STABILITY AND CONFORMATIONAL SPECIFICITY IN PROTEIN DESIGN: MODELS FOR BINARY PATTERNING AND ELECTROSTATICS

Thesis by

Shannon A. Marshall

In Parital Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2002

Defended 18 January 2002

Acknowledgements

I would like to thank the many people whose intellectual contributions, training, and support have made this thesis possible, First, I want to thank all of my collaborators. Barry Honig and his group, especially Emil Alexov and Walter Rocchia, helped in developing Poisson-Boltzmann methods for protein design calculations. Scott Ross has helped me with all aspects of determining the NMR structure of B6. Kevin Gardner and members of his group, especially Carlos Amezcua, helped Scott and I analyze our NMR data.

The Mayo lab members who preceeded me have taught me skills from molecular biology to programming, helped me define and execute the projects in this thesis, and helped me deal with the joys of graduate school. Alyce Su pointed me towards the binary patterning project as a way to address some of the problems presented by my first project. Julie Mayo helped me understand the mysteries of molecular biology, and now I never forget to sacrifice M&Ms before important experiments. Scott Ross has shared his enthusiasm for NMR, habeñaros, and travel.

I would especially like to thank Ben Gordon, Cathy Sarisky, and Chantal Morgan, for patiently answering my questions and helping me deal with the frustrations of graduate school. When Ben graduated, I was worried about what would become of the lab without him. Ben is a wonderful teacher who helped us all understand the theory behind protein design. Chantal was the first in the group to use homeodomain as an experimental system. The experimental work I have done has relied on her earlier work. Cathy, despite joining the group only six months before me, helped me with nearly all of the aspects of my work and was always willing to talk with me.

As the graduate students before me graduated and moved on with their lives, new people came into the group. Niles Pierce has been a wonderful resource for us all. The

conversations we had as I was starting to work on electrostatics were especially helpful. I would especially like to thank Shira Jacobson, Deepshika Datta, and Kirsten Lassilla for helping me maintain my sanity as the end has drawn near. Big thanks go to Cynthia Carlson, Rhonda Diguisto, and Marie Ary for making the lab a better place to work. You three take good care of us all- remembering our birthdays, helping to resolve conflicts, and providing a sympathetic ear to talk to.

Finally, I would like to thank Steve Mayo for being my advisor during graduate school. From the first time we talked, I knew that I wanted to study protein design. Steve has given me the freedom, support, and resources to pursue my projects successfully. I am very thankful for the professional support he has provided through the last few years.

I would also like to thank my family and friends in the outside world for helping me to and through graduate school. I'll share the mushy sentiments in person.

Abstract

Binary patterning (the arrangement of hydrophobic and polar amino acids) and electrostatics are important determinants of the stability and conformational specificity of designed proteins. We have developed methods to to select the optimal binary pattern and model electrostatics in protein design studies. The Genclass method of binary patterning uses a solvent accessible surface generated from backbone coordinates of the target fold and "generic" side chains, constructs whose size and shape are similar to an average amino acid. Each position is classified according to the solvent exposure of its generic side chain. The method was tested by analyzing several proteins in the Protein Data Bank and by experimentally characterizing homeodomain variants whose binary patterns were systematically varied. Selection of the optimal binary pattern results in a designed protein that is monomeric, well-folded, and hyperthermophilic. Homeodomain variants with fewer hydrophobic residues are destabilized, additional hydrophobic residues induce aggregation. The optimal variant was further characterized by nuclear magnetic resonance spectroscopy. Binary patterning, in conjunction with a force field that models folded state energies, appears sufficient to satisfy two basic goals of protein design: stability and conformational specificity.

Electrostatic interactions are critical determinants of protein structure and function. Computational protein design algorithms typically use fast methods based on Coulomb's law to model electrostatic interactions. These methods fail to accurately account for desolvation and solvent screening, which strongly attenuate electrostatic interactions in proteins. Using the current force field, we designed a 25-fold mutant with moderate stability similar to the wild type protein. Incorporating two classes of electrostatic interactions using simple rules yielded a nine-fold mutant of the initial design that is over 3 kcal mol^{-1} more stable. The simple electrostatic model used in the ORBIT force field is unable to predict the experimentally determined stabilities of the designed variants. Finite difference Poisson-Boltzmann (FDPB) methods have substantially better predictive power, but are far too slow for problems with high combinatorial complexity. We have developed new strategies for modeling electrostatics in protein design problems that utilize one- and two-body decomposable FDPB methods. Computational results indicate that this method has the accuracy and speed required for design calculations.

Table of Contents

List of Figures and Tables

Figures

Tables

Chapter I

Introduction to Protein Design, Binary Patterning, and Electrostatics

Introduction to Protein Design

The protein design problem asks which amino acid sequences are capable of forming a desired protein fold. More recently, the goals of protein design have expanded to the identification of amino acid sequences that will possess desired physical, chemical, and / or biological properties. Protein design can be driven by practical goals, such as developing catalysts for industrial processes and designing therapeutic agents. In addition, protein design has proved to be a valuable basic research tool for probing the links between protein sequence, structure, and function. A variety of approaches have been used to tackle the protein design problem. Heuristic, or rules-based, approaches have been used successfully to design highly symmetric coiled-coil structures. *In vitro* evolution procedures work quite well for modulating the activity of enzymes and identifying small peptides that bind desired ligands. Our research has focused on a third approach, computational protein design.

The number of possible amino acid sequences for even a small protein is extraordinarily large. It would take more matter than exists in the universe to generate all possible 100 amino acid sequences, and the average protein is more than twice as long. Using experimental methods, it is only possible to sample an insignificantly small fraction of sequence space. Computational protein design methods address this fundamental limitation by using computational rather than experimental procedures to identify protein sequences that are capable of folding to a target structure and possessing desired properties.

Computational protein design algorithms, such as ORBIT, typically comprise four steps[1]. First, the protein structure is modeled. The backbone structure is generally based on the crystal structure of a known protein, and a set of discrete amino acid conformations,

called rotamers, are used to describe the side chains. In addition, the list of amino acids that will be considered at each position is generated in the modeling stage. Next, side chain internal, side chain - backbone, and side chain - side chain energies are calculated using a force field, as discussed more thoroughly in Chapter II. Combinatorial search algorithms such as dead-end elimination and branch and bound are used to identify the optimal amino acid sequence. Finally, the selected sequences are characterized experimentally and the results are used to improve the protein design methodology.

Unsolved Problems in Protein Design

Several groups have successfully used computational protein design methods to redesign the hydrophobic cores of a variety of small proteins[2-5]. Accurate modeling of packing interactions seems to be the key to core design. Designing the solvent exposed surface and partially exposed boundary residues has proved more challenging. With the exception of highly symmetric helical bundle and coiled-coil domains, there was only one successful computational full sequence design reported at the start of my graduate studies[6]. In addition, a large fraction of the protein G boundary residues had been redesigned, yielding a hyperthermophilic variant[7].

At the time I began graduate school, there were four main unsolved problems in protein design that limited our ability to select sequences that would fold to the target structure and exhibit reasonable stability. These questions were: (1) how to ensure that selected sequences will fold to the target structure, rather than a misfolded or aggregated state (or the negative design problem), (2) how to account for flexibility in each protein sequence, as well as changes in backbone structure that result from changes in sequence, (3) how to select sequences for beta sheet surfaces, and (4) how to model electrostatic interactions in design calculations. In addition to being important for the proximal goal of designing stable, well

folded proteins, finding solutions to these questions is likely to be critical for designing proteins with desired functional properties.

Using Binary Patterning as a Negative Design Tool

At the start of graduate school, I worked on designing the core and boundary residues of a SH3 domain, as described in Appendix A. The designed SH3 variants were often destabilized and sometimes not well-folded. Although this project was not directly successful, it did suggest a direction for a second project. In the boundary calculations, we considered both hydrophobic and polar residues at the variable positions. The calculated sequences tended to either be overly polar and unstable or overly hydrophobic and not well-folded. The binary patterning project, described in Chapter III, arose from an attempt to determine the optimal pattern of hydrophobic and polar residues for a target structure at the start of a protein design calculation.

During the course of the project, we realized that binary patterning was the answer to a bigger problem than designing SH3 domains. Binary patterning can also be used to help ensure that designed proteins fold to the target structure rather than an alternate fold or misfolded state. One criticism that ORBIT and other computational design methods have faced is that their force fields only consider folded state energy, while protein stability is determined by the energy difference between the folded and unfolded states. According to the Random Energy Model developed by Shaknovich and coworkers[8], unfolded state energies are determined by the hydrophobic versus polar composition of the protein chain. Since all sequences with the same binary pattern have roughly the same composition, comparing the folded state energies of sequences with the same binary pattern should be sufficient to identify stable sequences.

A second project that arose from the binary patterning project was structure determination of a fully designed homeodomain variant, discussed in Chapter IV. In the experimental segment of the binary patterning project, I selected sequences for the 11 boundary positions and used the core and boundary sequences optimized previously by Chantal Morgan, a former graduate student in the Mayo group. As a result, the proteins in the binary patterning project were fully designed. Only a small handful of high resolution structures have been solved for fully designed proteins, so one goal of the project was to contribute to this short list. In addition, examination of the structure should allow us to assess the accuracy of several components of the design process.

Modeling Electrostatics in Protein Design Calculations

The results of the binary patterning study allow us to determine which positions should be hydrophobic versus polar, and previous design studies have established methods for selecting among the hydrophobic residues. However, a general solution for surface design problems had not yet been found. Electrostatic interactions, including hydrogen bonds, can make a significant contribution to the folded state interactions among surface side chains. However, protein design force fields have used very simplistic electrostatic models.

In Chapter V, we use the surface of the engrailed homeodomain as an experimental system to determine the effects of using a highly approximate electrostatic model in computational design studies. The initial calculations and experimental work on this project were performed by Chantal Morgan. She found that restricting sequence composition to select for helix N-capping interactions and to select against unfavorable side chain - helix dipole interactions yielded a protein that was significantly more stable than one designed allowing all polar residues at all of the surface positions. Since the two designed proteins differed at nine positions, it was difficult to identify the source of the stability difference

between the two proteins. I characterized four additional proteins to demonstrate that both helix dipole and N-capping interactions can contribute significantly to the stability of designed proteins. In addition, I used the finite difference Poisson-Boltzmann (FDPB) model, which is a well respected continuum electrostatic model, to analyze the limitations of the electrostatic model that had been used in the initial design calculations.

The electrostatic models that have been used for design calculations underestimate the importance of side chain - backbone interactions relative to side chain - side chain hydrogen bond and salt bridge interactions. Another electrostatic effect that would not be captured in the simple models used for design is cation-π interactions. Gallivan and Dougherty proposed that cation-π interactions may stabilize proteins more than salt bridge interactions[9], but the contribution of cation-π interactions to protein stability had not yet been experimentally determined. We used double mutant cycle analysis to measure the interaction between an (i, i+4) Arg-Trp pair on the helical surface of both protein G and homdeodomain, as described in Appendix B. The studies were inconclusive, but improvements in the electrostatic model used in design calculations, described below, should capture the energetic benefit of cation-π interactions.

The FDPB calculations used to analyze electrostatic interactions in the designed homeodomain surfaces could predict relative protein stabilities significantly better than the original ORBIT calculations. As a result, improvements to the ORBIT electrostatic model could be obtained by maximizing the agreement between the energies produced by ORBIT and FDPB energies. Using this approach, we optimized the values for dielectric constants and solvation parameters used in design calculations, as described in Chapter VI. However, even using these optimized parameters, a simple electrostatic model based on Coulomb's law and surface area based solvation parameters does not recapitulate the FDPB results.

So, we next worked to develop FDPB methods that are compatible with the requirements of design calculations.

As typically implemented, FDPB calculations would be far too slow to use for design. The full three-dimensional structure of the protein is used in FDPB calculations to define the boundary between the high dielectric solvent and the low dielectric protein. Since each possible arrangement of protein side chains will produce a slightly different dielectric boundary, FDPB calculations would need to be run for each rotameric sequence, which would require about 10^{60} years for the homeodomain surface design case. As described in Chapter VII, we have developed simplified surface descriptions that only require knowledge of the identity and conformation of one or two side chains at a time and allow rapid calculation of energies that correlate quite well with the results of FDPB calculations.

Conclusions

In protein core design, a force field that maximizes packing interactions and hydrophobic burial is generally sufficient for the design of stable, well-folded proteins. Designing the boundary and surface positions requires careful balancing of competing forces instead. While the burial of hydrophobic atoms in boundary residues can confer stability, incorporating too many hydrophobic residues results in protein aggregation. Similarly, maximizing salt bridge interactions between surface side chains does not necessarily optimize stability, as the effects of desolvation and side chain - backbone electrostatic interactions must also be considered. As protein design efforts begin to focus increasingly on activity, finding the right balance between the forces that contribute to structure, stability, and function is likely to become increasingly important.

References

1. Street, A. G. and Mayo, S. L. (1999). Computational protein design. *Structure*, **7**, R105-R109.
2. Dahiyat, B. I. and Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci., USA*, **94**, 10172-10177.
3. Desjarlais, J. R. and Handel, T. M. (1995). De novo design of the hydrophobic cores of proteins. *Protein Sci.*, **4**, 2006-2018.
4. Jiang, X., Bishop, E. J. and Farid, R. S. (1997). A de novo designed protein with properties that characterize natural hyperthermophilic proteins. *J. Am. Chem. Soc.*, **119**, 838-839.
5. Hellinga, H. W. and Richards, F. M. (1994). Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci., USA*, **91**, 5803-5807.
6. Dahiyat, B. I. and Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science*, **278**, 82-87.
7. Malakauskas, S. M. and Mayo, S. L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.*, **5**, 470-475.
8. Shakhnovich, E. I. and Gutin, A. M. (1993). A new approach to the design of stable proteins. *Protein Eng.*, **6**, 793-800.
9. Gallivan, J. P. and Dougherty, D. A. (2000). A computational study of cation-π interactions vs. salt bridges in aqueous media: implications for protein engineering. *J. Am. Chem. Soc.*, **122**, 870-874.

Chapter II

Energy Functions for Protein Design

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

D. B. Gordon, S. A. Marshall, and S. L. Mayo. (1999). *Curr. Opin. Struct. Biol.*, **9**, 509-513.

Introduction

Computational protein design is a general, closed-loop approach for finding the optimal sequence of amino acids for a desired protein fold[1]. A potential energy function that represents the dominant factors, as well as the subtleties, of protein stability is used to predict the energy of each possible amino acid sequence on a target protein structure. Current design efforts have used fixed protein backbones as target structures, with two notable exceptions[2-4]. Atomic level detail is introduced by using statistically significant sidechain conformations, called rotamers[5], to represent the flexibility of each amino acid. A variety of stochastic and deterministic search algorithms are then used to find the optimal combination of amino acid sidechain rotamers on the target structure as ranked by the potential energy function. Finally, the experimentally determined stability and structure of designed proteins are analyzed and rational improvements to the potential function are implemented.

The purpose of this review is to discuss the development of protein design force fields and to survey the potential energy terms that have been used thus far. The terms fall into five broad categories. First, we discuss the energies describing packing between atoms that are not covalently bonded. Nonbonded polar interactions are considered next. We briefly

survey internal coordinate energies, and finally examine solvation and entropy, which are computed differently than in typical molecular mechanics force fields.

Force Field Requirements

Protein design presents a demanding task for a potential energy function. Design potentials must be sensitive to subtle changes in amino acid identity that are known to perturb the experimental stability of proteins. However, design force fields should not be overly sensitive to small variations in rotamer geometry, since discrete rotamers are used to model sidechain conformations. The force field also must be compatible with the computational requirements of protein design. For example, most search algorithms demand that energy terms be pairwise decomposable, and design problems with large combinatorial complexity require energy terms that can be calculated quickly.

Because the energies produced by design potentials are intended to correlate with the free energy of folding, the force field must also model the unfolded state as well as the folded state. Experimental and theoretical studies[7] indicate that unfolded proteins can sometimes have residual structure, and mutations may alter the properties of the unfolded state ensemble. However, in design calculations, the unfolded state is commonly assumed to have no residual structure: nonbonded interactions between sidechains are considered to be insignificant, the sidechains are assumed to be fully solvated, all rotamers are modeled as being equally probable, and all sequences in the unfolded state are isoenergetic.

Due to the demands posed by protein design, force fields that are widely used to perform molecular mechanics calculations, such as CHARMM[8], AMBER[9-10], and DREIDING[11], are not necessarily appropriate for design. Similarly, statistically derived pair potentials that are quite effective in structure compatibility studies[12] do not manifest the structural sensitivity necessary for protein design. Instead, new force fields must be developed for protein design

that properly balance each factor described by the potential energy function. Over the past few years, the first force fields tailored for design have been constructed. However, very few potential energy terms have been used in these force fields, and even fewer have been evaluated through comparison of design predictions and experimental results. Future progress in protein design force fields will be realized by continued systematic experimental validation of the terms comprising the potential function.

van der Waals

Packing specificity is critical for protein design. For protein core calculations, which comprise the majority of design studies, a force field that models only packing specificity is sufficient to design well-folded proteins[13-16]. Although packing can be evaluated exclusively with interatomic distance restraints[17], most design programs utilize a van der Waals potential. This potential provides a physical basis for sidechain packing specificity, thereby favoring native-like folded states with well-organized cores and selecting against disordered or molten globule states. The van der Waals energy is typically calculated with a Lennard-Jones 12—6 expression.

$$E_{vdW} = D_0\left[\left(\frac{R_0}{R}\right)^{12} - 2\left(\frac{R_0}{R}\right)^{6}\right] \tag{1}$$

The interatomic distance, R, is computed from atomic coordinates. The equilibrium radii, R_0, and well-depths, D_0, are parameters that are defined within each force field.

Two examinations of van der Waals parameters underscore the need to tune molecular mechanics potential functions for protein design. Lazar and coworkers[16] compared the predictive ability of variations of Hagler and AMBER van der Waals parameters for a set of ubiquitin variants with redesigned cores. United atom parameters from AMBER95 were

markedly superior to the other variations when used in conjunction with a detailed rotamer library. Dahiyat and Mayo[15] generated sequences by systematically varying the scale of the atomic radii, based on the DREIDING parameter set and using rotamers with explicit hydrogen atoms. Scaling the radii by a factor of 0.90 achieved the optimal balance between packing specificity and hydrophobic collapse, as represented by a solvation term (discussed in a later section).

Hydrogen Bonding

Because the majority of computational protein design studies have focused on protein cores, electrostatic and hydrogen bonding terms have not been as thoroughly validated by experiment. Nevertheless, initial forays have proven these terms useful for the design of helical surfaces[18] and for full sequence design[19].

Hydrogen bonds are typically represented with an angle-dependent, 12-10 hydrogen bond potential,

$$E_{HB} = D_0\left[5\left(\frac{R_0}{R}\right)^{12} - 6\left(\frac{R_0}{R}\right)^{10}\right]F(\theta), \qquad (2)$$

where R_0 is the equilibrium distance, D_0 is the well depth, and R is the interatomic distance between donor and acceptor heavy atoms. The angle dependence term, $F(\theta)$, is typically $\cos^4\theta$, where θ is the donor-hydrogen-acceptor angle.

We have observed that calculations performed with the above potential will allow rotameric arrangements with non-physical hydrogen bond geometries, as shown in Figure I-1. To circumvent this problem, we employ more restrictive hybridization-dependent angle-dependence terms that enforce reasonable geometries[18].

sp^3 donor – sp^3 acceptor	$F = \cos^2\phi\cos^2(\theta\text{-}109.5°)$;	$\theta > 90°$, $\phi\text{-}109.5° < 90°$	(3)
sp^3 donor – sp^2 acceptor	$F = \cos^2\theta$;	$\phi > 90°$	(4)
sp^2 donor – sp^3 acceptor	$F = \cos^4\theta$		(5)
sp^2 donor – sp^2 acceptor	$F = \cos^2\phi\cos^2(\max[\phi,\varphi])$		(6)

The angles ϕ and φ refer to the hydrogen-acceptor-base angle (where the base is the atom covalently attached to the acceptor) and the angle of between the normals of the planes defined by the six atoms attached to the two sp^2 centers, respectively.

A potential energy term based on the above equations allows only physically reasonable side chain - side chain and side chain - backbone hydrogen bonds. Unfortunately, using a highly restrictive energy term in combination with a discrete rotamer library causes the force field to predict poor energies for some sequences that may actually form good hydrogen bond interactions.

Electrostatics

The role of electrostatics in protein stability is subject to debate. At moderate temperatures, favorable electrostatic interactions are not thought to be strong enough to compensate for the energy of desolvation[20]. In more extreme conditions, however, salt bridges may stabilize proteins[21-22]. Moreover, electrostatics may play a more significant role in defining the specificity, rather than the stability, of folding and of functional interactions[23-26].

Computational protein design efforts have not yet developed an electrostatic term intended to represent these considerations. Rather, electrostatics are used sparingly, primarily to guard against destabilizing interactions between like-charged residues. The simplest

treatment of electrostatic interactions is based on Coulomb's Law, which describes the energy of two charges, Q_i and Q_j, separated by distance, R, in a medium with dielectric constant, ε:

$$E_{elec} = 332.0637\left(\frac{Q_iQ_j}{\varepsilon R}\right) \quad (7)$$

We use a distance-attenuated version of Coulomb's law with an effective dielectric constant value of $40R$ and partial atomic charges that give a total coulombic energy of approximately ±1 kcal mol^{-1} for the interaction between juxtaposed charged residues. Thus, electrostatic contributions to the total energy are only significant when charged atoms are in close proximity. In sharp contrast, electrostatic energy is often the largest contributor to the total energy in potentials used for molecular mechanics and dynamics calculations.

Internal Coordinate Terms

Typical molecular mechanics force fields have terms that evaluate bonds, angles, torsions, and inversions among atoms that are covalently attached. These internal coordinate or "bonded" energies must be considered when generating rotamers or modifying the protein backbone, and have been used for protein design in some cases[4,16]. The usefulness of these terms for design, however, has not been rigorously demonstrated. Since rotamers derived from statistical analysis of protein structure databases generally have good internal coordinate energies, many design potential functions do not include them at all.

Solvation

Because the hydrophobic effect drives protein folding[27], modeling solvation effects is critical for a protein design force field. However, the computational expense of explicitly modeling protein/solvent interactions for all sequences under consideration is prohibitively expensive. Therefore, several groups have employed approximate methods utilizing octanol-

water and gas-water free energy of transfer data for each amino acid[28-29]. The experimentally measured free energies of transfer are correlated with the molecular surface area[30], shown in Figure I-2. These energies are either used directly for residues in the protein core[31] or they are scaled by the change in solvent exposed surface area associated with protein folding[14,32].

The energy required to transfer a sidechain from a solvated, unfolded protein to a partially or completely desolvated position in the folded protein is not necessarily the same as the transfer energy from water to gas or a nonpolar solvent. But, the approximate linear relationship between transfer energy and change in surface area should be correct for both cases. Dahiyat and Mayo[14] determined the optimal values for polar and nonpolar atomic solvation parameters by fitting to the experimentally determined stability of designed proteins. Inclusion of a hydrophobic burial benefit and a polar burial penalty in the protein design force field provides a significant improvement in predictive power compared to a force field with only a van der Waals term.

Two other considerations have affected the formulation of a protein design solvation potential. First, a negative design term that penalizes exposure of nonpolar surface area is sometimes used[15,33]. Although nonpolar exposure should not destabilize a protein, it can lead to aggregation or misfolding. Therefore, a nonpolar exposure penalty is required to limit the amount of exposed nonpolar surface area at boundary and surface positions[34]. Second, many optimization algorithms require that energy terms be pairwise decomposable, but pairwise calculation of buried surface areas leads to significant overcounting. Street and Mayo have developed a pairwise expression with one scalable parameter that closely reproduces both the true buried area and the true exposed solvent accessible surface areas[35].

Entropy

A simple entropy term is sometimes incorporated into protein design potential functions[31,32]. The change in sidechain entropy upon folding is modeled as the change in number of rotatable bonds, making the assumption that conformational freedom is completely restricted in the folded state. The unfolded state entropies are calculated either by assuming that all rotamers are equally populated or by fitting to semi-empirical estimate[36]. Inclusion of an entropy term based on the number of rotatable bonds did not significantly improve correlation between predicted and observed stabilities of the GCN4-p1 coiled coil core[14]. This simple model for entropy may have failed because it neglects residual sidechain entropy in folded proteins, as well as possible residual structure in the unfolded state.

Looking Forward

Protein design force fields have been successful, in part, because of their stringency. Restrictive functions such as the van der Waals and the hybridization-dependent hydrogen-bond potential, in particular, result in a very high rejection rate, and a significant false-negative rate. Fortunately, many design force fields also show a low false-positive rate. Therefore, sequences that are selected in protein design studies tend to fold properly, even though many other equally acceptable sequences are rejected.

Because of the high false-negative rate, potential functions derived through protein design efforts may not be suitable for folding studies. To gain a deeper understanding of the determinants of protein stability, it is therefore important to lower the false-negative rate. Softening of the restrictive potentials could result in design models that more accurately describe the fundamental relationship between sequence, structure, and stability.

Acknowledgements

We wish to thank A. G. Street for helpful comments on the manuscript. This work was supported by the Howard Hughes Medical Institute (S. L. M.), the Helen G. and Arthur McCallum Foundation (D. B. G.), an NIH NRSA Training Grant, and the Caltech Initiative in Computational Molecular Biology program awarded by the Burroughs Wellcome Fund (S. A. M.).

References

1. Street A. G. and Mayo S. L. (1999). Computational protein design. *Structure*, **7**, R105-R109.
2. Harbury P. B., Tidor B., and Kim P. S. (1995). Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc. Natl. Acad. Sci. U S A*, **92**, 8408-8412.
3. Su A. and Mayo S. L. (1997). Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci.*, **6**, 1701-1707.
4. 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.
5. Ponder J. W. and Richards F. M. (1987). Tertiary templates for proteins - use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.*, **193**, 775-791.
6. Desjarlais J. R. and Clarke N. D. (1998). Computer search algorithms in protein modification and design. *Curr. Opin. Struct. Biol.*, **8**. 471-475.
7. Dill K. A. and Shortle D. (1991). Denatured states of proteins. *Annu. Rev. Biochem.*, **60**, 795-825.
8. Brooks B. R., Bruccoleri R. E., Olafson B. D., States D. J., Swaminathan S., and Karplus M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, **4**, 187-217.
9. Weiner S. J., Kollman P. A., Case D. A., Singh U. C., Ghio C., Alagona G., Profeta S. J., and Weiner P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, **106**, 765-784.

