIT to design proteins that target specific DNA sequences. Using the yeast transcription factor GCN4 as a model system, we will first computationally generate sequences predicted to bind with high affinity to the wild-type DNA target. These proteins will be expressed in *E. coli* and experimentally characterized by gel shift electrophoresis and DNAse I footprinting. We will then simultaneously computationally optimize the DNA and protein sequences with a number of different docking configurations, to generate a library of proteins with predicted preferred DNA targets. Some of these proteins will be experimentally characterized to demonstrate the utility of this approach. During this work, we expect to elucidate some of the factors that are important for DNA binding with site-specific recognition as well as to develop a methodology for generating novel proteins with high affinity for target DNA sites.

Introduction

DNA binding proteins have received a great deal of interest in recent years. Proteins that bind to DNA can serve purposes such as regulation of transcription, maintenance of

cellular DNA, DNA repair, and control of replication. Thus, proteins that target specific DNA sequences are potential therapeutics for genetic diseases and cancers. We are interested in developing methods to design small proteins to target any given DNA sequence.

Attempts to redesign DNA binding proteins to bind alternate DNA sequences have enjoyed only modest success to date. Zinc fingers have been frequent targets for redesign. Although some investigators have argued for a simple "code" that relates protein sequence to recognized DNA sequence, if there is indeed a code, it must be highly dependent on context. Recent studies have used a database approach, in which common sequence patterns are used to select sequences for a target DNA binding site (Kim and Berg 1995). Even this approach does not entirely succeed at selection of sequences with the correct specificity. Other efforts at designing DNA binding proteins include the use of phage display techniques on zinc fingers (Greisman and Pabo 1997, Jamieson *et al.* 1994), fusion of known DNA binding domains (Pomerantz *et al.* 1995), and the "grafting" technique of Zondlo and Schepartz (1999) to introduce GCN4 monomer-like binding to avian pancreatic polypeptide.

We are interested in developing general methods to computationally select protein sequences that recognize a target DNA sequence. We will avoid the use of motif-specific knowledge, relying instead on a force field developed for use in the protein design process ORBIT. The force field used in ORBIT includes terms for van der Waals contacts, solvation (a benefit for burial of hydrophobic surface area, a penalty for burial of polar surface area or exposure of hydrophobic surface area), electrostatics, and

hydrogen bonding. Binary patterning, restriction of the identities of side-chains at certain positions to either hydrophobic or hydrophilic residues, results in improved search time and more uniquely folded sequences. Selection of the optimal sequence within the force field is performed using the Dead End Elimination (DEE) algorithm. This algorithm can quickly and rigorously find the optimal sequence for problems in the range of 10^{30} sequences. ORBIT has been previously used to design small proteins, including DNA-binding motifs such as a zinc finger (Dahiyat *et al.* 1997b, Dahiyat and Mayo 1997), a homeodomain (Morgan *et al.* in preparation), and the dimerization domain of a bZIP protein (Dahiyat *et al.* 1997a). However, such designs have not retained DNA binding; the calculations have been run without DNA and residues that are known to be critical to DNA binding have been altered.

An essential feature in the ORBIT process is the use of a design cycle (Dahiyat and Mayo, 1996). The computational results are verified by experimental work, with modification of the force field when the experimental results are not well-predicted by the force field. The force field has been extensively optimized for use in designing proteins for stability. We will now test the existing force field for utility in designing proteins to specifically bind DNA, making modifications to the force field as the need is indicated by experimental results.

Goals

The eventual goal of this work is to develop a force field and design methodology that allow generation of protein sequences to bind to any target DNA sequence. In the

process, we will gain insight into the forces that are important for docking proteins to DNA. It may also be possible to use this knowledge to make predictions about the DNA binding sites of DNA binding proteins that have not yet been experimentally characterized.