10. Cornell W. D., Cieplak P., Bayly C. I., Gould I. R., Merz K. M., Jr., Ferguson D. M., Spellmeyer D. C., Fox T., Caldwell J. W., and Kollman P. A. (1995). A second-generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, **117**, 5179-5197.

11. Mayo S. L., Olafson B. D., and Goddard W. A., III (1990). Dreiding - a generic force-field for molecular simulations. *J. Phys. Chem.*, **94**, 8897-8909.

12. Bowie J. U., Luthy R., and Eisenberg D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164-170.

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

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

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

16. Lazar G. A., Desjarlais J. R., and Handel T. M. (1997). *De novo* design of the hydrophobic core of ubiquitin. *Protein Sci.*, **6**, 1167-1178.

17. Jiang X., Bishop E. J., and Farid R. S. (1997). A *de novo* designed protein with properties that characterize natural hyperthermophilic proteins. *J. Am. Chem. Soc.*, **119**, 838-839.

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

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

20. Hendsch Z. S. and Tidor B. (1994). Do salt bridges stabilize proteins- a continuum electrostatic analysis. *Protein Sci.*, **3**, 211-226.

21. Elcock A. H. (1998). The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. *J. Mol. Biol.*, **284**, 489-502.

22. de Bakker P. I. W., Hunenberber P. H., and McCammon J. A. (1999). Molecular dynamics simulations of the hyperthermophilic protein Sac7d from *Sulfolobus acidocaldarius*: contribution of salt bridges to thermostability. *J. Mol. Biol.*, **285**, 1811-1830.

23. Lumb K. J. and Kim P. S. (1995). A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry*, **34**, 8642-8648.

24. Schneider J. P., Lear J. D., and DeGrado W. F. (1997). A designed buried salt bridge in a heterodimeric coiled coil. *J. Am. Chem. Soc.*, **119**, 5742-5743.

25. Sindelar C. V., Hendsch Z. S., and Tidor B. (1988). Effects of salt bridges on protein structure and design. *Protein Sci .*, **7**, 1898-1914.

26. Spek E. J., Bui A. H., Lu M., and Kallenbach N.R. (1998). Surface salt bridges stabilize the GCN4 leucine zipper. *Protein Sci.*, **7**, 2431-2437.

27. Dill K. A. (1990). Dominant forces in protein folding. *Biochemistry*, **29**, 7133-7155.

28. Fauchère J-L. and Plicska V. (1983). Hydrophobic parameters of amino-acid side-chains from the partitioning of n-acetyl-amino-acid amides. *Eur. J. Med. Chem.*, **18**, 369-375.

29. Ooi T., Oobatake M., Nementhy G., and Scheraga H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. U S A* **84**, 3086-3090.

30. Wesson L. and Eisenberg D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci.*, **1**, 227-235.

31. Hellinga H. W. and Richards F. M. (1994). Optimal sequence selection in proteins of known structure by simulated evolution. *Proc. Natl. Acad. Sci. U S A*, **91**, 5803-5807.

32. Kono H., Nishiyama M., Tanokura M., and Doi J. (1998). Designing the hydrophobic core of *Thermus flavus* malate dehydrogenase based on Side-chain packing. *Protein Eng.*, **11**, 47-52.

33. Sun S., Brem R., Chan H. S., and Dill K. A. (1995). Designing amino acid sequences to fold with good hydrophobic cores. *Protein Eng.*, **8**, 1205-1213.

34. Malakauskas S. M. and Mayo S. L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nat. Struct. Biol.*, **5**, 470-475.

35. Street A. G. and Mayo S. L. (1998). Pairwise calculation of protein solvent accessible surface areas. *Fold. Des.*, **3**, 253-258.

36. Sternberg M. J. E. and Chickos J. S. (1994). Protein side-chain conformational entropy derived from fusion data - comparison with other empirical scales. *Protein Eng.*, **7**, 149-155.

Figure II-1. An example of a non-physical hydrogen bond geometry that can be selected when a hydrogen bond potential dependent only on θ is used for protein design. A more restrictive hydrogen bond potential, described in Equations 2 through 6, correctly predicts that no favorable interaction is present because $\phi = 90°$.

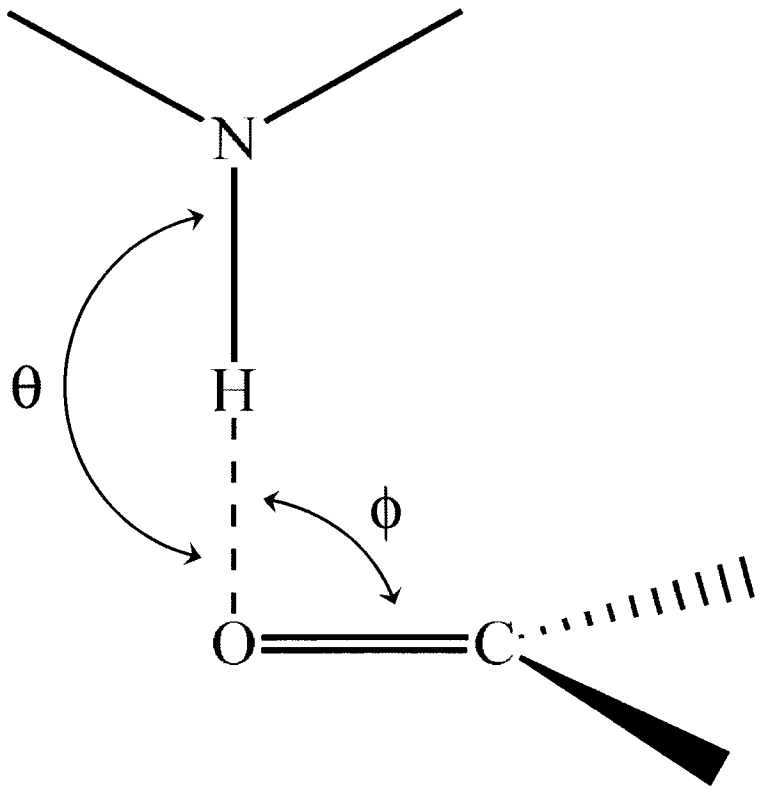
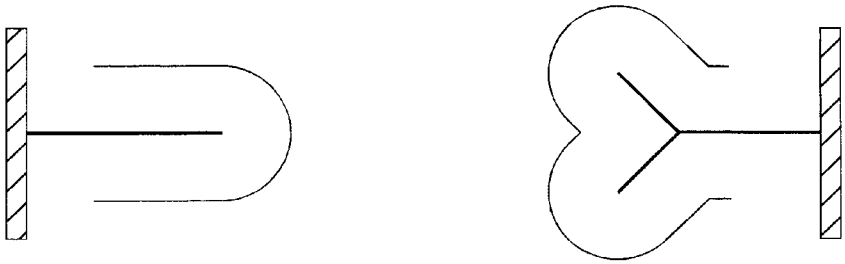
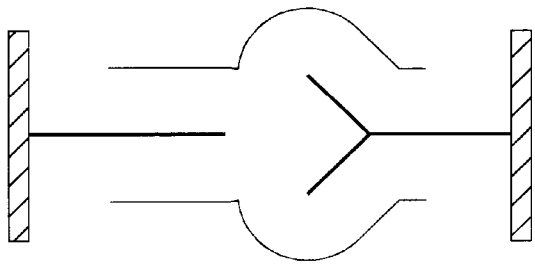

N
H
θ
ϕ
O
C

Figure II-2. Method to calculate buried surface area for a rotamer pair. (a) Unfolded or reference exposed surface areas for two sidechain rotamers. (b) Folded exposed surface area for the rotamer pair. (c) Buried surface area for the rotamer pair, which is calculated by subtracting (b) from (a).

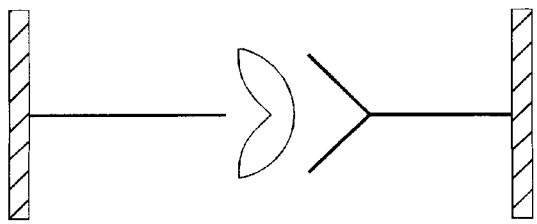
(a)



(b)



(c)



Chapter III

Achieving Stability and Conformational Specificity in Designed Proteins via Binary Patterning

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

S. A. Marshall and S. L. Mayo. (2001) *J. Mol. Biol.*, **305**, 619-631.

Abstract

We have developed a method to determine the optimal binary pattern (arrangement of hydrophobic and polar amino acids) of a target protein fold prior to amino acid sequence selection in protein design studies. A solvent accessible surface is generated for a target fold using its backbone coordinates and "generic" side chains, which are constructs whose size and shape are similar to an average amino acid. Each position is classified as hydrophobic or polar according to the solvent exposure of its generic side chain. The method was tested by analyzing a set of proteins in the Protein Data Bank and by experimentally constructing and analyzing a set of engrailed homeodomain variants whose binary patterns were systematically varied. Selection of the optimal binary pattern results in a designed protein that is monomeric, well-folded, and hyperthermophilic. Homeodomain variants with fewer hydrophobic residues are destabilized, while additional hydrophobic residues induce aggregation. Binary patterning, in conjunction with a force field that models folded state energies, appears sufficient to satisfy two basic goals of protein design: stability and conformational specificity.

Introduction

The location of hydrophobic and polar amino acids along a protein chain, called the binary pattern, and the physics of protein folding together define the topology of a protein fold[1]. Hydrophobic (H) residues are most commonly found in the core of proteins. Burial of these hydrophobic residues is one of the main sources of stability in folded proteins and one of the dominant forces driving protein folding[2]. Polar (P) residues, typically located on the surface of proteins, can play an important role in determining fold specificity[3, 4], controlling protein-protein interactions[5, 6], and promoting solubility. Binary patterning is a major determinant of secondary structure. In amphipathic α-helices and β-sheets, which are commonly observed in naturally occurring proteins, the pattern of hydrophobic and polar residues often matches the periodicity of the secondary structure element. Furthermore, patterns that contain more polar residues tend to encode helices, while patterns dominated by hydrophobic residues typically form β-strands[7].

Binary patterning has been successfully incorporated into several distinct approaches to protein design, including combinatorial design, lattice model simulations, and computational design. Choosing the binary pattern prior to selecting the exact amino acid sequence results in a tremendous reduction in the number of possible sequences. More importantly, correct implementation of binary patterning should help to ensure protein stability and conformational specificity. By restricting buried positions to be nonpolar, hydrophobic burial in the folded protein can be maximized. Selecting polar residues for solvent exposed positions helps to control oligomerization and aggregation behavior. Finally, selecting a pattern of hydrophobic and polar residues that closely matches the pattern of buried and exposed positions in the target protein fold helps to ensure that the designed sequence is compatible with the target structure and incompatible with alternative folds.

Hecht and coworkers have demonstrated that binary patterning alone can be used to design proteins with simple folds[8]. They generated a combinatorial library of proteins using a fixed binary pattern that is compatible with a four helix bundle fold. A significant number of the proteins in this library were demonstrated to unfold cooperatively[9]. Four helical bundles are somewhat unique: they are highly symmetric, their backbone structure can be described with a small number of parameters, and the optimal binary pattern of helical bundle proteins is defined by the helical repeat. Thus far, it has not been clear how to identify the optimal binary pattern for proteins with more complex topologies.

Theoretical protein design studies frequently use binary patterning by approximating protein sequences with a simplified amino acid alphabet consisting of only two monomers: H and P. Dill and coworkers developed two methods for selecting the sequence of H and P monomers that would fold to a desired tertiary structure[10]. The first method, called the Burial Algorithm, measures the solvent exposure of a side chain centroid with a 2 Å radius that is located 3 Å from the α carbon along the α carbon – β carbon bond vector. A 10 $Å^2$ cutoff is used to separate buried hydrophobic positions from exposed polar positions. In the second method, called the Grand Canonical Sequence Evolution Algorithm (GCSE), the optimal binary pattern was identified by maximizing a fitness function which benefits nonlocal contacts between hydrophobic residues and penalizes contacts between solvent and hydrophobic residues. Both algorithms were tested using lattice models for which the folding problem is tractable and then were applied to a set of 20 structures from the Protein Data Bank.

The binary patterns selected by the Burial Algorithm matched the naturally occurring sequences for 67% of the hydrophobic residues and 75% of the polar residues in the database. GCSE-predicted patterns matched the native patterns with an accuracy ranging from 58 to 81%, depending on the protein. GCSE often assigned exposed positions to be hydrophobic,

while the Burial Algorithm assigned all exposed positions to be polar. Database analysis does not reveal the source of the discrepancies between the GCSE patterns, the patterns predicted with the Burial Algorithm, and the actual patterns. Dill argues that the physics underlying the Burial Algorithm is different than that of GCSE, but it is not obvious which algorithm's underpinnings best match the physics underlying protein stability and conformational specificity. Discrepancies between the predicted and actual patterns may reflect the contributions of additional factors, such as folding kinetics and function, to binary patterning. It is possible that the naturally occurring binary patterns are not optimal for stability, and the patterns identified by Dill and coworkers are more optimized for this feature. Without further experimental evidence, it is difficult to ascertain the relative and absolute merits of the two methods.

Previous computational protein design studies conducted by this group have implemented partial binary patterning using a program called Resclass[11]. This program, which is run prior to sequence selection, uses the procedure described in Figures III-1 and III-2 to restrict positions that are clearly in the protein core to be hydrophobic and positions that are clearly on the protein surface to be polar. Between one quarter and one third of the positions, however, do not meet the criteria for either core or surface; these are referred to as boundary positions. During sequence selection, both hydrophobic and polar residues are considered at the boundary positions. Since the dead end elimination (DEE) algorithm used for sequence optimization has a fifth-order dependence on the number of rotamers per residue position[12], boundary calculations are often quite demanding and sometimes prove intractable. A more serious problem is that the sequences selected for boundary residues often lack an appropriate balance between hydrophobic and polar residues. Proteins with excess hydrophobic boundary residues are prone to aggregation, while excess polar boundary residues destabilize the protein.

A new method to assign binary pattern

To address the limitations of the binary patterning procedures that are currently available, we have used theory coupled with computation and experiment to develop and validate a new classification algorithm, Genclass. The Genclass method, described in Figure IIII-3, determines whether each position along a protein backbone should be hydrophobic or polar according to its inherent solvent exposure. Like Resclass, Genclass can be used to assign binary patterns to proteins with either simple or complex topologies, and it does not rely on prior sequence information. Since protein design aims to isolate the determinants of protein stability and conformational specificity, it is desirable to develop methods that do not rely on known sequences. Finally, the Genclass method was designed so that its predictions could be tested experimentally.

If a protein's sequence and structure are known, its residues can be classified as buried or exposed by generating a solvent accessible surface about the protein, measuring the solvent exposure of each residue, and selecting a surface area cutoff, SA_{cut}, that separates buried and exposed positions. At the start of a protein design problem, the sequence and hence the exact surface of the protein are not known. However, it is possible to construct a surface from the protein backbone and "generic" side chains, where the generic side chain is a construct whose size and shape is similar to an average amino acid. In this study, we have used a methyl acetylene-like construct. Other generic side chains such as valine also work, but the side chain must be larger than alanine in order to best distinguish buried and exposed positions. To classify a position, the solvent exposure of its generic side chain is calculated and compared to SA_{cut}. If the solvent exposure of the generic side chain is less than SA_{cut}, the position is classified as hydrophobic and if its exposed area is greater than SA_{cut} the position is classified as polar. We have used both database analysis and experimental studies, described below, to identify the proper value of SA_{cut}.

Genclass uses a "hydrophobic-in, polar-out" metric to assign binary pattern. As has been observed previously, solvent accessibility does not correlate perfectly with residue hydrophobicity, although they are highly coupled[10]. Most naturally occurring proteins contain some buried polar and exposed hydrophobic amino acids. In some cases, these residues are necessary for protein stability; for instance, many turns contain buried polar residues which form hydrogen bonds to main chain amides[13]. Buried polar and exposed hydrophobic residues are often found in binding sites and enzyme active sites, where they may be necessary for activity. In other cases, mutating exposed hydrophobic and buried polar residues improves protein stability[14, 15].

Genclass is intended to provide a reasonable first approximation to the optimal binary pattern. Further protein design studies can be conducted to identify structural contexts in which exposed hydrophobic and buried polar residues contribute to protein stability and conformational specificity. For instance, studies of coiled-coil proteins have shown that replacing a buried Asn with Leu results in a gain in stability but a loss of conformational specificity: both the oligomerization state and the relative orientation of the helices are heterogeneous in the absence of the buried polar residue[16]. Furthermore, as functional properties are introduced into designed proteins, it will be necessary to understand how perturbations in the optimal binary pattern required for the construction of active sites and binding sites impact protein stability and conformational specificity.

Database analysis

Genclass was initially validated by comparing the predicted and actual patterns of hydrophobic and polar residues in a set of 29 water soluble proteins of known structure. The solvent accessible surface area of the generic side chain was calculated for each non-glycine position in the 29 proteins. For each protein, the fraction of residues predicted

correctly (that is, hydrophobic residues whose generic surface area is less than SA_{cut} and polar residues whose surface area is greater than SA_{cut}) was calculated for each value of SA_{cut}. The results for the 29 proteins are shown in Figure III-4.

The fraction of residues whose binary pattern is predicted correctly does not depend strongly on the value of SA_{cut}. If SA_{cut} is set to 0, all residues are predicted to be polar. 60% of the residues are predicted correctly, since 60% of the residues in the proteins analyzed are polar. The optimal value of SA_{cut}, defined as the value at which the agreement between the predicted and actual binary patterns is maximized, is 23.9 $Å^2$, which yields 76% agreement to the database binary patterns. As SA_{cut} is increased beyond the optimal value, the fraction correct slowly decreases and finally plateaus at 40% correct at 136.8 $Å^2$.

Experimental validation of Genclass

In order to test the Genclass method and to more precisely determine the optimal value of SA_{cut}, we have constructed and analyzed a series of engrailed homeodomain variants. Homeodomain is a small fold that is minimally defined by 51 amino acids[17]. Using the Resclass program, 10 of the 51 positions in the engrailed homeodomain fragment are classified as core, 30 as surface, and 11 as boundary. We have systematically varied the binary pattern at the 11 boundary positions in order to determine the best experimental setting of SA_{cut}.

The engrailed homeodomain is an attractive target for protein design studies and for the validation of Genclass. The homeodomain fold has a nontrivial topology so its optimal binary pattern is not immediately apparent. Furthermore, the engrailed homeodomain has been the target of earlier successful design studies. In this study, we use the SC1 variant as a background for all further mutations[18]. SC1 is a 29-fold mutant that was generated by computationally optimizing the core and surface positions. The melting temperature (T_m)

of SC1 is 92 °C, compared to 50 °C for the wild type protein. The modestly high stability of SC1 allows both stabilizing and destabilizing mutations to be readily expressed and thermodynamically characterized. More significantly, using the SC1 variant as a background for boundary design results in the production of fully redesigned homeodomain variants and ensures that the Genclass binary patterning procedure is compatible with the designed protein core and surface.

Genclass agrees with the Resclass results for the homeodomain fold, as shown in Figure III-5. Positions classified as core by Resclass have little or no exposure, surface residues exhibit significant exposure, and boundary residues have intermediate exposure. Using the results of Genclass, the boundary residues were rank-ordered according to their intrinsic solvent accessibility, as shown in Figure III-6. To determine the optimal balance between polar and nonpolar residues, ten binary patterns were selected. In the first pattern, B1, the most buried position was assigned to be hydrophobic and the ten most exposed positions were assigned to be polar, as shown in Figure III-7. B2 assigns the two most buried positions to be hydrophobic and the nine most exposed to be polar. Patterns B3 through B10 are assigned in a similar manner, so that the ten most buried positions in B10 are hydrophobic and only the most exposed position is polar.

The computational protein design algorithm ORBIT (Optimization of Rotamers by Iterative Techniques)[11] was then used to select the optimal amino acid sequence and rotameric configuration for each binary pattern. The sequences selected for the proteins, denoted B1 through B10 according to the number of hydrophobic boundary residues, are shown in Figure III-7. B4 and B5 are identical, since alanine is allowed at both polar and hydrophobic positions. The B1 through B10 variants were compared by experimental analysis to SC1 and to each other in order to assess the effects of varying the binary pattern and to determine the value of SA_{cut} that optimally separates hydrophobic and polar residues.

Stability and conformational specificity of designed variants

The designed proteins were judged according to two criteria: stability and conformational specificity. In order to exhibit conformational specificity, a protein must satisfy three criteria. First, the protein must fold to a unique tertiary structure rather than exhibiting the conformational heterogeneity that is characteristic of molten globule and gemisch states[1]. The protein must possess the desired oligomerization state; in the case of homeodomain, all designed variants should be monomeric. Finally, the designed variants must assume the target fold rather than assuming an alternate fold. In this paper, we focus on the first two criteria to determine the optimal value of SA_{cut}, defined as that value which yields the most stable protein with uncompromised conformational specificity.

The homeodomain fold is remarkably tolerant to perturbations in its binary pattern. The far UV circular dichroism (CD) spectra such as those shown in Figure III-8 all have minima at 208 nm and 222 nm, indicating that the entire series of homeodomain variants is helical at 25 °C. The relative intensity of the two minima varies somewhat and is correlated with the number of tryptophan residues. As tryptophan is known to contribute to ellipticity in this region, the observed variations are not thought to reflect changes in secondary structure[19]. Thermal denaturation experiments indicate that B1 is destabilized relative to SC1; the T_m of B1 is 70 °C while SC1 denatures at 92 °C. Variants B2 through B7 are all hyperthermophilic. At 99 °C these proteins retain significant helical content. Variants B8 through B10 undergo irreversible unfolding transitions, which prohibit thermodynamic analysis. The free energy of unfolding, ΔG_u, determined from guanidinium chloride denaturation at 25 °C increases from B2 through B7 as polar boundary residues are replaced by hydrophobic residues, as shown in Figure III-9 and Table III-1. The increased stability in this series correlates with increased burial of hydrophobic surface area and decreased burial of polar surface area as determined in the modeled protein structures.

Dynamic light scattering (DLS) experiments, shown in Figure III-10 and Table III-2, were used to assess the oligomerization state of the designed proteins. Variants B1 through B6 are all monomeric, as the concentrations of any minor components are thought to be within the experimental error of DLS studies conducted on very small proteins. B7 is primarily monomeric, but aggregated states are also substantially populated. By contrast, B8, B9, and B10 have lost the ability to specifically form monomeric structures. B8 predominantly populates low order oligomers; however, the light scattering data do not reveal whether these oligomers are well-defined. B9 and B10 exclusively form large aggregates.

One-dimensional proton nuclear magnetic resonance (1D ^{1}H NMR) spectra, shown in Figure III-11, were analyzed to determine the solution behavior of each protein. Well-folded proteins have relatively narrow linewidths in 1D ^{1}H NMR spectra, while conformational heterogeneity and increased internal mobility at the millisecond to microsecond time scale, which characterize molten globule and aggregated states, result in broad and/or heterogeneous line widths. Spectra of well-folded proteins are also characterized by pronounced chemical shift dispersion, which arises from the variety of unique magnetic environments that are present in a well-folded protein. The lineshape and dispersion in the spectra shown in Figure III-11 indicate that variants B1 through B6 are well-folded and do not significantly populate aggregated states. B7 was observed to aggregate during the course of data acquisition; as a result, its spectrum has reduced signal to noise and line broadening is observed in the presumptive tryptophan resonances. The pronounced line broadening and reduced chemical shift dispersion in the B8 spectrum, in conjunction with the light scattering results, suggest that B8 forms small, nonspecific aggregates rather than well-ordered oligomers.

On the basis of the experimental data, B6 is judged to be the best protein. According to differential scanning calorimetry results, shown in Figure III-12, the apparent T_m of B6 is 114 °C. By guanidinium denaturation, the ΔG_u of B6 is 6.3 kcal mol^{-1} at 25 °C and pH 4.5, significantly higher than SC1 and variants B1 through B4. To confirm that B6 is stable at more physiological pH, guanidinium denaturation experiments were also conducted at pH 6.0 and pH 7.5, as shown in Figure III-13. The ΔG_u of B6 decreases somewhat with increasing pH, to 5.4 kcal mol^{-1} at pH 6.0 and to 4.3 kcal mol^{-1} at pH 7.5, but the overall stability of B6 remains quite high. 1D ^1H NMR and DLS experiments indicate that B6 assumes a unique folded conformation, while the more hydrophobic variants B7 through B10 lack conformational specificity.

Two additional variants were generated in order to determine whether it is necessary to specify the exact arrangement of hydrophobic and polar boundary residues, or if fixing the absolute number of hydrophobic and polar residues is sufficient. The optimal binary pattern, B6, contains hydrophobic residues at the six most buried positions and polar residues at the five most exposed positions (HHHHHHPPPPP). The binary pattern of the control protein C1 is the reverse of the binary pattern B6. In C1, the five most buried positions are assigned to be polar and the six most exposed positions are hydrophobic (PPPPPHHHHHH). The second control protein, C2, alternates hydrophobic and polar residues (HPHPHPHPHPH) while retaining the same H/P composition as B6 and C1. Sequences were selected for patterns C1 and C2, and are shown in Figure III-7. The ORBIT force field predicts that the folded state energies for C1 and C2 are far less favorable than the energies of any of the other designed variants: the computed energy of C1 is –153.6 kcal mol^{-1}, the energy of C2 is –169.1 kcal mol^{-1}, and the energies of B1 through B10 range from –271.0 to –292.7 kcal mol^{-1}. The CD spectra of C1 and C2 (data not shown) indicate that both proteins are helical. Chemical denaturation experiments, shown in Figure III-14, demonstrate that C1 and C2

are significantly destabilized relative to B6 , as indicated in Table III-1. Variants C1 and C2 also lack conformational specificity. The 1D ^{1}H NMR spectrum of C2 exhibits broad lines and poor chemical shift dispersion, as shown in Figure III-11, and C1 forms an insoluble pellet at concentrations required for NMR.

Binary patterns of naturally occurring sequences provide moderate stability and resistance to aggregation

The optimal value of SA_{cut} identified by the experimental analysis above is 43 Å^2, as a surface area cutoff of 43 Å^2 is required to generate the binary pattern leading to B6. In contrast, the database survey predicted a cutoff of 23.9 Å^2, which yields 75.8% agreement between the predicted and actual binary patterns in a set of 29 proteins. While the 19 Å^2 difference between the optimal SA_{cut} value determined by database analysis and the optimal value identified by experimental analysis is certainly significant, the discrepancy also reflects the fact that, in both the database study and the experimental study, a variety of binary patterns are nearly equally successful. If an SA_{cut} of 43 Å^2 is applied to the proteins in the database study, the agreement between the predicted and actual patterns decreases only modestly, to 72.7%. Setting SA_{cut} to 23.9 Å^2 produces the binary pattern seen in B3. According to the criteria used to judge the designed variants, B3 is nearly as good as B6; both are well folded, monomeric, and hyperthermophilic, although B6 is significantly more resistant to chemical denaturation.

It is plausible that the cutoff identified in the database study is lower than the cutoff found for the homeodomain series because Nature's selection criteria are somewhat different than the criteria that were used to judge the designed variants. Most naturally occurring proteins are not maximally stable, as there is little or no selective pressure to be stable far beyond physiological temperatures and excess stability could compromise function. In

contrast, there is likely strong selective pressure against protein aggregation. Since B6 is only a binary pattern point mutation away from a protein with a significantly increased propensity for aggregation, it is perhaps not surprising that the database study predicted a somewhat lower cutoff corresponding to protein B3.

Binary patterning results in conformational specificity

In protein design, it is not sufficient to select the sequence that is predicted to be most stable; it is also necessary to ensure that the chosen sequence will specifically assume the target fold. The ORBIT force field captures the underlying physics that leads to protein stability, allowing selection of the sequence with the minimal free energy in the folded state. Since satisfactory methods for modeling all the possible unfolded states, aggregated states, partially folded states, and alternative folded states have not yet emerged, the energy terms in the ORBIT force field are not well-suited to the explicit modeling of conformational specificity. However, additional non-thermodynamic considerations, often referred to as negative design, have been incorporated into protein design procedures in order to ensure that selected sequences fold specifically as well as stably to the desired target structure[1, 20-22].

Without negative design, a force field that considers only the energetics of the folded state will tend to favor sequences that are extremely hydrophobic. This occurs because burial of hydrophobic surface area is benefited, while interactions involving polar residues can be either stabilizing or destabilizing. However, sequences that are overly hydrophobic are prone to aggregation and are predicted to have a smaller energy gap between a target structure and alternate states[1]. To select against excessively hydrophobic sequences, the ORBIT force field contains a term that penalizes the exposure of hydrophobic surface area in the folded state[20]. Theoretical studies indicate that incorporation of a hydrophobic exposure

penalty in protein design calculations significantly favors the selection of sequences with good conformational specificity[10]. Despite this, the ORBIT force field favors the excessively hydrophobic homeodomain boundary variants: B8, B9, and B10 have folded state energies between –285.6 and –292.7 kcal mol^{-1} while B2 through B6 have energies between –271.0 and –278.2 kcal mol^{-1}.

Why does the hydrophobic exposure penalty fail to select against the aggregation prone homeodomain variants? Analysis of the sequences and modeled structures of the variants reveals that the exposed hydrophobic surface areas of these proteins do not correlate with total number of hydrophobic residues in the proteins or with their aggregation behavior. In fact, B6, which is monomeric, is predicted to have more exposed hydrophobic surface area than variants B8, B9, or B10, which form aggregates. One possible explanation is that the aggregates may arise from partially folded states rather than the native state[23, 24], suggesting that it would be necessary to compare the exposed surface area of all partially folded states to predict the observed propensities towards aggregation.

While explicitly modeling the factors that govern protein conformational specificity is extremely challenging, the results of the homeodomain series suggest that well-folded proteins can be designed using binary patterning in conjunction with a force field that accurately models the folded state alone. This conclusion is compatible with the random energy model proposed by Shakhnovich and coworkers[25], which postulates that the energies of the vast majority of possible protein conformations are determined only by amino acid composition. Binary patterning ensures that the sequences which are considered in a protein design problem all have, at low resolution, the same composition and hence populate nearly isoenergetic unfolded and partially folded states. Therefore, once the binary pattern is fixed, comparison of folded state energies is sufficient to identify sequences with a large energy gap between the target fold and competing folds.

Conclusions

The engrailed homeodomain study demonstrates that well-folded, extremely stable proteins can be designed using the binary patterns selected by Genclass and the amino acid sequences generated by ORBIT. The optimal value of SA_{cut}, the surface area cutoff parameter, predicts that all positions previously classified as core by the Resclass program should be hydrophobic and all positions previously classified as surface should be polar. As desired, proper selection of SA_{cut} also determines whether each boundary residue should be hydrophobic or polar. Systematic variation of SA_{cut} was shown to affect protein stability and conformational specificity; furthermore, the exact arrangement of hydrophobic and polar residues was also found to be important for stability and conformational specificity. Selection of the best value of SA_{cut}, 43 $Å^2$, results in a designed protein that is monomeric, hyperthermophilic, and well-folded. Without further study, it is difficult to assess whether the cutoff found in the homeodomain study will be optimal for all proteins, but the methods used to identify the proper cutoff should be generally applicable.

Identifying the binary pattern that is optimal for a target protein fold prior to sequence selection has proven to be advantageous for several reasons. First, the region of sequence space that must be searched is reduced. More importantly, selection of the proper binary pattern has proved to be an efficient way to introduce negative design considerations into computational protein design. Capturing global properties such as aggregation behavior using a pairwise potential describing only the folded state presents many difficulties, both theoretical and computational. The results of this study indicate that it is possible to model at least one global property, the binary pattern, by generating a protein surface that is independent of sequence. The results also suggest that it may be possible to similarly describe other global properties that rely on protein surfaces. Using binary patterning in conjunction

with an accurate force field that models the folded state has proven to be a simple, efficient, and effective means of designing proteins with good stability and conformational specificity.

Methods

Resclass and Genclass. Resclass identifies positions as core, boundary, and surface using simple geometric criteria, as shown in Figures III-1 and III-2. In protein design calculations, core residues are typically restricted to Ala, Val, Leu, Ile, Phe, Tyr and Trp, surface residues are restricted to Ala, Ser, Thr, Asp, Asn, His, Glu, Gln, Arg, and Lys, and both sets are considered at boundary positions.

In Genclass, the generic side chains are added to each non-glycine α carbon in the target protein backbone, as shown in Figure III-3. A methyl acetylene-like construct comprised of three carbon atoms is used as the generic side chain. The first atom in each generic side chain is located at the crystal lographic coordinates for the β carbon. The second and third atoms lie along the Cα-Cβ bond vector at a distance equal to two and three times the crystallographically determined Cα-Cβ bond length from the α carbon. All atoms in the generic side chain have the atomic radius of carbon. A surface is generated using the Lee and Richards[26] definition and by applying the Connolly algorithm[27] to the protein backbone, including explicit backbone hydrogen atoms, and generic side chains using a carbon radius of 1.95 Å and an add-on radius of 1.4 Å.

Database Survey. The proteins selected for the database study are water soluble and monomeric with crystal structures solved to a resolution of at least 2.3 Å. The PDB codes of the proteins used in the database survey are: 154l, 1a45, 1a8p, 1ab1, 1af7, 1agi, 1ah4, 1aii, 1ajz, 1aky, 1bhp, 1bzm, 1cka, 1fel, 1iuz, 1jlm, 1lec, 1mai, 1omd, 1onc, 1pga, 1rcy, 1rec, 1uke, 1wod, 1who, 2chf, 2phy, and 3cbp. The naturally occurring residues were

classified as hydrophobic (Val, Leu, Ile, Phe, Tyr, Trp, and Met), polar (Ser, Thr, Asp, Asn, His, Glu, Gln, Arg, and Lys) or other (Ala, Cys, Gly, Pro). Alanine is not classified because it is included in both the hydrophobic and polar groups in the design calculations. Cysteine, glycine, and proline are excluded because they play special structural roles that are not well described by binary patterning and are not currently included in ORBIT.

The solvent exposed surface area of each generic side chain on each protein was determined as described above. The fraction of residues predicted correctly (that is, hydrophobic residues whose generic surface area is less than SA_{cut} and polar residues whose generic surface area is greater than SA_{cut}) was calculated for each protein using values for SA_{cut} ranging from 0.0 to 150.0 $Å^2$ with a step size of 0.1 $Å^2$. The relationship between SA_{cut} and the fraction of residues predicted correctly was found by averaging over the set of proteins.

Modeling. Structural coordinates for the engrailed homeodomain were obtained from PDB entry 1enh[17]. Residues 1-5, which are disordered in the absence of DNA binding, were removed from the structure and explicit hydrogens were added to the remaining 51 residues using BIOGRAF (Molecular Simulations, Inc., San Diego). The resulting structure was minimized for 50 steps using the DREIDING force field[28]. Side chains were represented as discrete rotamers from the backbone dependent rotamer library developed by Dunbrak and Karplus[29], as previously described[30].

Sequence selection. For each calculation, the hydrophobic boundary residues were restricted to be Ala, Val, Leu, Ile, Phe, Tyr, or Trp and the polar boundary residues were restricted to be Ala, Ser, Thr, Asp, Asn, His, Glu, Gln, Arg, or Lys. The sequence for the core and surface positions was held constant, but the rotameric conformations at these positions were allowed

to vary. Pairwise rotamer-template and rotamer-rotamer energies were calculated using a force field containing terms describing van der Waals interactions, hydrogen bonding, electrostatics, and solvation[31]. The optimal amino acid sequence and rotamer conformations were determined using the Dead End Elimination (DEE) theorem[12, 32-34].

The combinatorial complexity of the resulting rotamer space optimization problems was as high as 4.0×10^{71}. DEE could reduce the size of the problem by over 30 orders of magnitude, but failed to converge to a single solution. To obtain a sequence, the boundary resides were divided into three minimally interacting groups: a) 1, 3; b) 10, 14, 21, 25, 26, 30; and c) 19, 47, 51. For each group, a set of calculations was run to determine the optimal amino acid sequence for each of the desired binary patterns. The wild type sequence at the remaining positions was held constant but rotameric conformation was allowed to vary. Dividing the boundary residues into groups reduced the maximum combinatorial complexity to 8.1×10^{60} and enabled DEE to converge to a single solution. After sequences were selected for all three sets of boundary residues, a second set of calculations was run to find the optimal rotameric conformation of all the residues for each desired binary pattern. All calculated energies and surface areas are based on the structures predicted in this second set of calculations.

Protein Expression. Synthetic genes encoding SC1, B3, B8, B9, C1, and C2 were constructed using recursive PCR[35] and cloned into a pET-11a (Novagen) variant. The remaining genes were obtained by site directed mutagenesis using inverse PCR. Sequences for all constructs were confirmed by DNA sequencing. Recombinant proteins were expressed in BL21 (DE3) *Escherichia coli* cells (Stratagene) and isolated using either the freeze-thaw method[36] or sonication in 1 M urea. The proteins were purified by reverse-phase HPLC using a C8 prep column (Zorbax) and a linear acetonitrile-water gradient with 0.1 % TFA.

Protein masses were determined by MALDI-TOF or electrospray mass spectrometry; all masses were within one mass unit of the predicted molecular weight.

Solution Conditions. pH 4.5 was used in the following experiments unless otherwise noted because these solution conditions were compatible with all proteins and experiments. Variants B8, B9, and B10 were observed to form gels at higher pH at the concentrations required for light scattering studies.

Circular Dichroism Studies. CD data were obtained on an Aviv 62A DS spectropolarimeter equipped with a thermoelectric cell holder and an autotitrator. Samples for wavelength scans and thermal denaturation experiments contained between 5 and 50 μM protein and 50 mM sodium phosphate buffer at pH 4.5. For wavelength scans, ellipticity was measured from 200 to 250 nm at 25 °C. Thermal denaturation data were obtained from 1 °C to 99 °C with a step size of 1 °C, an equilibration time of 90 sec, and an averaging time of 30 sec. Melting temperatures for B1 and SC1 were determined by fitting to a two state transition as previously described[37]. Guanidinium chloride denaturation data were obtained from samples containing 5 μM protein and 50 mM sodium phosphate buffer at pH 4.5 at 25 °C. To maintain constant pH, the guanidinium chloride stock solution also contained 50 mM sodium phosphate buffer at pH 4.5. Initial and final denaturant concentrations were determined by refractometry[38]. Data were acquired every 0.2 M from 0.0 M to 8.2 M GdmCl using a mixing time of 9 min and an averaging time of 100 sec. ΔG_u was calculated from the chemical denaturation data assuming a two-state transition and using the linear extrapolation model[39]. Guanidium chloride denaturation data were also obtained for variant B6 at pH 6.0 and 7.5. Both thermal and chemical denaturation were monitored by CD ellipticity at 222 nm.

Dynamic Light Scattering Studies. DLS data were obtained using a Protein Solutions Dyna Pro 801 molecular sizing instrument. Samples contained 1 mg ml^{-1} protein and 50 mM sodium phosphate buffer at pH 4.5. Residual dust was removed using a 0.02 μm filter (Whatman). The radius of hydration for each protein was obtained by averaging over at least 20 measurements. Molecular weights were obtained by fitting to a bimodal distribution and assuming a globular protein shape.

Nuclear Magnetic Resonance Studies. 1D ^1H NMR spectra were obtained using a Varian 600 MHz spectrometer using a Varian triple resonance probe. Samples contained 1 mM protein and 50 mM sodium phosphate in a 10% D_2O buffer at pH* 4.5.

Differential Scanning Calorimetry Studies. DSC data were obtained using an Applied Thermodynamics N-DSC II calorimeter. The sample contained 4.5 mg ml^{-1} protein and 50 mM sodium phosphate buffer at pH 4.5 and was thoroughly dialyzed against the buffer. Data scans were obtained for the buffer and the protein solution at a rate of 1 °C min^{-1} from 1 °C to 130 °C at a pressure of 4 atm. Due to the high stability and small size of B6, the unfolding transition is not completed by 130 °C. The apparent thermal denaturation temperature was defined to be the maximum of the scan.

Acknowledgements

This work was supported by the Howard Hughes Medical Institute (S. L. M.), the National Institutes of Health, and the Caltech Initiative in Computational Molecular Biology, which is funded by a Burroughs Wellcome Fund Interfaces Award (S. A. M.).

References

1. Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D. and Chan, H. S. (1995). Principles of protein folding - a perspective from simple exact models. *Protein Sci.*, **4**, 561-602.
2. Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, 29, 7133-7155.
3. O'Shea, E. K., Lumb, K. J. and Kim, P. S. (1993). Peptide 'velcro': design of a heterodimeric coiled coil. *Curr. Biol.*, **3**, 658-667.
4. O'Shea, E. K., Klemm, J. D., Kim, P. S. and Alber, T. (1991). X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. *Science*, **254**, 539-544.
5. Xu, D., Lin, S. L. and Nussinov, R. (1997). Protein binding versus protein folding: The role of hydrophilic bridges in protein associations. *J. Mol. Biol.*, **265**, 68-84.
6. Xu, D., Tsai, C. J. and Nussinov, R. (1997). Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng.*, **10**, 999-1002.
7. West, M. W. and Hecht, M. H. (1995). Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci.*, **4**, 2032-2039.
8. Kamtekar, S., Schiffer, J. M., Xiong, H., Babik, J. M. and Hecht, M. H. (1993). Protein design by binary patterning of polar and nonpolar amino acids. *Science*, **262**, 1680-1685.
9. Roy, S. and Hecht, M. H. (2000). Cooperative thermal denaturation of proteins designed by binary patterning of polar and nonpolar amino acids. *Biochemistry*, **39**, 4603-4607.
10. Sun, S., Brem, R., Chan, H. S. and Dill, K. A. (1995). Designing amino acid sequences to fold with good hydrophobic cores. *Protein Eng.*, **8**, 1205-1213.
11. Dahiyat, B. I. and Mayo, S. L. (1997). *De novo* protein design: fully automated sequence selection. *Science*, **278**, 82-87.

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

13. Eswar, N. and Ramakrishnan, C. (2000). Deterministic features of side-chain main-chain hydrogen bonds in globular protein structure. *Protein Eng.*, **13**, 227-238.

14. Pakula, A. A. and Sauer, R. T. (1990). Reverse hydrophobic effects relieved by amino-acid substitutions at a protein surface. *Nature*, **344**, 363-364.

15. Waldburger, C. D., Schildbach, J. F. and Sauer, R. T. (1995). Are buried salt bridges important for protein stability and conformational specificity. *Nat. Struct. Biol.*, **2**, 122-128.

16. Lumb, K. J. and Kim, P. S. (1995). A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry*, **34**, 8642-8648.

17. Clarke, N. D., Kissinger, C. R., Desjarlais, J., Gilliland, G. L. and Pabo, C. O. (1994). Structural studies of the engrailed homeodomain. *Protein Sci.*, **3**, 1779-1787.

18. Morgan, C. S. (2000). Ph.D. Thesis. California Institute of Technology, Pasadena, CA.

19. Woody, R. W. and Dunker, A. K. (1996). Aromatic and cystine side-chain circular dichroism in proteins. In *Circular dichroism and the conformational analysis of biomolecules* (G. D. Fasman, ed) Plenum Press, New York.

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

21. Hellinga, H. W. (1997). Rational protein design: combining theory and experiment. *Proc. Natl. Acad. Sci.*, USA, **94**, 10015-10017.

22. Street, A. G., Datta, D., Gordon, D. B. and Mayo, S. L. (2000). Designing protein b-sheet surfaces by z-score optimization. *Phys. Rev. Lett.*, **84**, 5010-5013.

23. Broglia, R. A., Tiana, G., Pasquali, S., Roman, H. E. and Vigezzi, E. (1998). Folding and aggregation of designed proteins. *Proc. Natl. Acad. Sci.*, USA, **95**, 12930-12933.

24. Fink, A. L. (1998). Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold. Des.*, **3**, R0-R23.

25. Shakhnovich, E. I. and Gutin, A. M. (1993). A new approach to the design of stable proteins. *Protein Eng.*, **6**, 793-800.

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

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

28. Mayo, S. L., Olafson, B. D. and Goddard, W. A., III. (1990). Dreiding - a generic force-field for molecular simulations. *J. Phys. Chem.*, **94**, 8897-8909.

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

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

31. Gordon, D. B., Marshall, S. A. and Mayo, S. L. (1999). Energy functions for protein design. *Curr. Opin. Struct. Biol.*, **9**, 509-513.

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

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

34. Gordon, D. B. and Mayo, S. L. (1998). Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J. Comput. Chem.*, **19**, 1505-1514.

35. Prodromou, C. and Pearl, L. H. (1992). Recursive PCR: a novel technique for total gene synthesis. P*rotein Eng.*, **5**, 827-829.

36. Johnson, B. H. and Hecht, M. H. (1994). Recombinant proteins can be isolated from *E. coli* cells by repeated cycles of freezing and thawing. *Biotechnology*, **12**, 1357-1360.

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

38. Pace, N. C. (1986). Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol.*, **131**, 266-280.

39. Santoro, M. M. and Bolen, D. W. (1988). Unfolding free-energy changes determined by the linear extrapolation method .1. unfolding of phenylmethanesulfonyl α-chymotrypsin using different denaturants. *Biochemistry*, **27**, 8063-8068.

40. Koradi, R., Billeter, M. and Wüthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics*, **14**, 51.

Table III-1: Guanidinium chloride denaturation data

	ΔG_u[1] (kcal mol^{-1})	C_m[2](M)	m[3] (kcal mol^{-1}M^{-1})
SC1	2.47 ± 0.27	1.8	1.35 ± 0.11
B2	2.74 ± 0.10	2.6	1.06 ± 0.04
B3	4.15 ± 0.17	3.8	1.08 ± 0.06
B4	4.84 ± 0.21	4.9	0.99 ± 0.05
B6	6.30 ± 0.41	5.3	1.19 ± 0.09
B7	6.19 ± 0.28	5.3	1.17 ± 0.06
C1	1.88 ± 0.27	1.6	1.04 ± 0.10
C2	1.61 ± 0.23	1.7	1.00 ± 0.08

[1] Free energy of unfolding at 25 °C

[2] Midpoint of unfolding transition

[3] Slope of ΔG_u vs. denaturant concentration

Table III-2: Dynamic light scattering data

	MW[1] (kDa)	%[2]	MW[3] (kDa)	%[4]
B1	5.5	99	24000	1
B2	7.1	93	220	7
B3	5.3	97	1300	3
B4	7.0	93	1700	7
B6	5.2	100	-	-
B7	6.1	81	170,000	19
B8	21.8	87	350,000	13
B9	>100	100	-	-
B10	>100	100	-	-

[1] Molecular weight of dominant component

[2] Percent of dominant component

[3] Molecular weight of minor component

[4] Percent of minor component

Figure III-1. Procedure for the Resclass binary patterning method applied to a three helical bundle protein, Step 1. A Connolly surface[27] is generated about the α carbon atoms in the target protein fold using an 8 Å probe radius and a 1.95 Å atomic radius.

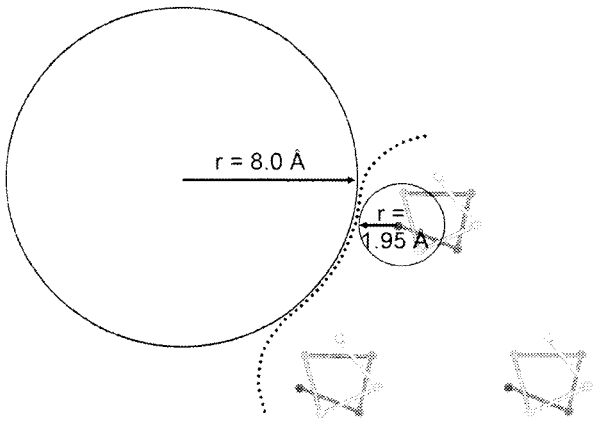

r = 8.0 Å
r =
1.95 Å

Figure III-2. Procedure for the Resclass binary patterning method applied to a three helical bundle protein, Step 2. Two distances are measured for each residue: D1 is the distance from the a carbon to the surface along the vector connecting the a and b carbons and D2 is the distance from the b carbon to the closest point on the surface. Positions are classified as core if D1 ≥ 5 Å and D2 ≥ 2 Å, positions at which D1 + D2 ≤ 2.7 Å are classified as surface, and residues which do not meet the criteria for either surface or core are classified as boundary.

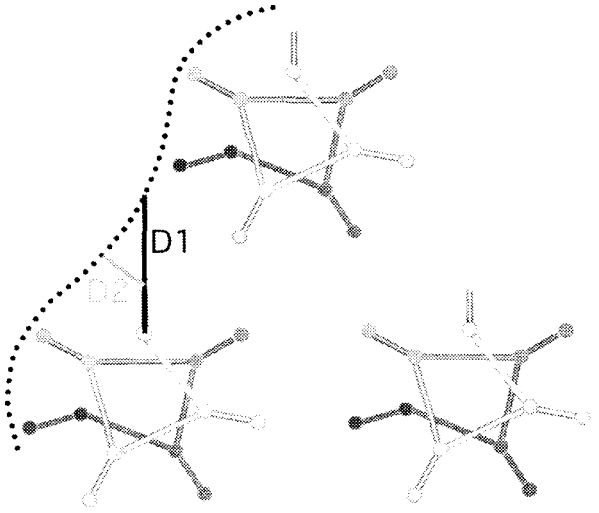

D1
D2

Figure III-3. Procedure for the Genclass binary patterning method. (a) A target backbone is selected and the naturally occurring side chains are removed. (b) Generic side chains are added to each position. In this study, the generic side chains consist of three carbon atoms located along the C□-Cβ bond vector at distances equal to one, two, and three times the C□-Cβ bond length. (c) A solvent accessible surface is generated about the backbone and generic side chains using the Lee and Richards[26] definition and by applying the Connolly algorithm[27] using a carbon radius of 1.95 Å and an add-on radius of 1.4 Å. (d) Each position is classified as hydrophobic or polar according to whether the solvent exposure of its generic side chain is above or below the surface area cutoff, SA_{cut}, shown schematically as a horizontal line.

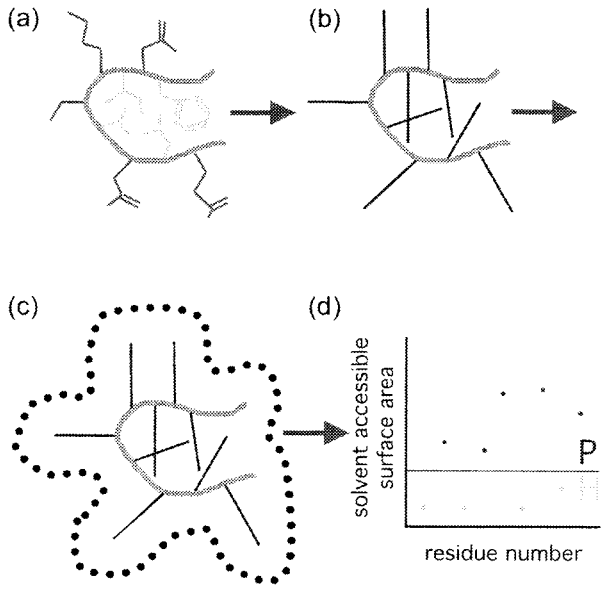

(a)
(b)
(c)
(d)
solvent accessible surface area
P
residue number

Figure III-4. Results of Genclass analysis of 29 proteins in the Protein Data Bank. The average fraction of residues predicted correctly (that is, hydrophobic residues whose generic surface area is less than the surface area cutoff, SA_{cut}, and polar residues whose generic surface area is greater than SA_{cut}) is shown for values of SA_{cut} between 0.0 and 130.0 $Å^2$.

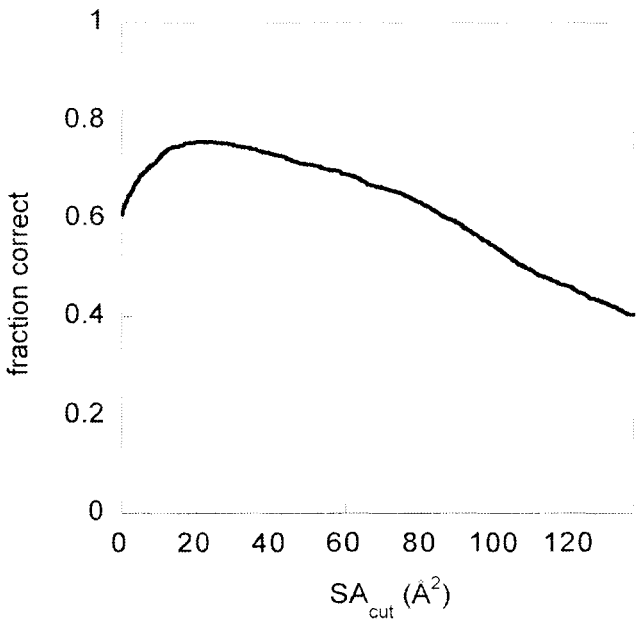

1
0.8
0.6
0.4
0.2
0
fraction correct
0
20
40
60
80
100
120
SA cut (Å2)

Figure III-5. A comparison of Genclass and Resclass binary pattern assignments for the engrailed homeodomain. Residues are plotted along the x-axis from N to C termini. The solvent accessibility of the generic side chain located at each position is plotted along the y-axis. The Resclass categorization of core (red), boundary (green), or surface (blue) is indicated for each position. The Genclass categorization is obtained by drawing a horizontal line at the desired value of SA_{cut}. The residue number of each of the eleven boundary positions is also shown.