The GCN4 bZIP transcription factor element is an attractive target for computational design. GCN4 is a parallel homodimer consisting of two long helices that form a leucine zipper at the C-terminal ends, but separate at the N-terminus to bind DNA in the major groove, as shown in Figure 1. GCN4 and other bZIP proteins bind to palindromic or pseudopalindromic DNA sequences, recognizing seven or eight base pairs in a sequence-specific fashion (Hope and Struhl 1985). Several crystal structures of GCN4 bound to DNA recognition sites are available, most of the contacts to the base pairs are direct rather than water mediated, and minimal distortion of the DNA occurs upon protein binding (Ellenberger *et al.* 1992, Keller *et al.* 1995). The bZIP element of GCN4 is also a desirable target because its small size, absence of cysteines, and absence of cofactors makes it a good candidate for over-expression in *E. coli* and subsequent purification steps.

As a first step, we retained the native docking and target DNA sequence of GCN4. We used ORBIT to select side-chains for positions that make base-specific contacts with the DNA. Although it will eventually be interesting to change the side-chains that contact the DNA phosphate backbone, we have retained wild-type side-chains at those positions. In the absence of sequence-specific bending of the DNA (indirect readout), contacts to the phosphate backbone will favor binding of the protein to all DNA sequences with no

specificity. Thus, this first attempt is a test of the ability of ORBIT to select side-chains that make good contacts with the DNA bases when the docking is held in the wild-type conformation. Specificity is not considered at this point, and so it is possible that the designed sequences will have higher affinities for non-target sequences.

Using standard force field parameters, we have selected a sequence for experimental characterization. This sequence, shown in Table 1, is a double mutant ("TQ"). An additional sequence has been selected using slightly altered force field parameters. Because it appears that favorable van der Waals terms are overpowering the hydrogen bonding terms, we have also performed the calculation using only repulsive van der Waals terms. It is likely that specific hydrogen bonds are more valuable than non-specific van der Waals contacts for specific binding. The resulting sequence ("SSA") is a triple mutant, but this sequence closely resembles known bZIP sequences. In both cases, highly conserved residues N109 and R117 were selected in ORBIT. The selected rotamers are highly similar to the orientation observed in the crystal structure, as shown in Figure 4. We are also performing calculations where we modify the penalty for burial of polar hydrogen atoms on the DNA bases to encourage formation of hydrogen bond contacts to the bases. To determine which force field parameters result in protein sequences with the correct specificity, we will experimentally characterize each protein.

The next step after characterization of designed proteins that bind the wild-type DNA sequence is the design of a protein that binds an altered DNA sequence. For this part of the project, we will simultaneously design both the DNA binding residues on the protein and the DNA base pairs potentially contacted by those residues. To prevent the selection

of all G-C pairs due to the stronger interaction between these bases, we are using a modified guanine in the computational work. We have removed the exocyclic (N2) amino group from guanine, as shown in Figure 2. As this group can only be contacted via the minor groove, this change is unlikely to cause a change in the way the base interacts with proteins that bind in the major groove. In initial trials, we find that the interactions within this modified G-C base pair are equivalent to those within an A-T pair.

Simultaneous optimization of the DNA and protein may result in a new DNA target sequence, or may instead reproduce the wild-type DNA target. If the calculation yields a different DNA sequence from the wild-type, we will express the new protein and assay it for binding to the new DNA sequence. If the wild-type DNA sequence is selected, this suggests that too much information is contained in the docking orientation, and it will be necessary to alter the docking conformation to generate new DNA target sequences.

The docking conformation between the protein and DNA and the sequence of the protein are highly coupled. Because the lengths of the side-chains vary, different side-chains require different distances between the alpha carbons of the protein helix and the edges of the base pairs. Suzuki and Gerstein (1995) have studied a number of proteins that bind via helices in the major groove and provide information on the areas of conformational space that are commonly sampled by these complexes. We will generate a number of docking conformations within the constraints set by Suzuki and Gerstein, and then optimize the protein and DNA sequences for each conformation. This method will allow the generation of protein sequences that target a number of different DNA sequences.

With sufficient coverage of conformational space, we will be able to predict which sequences can be recognized with good selectivity by the GCN4 motif, and can in addition predict which sequences cannot be recognized by this motif. A possible further extension of this work would be to verify that the sequences that we predict are unrecognizable cannot in fact be recognized specifically by GCN4 variants created by phage display or other screening techniques.