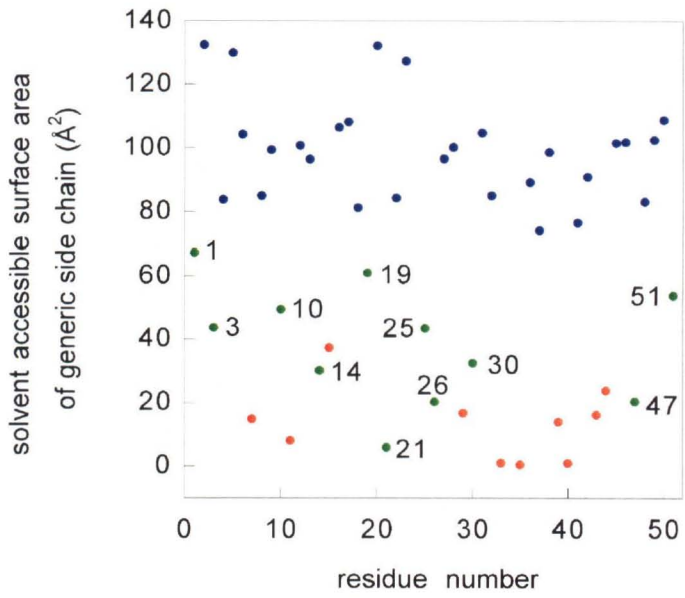

solvent accessible surface area
of generic side chain (Å²)
140
120
100
80
60
40
20
0
0
10
20
30
40
50
residue number
1
3
10
14
19
25
26
30
21
51
47

Figure III-6. Structure of the engrailed homeodomain. Boundary residues are colored along the spectrum according to the solvent accessibility of their generic side chains. The most buried position is red and the most exposed is blue. The ribbon diagram was generated using MOLMOL[40].

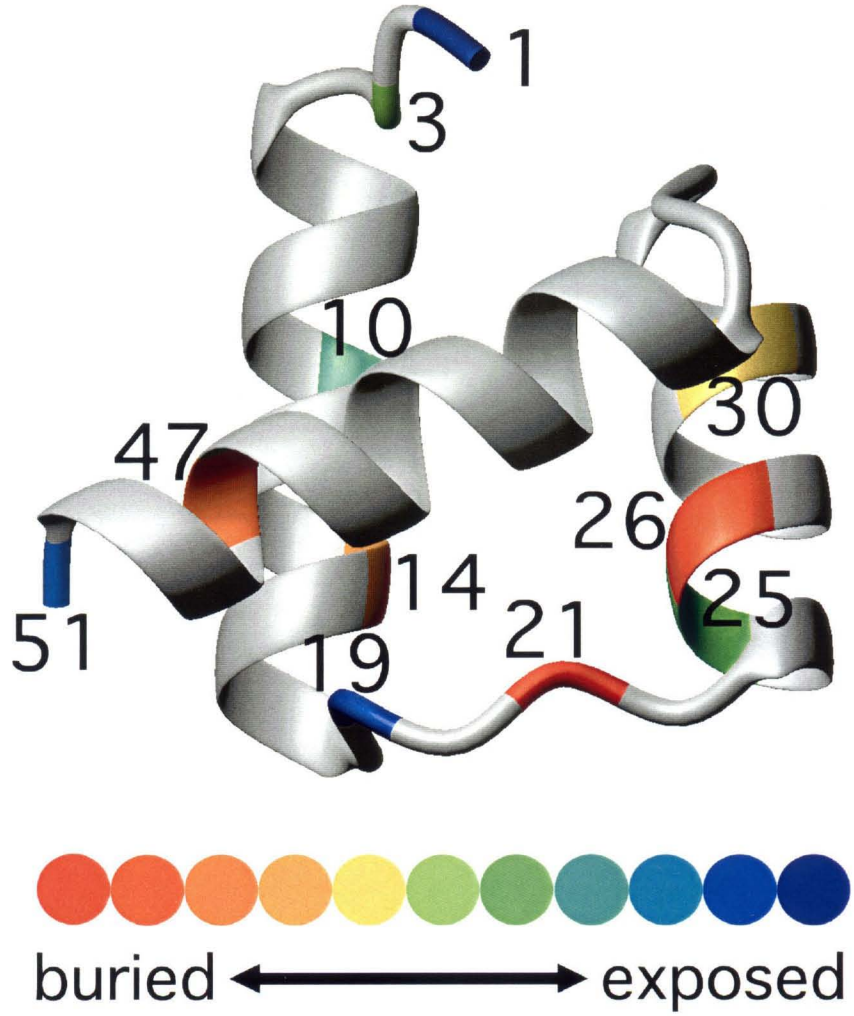

1
3
10
30
47
26
14
25
51
21
19
buried
exposed

Figure III-7. Binary patterns and sequences of the boundary residues in SC1 and the designed homeodomain variants. Boundary positions are colored as in Figure III-6 and are ordered so that the most buried position is on the left and the most exposed position is on the right. The binary pattern of each protein is shown by the arrangement of red and blue beads, which refer to positions that were restricted to be hydrophobic and polar, respectively. The boundary sequence for SC1, which matches wild type, is labeled "SC1." Proteins B1 through B10 are named according to the number of hydrophobic boundary residues in each sequence. The binary patterns and sequences for the control proteins C1 and C2 are shown at the bottom.

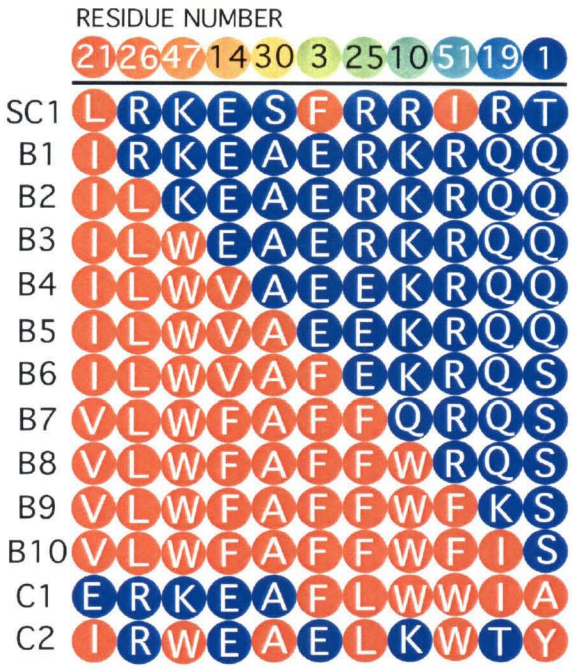

RESIDUE NUMBER
21 26 47 14 30 3 25 10 51 19 1
SC1 L R K E S F R R I R T
B1 I R K E A E R K R Q Q
B2 I L K E A E R K R Q Q
B3 I L W E A E R K R Q Q
B4 I L W V A E E K R Q Q
B5 I L W V A E E K R Q Q
B6 I L W V A F E K R Q S
B7 V L W F A F F Q R Q S
B8 V L W F A F F W R Q S
B9 V L W F A F F W F K S
B10 V L W F A F F W F I S
C1 E R K E A F L W W I A
C2 I R W E A E L K W T Y

Figure III-8. CD wavelength scans of B6 (—) and SC1 (- - -) measured at 25 °C.

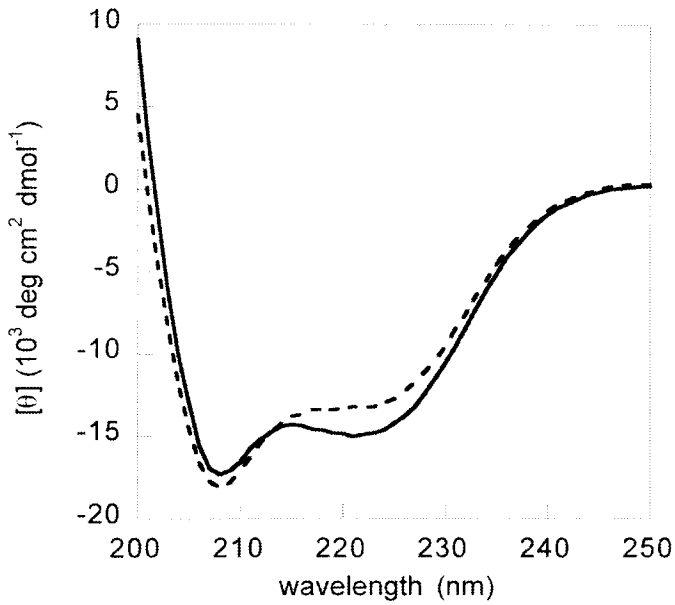

10
5
0
-5
-10
-15
-20
200
210
220
230
240
250
wavelength (nm)
[θ] (10^3 deg cm^2 dmol^-1)

Figure III-9. Guanidinium chloride denaturation at 25 °C monitored by CD of (from left to right) SC1 (black), B2 (orange), B3 (green), B4 (turquoise), B7 (purple), and B6 (blue).

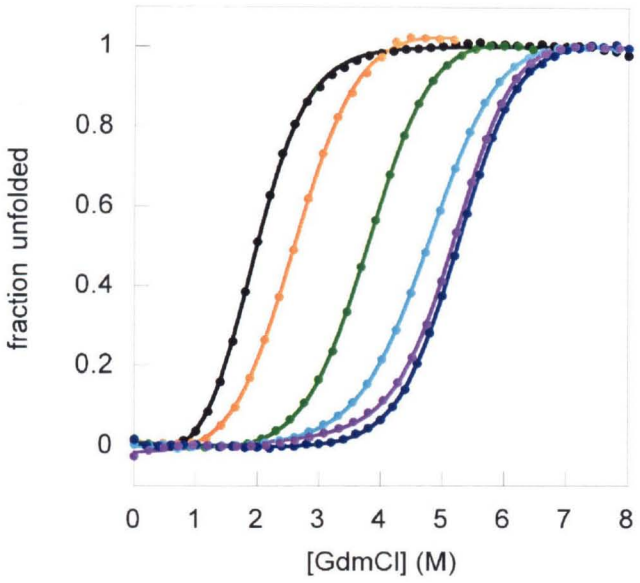

fraction unfolded
1
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
6
7
8
[GdmCl] (M)

Figure III-10. Dynamic light scattering results for homeodomain boundary variants. The proteins are ordered along the x-axis from the most polar variant, B1, on the left to the most hydrophobic variant, B10, on the left. Percent monomer (—), low order oligomer (– – –) and aggregate (- - - -), calculated by fitting to a bimodal distribution and assuming a globular protein shape, are indicated for each variant.

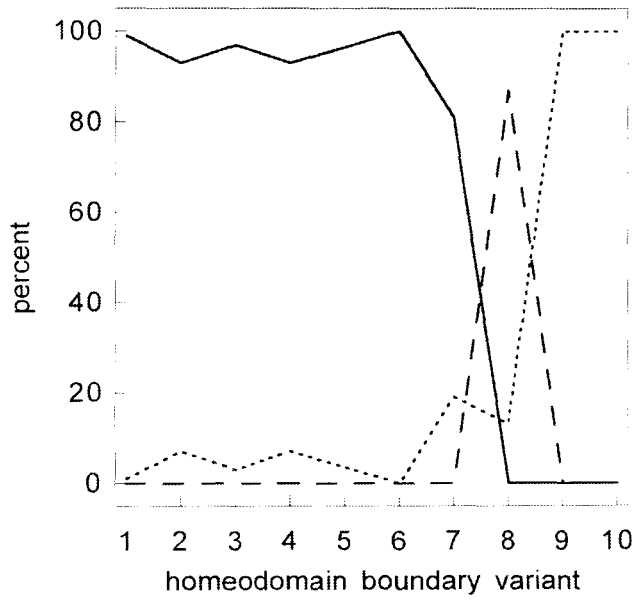

100
80
60
40
20
0
percent
1
2
3
4
5
6
7
8
9
10
homeodomain boundary variant

Figure III-11. Aromatic and amide region of the 1D ^{1}H NMR spectra of (a) B1, (b) B2, (c) B3, (d) B4, (e) B6, (f) B7, (g) B8, (h) C2, and (i) SC1. A * indicates that the peak was cropped.

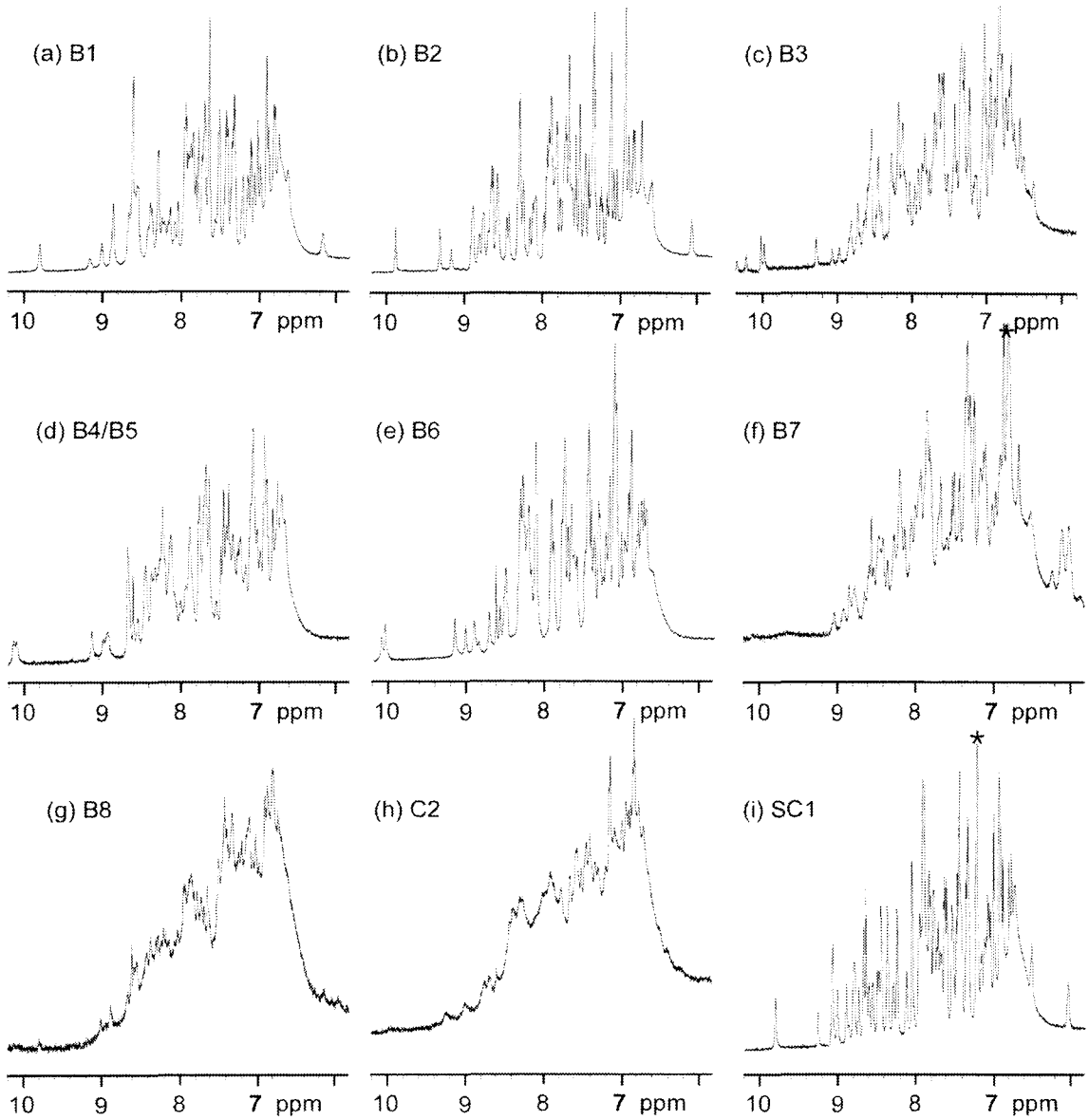

(a) B1
(b) B2
(c) B3
(d) B4/B5
(e) B6
(f) B7
(g) B8
(h) C2
(i) SC1
10
9
8
7 ppm

Figure III-12. Thermal denaturation of variant B6 monitored by differential scanning calorimetry. The maximum of the thermogram, which is approximately equal to the thermal denaturation temperature, is at 114 °C.

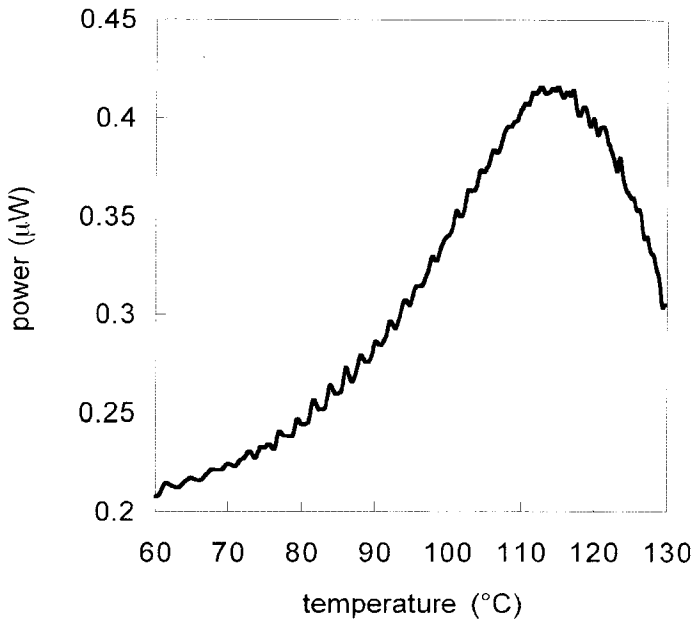

0.45
0.4
0.35
0.3
0.25
0.2
power (μW)
60
70
80
90
100
110
120
130
temperature (°C)

Figure III-13. Guanidinium chloride denaturation of homeodomain variant B6 at 25 °C monitored by CD at (from left to right) pH 7.5 (blue), pH 6.0 (green), and pH 4.5 (red).

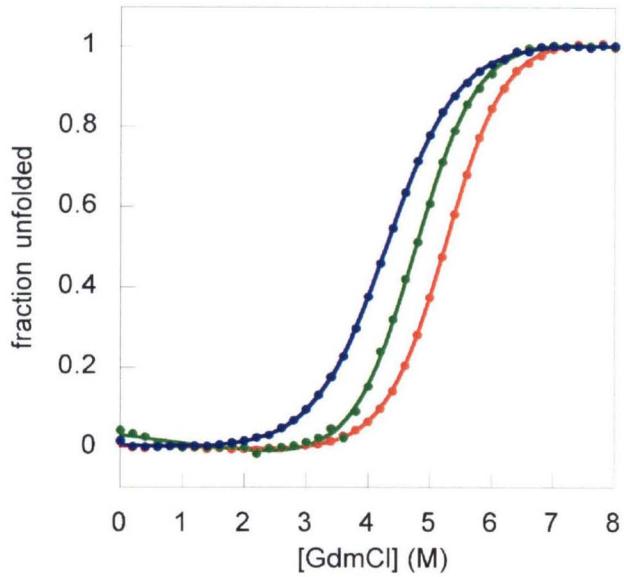
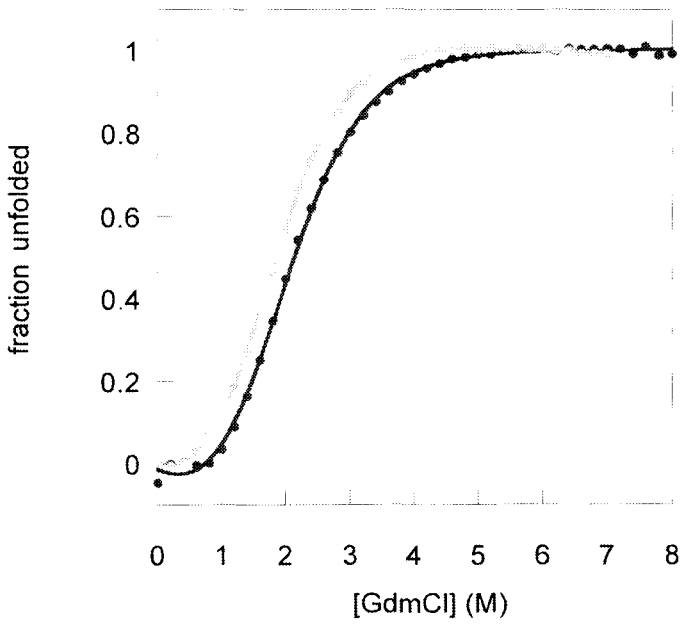

fraction unfolded
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7 8
[GdmCl] (M)

Figure III-14. Guanidinium chloride denaturation at 25 °C monitored by CD of (from left to right) C1 (light gray) and C2 (dark gray).


fraction unfolded
1
0.8
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7 8
[GdmCl] (M)

Chapter IV

Towards the Solution Structure of a Fully Designed Homeodomain Variant

The text of this chapter presents work that was conducted in collaboration with Scott Ross, Carlos Amezcua (University of Texas Southwestern Medical Center), Professor Kevin Gardner (University of Texas Southwestern Medical Center), and Professor Stephen L. Mayo.

Introduction

One of the most demanding experimental tests for a designed protein is structure determination by NMR or x-ray crystallography. Many of the structural predictions made during a protein design calculation can be tested through analysis of the experimentally determined structure. The structure of the protein backbone can be compared to the template to confirm that the designed sequence has assumed the desired fold. The predicted rotameric conformations of side chains can be compared to the observed conformations to test the accuracy of the side chain models currently used in design calculations. In addition, the structure can be examined to determine whether specific side chain - side chain and side chain - backbone contacts that were predicted to be energetically favorable are in fact significantly populated.

Here, we present work towards the determination of the structure of a fully designed homeodomain variant, B6. The amino acid sequence for the core and surface residues was selected in one design calculation[1], using the helix dipole and capping restrictions described in Chapter V. The amino acid sequence for the boundary residues was selected in a second design calculation, described in Chapter III, in which the previously selected sequence was included at the core and surface positions.

Most designed proteins have not been amenable to high resolution structure determination because they exhibit conformational heterogeneity. As a result, structures have been determined for only a handful of fully designed proteins, including a zinc finger domain[2] and symmetric helical bundle proteins[3, 4]. Obtaining a high resolution structure of a fully designed homeodomain variant would add to the small database of high resolution structures resulting from protein design calculations.

Initially, both crystallographic and NMR approaches were pursued to determine the structure of B6. Small, thin crystals were obtained for B6; however, the crystals diffracted poorly. To date, it has not been possible to determine the structure of B6 by x-ray crystallography. Using NMR approaches, we have determined that B6 assumes the topology of the target fold but have not yet obtained a high resolution structure. Work is ongoing to obtain a high resolution structure of B6.

Results

Chemical shift assignments, enumerated in Table IV-1, were obtained for the majority of the ^{1}H, ^{13}C, and ^{15}N atoms in the protein. The overall chemical shift dispersion is modest, as is typical for helical proteins[5]. Nonetheless, examination of the ^{15}N HSQC spectrum, shown in Figure IV-1, demonstrates that distinct signals can be observed for almost all of the residues. Representative strip plots from the backbone assignment experiments, HNCACB and CBCA(CO)NH, shown in Figure IV-2, are included to provide an indication of the spectral quality realized in the triple resonance experiments. Analysis of the spectra and chemical shift assignments confirms that B6 folds to a unique native state.

In addition to considering the overall chemical shift dispersion, it can also be useful to compare chemical shifts for each of the commonly used amino acid types. B6 has somewhat lower sequence complexity than a naturally occurring protein: the 52 residue

protein contains seven Glu residues, and six residues each of Arg, Lys, Gln, and Leu. The chemical shifts among residues with the same amino acid often exhibit little variation. Exceptions to this general trend may result from interesting structural features. For example, the chemical shifts observed for Lys 9 are quite distinct from the other lysines. In the ORBIT-generated structure of B6, Lys 9 is located close to Trp 44 and Trp 48, and NOEs are observed from both tryptophans to resonances whose chemical shifts are consistent with the Lys 9 assignments. The data raise the possibility that a cation-π interaction, discussed in Appendix B, may be present in B6.

Another observation regarding the assigned chemical shifts of B6 is that many of the methylene protons are not degenerate, indicating that their side chains populate a specific rotameric state rather than rotating freely. It is interesting to note that even long, polar side chains, such as Lys and Arg, sometimes have nondegenerate methylene chemical shifts. In the case of Asn and Gln, there are sometimes large chemical shift differences between the two side chain amide protons. Finally, the two methyl groups of several of the leucine residues are not degenerate.

Following completion of the chemical shift assignments, we used TALOS[11] to obtain restraints for the backbone ϕ and ψ angles. TALOS compares the local sequence and the chemical shifts of the CA, CB, HA, and N atoms to a database of NMR and crystal structures to generate backbone angle restraints. The TALOS predictions generally match the backbone angles of the target structure to within ± 15°, as shown in Table IV-2. The TALOS results clearly indicate that B6 contains three helices whose locations closely match the target structure. The most significant deviations between the TALOS predictions and the target backbone structure occur at the ends of helices and in turn regions. In addition, TALOS does not make predictions for a small fraction of the backbone angles.

Since the chemical shift dispersion of B6 is modest, the vast majority of the NOE crosspeaks can not be assigned unambiguously. In almost all cases, the heteroatom and covalently attached proton can be assigned unambiguously. However, in a typical case, about ten protons will have a chemical shift that is within 0.03 ppm of the second proton participating in the NOE. Aria[12, 13], in conjunction with CNS[14], was then used to assign ambiguous NOEs and to determine the structure of B6. Thus far, the structures have not fully converged. However, the top structures are observed to adopt, at low resolution, the correct target fold. An example of a structure predicted using Aria versus the target backbone structure is shown in Figure IV-3.