Experimental progress

The gene for wild-type GCN4 was constructed by recursive PCR and ligated into pET-11a. Site directed mutagenesis was used to create the TQ and SSA mutants. No significant protein expression for the wild-type or either of the two mutants was observed following induction with IPTG. Mutation of residue 2 to the Lysine AAA codon improved protein expression in the wild-type, as suggested by Tom Ellenberger (personal communication, 2000). Satisfactory expression levels of the wild-type were obtained following a site directed mutagenesis to introduce the 2K mutation. The TQ and SSA mutants were subjected to another round of site directed mutagenesis to introduce the 2K mutation. The correct PCR products were confirmed by DNA sequencing, but over-expression has not yet been attempted.

Conclusions

We have described a general methodology for generation of novel proteins that bind site-specifically to DNA. This approach will allow us to elucidate the relative importance of various forces (such as hydrogen bonding, solvation, and electrostatics) to the

formation of protein-DNA complexes. The ability to target specific DNA sequences with small proteins may also prove useful for therapeutic or diagnostic purposes where binding of a specific DNA sequence is necessary.

References

Dahiyat, B. I. and Mayo, S. L. (1997). *De novo* protein design: Fully automated sequence selection. Science *278*, 82–87.

Dahiyat, B. I. And Mayo, S. L. (1996). Protein design automation. Prot. Sci *5*, 895–903.

Dahiyat, B. I., Gordon, D. B., and Mayo, S. L. (1997a). Automated design of the surface positions of protein helices. Prot. Sci *6*, 1333–1337.

Dahiyat, B. I., Sarisky, C. A. and Mayo, S. L. (1997b). De novo protein design: towards fully automated sequence selection. J. Mol. Biol. *273*, 789–796.

Ellenberger, T. E., Brandl, C. J., Struhl, K., and Harrison, S. C. (1992). The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted α helices: Crystal structure of the protein-DNA complex. Cell *71*, 1223–1237.

Greisman, H. A. and Pabo, C. O. (1997). A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. Science *275*, 657–661.

Hope, I. A. and Struhl, K. (1985). GCN4 protein, synthesized in vitro, binds *HIS3*

regulatory sequences: implications for general control of amino acid biosynthetic genes in yeast. Cell *43*, 177–188.

Jamieson, A. C., Kim, S.-H., and Wells, J. A. (1994). *In vitro* selection of zinc fingers with altered DNA-binding specificity. Biochemistry *33*, 5689–5695.

Keller, W., Konig, P. and Richmond, T. J. (1995). Crystal structure of a bZIP/DNA complex at 2.2 Å: Determinants of DNA specific recognition. J. Mol. Biol. *254*, 657–667.

Kim, C.A. and Berg, J. M. (1995). Serine at Position 2 in the DNA recognition helix of a Cys_2-His_2 zinc finger peptide is not, in general, responsible for base recognition. J. Mol. Biol. *252*, 1–5.

Morgan, C. S., Marshall, S. A., and Mayo, S. L. Incorporating helix dipole and N-capping effects into the surface design of an α-helical protein. *In preparation.*

Pomerantz, J. L., Sharp, P. A., and Pabo, C. O. (1995). Structure-based design of transcription factors. Science *267*, 93–96.

Suzuki, M. and Gerstein, M. (1995). Binding Geometry of α-helices that recognize DNA. Proteins *23*, 525–535.

Weiss, M. A., Ellenberger, T. E., Wobbe, C. R., Lee, J. P., Harrison, S. C., and Struhl, K. (1990). Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA. Nature *347*, 575–578.

Zondlo, N. J. and Schepartz, A. (1999). Highly specific DNA recognition by a designed miniature protein. J. Am. Chem. Soc. *121*, 6938–6939.