In the absence of a final structure, ambiguous NOEs can be analyzed for compatibility with the structure of B6 generated using ORBIT, described in Chapter III. For each NOESY crosspeak, the distance between the assigned proton and each possible NOE partner was calculated using the ORBIT structure. If at least one possible partner was within 6 Å of the assigned proton, the restraint was classified as compatible with the model structure. The number of short, medium, and long range restraints that are compatible with the model structure, as well as the number of restraints that are not compatible, are indicated for each residue in Figure IV-4. The pattern of NOE contacts between different residues is shown in Figure IV-5. Using this approach, a total of 336 inter-residue distance restraints were obtained: 99 short range (i to i ± 1), 72 medium range (i to i ± 2, 3, or 4), 69 long range (i to i ± >4), and 96 incompatible.

Discussion

Several factors likely account for the difficulties we have encountered in determining the solution structure of B6. First, the protein has fairly modest chemical shift dispersion, since it is helical and has fairly low sequence complexity. As a result, almost no NOEs can

be assigned unambiguously. Aria was designed to use ambiguous NOEs in protein structure calculations. The main barrier we have faced using Aria is that there are systematic differences between the chemical shifts observed in the different spectra. When handling an ambiguous NOE, Aria considers all protons within a small chemical shift tolerance of the observed peak location. If the assigned chemical shifts are not sufficiently accurate, the set of potential partners considered for each ambiguous NOE may not include the correct proton.

Acquiring additional experimental data would reduce some of the ambiguity in the chemical shift assignments. The backbone amide proton and nitrogen assignments for the end of helix one and in the turn between helix one and helix two may contain some errors. Several peaks in this region were weak or absent from the CBCA(CO)NH and HNCACB experiments that were used to determine sequential backbone assignments. Additional data could be obtained from a complementary pair of backbone assignment experiments, such as the HNCO and HCACO experiments. The aromatic side chain assignments may also contain errors. NOEs were used in some cases to assign chemical shifts to a particular Phe or Trp residue. However, the aromatic residues in homeodomains are tightly clustered, so NOEs could also have arisen from inter-residue contacts. The aromatic TOCSY experiment could be used to obtain unambiguous assignments for the aromatic side chains.

A more serious difficulty is that systematic variation of chemical shifts among the spectra was observed, suggesting that the solution conditions were not constant. The spectra were acquired over a period of many months. In addition, it was necessary to exchange buffers, as some experiments were conducted in 90:10 H_2O : D_2O and other experiments were conducted in 99.9% D_2O. It is likely that buffer choice is partially responsible for the observed chemical shift variations, as phosphate is a poor buffer at pH 4.5. It would be advisable to use an alternate buffer system such as acetate in the future.

The chemical shift variation between spectra and assignment errors are likely responsible for many of the observed NOESY crosspeaks that appear incompatible with the ORBIT structure. Two other factors may also be important. First, several of the crosspeaks classified as incompatible can be satisfied by a proton pair located 6 to 7 Å apart in the ORBIT structure. While such a distance is too large to yield a NOE, slight conformational adjustments could bring such proton pairs closer together. Also, a small fraction of the protons in B6 could not be assigned. NOEs to any unassigned proton would also be classified as incompatible, even if the restraint is actually compatible with the ORBIT structure.

Conclusions

Despite the difficulties encountered, it appears likely that B6 assumes the target backbone fold rather than an alternate conformation. The backbone angle restraints generated using TALOS indicate that B6 contains three helices whose locations are very similar to the positions of the helices in the target backbone. The majority of the inter-residue NOEs are compatible with the structure generated using ORBIT. An even greater number of NOEs would be compatible with slight conformational changes or using a slightly larger chemical shift tolerance to generate lists of potential NOE partners. Even in the absence of a high resolution structure, it is probable that B6 meets all of the tests of conformational specificity described in Chapter III: it is monomeric, adopts a well-defined folded state, and assumes the target fold.

Materials and Methods

Protein expression and purification. A synthetic gene encoding B6 was constructed as described previously[6]. $^{13}C,^{15}N$-labeled protein was expressed in BL21 (DE3) *Escherichia coli* cells (Stratagene) using minimal media supplemented with Basal medium eagle vitamin

solution (Gibco). ^{13}C-glucose was used as the carbon source and ^{15}N-ammonium sulfate was used as the nitrogen source (Isotec). The protein was isolated using the freeze-thaw method[7] and purified by reversed-phase HPLC using a C8 prep column (Zorbax) and linear water-acetonitrile gradient with 0.1% trifluoroacetic acid.

NMR spectroscopy. All samples were dissolved in 50 mM sodium phosphate and 50 μM sodium fluoride in either 90:10 H_2O:D20 or 99.9% D_2O adjusted to pH* 4.5. Final protein concentration was 1 mM. (HB)CB(CGCD)HD and (HB)CB(CGCDCE)HE experiments were performed using a Varian INOVA 500 MHz spectrometer and all other experiments were acquired on a Varian INOVA 600 MHz spectrometer. All spectra were acquired at 20 °C. Typical spectral widths for data acquired at 600 MHz were 6982 Hz (^{1}H), 1300 Hz (^{15}N), and 4500 - 8000 Hz (^{13}C). Protein Pack (Varian, Inc.) pulse sequences were used for all experiments other than the (HB)CB(CGCD)HD and (HB)CB(CGCDCE)HE experiments.

Sequential backbone assignments were obtained using CBCA(CO)NH and HNCACB experiments. ^{15}N edited TOCSY, C(CO)NNH-TOCSY and HCCH-TOCSY experiments were used to obtain side chain assignments. Aromatic ring assignments were obtained using 2D (^{13}C,1H)-HSQC, (HB)CB(CGCD)HD, (HB)CB(CGCDCE)HE8, and 3D aromatic ^{13}C-edited NOESY experiments.

Distance restraints were obtained from NOE cross peak volumes in ^{13}N edited NOESY and in aliphatic and aromatic ^{13}C edited NOESYspectra acquired using a 75 ms mixing time.

NMR data processing and analysis. Data were processed initially in VNMR (Varian, Inc.). Data were extended by linear prediction and processed in NMRPipe[9].

NMR assignments. All chemical shift assignments were made using NMRView version 5.0.4[10] using the experiments described above. For each NOESY crosspeak, the carbon or nitrogen and its covalently bonded proton were assigned based on chemical shift. In addition, intra-residue NOEs were assigned based on chemical shift. NOESY crosspeaks that did not correspond to previously assigned chemical shifts were not used. The chemical shifts in the aliphatic ^{13}C-edited NOESY were observed to be slightly different than the previously assigned chemical shifts. To improve the accuracy of subsequent structure calculations, a separate set of assignments was made for the ^{13}C(aliphatic) HSQC-NOESY based on the location of intra-residue crosspeaks.

Angle restraints. ϕ and ψ angle restraints were obtained using the computer program TALOS[11]. Statistically significant predictions (that is, all ten tripeptides in the same region of ϕ,ψ space, or nine of ten in the same region of ϕ,ψ space and the tenth also has $\phi < 0$, or nine of ten in the same region of ϕ,ψ space with $\phi > 0$) were made for 40 of the 52 residues in the protein. Tolerances of ± 30° about each predicted angle were used during structure calculations.

Structure calculations. Ambiguous NOE assignments and structure calculations were conducted using Aria[12, 13] coupled with CNS[14]. Integrated NOE peak lists, chemical shift assignments, and torsion angle restraints were input to Aria. Chemical shift tolerances of ± 0.03 ppm were used to determine the set of protons that might be participating in each ambiguous NOE. Eight rounds of torsion angle simulated annealing followed by one round of water refinement were carried out. In all cases, the assignment of the carbon or nitrogen and the covalently attached proton were held fixed. For intra-residue NOEs, the second

proton involved in the NOE was also assigned, while for inter-residue NOEs Aria was used to determine the identity of the second proton.

The best results were obtained using two sets of Aria calculations. In the first Aria run, backbone angle restraints and NOEs to methyl groups (chemical shift < 1.0 ppm) or aromatic groups (chemical shift > 6.0 ppm for carbon-edited experiments conducted in D_2O, so amide protons are exchanged away) only are used. Aromatic - methyl contacts are typically long range, and can be used to define the topology of a protein fold. Next, a second Aria run was conducted starting from the lowest energy structure generated in the final iteration of the first run. In the second round, all of the inter-residue and intra-residue distance restraints and the backbone angle restraints were used.

Compatibility with ORBIT structure. Each observed inter-residue NOE was analyzed for compatibility with the structure of B6 generated using ORBIT. The distance between each pair of protons that could satisfy each ambiguous NOE, using a chemical shift tolerance of ± 0.03 ppm, was calculated based on the ORBIT structure. For Phe residues, distances were calculated for both HD1 and HD2 or HE1 and HE2 atoms, as these pairs are degenerate. Distances to methyl groups were calculated using the averaged location of the three methyl protons, and distances to both methyl groups in Leu and Val were calculated. Distances involving methylene pairs were calculated to the atom numbered "1" for degenerate pairs and as numbered for non-degenerate pairs.

NOEs were classified as compatible with the ORBIT structure if at least one possible assignment consisted of a proton pair separated by less than 6 Å in the ORBIT structure. The compatible NOEs were further classified as short range (residue i to residue i ± 1), medium range (i to i ± 2, 3, or 4), or long range (i to i ± > 4). In cases where a NOE could be satisfied by more than one pair, it was binned with the shortest range restraint possible.

That is, a NOE that could be satisfied by a medium or a long range contact is classified as a medium range contact.

References

1. Morgan, C. S. (2000) Ph.D. Thesis. California Institute of Technology, Pasadena, CA.
2. Dahiyat, B. I. and Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science*, **278**, 82-87.
3. 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.
4. Harbury, P. B., Kim, P. S. and Alber, T. (1994). Crystal structure of an isoleucine-zipper trimer. *Nature*, **371**, 80-83.
5. Cavanagh, J., Fairbrother, W. J., Palmer, A. G. I. and Skelton, N. J. (1996). In *Protein NMR spectroscopy: principles and practice.* Academic Press, San Diego.
6. Marshall, S. A. and Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol.*, **305**, 619-631.
7. Johnson, B. H. and Hecht, M. H. (1994). Recombinant proteins can be isolated from E. coli cells by repeated cycles of freezing and thawing. *Biotechnology*, **12**, 1357-1360.
8. Yamazaki, T., Forman-Kay, J. D. and Kay, L. E. (1993). Two-dimensional NMR experiments for correlating $^{13}C\beta$ and $^{1}H\delta/$[illegible]chemical shifts of aromatic residues in ^{13}C-labeled proteind via scalar couplings. *J. Am. Chem. Soc.*, **115**, 11054-11055.
9. Delaglio, F., Grzesiek, S., Vuister, G., Zhu, G., Pfeifer, J. and Bax, A. (1995). NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR*, **6**, 227-293.
10. Johnson, B. A. and Blevins, R. A. (1994). NMRView- A computer program for the visualization and analysis of NMR data. *J. Biomol. NMR*, **4**, 603-614.
11. Cornilescu, G., Delaglio, F. and Bax, A. (1999). Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR*, **13**, 289-302.

12. Nilges, M. and O'Donoghue, S. I. (1998). Ambiguous NOEs and automated NOE assignment. *Prog. Nucl. Mag. Res. Sp.*, **32**, 107-139.

13. Linge, J. P., O'Donoghue, S. I. and Nilges, M. (2001). Automated assignment of ambiguous nuclear Overhauser effects with Aria. *Methods Enzymol.*, **339**, 71-90.

14. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. and Warren, G. L. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D*, **54**, 905-921.

Table IV-I: Chemical shift assignments for B6

Met 1	HA: 3.78
Ser 2	N: 116.84, HN: 8.42, CA: 57.25, HA: 4.25, CB: 63.41, HB: 4.05
Lys 3	N: 122.12, HN: 8.88, CA: 57.30, HA: 4.30, CB: 32.87, HB: 1.79, CG: 24.70, HG: 1.42, CD: 28.97, HD: 1.69, CE: 42.17, HE: 2.99
Phe 4	N: 121.71, HN: 8.49, CA: 58.29, HA: 4.57, CB: 38.76, HB2: 3.12, HB1: 2.98, CD: 131.73, HD: 7.15, CE: 131.29, HE: 7.21, CZ: 128.71, HZ: 7.09
Asp 5	N: 122.53, HN: 7.75, CA: 54.46, HA: 4.50, CB: 41.48, HB2: 2.92, HB1: 2.69
Glu 6	N: 199.79, HN: 8.31, CA: 59.18, HA: 3.89, CB: 29.25, HB: 2.04, CG: 35.90, HG: 2.38
Gln 7	N: 118.15, HN: 8.10, CA: 58.70, HA: 4.00, CB: 27.99, HB: 2.05, CG: 34.16, HG2: 2.46, HG1: 2.38
Leu 8	N: 122.06, HN: 7.88, CA: 57.60, HA: 3.97, CB: 41.31, HB2: 1.58, HB1: 1.52, CG: 27.35, HG: 1.53, CD1: 24.57, HD11: 0.864, CD2: 25.20, HD21: 0.83
Lys 9	N: 118.91, HN: 7.91, CA: 60.35, HA: 3.36, CB: 32.22, HB2: 1.66, HB1: 1.50, CG: 25.24, HG2: 0.81, HG1: 0.61, CD: 29.55, HD: 1.34, CE: 41.66, HE: 2.39

Arg 10	N: 116.76, HN: 7.11, CA: 58.95, HA: 4.04, CB: 30.01, HB: 1.89, CG: 27.36, HG2: 1.78, HG1: 1.69, CD: 43.38, HD: 3.21, NE: 84.72, HE: 7.45
Lys 11	N: 119.56, HN: 7.64, CA: 58.69, HA: 4.05, CB: 31.99, HB2: 1.95, HB1: 1.86, CG: 24.70, HG2: 1.56, HG1: 1.44, CD: 28.98, HD: 1.63, CE: 42.23, HE 2.95
Leu 12	N: 119.10, HN: 8.16, CA: 57.22, HA: 3.79, CB: 40.77, HB2: 1.59, HB1: 0.96, CG: 26.11, HG: 0.66, CD1: 23.50, HD11: 0.68, CD2: 25.66, HD21: 0.63
Glu 13	N: 118.01, HN: 8.13, CA: 59.53, HA: 3.46, CB: 28.95, HB: 2.01, CG: 36.64, HG2: 2.50, HG1, 2.14
Glu 14	N: 116.97, HN: 7.49, CA: 58.87, HA: 4.01, CB: 29.49, HB: 2.12, CG: 36.00, HG2: 2.45, HG1: 2.29
Val 15	N: 119.06, HN: 7.35, CA: 65.75, HA: 3.65, CB: 31.66, HB: 2.08, CG2: 21.50, HG21: 0.97, CG1: 21.50, HG11: 0.83
Phe 16	N: 119.42, HN: 8.19, CA: 58.56, HA: 4.27, CB: 37.18, HB: 2.36, CD: 132.14, HD: 7.25, CE: 131.31, HE: 7.35, CZ: 129.65, HZ: 7.24
Lys 17	N: 118.32, HN: 8.26, CA: 55.28, HA: 4.26, CB: 33.90, HB: 1.95, CG: 25.48, HG: 1.69, CD: 29.36, HD: 1.77, CE: 41.43, HE: 3.03

Arg 18	N: 123.70, HN: 8.00, CA: 60.72, HA: 4.26, CB: 30.83, HB: 1.91, CG: 27.06, HG2: 1.76, HG1: 1.64, CD: 43.04, HD: 3.17
Asp 19	N: 122,40, HN: 8.63, CA: 54.40, HA: 4.62, CB: 40.19, HB2: 2.83, HB1: 2.71
Gln 20	N: 118.77, HN: 8.17, CA: 57.05, HA: 4.27, CB: 29.06, HB: 2.20, CG: 34.12, HG: 2.38
Arg 21	N: 121.09, HN: 8.37, CA: 55.43, HA: 4.44, CB: 30.42, HB2: 1.88, HB1: 1.79, CG: 26.88, HG: 1.62, CD: 40.33, HD: 3.22, NE: 84.72, HE: 7.35
Ile 22	N: 122.11, HN: 8.26, CA: 60.75, HA: 4.28, CB: 38.63, HB: 1.64, CG1: 27.91, HG12: 1.38, HG11: 1.05, CD1: 13.31, HD11: 0.65, CG2: 17.04, HG21: 0.51
Thr 23	N: 116.96, HN: 8.15, CA: 60.37, HA: 4.52, CB: 71.28, HB: 4.67, CG2: 21.62, HG21: 1.30
Asn 24	N: 19.46, HN: 8.99, CA: 56.77, HA: 4.30, CB: 38.04, HB2: 2.83, HB1: 2.73, ND2: 111.80, HD2: 7.72
Gln 25	N: 120.03, N: 8.56, CA: 58.99, HA: 3.97, CB: 28.19, HB: 1.95, CG: 33.61, HG: 2.42
Glu 26	N: 119.77, HN: 7.73, CA: 58.95, HA: 4.13, CB: 29.23, HB2: 2.31, HB1: 1.96, CG: 35.87, HG: 2.38

Leu 27	N: 120.44, HN: 8.22, CA: 58.38, HA: 3.96, CB: 41.41, HB2: 1.88, HB1: 1.54, CG: 23.99, HG: 1.56, CD1: 25.19, HD11: 0.74, CD2: 25.19, HD21: 0.67
His 28	N: 117.16, HN: 8.06, CA: 59.18, HA: 4.24, CB: 27.82, HB: 3.38, CD2: 120.34, HD2: 7.36, CE1: 136.49, HE1: 8.61
Asp 29	N: 120.99, HN, 8.48, CA: 57.35, HA: 4.37, CB: 40.15, HB2: 2.82, HB1: 2.72
Leu 30	N: 122.21, HN: 8.34, CA: 57.76, HA: 4.07, CB: 42.43, HB2: 1.86, HB1: 1.61, CG: 26.88, HG: 1.71, CD: 24.49, HD: 0.88
Ala 31	N: 121.10, HN: 8.14, CA: 55.79, HA: 3.88, CB: 17.86, HB1: 1.56
Gln 32	N: 115.79, HN: 7.75, CA: 58.46, HA: 4.03, CB: 28.55, HB: 2.12, CG: 33.98, HG: 2.37
Lys 33	N: 120.13, HN: 8.09, CA: 58.99, HA: 4.04, CB: 32.66, HB2: 1.97, HB1: 1.92, CG: 25.10, HG2: 1.56, HG1: 1.44, CD: 29.13, HD2: 1.77, HD1: 1.64, CE: 42.14, HE: 2.96
Leu 34	N: 115.82, HN: 8.25, CA: 55.00, HA: 4.28, CB: 42.07, HB2: 1.72, HB1: 1.50, CG: 26.01, HG: 1.83, CD1: 22.28, HD11: 0.81, CD2: 26.07, HD21: 0.76,
Gly 35	N: 109.09, HN: 7.87, CA: 46.61, HA: 3.91

Ile 36	N: 115.69, HN: 7.66, CA: 55.52, HA: 4.50, CB: 41.52, HB: 1.62, CG1: 25.49, HG12: 1.32, HG11: 1.00, CD1: 13.87, HD11: 0.72, CG2: 17.87, HG21: 0.86
Asn 37	N: 121.02, HN: 8.26, CA: 53.94, HA: 4.47, CB: 39.21, HB2: 2.85, HB1: 2.81, ND2: 112.03, HD21: 7.79, HD22: 7.09
Glu 38	N: 124.92, N: 9.22, CA: 60.05, HA: 4.46, CB: 28.83, HB: 1.96, CG: 35.36
Glu 39	N: 118.12, HN: 8.80, CA: 59.48, HA: 4.01, CB: 28.47, HB2: 2.14, HB1: 2.01, CG: 36.08, HG2: 2.50, HG1: 2.39
Leu 40	N: 119.49, HN: 7.44, CA: 57.35, HA: 4.11, CB: 41.61, HB2: 1.64, HB1: 1.36, CG: 26.84, HG: 1.39, CD1: 24.39, HD11: 0.91, CD2: 24.39, HD21: 0.80
Ile 41	N: 117.82, HN: 7.13, CA: 64.85, HA: 3.72, CB: 37.41, HB: 1.92, CG1: 28.75, HG12: 1.61, HG11: 1.09, CD1: 13.33, HD11: 0.80, CG2: 18.19, HG21: 0.98
Glu 42	N: 119.43, HN: 8.24, CA: 59.81, HA: 4.00, CB: 28.52, HB: 2.16, CG: 35.46, HG2: 2.50. HG1: 2.39
Asp 43	N: 119.36, N: 8.05, CA: 57.49, HA: 4.45, CB: 41.24, HB2: 2.88, HB1: 2.77
Trp 44	N: 120.39, HN: 8.21, CA: 60.44, HA: 4.26, CB: 29.02, HB2: 3.46, HB1: 3.22, CD1: 125.31, HD1: 7.08, NE1: 128.81, HE1: 10.07, CZ2: 114.28, HZ2: 7.28, CH2: 124.27, HH2: 7.08, CZ3: 121.58, HZ3: 6.96, CE3: 120.75, HE3: 7.06

Phe 45	N: 118.52, HN: 8.25, CA: 59.66, HA: 4.02, CB: 40.26, HB: 3.88, CD: 131.31, HD: 6.77, CE: 131.11, HE: 7.09, CZ: 129.35, HZ: 6.95
Arg 46	N: 125.70, HN: 7.83, CA: 59.20, HA: 3.97, CB: 30.00, HB2: 2.06, HB1: 1.91, CG: 27.90, HG: 1.67, CD: 43.42, HD: 3.27
Arg 47	N: 118.30, HN: 7.75, CA: 58.96, HA: 3.95, CB: 29.96, HB: 1.73, CG: 27.76, HG2: 1.67, HG1: 1.47, CD: 43.24, HD2: 3.02, HD1: 2.96
Trp 48	N: 122.07, HN: 8.25, CA: 60.29, HA: 4.12, CB: 28.22, HB2: 3.03, HB1: 2.74, CD1: 126.35, HD1: 7.07, NE1: 128.11, HE1: 10.11, CZ2: 114.54, HZ2: 7.23, CH2: 124.27, HH2: 6.85, CZ3: 121.38, HZ3: 6.67, CE3: 120.75, HE3: 7.34
Glu 49	N: 118.04, HN: 8.34, CA: 58.17, HA: 3.60, CB: 29.09, HB2: 1.91, HB1: 1.85, CG: 35.16, HG2: 2.28, HG1: 2.12
Gln 50	N: 116.40, HN: 7.47, CA: 56.55, HA: 4.11, CB: 29.10, HB2: 2.14, HB1: 2.05, CG: 34.10, HG2: 2.49, HG1: 2.39
Gln 51	N: 118.65, HN: 7.59, CA: 55.85, HA: 4.36, CB: 29.01, HB2: 2.09, HB1: 1.96, CG: 33.90, HG2: 2.36, HG1: 2.31
Arg 52	N: 126.43, HN: 7.56, CA: 57.42, HA: 3.99, CB: 31.09, HB2: 1.63, HB1: 1.59, CG: 26.80, HG: 1.41, CD: 43.14, HD: 2.85

Table IV-II: Template versus TALOS-predicted backbone dihedral angles

residue	□(template)	□(TALOS)	ψ(template)	ψ(TALOS)
Phe 4	-77	-76	132	143
Glu 6	-57	-66	-42	-40
Gln 7	-68	-64	-46	-36
Leu 8	-57	-66	-52	-38
Lys 9	-55	-65	-51	-42
Arg 10	-52	-64	-53	-35
Lys 11	-60	-63	-47	-43
Leu 12	-55	-68	-49	-38
Glu 13	-59	-64	-46	-42
Glu 14	-56	-65	-50	-37
Val 15	-58	-68	-45	-43
Phe 16	-58	-65	-48	-31
Arg 18	-56	-64	-52	-27
Asp 19	-153	-74	131	21
Arg 21	-109	-90	128	136
Thr 23	-79	-110	159	160
Asn 24	-59	-62	-38	-33
Gln 25	-70	-66	-47	-40
Glu 26	-61	-68	-41	-37
Leu 27	-56	-64	-51	-39
His 28	-54	-67	-50	-38
Asp 29	-57	-65	-53	-34

Leu 30	-60	-66	-41	-38
Ala 31	-52	-66	-56	-37
Gln 32	-51	-65	-52	-39
Leu 34	-80	-91	-21	-1
Gly 35	76	78	45	20
Ile 36	-129	-118	154	148
Glu 38	-36	-68	-43	-38
Glu 39	-61	-65	-36	-37
Leu 40	-72	-65	-42	-41
Ile 41	-63	-69	-50	-38
Glu 42	-54	-62	-53	-40
Asp 43	-54	-65	-52	-40
Trp 44	-54	-67	-51	-39
Phe 45	-56	-69	-52	-35
Arg 46	-56	-69	-52	-35
Arg 47	-60	-66	-43	-40
Trp 48	-58	-64	-48	-41
Glu 49	-56	-67	-48	-34
Gln 51	-61	-81	-48	136

Figure IV-1. ^{15}N-HSQC spectrum of B6. The peak corresponding to Gly 35, with N = 109.09 ppm and HN = 7.87, is located outside of the cropped region of the spectrum that is shown. Assigned peaks are marked with their corresponding residue number.

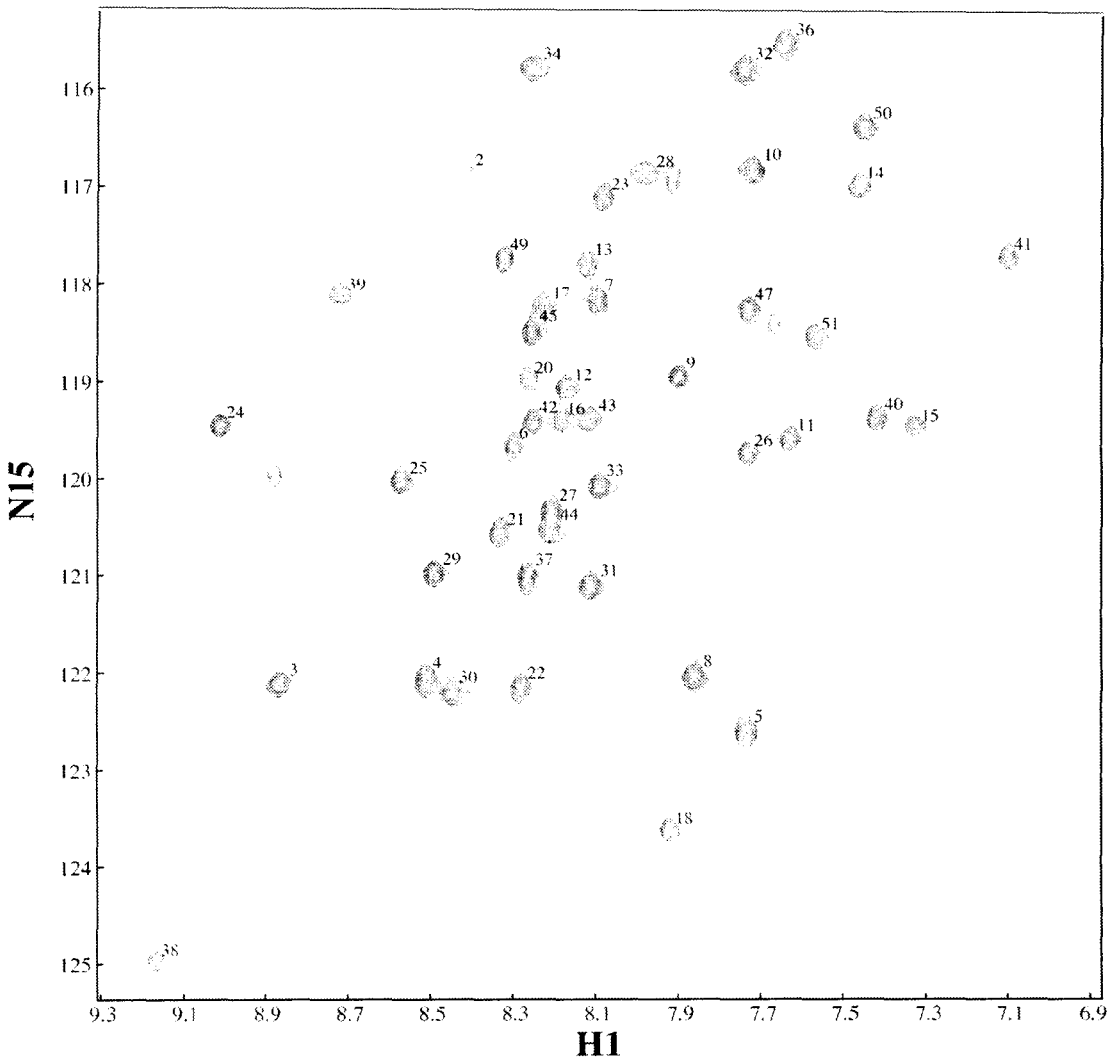

N15
H1
116
117
118
119
120
121
122
123
124
125
9.3
9.1
8.9
8.7
8.5
8.3
8.1
7.9
7.7
7.5
7.3
7.1
6.9

Figure IV- 2. Example of the strip plots obtained from the backbone assignment experiments HNCACB (the left strip in each pair) and CBCA(CO)NH (the right strip in each pair). Peaks with positive intensity are shown in black and peaks with negative intensity are shown in red. The x-axis corresponds to amide proton chemical shifts, the y-axis gives CA and CB chemical shifts, and the z-axis gives amide nitrogen chemical shifts.

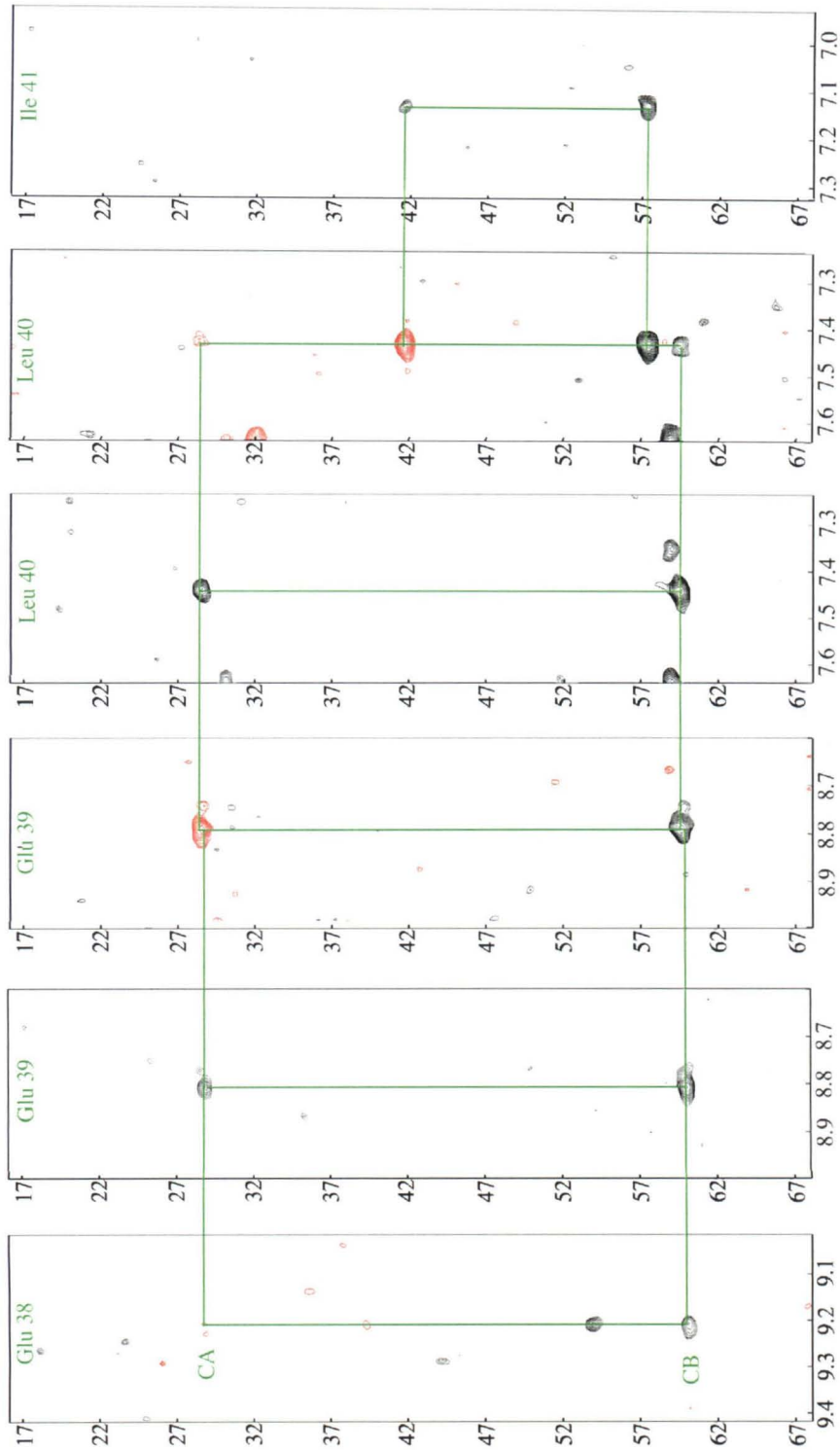

Glu 38
Glu 39
Glu 39
Leu 40
Leu 40
Ile 41
CA
CB

Figure IV-3. Structures generated using Aria (shown in red) versus the ORBIT-generated structure of B6 (shown in blue). For clarity, each individual structure as well as the superimposition of the two structures is shown.

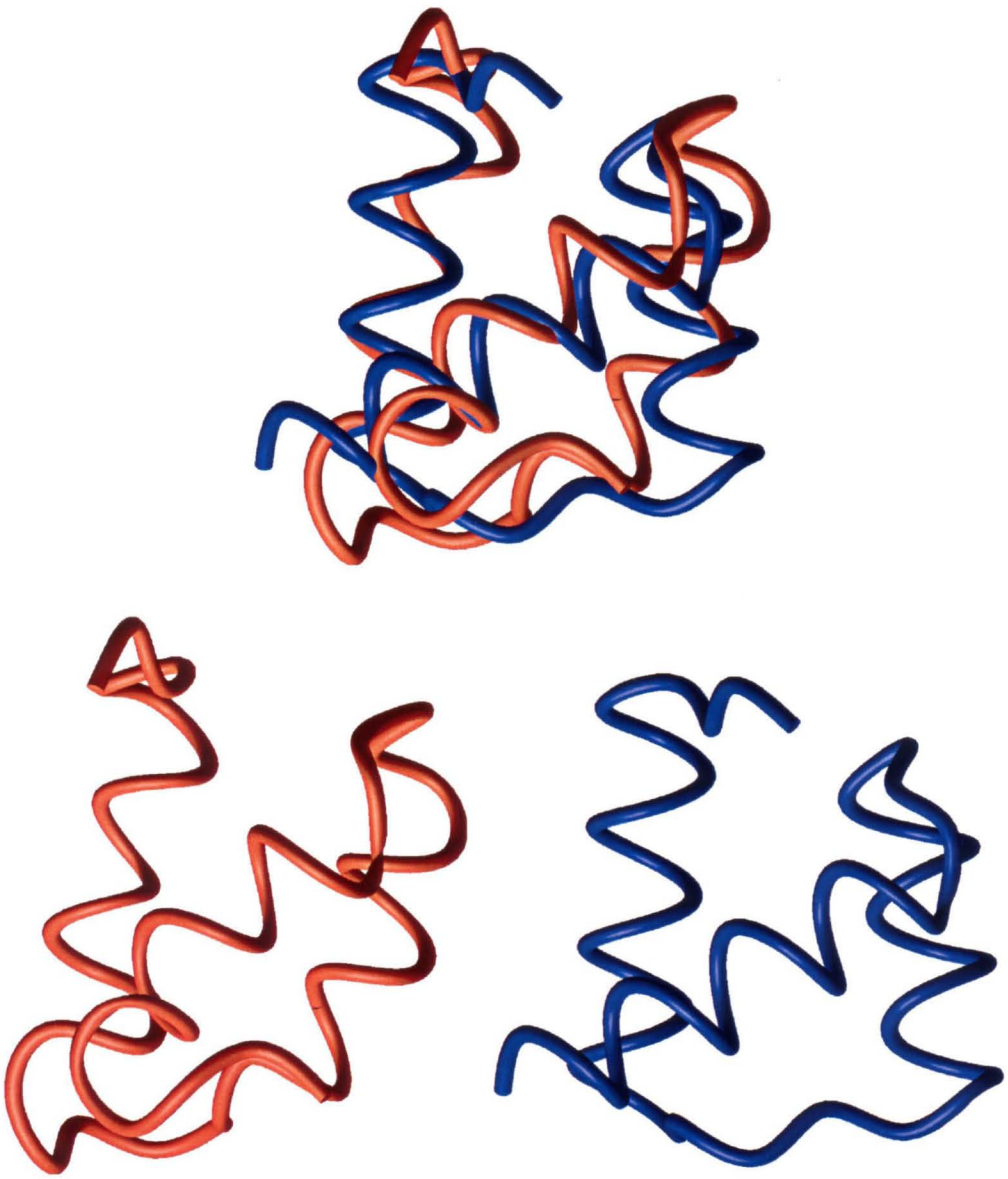


Figure IV-4. The number of short ($i \pm 1$, shown in red), medium ($i \pm 2$, 3, or 4, shown in green), and long ($i \pm > 4$, shown in blue) ambiguous NOEs for each residue that are compatible with the structure of B6 generated using ORBIT. Ambiguous NOEs for each residue that are not compatible with the ORBIT structure are indicated in black. In all cases, the residue number corresponds to the residue containing the heteroatom and covalently attached proton. If an ambiguous restraint can be satisfied by more than one partner, it is binned with the closest possible distance class. For example, if an ambiguous NOE can be satisfied by either a short or medium range contact, it is classified as a short range contact.

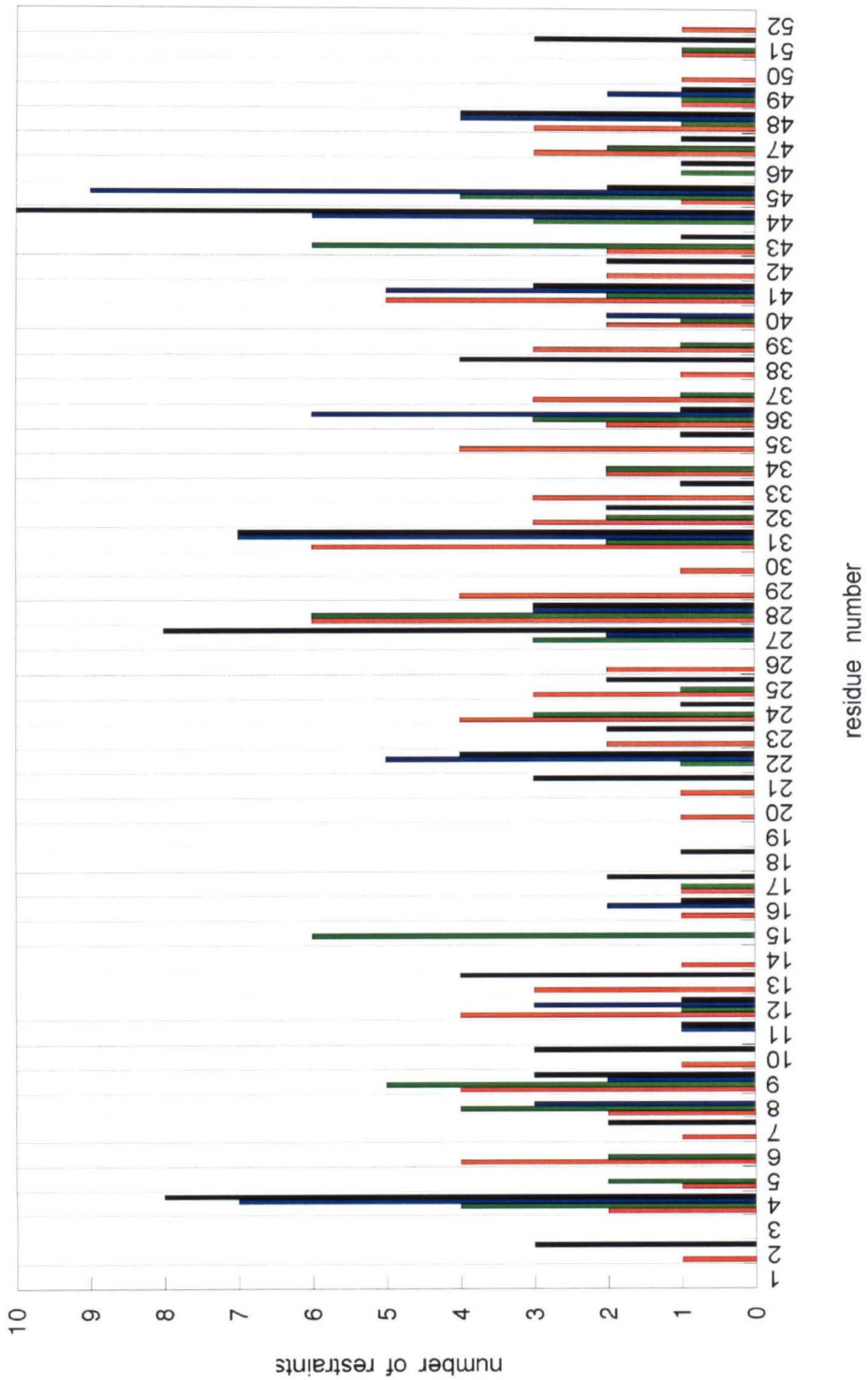
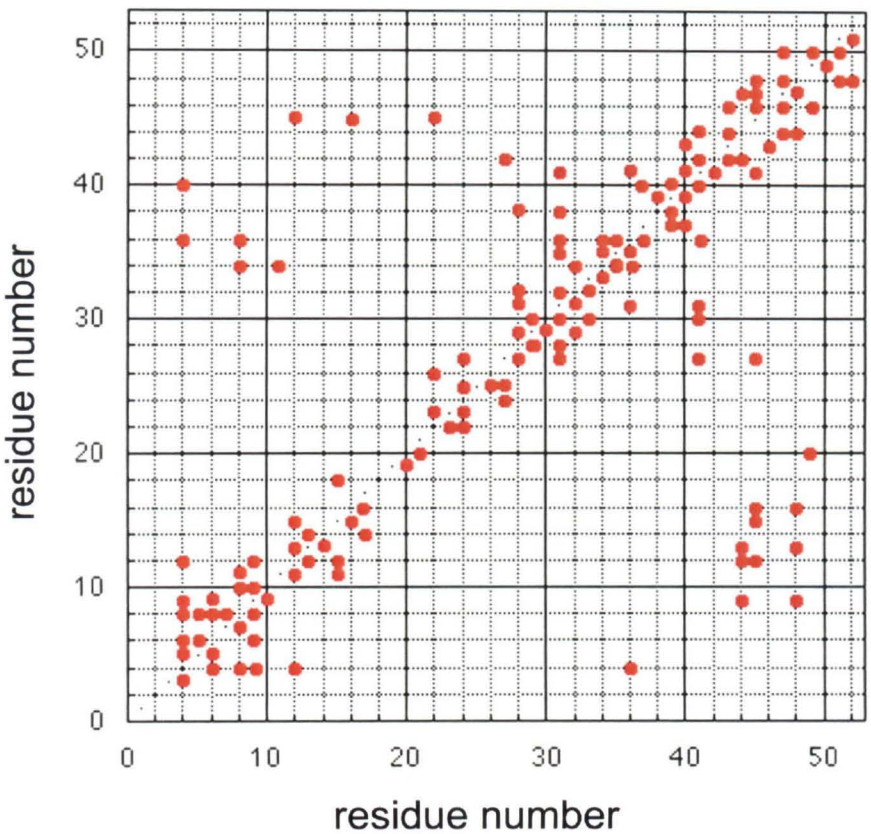

number of restraints
0
1
2
3
4
5
6
7
8
9
10
residue number
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52

Figure IV-5. NOE contact map for B6. Each red dot corresponds to at least one NOE between the two residues. The x-axis residue is covalently attached to the heteroatom and the y-axis corresponds to the NOE partner. Only ambiguous restraints that are compatible with the structure of B6 generated using ORBIT are considered. If an ambiguous restraint can be satisfied by more than one partner, it is binned with the closest possible distance class. Within a given distance class, the shortest interatomic distance in the model structure is used. For example, if an ambiguous NOE can be satisfied by either a short or medium range contact, it is classified as short, and if an ambiguous NOE can be satisfied by two short range contacts, one with a predicted interatomic distance of 3 Å and the other with a predicted distance of 5 Å, the 3 Å contact will be used.


residue number
0
10
20
30
40
50
0
10
20
30
40
50
residue number

Chapter V

Electrostatics Significantly Affect the Stability of Designed Homeodomain Variants

The text of this chapter is adopted from a manuscript that was coauthored with Chantal S. Morgan and Professor Stephen L. Mayo.

S. A. Marshall, C. S. Morgan, and S. L. Mayo. *J. Mol. Biol.* accepted.

Abstract

The role of electrostatic interactions in determining the stability of designed proteins was studied by constructing and analyzing a set of designed variants of the *Drosophila* engrailed homeodomain. Computational redesign of 29 surface positions results in a 25-fold mutant with moderate stability similar to the wild type protein. Incorporating helix dipole and N-capping considerations into the design algorithm by restricting amino acid composition at the helix termini and N-capping positions yields a nine-fold mutant of the initial design (a 23-fold mutant of wild type) that is over 3 kcal mol^{-1} more stable than the protein resulting from the unbiased design. Four additional proteins were constructed and analyzed to isolate the effects of helix dipole and N-capping interactions in each helix. Based on the results of urea denaturation experiments and calculations using the finite difference Poisson-Boltzmann (FDPB) method, both classes of interaction are found to significantly increase the stability of the designed proteins. The simple electrostatic model used in the ORBIT force field, which is similar to the electrostatic models used in other protein design force fields, is unable to predict the experimentally determined stabilities of the designed variants. The helix dipole and N-capping restrictions provide a simple but

effective method to incorporate two types of electrostatic interactions that significantly impact protein stability.

Introduction

Electrostatic interactions are often critical determinants of protein structure and function. Computational protein design algorithms[1-5] typically use fast, two-body methods based on Coulomb's law and/or explicit hydrogen bond terms to model electrostatic interactions. These methods are computationally efficient, but fail to accurately account for desolvation of polar protein groups and solvent screening of Coulombic interactions, which both strongly attenuate electrostatic interactions in proteins. Continuum models of electrostatics based on the finite difference Poisson-Boltzmann (FDPB) method[6] are thought to have substantially better predictive power, but are far too slow for computational methods that attempt to address a large combinatorial sequence space. The ORBIT (Optimization of Rotamers by Iterative Techniques) protein design force field intentionally de-emphasizes electrostatic interactions so that inaccurate electrostatic energies do not dominate the more accurate components of the force field. This is accomplished by using a high, distance-dependent dielectric of $40r$ and partial charges on polar groups that are somewhat smaller than those used in most other force fields[7]. To date, the role of electrostatics in protein design has not been the subject of extensive experimental testing and so the consequences of using a highly approximate model of electrostatics in design force fields are not well understood.

The engrailed homeodomain, shown in Figure V-1, was selected to test the importance of electrostatic interactions in the design of protein surfaces. Alpha-helical proteins provide an excellent model system to examine a variety of electrostatic interactions. Polar side chains on an α-helix can form hydrogen bonds and salt bridges with each other when

appropriately spaced. Charged side chains may also interact with the helix dipole, which consists of partial positive and negative charges located at the N and C termini, respectively, of each α-helix. This dipole arises from the three unsatisfied hydrogen bonds at each of the helix termini[8-11]. Interactions between side chains and the helix dipole have been demonstrated to impact the stability of both model peptides and proteins[8, 12, 13]. N-capping interactions, which are hydrogen bonds between the side chain at the position immediately preceding the helix and one of the three most N-terminal backbone amides in the helix, also can confer stability to both model peptides and proteins[14, 15]. C-capping residues have a more modest effect on helical stability[14]. Helix dipole and N-capping interactions can be incorporated into computational protein design using simple rules, thus providing a readily implemented method to test the importance of electrostatic interactions in designed proteins.

Initial design results

In order to test the effectiveness of our current design methodology, we calculated the optimal amino acid sequence and rotameric conformations for twenty-nine surface positions of the engrailed homeodomain. The sequence of the resulting protein, called NC0, is a twenty-five fold mutant from wild type, as shown in Figure V-2. The NC0 sequence is strongly biased towards large, charged amino acids: Arg, Glu, or Lys was selected at twenty-two of the twenty-nine variable positions, while they are present at only eleven of the twenty-nine positions in the wild type protein. NC0 has only one predicted N-capping interaction, compared to three in wild type, and NC0 has seven violations of helix dipole "rules" (positive charges in the three most N-terminal helical positions or negative charges in the three most C-terminal helical positions) while wild type has only three. NC0 is charge neutral, while the wild type protein has a +7 charge, and NC0 is predicted to contain many more side chain

- side chain hydrogen bonds and salt bridges than were observed in the wild type crystal structure[16], as shown in Figures V-3 through V-5.

Despite these differences, wild type homeodomain and NC0 have similar, modest stability as measured by thermal denaturation experiments, as indicated in Figure V-6 and Table V-1. While matching the stability of the wild type protein after changing nearly 50 % of the residues would typically be considered a positive result in protein design studies, it is likely that the electrostatic interactions on the surface of both wild type homeodomain and the designed NC0 variant are not optimal for stability. The electrostatic interactions on the surface of NC0 appear suboptimal for stability because favorable helix dipole and capping interactions that are present in wild type homeodomain are not present in the designed protein. The arrangement of charged residues on the wild type protein likely reflects the functional role of homeodomains – DNA binding. While it is likely that the high net charge of wild type homeodomain contributes to its DNA binding affinity, it would be expected to destabilize the isolated protein.

Incorporating N-capping and helix dipole interactions

To prevent unfavorable helix dipole interactions and to ensure favorable N-capping interactions, we performed a second protein design calculation in which we limited the allowed residues at the N-capping positions and the three most N- and C- terminal positions of each of the three helices. N-capping positions were restricted to the four amino acids with the highest N-capping propensity (Ser, Thr, Asn, and Asp)[17], positively charged amino acids (His, Lys, and Arg) were disallowed at the N-terminal positions, and negatively charged amino acids (Asp and Glu) were disallowed at the C-terminal positions. The resulting sequence, called NC3-Ncap, is a 23-fold mutant from wild type and a nine-fold mutant

from NC0. Seven of the nine mutations relative to NC0 are required to fix violations of the helix dipole and capping rules. Two additional mutations (H27R and K12R) arise as the effects of the required mutations propagate through the protein. Like NC0, the NC3-Ncap sequence is heavily biased towards large, charged amino acids. NC3-Ncap is predicted to make many more favorable side chain - side chain salt bridges than wild type, but slightly fewer than NC0. Chemical and thermal denaturation experiments, shown in Figures V-6 and V-7 and Table V-1, indicate that NC3-Ncap is significantly more stable than both NC0 and wild type. The dramatic increase in stability is notable given the crude manner in which helix dipole and capping interactions were introduced, and suggests that electrostatic interactions can significantly modulate the stability of designed proteins.

Since NC3-Ncap is a nine-fold mutant from NC0, it is difficult to identify which of the mutations are responsible for the differences in stability between the two proteins. To further elucidate the role of helix dipole and capping interactions in NC3-Ncap, we constructed four additional proteins called H1, H2, H3, and CAP. Relative to NC0, H1 contains two mutations (K6E and E16R) that fix interactions with the helix one dipole, H2 contains two mutations (R24E and E31Q) that fix interactions with the helix two dipole, H3 contains one mutation (E50Q) that fixes interactions with the helix three dipole, and CAP contains two mutations (E22T and R36N) that fix N-capping interactions, as shown in Figure V-2. In all cases, the residues were mutated from the amino acid selected in the NC0 calculation to the amino acid selected in the NC3-Ncap calculation. The stabilities of H1 and CAP determined by thermal and urea denaturation are between NC0 and NC3-Ncap. H2 and H3 are slightly destabilized relative to NC0 in urea denaturation experiments and have similar stability to NC0 in thermal denaturation experiments, as indicated in Table V-1 and Figures V-6 and V-7.

Comparing the homeodomain variants using Poisson-Boltzmann electrostatics

The total energies predicted by the ORBIT force field, shown in Table V-2, are not properly correlated with the measured stabilities. When the contributions of each energy term are considered separately, the side chain - backbone Coulombic energies and the side chain - backbone hydrogen bond energies are observed to improve as the helix dipole and capping restrictions, respectively, are incorporated. However, the magnitude of these changes- are quite different: the range in side chain - backbone Coulombic energies in the designed variants is 1.4 kcal mol^{-1}, while the range in side chain - backbone hydrogen bond energies is 8.8 kcal mol^{-1}. In contrast with the side chain - backbone energies, the side chain - side chain hydrogen bond and Coulombic energies are anticorrelated with experimentally determined stability among the designed variants. While they are a poor predictor of stability, the side chain - side chain hydrogen bond energies are large in magnitude and therefore can dominate sequence selection.