Table 1

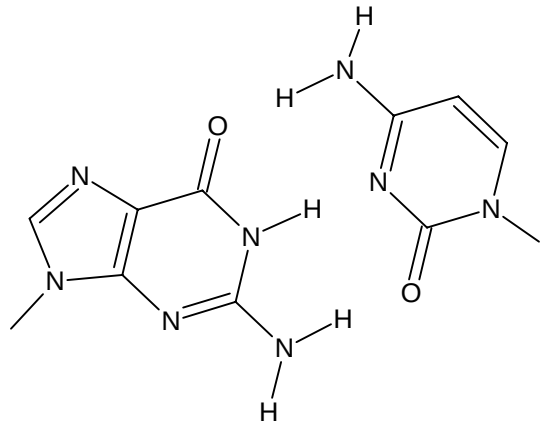
Comparison of two designed sequences to the wild-type GCN4 sequence and other bZIP sequences. The two residues (Asn 109 and Arg 117) that are conserved among bZIP proteins are reproduced by the calculations. The other three positions are more variable in the bZIP family and in the calculations. (bZIP proteins recognize a variety of DNA targets, so some of the variability in the family results from differences in DNA recognition sites.)

	109	112	113	116	117
Wild-type GCN4	Asn	Ala	Ala	Ser	Arg
"Typical" parameters	Asn	Thr	Gln	Ser	Arg
Repulsive vdW	Asn	Ser	Ser	Ala	Arg
Other bZIP	Asn	Ala	Ala	Cys	Arg
		Ser (20%)	Gln (8%)	Ser (36%)	
			Val (8%)	Phe (16%)	
			Ser (4%)		

Figures

Figure 1. Crystal structure of the bZIP element of GCN4 bound to the AP1 recognition site (Ellenberger *et al.* 1992). The N-terminus of each helix is at the left side of the figure.

Figure 2. (top) A G-C pair. (bottom) The G-C pair with the exocyclic amine of G removed.


H
N
H
O
N
N
H
N
N
N
H
N
H
H

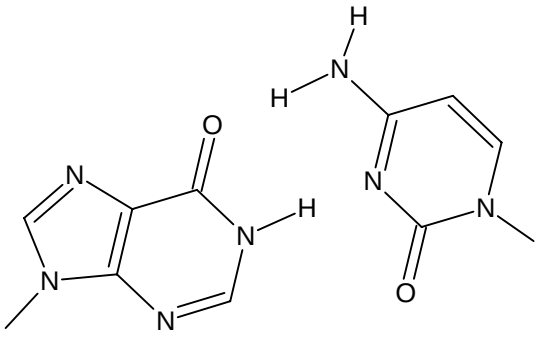
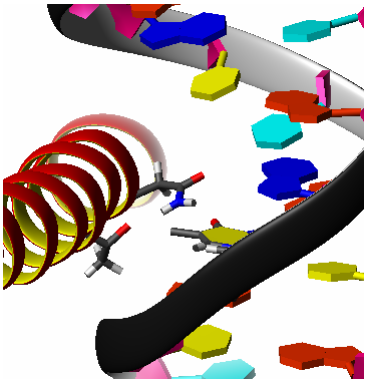
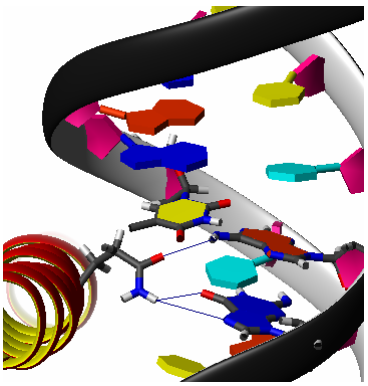

H
N
H
O
N
N
H
N
N
N
O

Figure 3. Interaction between designed side-chains and the DNA bases. Part (a) shows the new hydrogen bonding network formed between T112 in "TQ" and N109. (S112 in "SSA" forms a similar network.) Part (b) shows the contacts between side-chain Q113 and the recognition site for the double mutant "TQ". (c) shows the orientation of the designed serine 113 in "SSA."

(a)



(b)



(c)

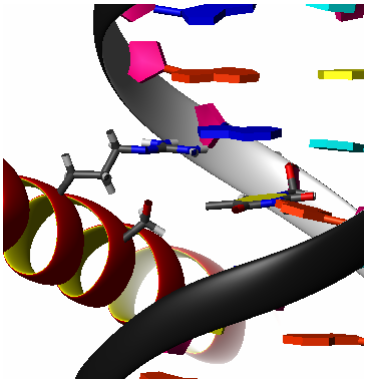


Figure 4. Retention of important contacts. The side-chain orientations present in the crystal structure and the side-chains selected in ORBIT are shown for N109 and R117.