Earlier design studies have suggested that predicted side chain - side chain salt bridges and hydrogen bonds in designed proteins do not necessarily contribute to stability and may not be significantly populated[1, 18]. Since desolvation and loss of side chain entropy, which oppose salt bridge and hydrogen bond formation, are not included in the force field, it is not surprising that the energetic benefit of side chain - side chain salt bridges and hydrogen bonds is overemphasized. The polar hydrogen burial penalty is unlikely to accurately capture the desolvation energy for several reasons, including its failure to penalize burial of carboxylates or to account for desolvation of polar groups that form hydrogen bonds.

To better understand the effects of the helix dipole and capping rules and the limitations of the current electrostatic potential, electrostatic energies for each protein were calculated by the FDPB method using the computer program DelPhi[6,19-22]. As shown in

Table V-3, three classes of interactions were considered: the desolvation energy of each side chain, the screened Coulombic interaction between each side chain and the protein backbone, and the screened Coulombic interaction between each pair of side chains.

The FDPB calculations indicate that the distribution of electrostatic energies in the designed homeodomain variants is quite different than in wild type. Wild type pays a lower side chain desolvation penalty than any of the designed variants, largely because it has significantly fewer charged residues, hydrogen bonds, and salt bridges. Its side chain - backbone interactions are more favorable than all of the designed variants except NC3-Ncap. Side chain - side chain interactions are predicted to be slightly favorable in wild type and significantly more favorable in the designed variants.

Similarly, the FDPB results can be used to directly compare the designed proteins. NC0 is predicted to have a high desolvation penalty, the least favorable side chain - backbone interactions, and the most favorable interactions between side chains. The other designed proteins, H1, H2, H3, CAP, and NC3-Ncap, are all predicted to have more favorable side chain - backbone electrostatic interactions than NC0. In the set of designed variants, the positions that were mutated relative to NC0 experience the largest changes in side chain - backbone electrostatic energies. In addition, the desolvation energy changes significantly only at the positions that were mutated relative to NC0 and at their hydrogen bond and salt bridge partners. Incorporation of the helix dipole and capping rules is observed to reduce the number of predicted side chain - side chain salt bridges and hydrogen bonds. As a result, the desolvation energy is highest for NC0 and H3 and lowest for NC3-Ncap while side chain - side chain electrostatic interactions are predicted to be most favorable for NC0 and H3 and least favorable for NC3-Ncap.

The results of the FDPB calculations suggest that electrostatic interactions are the primary source of the stability differences among the designed proteins. Another possible

source of the variation in stability is helix propensity, which was found to be an important predictor of stability in surface designs of GCN4[1]. A rough estimate of relative helix propensity was obtained by summing the standard free energies of helix propagation[23] of the amino acids at surface helical positions. The wild type protein has somewhat better helical propensity than the designed variants, as shown in Table V-3, which largely results from the presence of three Ala residues in wild type. Nonetheless, the designed proteins all have fairly good helix propensities, as the long, charged amino acids that are systematically favored by the design calculations also have among the best helix propensities. NC3-Ncap has slightly greater helix propensity than the other designed variants as a result of the H27R mutation. The other designed variants have very similar calculated helix propensities. This suggests that differences in helix propensity may be important, but are not sufficient to account for the observed stability trends in the homeodomain surface variants.

Analysis of electrostatic interactions in the homdeodomain variants

Interactions between a side chain and the helix dipole can contribute to protein stability. However, the strength of a single side chain - helix dipole or N-capping interaction depends on the identity of the side chain and the conformation of the side chain and the backbone. As a result, the number of rules violations is not sufficient to predict protein stability. For instance, while all negative and neutral amino acids are considered at the N1 position of each helix, the free energy of the helix - coil transition varies by over 1 kcal mol^{-1} among the allowed residues[24]. In a design calculation, it is important to consider all of the interactions that the side chain forms with the backbone and with other side chains in order to select the optimal sequence. In some instances, the energetic benefit that is gained by forming a favorable side chain - backbone electrostatic interaction may be overshadowed by unfavorable steric or electrostatic interactions with other side chains. While including the helix dipole

and capping rules is an improvement over the current electrostatic model, it would be desirable to have a more sophisticated model that captures additional context effects.

Trends in the FDPB desolvation and side chain - backbone energies mirror the observed trends in protein stability. The sum of the FDPB desolvation, side chain - backbone, and side chain - side chain electrostatic energies and the ORBIT van der Waals energy correctly predicts that NC3-Ncap is the most stable variant, but significantly underestimates the stability of the wild type protein. A significant limitation of these calculations is that they rely on protein structures with side chain conformations predicted using ORBIT. In all of the design calculations, the backbone conformation was held fixed and all of the side chains were modeled using discrete rotamers. Since the hydrogen bond energies are large in the ORBIT force field, many of the side chains were positioned to form hydrogen bonds with other side chains. It is likely that the modeled structures contain subtle errors in backbone conformation. More importantly, a significant fraction of the side chains may sample a variety of conformations rather than assuming only the modeled rotameric state.

Electrostatic energies can be very sensitive to small changes in conformation. For example, the electrostatic energy of each pair and network of charged residues in the 40 NMR conformers of a leucine zipper fluctuates from net stabilizing to net destabilizing depending on which conformer is examined[25]. The conformations selected by ORBIT are likely to be heavily biased towards the most stable possible conformation. Previous results suggest that the side chain - side chain hydrogen bonds and salt bridges predicted by ORBIT are not actually significantly populated, further suggesting that the FDPB calculations were not performed using a sufficiently accurate model of the protein structures[1, 18].

Examination of the side chain - side chain electrostatic interaction energies in the wild type versus designed proteins reveals that the designed variants contain a small number of pairs that are predicted to have extremely favorable interactions, as shown in Figures V-

10 and V-11. Some of the side chain - side chain electrostatic energies are predicted to stabilize the folded state by as much as 5 kcal mol^{-1}. However, it is unlikely that these surface salt bridge pairs could contribute so significantly to protein stability. Many studies have investigated whether surface salt bridges can stabilize proteins. In experimental studies, the stability conferred by single surface salt bridges ranges from 0.0 kcal mol^{-1} [26] to 1.25 kcal mol^{-1} [27]; exposed salt bridges in helical proteins generally contribute no more than 0.5 kcal mol^{-1} [28].

One side chain - backbone contact, which is present in all of the designed variants, also appears to be unreasonably large. The interaction energy between Glu 2 and the backbone ranges from -4.7 to -5.1 kcal mol^{-1} and arises primarily from the +1 net charge of the N-terminus. As the N-terminal methionine was retained on all of the designed variants, the actual position of the N-terminus will be somewhat different than in the modeled structures. Additionally, very favorable contacts that are so close in primary sequence may be populated in the unfolded state, reducing their contribution to protein stability.

The salt bridge pairs with predicted energies larger than 1.5 kcal mol^{-1} dominate the total side chain - side chain electrostatic energies of the designed variants. A simple way to minimize the effect of such unreasonably large interactions is to truncate the side chain - side chain and side chain - backbone electrostatic energies at $\pm$ 1.5 kcal mol^{-1}. The truncation affects the energies of at most 29 side chain - side chain interactions and two side chain - backbone interactions in each of the designed variants out of the approximately 1900 total interactions. Using this simple modification, NC3-Ncap is predicted to be most stable and the wild type protein is predicted to have similar stability to NC0, as observed. However, the rank order of the designed variants with intermediate stability is poorly reproduced, as shown in Figure V-12.

Thresholding extremely large electrostatic interactions is unlikely to be the optimal method to compensate for the limitations of using modeled structures with fixed side chain conformations. However, it is clear that the structures generated using ORBIT contain side chain - side chain salt bridge and hydrogen bond pairs that are not actually significantly populated and that the results of FDPB calculations are sensitive to the exact locations of the side chains. As further research is conducted to improve the accuracy of electrostatic models for protein design calculations, it may be necessary to also consider further refinements in the structural models used in protein design methods.

Conclusions

Use of a simple electrostatic model in the design of protein surfaces results in the selection of sequences with moderate but suboptimal stability. The current ORBIT potential succeeds in selecting sequences with no or small net charge that have good helix propensity. However, the relative energetic contribution of various types of electrostatic interactions is not captured accurately. The stability conferred by side chain - side chain hydrogen bonds and salt bridges is overestimated, as the competing factors of desolvation and side chain entropy loss are not currently modeled. At the same time, the ORBIT force field slightly underestimates the relative importance of side chain - backbone hydrogen bonds and substantially underestimates the contribution of longer range side chain - backbone electrostatic interactions. In the context of the current electrostatic model, incorporating simple rules to account for two classes of side chain - backbone interactions, helix dipole and N-capping interactions, has been shown to significantly improve the stability of designed homeodomain variants.

Recent experimental results from several groups support the idea that optimization of the electrostatic interactions on protein surfaces can significantly increase protein stability[29-

[32]. In each study, solvent exposed residues that made unfavorable electrostatic interactions with the rest of the protein were identified; mutating these residues to neutral or oppositely charged amino acids stabilized the protein in all cases. Both the proteins studied and the methods used to identify residues that make unfavorable interactions differed significantly in the four studies, suggesting that optimization of global electrostatics is a robust and effective general strategy for increasing protein stability. Havranek and Harbury recently developed a method that calculates electrostatic energies with an accuracy comparable to continuum methods, and that is efficient enough to apply to problems with high combinatorial complexity[33]. Further development and validation of accurate electrostatic models that are compatible with the demands of protein design will significantly enhance the ability of future design studies to elucidate the relationship between sequence, structure, stability, and function.

Methods

Modeling. The engrailed homeodomain structure coordinates were obtained from PDB entry 1enh[16]. Residues 1-5 of the 56 residue domain are disordered in the absence of DNA. These residues were removed from the structure prior to performing any calculations and were not included in the proteins that were studied experimentally. The remaining residues were renumbered from 1 to 51. Explicit hydrogens were added using the program BIOGRAF (Molecular Simulations, Inc., San Diego) and the resulting structure was minimized for 50 steps using the DREIDING force field[34]. Positions along the homeodomain backbone were classified as core, boundary, or surface as previously described4.

Sequence Selection. Amino acid identities and conformations were optimized at the following 29 surface positions: 2, 4, 5, 6, 8, 9, 12, 13, 16, 17, 18, 20, 22, 23, 24, 27, 28, 31,

32, 36, 37, 38, 41, 42, 45, 46, 48, 49, and 50. Position 34, although classified as surface, was fixed to wild type Gly because of its positive f angle. For the NC0 calculation, Ala, Ser, Thr, Asp, Asn, His, Glu, Gln, Arg, and Lys were considered at each variable position. For the NC3-Ncap calculation, positively charged amino acids (His, Lys, and Arg) were disallowed at the three most N-terminal positions of each helix (positions 5, 6, 23, 24, 37, and 38), negatively charged amino acids (Asp and Glu) were disallowed at the three most C-terminal positions of each helix (positions 16, 17, 31, 32, 49, and 50), and the three N-capping positions (positions 4, 22, 36) were restricted to the four residues with the highest N-capping propensity (Ser, Thr, Asn, and Asp). At the remaining surface positions, all the amino acid types allowed in the NC0 calculation were considered. Note that several positions at helical termini (positions 7, 15, 25, 20, 39, and 51) are classified as core or boundary and therefore are not included in the calculations. In all calculations, the variable side chains were represented as discrete rotamers from the Dunbrak and Karplus backbone dependent rotamer library[35].

Pairwise side chain - backbone and side chain - side chain energies were calculated using a force field containing van der Waals, Coulombic, hydrogen bond, and polar hydrogen burial penalty terms[7]. Electrostatics were modeled using Coulomb's law with a distance-dependent dielectric of $40r$, hydrogen bonds were modeled using a 10-12 angle- and hybridization-dependent potential, and a 2.0 kcal mol^{-1} penalty was given for each buried polar hydrogen not participating in a hydrogen bond.

The optimal amino acid sequence and rotamer configuration for NC0 and NC3-Ncap were determined using the Dead-End Elimination (DEE) theorem[36-39]. The NC0 calculation considered 10^{29} amino acid sequences corresponding to 5.3×10^{68} rotamer sequences, and the NC3-Ncap calculation considered 2.0×10^{26} amino acid sequences corresponding to 2.9×10^{64} rotamer sequences. The NC3-Ncap calculation required only

1.6 CPU hours versus 5.7 CPU hours for the NC0 calculation. This 3.5-fold reduction in computational time can be largely attributed to the fifth-order dependence that DEE has on the average number of rotamers per residue position[39]. Calculations were performed using a Silicon Graphics Origin 2000 with 32 R10000 processors running at 195 MHz.

The remaining sequences were generated by changing subsets of the amino acids that violate helix dipole or capping "rules" from the NC0 sequence to the NC3-Ncap sequence. H1 corrects helix dipole interactions in helix 1, H2 corrects helix dipole interactions in helix 2, H3 corrects helix dipole interactions in helix 3, and CAP corrects the N-capping interactions. The rotameric conformations of the surface residues in variants H1, H2, H3, and CAP were determined using the same force field and optimization methods as were used for the NC0 and NC3-Ncap calculations. The optimal rotameric conformation was also calculated for the wild type sequence.

Electrostatic Calculations. Finite difference solutions to the linearized Poisson-Boltzmann equation were obtained using the computer program DelPhi (Version II, May 1998 release) with a grid size of 2.0 grids / Å, 80 % box fill, an interior dielectric of 4.0, an exterior dielectric of 80.0, and 0.050 M salt, and a probe radius of 0.0. The PARSE[41] parameter set charges and atomic radii were used in all calculations. Three classes of electrostatic energies were calculated: (1) the desolvation energy of each side chain, (2) the side chain - backbone screened Coulombic interaction energy for each side chain, and (3) the screened Coulombic interaction energy for each pair of side chains. The structures generated using ORBIT were used for calculations on the wild type and designed proteins. Additional calculations were performed using the crystallographic structure for wild type homeodomain.

The solvation energy of an individual side chain was calculated in a manner similar to that used by Hendsch and Tidor[42]. The folded state energy was calculated using all of the

low dielectric protein. Charges were included for only the side chain of interest. The unfolded state energy was calculated using only the atoms and charges of the side chain of interest. The side chain atoms were mapped onto the grid exactly as in the folded state calculation. The desolvation energy of each side chain was obtained by subtracting its unfolded state electrostatic solvation energy (previously referred to as the reaction field energy) from its folded state electrostatic solvation energy.

Side chain - backbone screened Coulombic interaction energies were obtained using three calculations. The internal Coulombic energy and folded state electrostatic solvation energy of the side chain only were obtained from the folded state calculation described above. The internal Coulombic and electrostatic solvation energies of the backbone only were obtained using all of the protein atoms to define the dielectric boundary and including charges for the backbone atoms only. The total Coulombic and electrostatic solvation energies of the side chain and backbone were obtained using all of the protein atoms to define the dielectric boundary and including charges for the backbone and side chain of interest. The internal side chain and backbone Coulombic and electrostatic solvation energies were subtracted from the sum of the total side chain - backbone Coulombic and electrostatic solvation energies to obtain the side chain - backbone screened Coulombic interaction energy.

Side chain - side chain screened Coulombic energies were also obtained using three calculations. The internal Coulombic and electrostatic solvation energies of each side chain were obtained using all of the protein atoms to define the dielectric boundary and including charges for the side chain of interest only. The total Coulombic and electrostatic solvation energies for each pair of side chains were obtained using all of the protein atoms to define the dielectric boundary and including charges only for the two side chains of interest. The internal Coulombic and electrostatic solvation energies of each side chain were subtracted

from the sum of the total Coulombic and electrostatic solvation energies to obtain the screened Coulombic interaction energy for the pair of side chains.

Protein Expression and Purification. Synthetic genes encoding wild type, NC0, and NC3-Ncap were prepared by recursive PCR[43] and cloned into a pET-11a (Novagen) variant. Synthetic genes encoding H1, H2, H3, and CAP were obtained by site-directed mutagenesis of NC0 using inverse PCR. Sequences for all constructs were confirmed by DNA sequencing. Recombinant proteins were expressed in BL21(DE3) Escherichia coli cells (Stratagene) at room temperature (wild-type) or 37 °C (all designed variants). The proteins were isolated using the freeze-thaw method[44]. All proteins were purified by HPLC using a reverse-phase C8 prep column (Zorbax) and linear acetonitrile-water gradients containing 0.1 % TFA. Protein masses were determined by MALDI-TOF or electrospray mass spectrometry and were found to be within one mass unit of the expected values.

Circular Dichroism Studies. Circular dichroism (CD) data were obtained on an Aviv 62A DS spectropolarimeter equipped with a thermoelectric cell holder and an autotitrator. pH 5.5 was used for all experiments to maximize the likelihood that all amino acids were in their modeled charge state. Thermal denaturation data were obtained from samples containing 50 mM protein and 50 mM sodium phosphate at pH 5.5. Data were collected every 1 °C from 1 °C to 99 °C using an equilibration time of 90 s and an averaging time of 30 s. Melting temperatures were determined by fitting to a two state transition as previously described45. Urea denaturation data were obtained from samples containing 5 mM protein and 50 mM sodium phosphate, pH 5.5, at 20 °C. To maintain constant pH, the urea stock solution also contained 50 mM sodium phosphate at pH 5.5. Data were collected every 0.2 M from 0.0 M to 9.0 M urea. Initial and final denaturant concentrations were determined

by refractometry[46]. ΔG_u was calculated from the chemical denaturation data assuming a two-state transition and using the linear extrapolation method[47]; nonlinear regression calculations were performed using KaleidaGraph (Synergy Software). Both chemical and thermal denaturation were followed by monitoring CD ellipticity at 222 nm.

Acknowledgments

We would like to thank Barry Honig for helpful conversations. This work was supported by the Howard Hughes Medical Institute, the Ralph M. Parsons Foundation, an IBM Shared University Research Grant (S. L. M.), the James Irvine Foundation Fellowship (C. S. M.), a National Institutes of Health training grant, and the Caltech Initiative in Computational Molecular Biology program, awarded by the Burroughs Wellcome Fund (S. A. M.).

References

1. Dahiyat, B. I., Gordon, D. B. and Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci.*, **6**, 1333-1337.
2. 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.
3. Koehl, P. and Levitt, M. (1999). *De novo* protein design. I. In search of stability and specificity. *J. Mol. Biol.*, **293**, 1161-1181.
4. Dahiyat, B. I. and Mayo, S. L. (1997). *De novo* protein design: fully automated sequence selection. *Science*, **278**, 82-87.
5. Street, A. G. and Mayo, S. L. (1999). Computational protein design. *Structure*, **7**, R105-R109.
6. Honig, B. and Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, **268**, 1144-1149.
7. Gordon, D. B., Marshall, S. A. and Mayo, S. L. (1999). Energy functions for protein design. *Curr. Opin. Struct. Biol.*, **9**, 509-513.
8. Walter, S., Hubner, B., Hahn, U. and Schmid, F. X. (1995). Destabilization of a protein helix by electrostatic interactions. *J. Mol. Biol.*, **252**, 133-143.
9. Hol, W. G. J., van Duijnen, P. T. and Berendsen, H. J. C. (1978). The α-helix dipole and the properties of proteins. *Nature*, **273**, 443-446.
10. Lockhart, D. J. and Kim, P. S. (1992). Internal stark effect measurement of the electric field at the amino terminus of an α-helix. *Science*, **257**, 947-951.
11. Lockhart, D. J. and Kim, P. S. (1993). Electrostatic screening of charge and dipole interactions with the helix backbone. *Science*, **260**, 198-202.

12. Huyghues-Despointes, B. M. P., Scholtz, J. M. and Baldwin, R. L. (1993). Effect of a single aspartate on helix stability at different positions in a neutral alanine based peptide. *Protein Sci.*, **2**, 1604-1611.

13. Nicholson, H., Becktel, W. J. and Matthews, B. W. (1988). Enhanced protein thermostability from designed mutations that interact with alpha-helix dipoles. *Nature*, **336**, 651-656.

14. Doig, A. J. and Baldwin, R. L. (1995). N- and C- capping preferences for all 20 amino acids in alpha-helical peptides. *Protein Sci.*, **4**, 1325-1336.

15. Bell, J. A., Becktel, W. J., Sauer, J., Baase, W. A. and Matthews, B. W. (1992). Dissection of helix capping in T4 lysozyme by structural and thermodynamic analysis of six amino acid substitutions at Thr 59. *Biochemistry*, **31**, 3590-3596.

16. Clarke, N. D., Kissinger, C. R., Desjarlais, J., Gilliland, G. L. and Pabo, C. O. (1994). Structural studies of the engrailed homeodomain. *Protein Sci.*, **3**, 1779-1787.

17. Aurora, R. and Rose, G. D. (1998). Helix capping. *Protein Sci.*, **7**, 21-38.

18. Strop, P., Marinescu, A. and Mayo, S. L. (2000). Structure of a protein G helix variant suggests the importance of helix propensity and helix dipole interactions in protein design. *Protein Sci.*, **9**, 1391-1394.

19. Gilson, M. K., Sharp, K. A. and Honig, B. H. (1987). Calculating the electrostatic potential of molecules in solution: method and error assessment. *J. Comput. Chem.*, **9**, 327-335.

20. Gilson, M. and Honig, B. (1988). Calculation of the total electrostatic energy of a macromolecular system: Solvation energies, binding energies, and conformational analysis. *Proteins*, **4**, 7-18.

21. Sharp, K., Gilson, M., Fine, R. and Honig, B. (1987). Electrostatic interactions in proteins. *Proteins*, **2**, 235-244.

22. Sharp, K. and Honig, B. (1990). Electrostatic interactions in macromolecules: Theory and applications. *Annu. Rev. Biophys. Biophys. Chem.*, **19**, 301-332.

23. Chakrabartty, A., Kortemme, T. and Baldwin, R. L. (1994). Helix propensities of the amino-acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Sci.*, **3**, 843-852.

24. Cochran, D. A. E., Penel, S. and Doig, A. J. (2001). Effect of the N1 residue on the stability of the α-helix for all 20 amino acids. *Protein Sci.*, **10**, 463-470.

25. Kumar, S. and Nussinov, R. (2000). Fluctuations between stabilizing and destabilizing electrostatic contributions of ion pairs in conformers of the c-Myc-Max leucine zipper. *Proteins*, **41**, 485-497.

26. Strop, P. and Mayo, S. L. (2000). Contribution of surface salt bridges to protein stability. *Biochemistry*, **39**, 1251-1255.

27. Horovitz, A., Serrano, L., Avron, B., Bycroft, M. and Fersht, A. R. (1990). Strength an co-operativity of contributions of surface salt bridges to protein stability. *J. Mol. Biol.*, **216**, 1031-1044.

28. Olson, C. A., Spek, E. J., Shi, Z., Vologodskii, A. and Kallenbach, N. R. (2001). Cooperative helix stabilization by complex Arg-Glu salt bridges. *Proteins*, **44**, 123-132.

29. Spector, S., Wang, M., Carp, S. A., Robblee, J. and Hendsch, Z. S. (2000). Rational modification of protein stability by the mutation of charged surface residues. *Biochemistry*, **39**, 872-879.

30. Grimsley, G. R., Shaw, K. L., Fee, L. R., Alston, R. W., Huyghues-Despointes, B. M. P., Thurlkill, R. L., Scholtz, J. M. and Pace, C. N. (1999). Increasing protein stability by altering long-range Coulombic interactions. *Protein Sci.*, **8**, 1843-1849.

31. Loladze, V. V., Ibarra-Molero, B., Sanchez-Ruiz, J. M. and Makhatadze, G. I. (1999). Engineering a thermostable protein via optimization of charge-charge interactions on the protein surface. *Biochemistry*, **38**, 16419-16423.

32. Perl, D., Mueller, U., Heinemann, U. and Schmid, F. X. (2000). Two exposed amino acid residues confer thermostability to a cold shock protein. *Nat. Struct. Biol.*, **7**, 380-383.

33. Havranek, J. J. and Harbury, P. B. (1999). Tanford-Kirkwood electrostatics for protein modeling. *Proc. Natl. Acad. Sci., USA*, **96**, 11145-11150.

34. Mayo, S. L., Olafson, B. D. and Goddard, W. A., III. (1990). Dreiding - a generic force-field for molecular simulations. *J. Phys. Chem.*, **94**, 8897-8909.

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

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

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

38. Gordon, D. B. and Mayo, S. L. (1998). Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J. Comput. Chem.*, **19**, 1505-1514.

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

40. Vijayakumar, M. and Zhou, H.-X. (2001). Salt bridges stabilize the folded structure of barnase. *J. Phys. Chem.*, **105**, 7334-7340.

41. Sitkoff, D., Sharp, K. and Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.*, **98**, 1978-1988.

42. Hendsch, Z. S. and Tidor, B. (1994). Do salt bridges stabilize proteins- a continuum electrostatic analysis. *Protein Sci.*, **3**, 211-226.

43. Prodromou, C. and Pearl, L. H. (1992). Recursive PCR: a novel technique for total gene synthesis. *Protein Eng.*, **5**, 827-829.

44. Johnson, B. H. and Hecht, M. H. (1994). Recombinant proteins can be isolated from *E. coli* cells by repeated cycles of freezing and thawing. *Biotechnology*, **12**, 1357-1360.

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

46. Pace, N. C. (1986). Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol.*, **131**, 266-280.

47. Santoro, M. M. and Bolen, D. W. (1988). Unfolding free-energy changes determined by the linear extrapolation method .1. unfolding of phenylmethanesulfonyl α-chymotrypsin using different denaturants. *Biochemistry*, **27**, 8063-8068.

48. Koradi, R., Billeter, M. and Wüthrich, K. (1996). MOLMOL: a program for display and analysis of macromolecular structures. *J. Mol. Graphics*, **14**, 51.

Table V-1: Thermal and urea denaturation data

	T_m (°C)[1]	ΔG_u (kcal mol^{-1})[2]	C_m (M)[3]	*m* (kcal $mol^{-1}M^{-1}$)[4]
wild type	49	-	-	-
NC0	53	2.3	3.4	0.7
H1	72	4.6	5.0	0.9
H2	50	2.0	3.7	0.7
H3	53	1.4	2.1	0.7
CAP	68	3.5	4.7	0.7
NC3-Ncap	88	5.9	6.2	1.0

[1] Midpoint of the thermal denaturation transition ΔG_u

[2] Free energy of unfolding at 20 °C determined by urea denaturation

[3] Midpoint of the unfolding transition determined by urea denaturation

[4] Slope of ΔG_u versus denaturant concentration

Table V-2: Electrostatic energies calculated using ORBIT[1]

	vdW[2]	pHb[3]	sc-bb Coul[4]	sc-bb Hbond[5]	sc-sc Coul[6]	sc-sc Hbond[7]	Total
wt crystal[8]	-181.9	80	-2.7	-53.0	-2.9	-33.2	-193.7
wt ORBIT[9]	-178.9	54	-2.3	-44.9	-4.2	-31.7	-208.0
NC0	-181.0	88	-1.3	-22.6	-15.4	-89.2	-221.5
H1	-181.1	90	-1.7	-22.6	-13.6	-78.8	-207.8
H2	-182.4	86	-1.6	-22.6	-13.4	-78.1	-212.1
H3	-180.1	90	-1.4	-22.6	-14.8	-89.9	-218.8
CAP	-181.4	90	-1.2	-27.4	-13.8	-74.2	-208.0
NC3-Ncap	-183.7	76	-2.6	-31.4	-11.4	-67.4	-220.5

[1] All energies reported in kcal mol^{-1}

[2] van der Waals energy

[3] 2.0 kcal mol^{-1} penalty for each buried polar hydrogen not participating in a hydrogen bond

[4] Side chain - backbone Coulombic energy calculated using a dielectric of 40*r*

[5] Side chain - backbone hydrogen bond energy

[6] Side chain - side chain Coulombic energy calculated using a dielectric of 40*r*

[7] Side chain - side chain hydrogen bond energy

[8] Energies calculated using the minimized crystallographic coordinates for wild type

[9] Energies calculated using the wild type side chain coordinates selected using ORBIT

Table V-3: Electrostatic energies calculated using DelPhi

	desolvation[2]	side chain - backbone[3]	side chain - side chain[4]
wt crystal[5]	10.2	-12.4	-5.2
wt ORBIT[6]	9.8	-12.8	-6.7
NC0	18.0	-8.0	-36.8
H1	17.0	-10.0	-32.7
H2	17.1	-9.3	-33.5
H3	18.1	-8.6	-35.5
CAP	16.4	-8.4	-33.0
NC3-Ncap	15.1	-13.9	-29.2

[1] All energies are in kcal mol^{-1}

[2] Sum of the side chain desolvation energies of the designed surface residues

[3] Side chain - backbone screened Coulombic energy

[4] Side chain - side chain screened Coulombic energy

[5] Energies calculated using the minimized crystallographic coordinates for wild type

[6] Energies calcualted using the wild type side chain coordinates selected using ORBIT

Figure V-1. Structure of the 51-residue engrailed homeodomain fragment[16]. N-capping positions are highlighted in blue, and the three most C- and N-terminal positions of each helix are highlighted in yellow. The ribbon diagram was generated using MOLMOL[48].

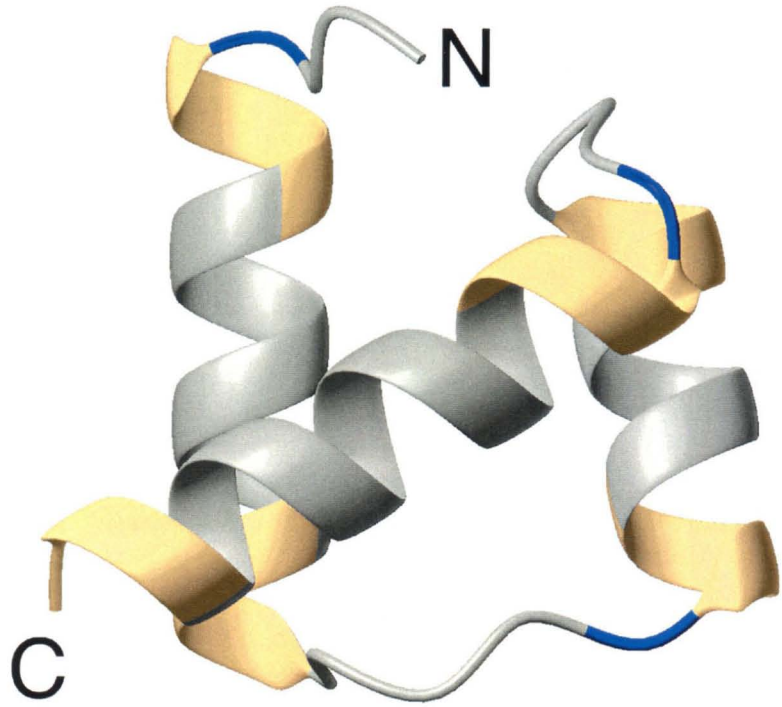

N
C

Figure V-2. Sequences of the wild type and designed homeodomain variants. N-capping positions are highlighted in blue, and the three most C- and N-terminal positions of each helix are highlighted in yellow. Classification of residues as core (c), boundary (b), or surface (s) is denoted below the NC3-Ncap sequence, and the location of the helices is indicated at the bottom of the figure. Core and boundary residues, marked " | " in the sequence alignment, were held fixed in the design calculations. The number of mutations relative to the wild type sequence and the number of violations of the helix dipole and capping "rules" in the surface residues is indicated at the right of each sequence.

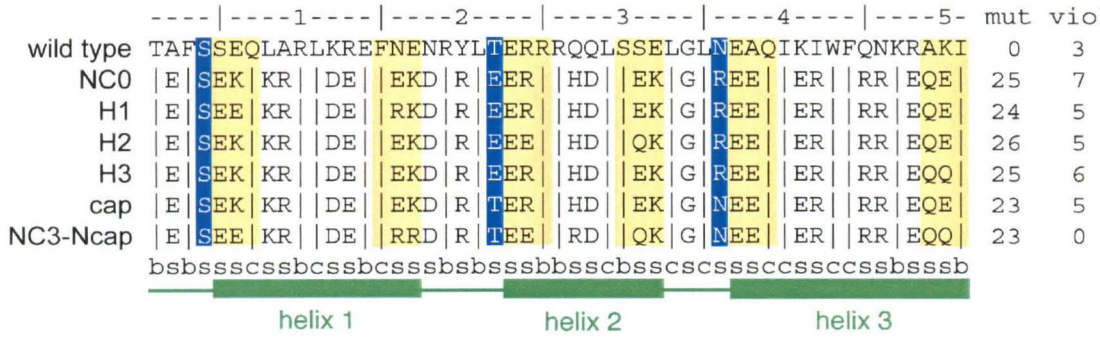

----|----1----|----2----|----3----|----4----|----5- mut viol
wild type TAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKI 0 3
NC0 |E|SEK|KR||DE||EKD|R|EER||HD||EK|G|REE||ER||RR|EQE| 25 7
H1 |E|SEE|KR||DE||RKD|R|EER||HD||EK|G|REE||ER||RR|EQE| 24 5
H2 |E|SEK|KR||DE||EKD|R|EEE||HD||QK|G|REE||ER||RR|EQE| 26 5
H3 |E|SEK|KR||DE||EKD|R|EER||HD||EK|G|REE||ER||RR|EQQ| 25 6
cap |E|SEK|KR||DE||EKD|R|TER||HD||EK|G|NEE||ER||RR|EQE| 23 5
NC3-Ncap |E|SEE|KR||DE||RRD|R|TEE||RD||QK|G|NEE||ER||RR|EQQ| 23 0
bsbsssscssbcssbcsssbsbsssbbsscbsscscsssccssccssbsssb
helix 1
helix 2
helix 3

Figure V-3. Side chain identities and conformations in helix one. Wild type homeodomain is at the left, NC0 is in the middle, and NC3-Ncap is on the right. Red lines indicate hydrogen bonds and salt bridges whose interaction energy is predicted to be at least 1 kcal mol^{-1} by the ORBIT force field. Wild type side chain conformations were obtained from the minimized crystal structure[16] and the side chain conformations shown for the NC0 and NC3-Ncap were predicted by ORBIT. Note the abundance of charged residues and putative hydrogen bonds and salt bridges in the designed variants.

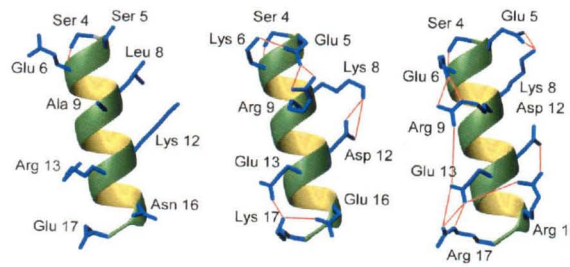

Ser 4
Ser 5
Glu 6
Leu 8
Ala 9
Lys 12
Arg 13
Asn 16
Glu 17
Ser 4
Lys 6
Glu 5
Lys 8
Arg 9
Asp 12
Glu 13
Glu 16
Lys 17
Ser 4
Glu 5
Glu 6
Lys 8
Asp 12
Arg 9
Glu 13
Arg 16
Arg 17

Figure V-4. Side chain identities and conformations in helix two. Wild type homeodomain is at the left, NC0 is in the middle, and NC3-Ncap is on the right. Red lines indicate hydrogen bonds and salt bridges whose interaction energy is predicted to be at least 1 kcal mol^{-1} by the ORBIT force field. Wild type side chain conformations were obtained from the minimized crystal structure[16] and the side chain conformations shown for the NC0 and NC3-Ncap were predicted by ORBIT. Note the abundance of charged residues and putative hydrogen bonds and salt bridges in the designed variants.

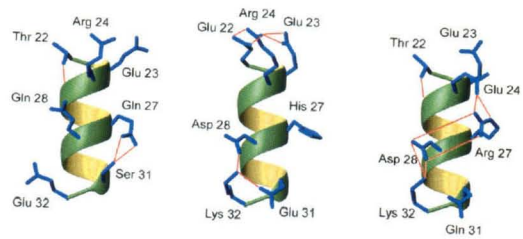

Arg 24
Thr 22
Glu 23
Gln 28
Gln 27
Ser 31
Glu 32
Arg 24
Glu 22
Glu 23
His 27
Asp 28
Lys 32
Glu 31
Glu 23
Thr 22
Glu 24
Arg 27
Asp 28
Lys 32
Gln 31

Figure V-5. Side chain identities and conformations in helix three. Wild type homeodomain is at the left, NC0 is in the middle, and NC3-Ncap is on the right. Red lines indicate hydrogen bonds and salt bridges whose interaction energy is predicted to be at least 1 kcal mol^{-1} by the ORBIT force field. Wild type side chain conformations were obtained from the minimized crystal structure[16] and the side chain conformations shown for the NC0 and NC3-Ncap were predicted by ORBIT. Note the abundance of charged residues and putative hydrogen bonds and salt bridges in the designed variants.

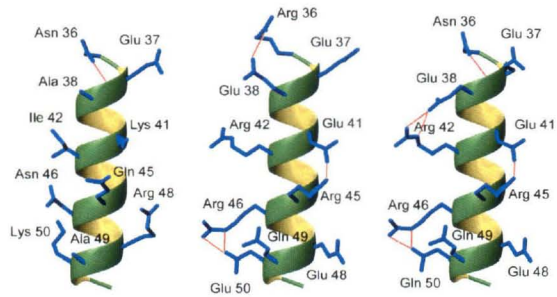

Asn 36
Glu 37
Ala 38
Ile 42
Lys 41
Asn 46
Gln 45
Arg 48
Lys 50
Ala 49
Arg 36
Glu 37
Glu 38
Arg 42
Glu 41
Arg 46
Arg 45
Gln 49
Glu 48
Glu 50
Asn 36
Glu 37
Glu 38
Arg 42
Glu 41
Arg 46
Arg 45
Gln 49
Glu 48
Gln 50

Figure V-6. Thermal denaturation data monitored by CD of (from left to right) wild type (black), H2 (purple), H3 (gray), NCO (red), CAP (orange), H1 (blue), and NC3-Ncap (green).

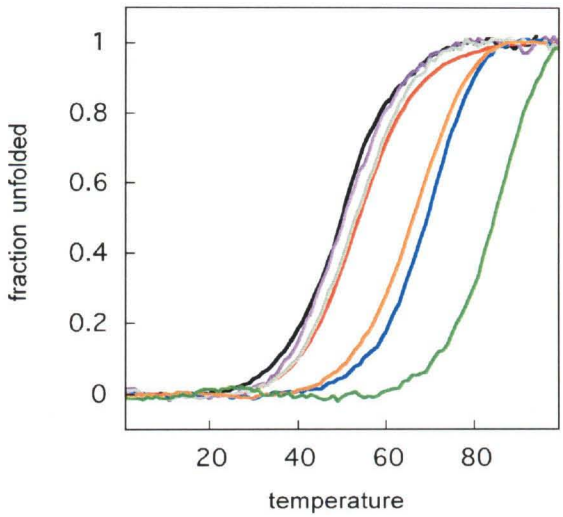

fraction unfolded
1
0.8
0.6
0.4
0.2
0
20
40
60
80
temperature

Figure V-7. Urea denaturation at 20 °C monitored by CD of (from left to right) H3 (gray), H2 (purple), NC0 (red), CAP (orange), H1 (blue), and NC3-Ncap. (green)

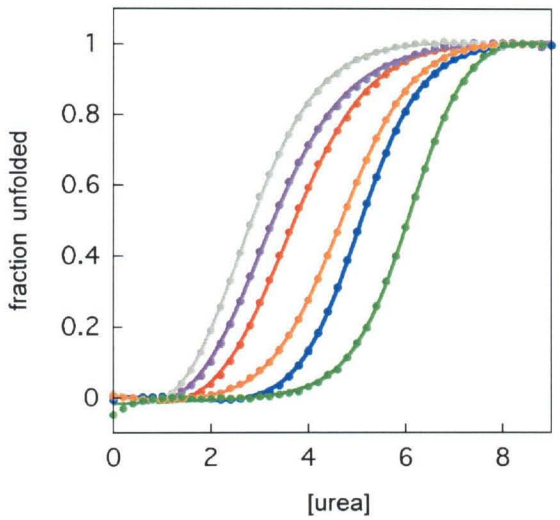

fraction unfolded
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
[urea]

Figure V-8. Energy predicted using the ORBIT force field versus the experimentally determined stability of each designed variant.

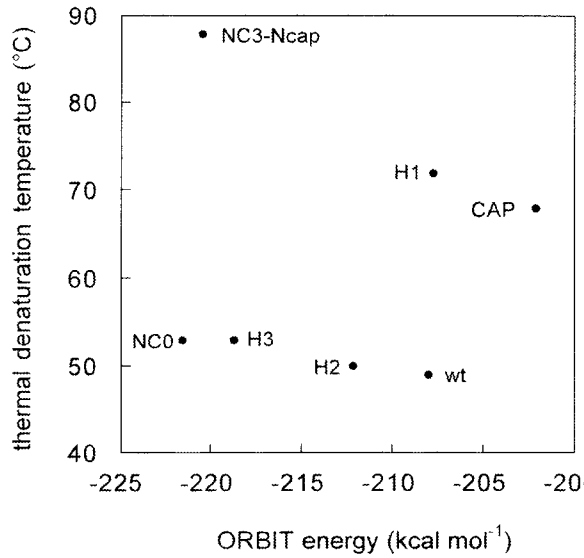

90
80
70
60
50
40
thermal denaturation temperature (°C)
NC3-Ncap
H1
CAP
NC0
H3
H2
wt
-225
-220
-215
-210
-205
-200
ORBIT energy (kcal mol-1)

Figure V-9. Sum of the ORBIT van der Waals energy and the DelPhi desolvation energy, side chain - backbone screened Coulombic energy, and side chain - side chain screened Coulombic energy versus the experimentally determined stability of each homeodomain variant.

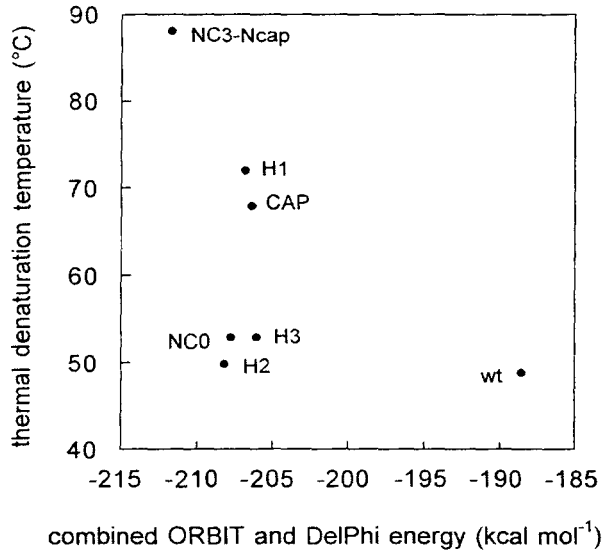

90
80
70
60
50
40
thermal denaturation temperature (°C)
NC3-Ncap
H1
CAP
NC0
H3
H2
wt
-215
-210
-205
-200
-195
-190
-185
combined ORBIT and DelPhi energy (kcal mol-1)

Figure V-10. Histogram of the side chain - side chain screened Coulombic energies calculated using DelPhi for the wild type homeodomain. The inset highlights interactions with unreasonably large predicted energies.

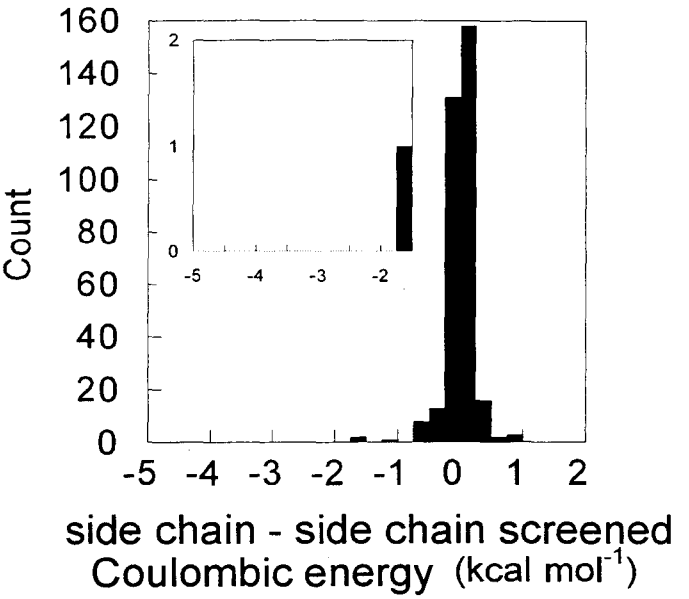

160
140
120
100
80
60
40
20
0
Count
2
1
0
-5
-4
-3
-2
-5 -4 -3 -2 -1 0 1 2
side chain - side chain screened
Coulombic energy (kcal mol-1)

Figure V-11. Histogram of the side chain - side chain screened Coulombic energies calculated using DelPhi for the designed homeodomain variants. The inset highlights interactions with unreasonably large predicted energies; note that the designed variants are predicted to have a larger number of extremely favorable side chain - side chain electrostatic

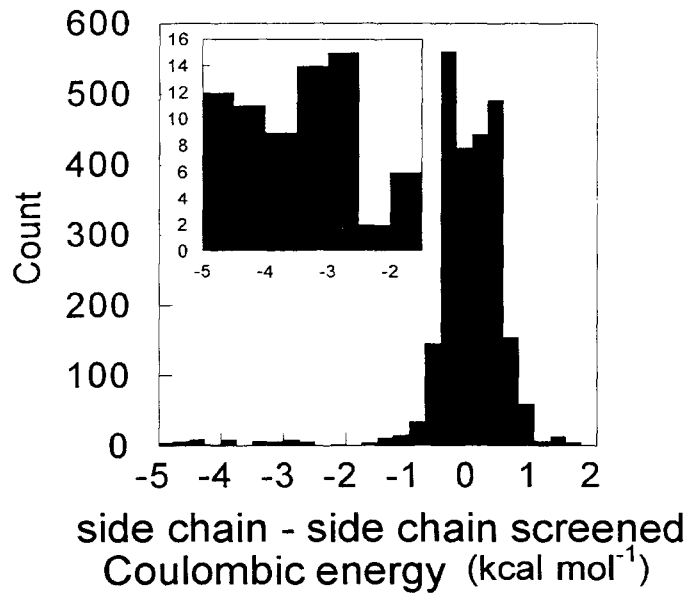
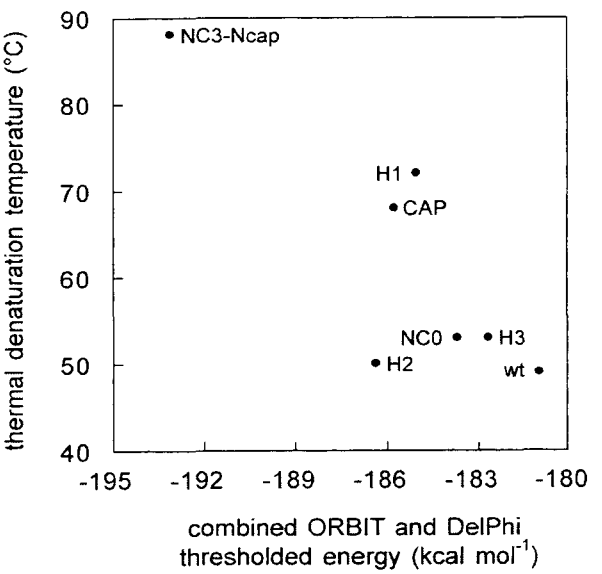

600
500
400
300
200
100
0
Count
-5 -4 -3 -2 -1 0 1 2
16
14
12
10
8
6
4
2
0
-5 -4 -3 -2
side chain - side chain screened
Coulombic energy (kcal mol-1)

Figure V-12. Sum of the ORBIT van der Waals energy and the DelPhi desolvation energy, side chain - backbone screened Coulombic energy, and side chain - side chain screened Coulombic energy, incorporating a $\pm$ 1.5 kcal mol^{-1} cutoff on the side chain - backbone and side chain - side chain screened Coulombic energies, versus the experimentally determined stability of each homeodomain variant.


90
80
70
60
50
40
thermal denaturation temperature (°C)
NC3-Ncap
H1
CAP
NC0
H3
H2
wt
-195
-192
-189
-186
-183
-180
combined ORBIT and DelPhi
thresholded energy (kcal mol-1)

Chapter VI:

Electrostatic Models for Protein Design Calculations. I. Optimized Dielectrics and Polar Solvation Parameters

The text of this chapter is adopted from an unpublished manuscript that was coauthored with Professor Stephen L. Mayo.

Abstract

Computational protein design algorithms have typically used fast methods based on Coulomb's law and/or geometry dependent hydrogen bond terms to model electrostatic interactions. Desolvation effects have either been neglected in design calculations or approximated using surface area based or per-atom desolvation penalties. The results of a previous protein design study indicate that the balance among the electrostatic components of the current ORBIT protein design force field requires optimization. In this chapter, we assess the accuracy of the electrostatic terms in the ORBIT force field by comparing ORBIT energies to energies calculated using the finite difference Poisson-Boltzmann (FDPB) method. Next, we identify optimal electrostatic parameters for the ORBIT force field by maximizing agreement between ORBIT and FDPB energies. The new parameters include a set of solvation parameters for polar functional groups and dielectric values for different classes of electrostatic interactions. For problems with extremely high combinatorial complexity such as protein design, fast, approximate methods can prove valuable so long as the errors are tolerably small. The optimized parameters dramatically increase the accuracy of the electrostatic energies calculated using a design force field while maintaining excellent computational efficiency.

Introduction

One of the challenges in developing force fields for protein design calculations is identifying the optimal balance between terms describing various covalent and noncovalent interactions. The ORBIT force field currently contains five terms that are electrostatic in nature: the polar hydrogen burial penalty (typically applied only to side chains), side chain - backbone and side chain - side chain hydrogen bond terms, and side chain - backbone and side chain - side chain Coulombic terms. Results of an earlier protein design study indicate that the current electrostatic parameters systematically overestimate the importance of hydrogen bonds relative to longer range interactions and underestimate the importance of side chain - backbone interactions relative to side chain - side chain interactions. Consequently, the stability of proteins designed using these parameters is far from optimal[1]. Simple electrostatic models such as the one currently included in the ORBIT force field are unlikely to capture all of the subtlties of protein electrostatics. However, simple models have excellent computational efficiency and may be sufficiently accurate to select stable, well-folded proteins.

Many of the terms in the ORBIT force field have been successfully parametrized using the results of protein design experiments. This approach works quite well for optimizing a single parameter, but becomes increasingly challenging as the number of variables increases. An alternative approach, taken here, is to first parametrize the electrostatic components of the ORBIT force field by maximizing the agreement to the results of a more sophisticated electrostatic model. Results of an earlier design study indicate that FDPB electrostatic energies are a significantly better predictor of the stability of designed protein variants than the energies predicted using the current ORBIT force field[1]. Experimental testing can then be conducted to fine tune the electrostatic terms and to determine the proper balance between electrostatic and nonelectrostatic force field terms.

In this study, we have calculated electrostatic energies using ORBIT and DelPhi[3] for eight proteins that have been targets for design and other biophysical studies: fragment B of Streptococcal protein A, the c-Crk SH3 domain, hen egg white lysozyme, the β1 domain of Streptococcal protein G, ubiquitin, plastocyanin, bovine pancreatic trypsin inhibitor, and rubredoxin. For compatibility with the current design procedure, we have calculated side chain internal energies, side chain - backbone energies, and side chain - side chain energies separately. Both desolvation energies and screened Coulombic energies, described below, were considered.

Polar protein groups can form favorable electrostatic interactions with the solvent; we refer to the resulting energies as electrostatic solvation energies. The desolvation energy of a side chain is defined as the difference in the electrostatic solvation energy of the side chain alone (a simple model of the unfolded state) versus the electrostatic solvation energy of the side chain in the context of the folded protein, as shown in Figure VI-1. Electrostatic interactions between polar protein groups and the solvent also act to screen Coulombic interactions within a protein. The screening energy is always opposite in sign and weaker in magnitude than the Coulombic energy for a given interaction. The procedures used to calculate side chain - backbone and side chain - side chain screening energies are shown in Figures VI-2 and VI-3, respectively. In all cases, the screening energies and Coulombic energies are added to yield screened Coulombic energies and the screened Coulombic energies predicted by the different electrostatic models are then compared.

Assessment of the current ORBIT electrostatic model

The ORBIT (Optimization of Rotamers by Iterative Techniques) protein design force field[3, 4] currently uses three terms to describe electrostatic interactions: a distance, angle, and hybridization-dependent hydrogen bond term, a Coulombic term calculated using either

a distance-dependent dielectric or a constant dielectric, and a penalty for burying polar hydrogens that are not participating in a hydrogen bond. Alternatively, desolvation effects can be modeled by penalizing buried polar surface area[5]. A polar hydrogen burial penalty can also be used to describe backbone desolvation, although in practice backbone desolvation terms have not been used. Hydrogen bond and Coulombic interactions are calculated for both side chain - side chain and side chain - backbone interactions.

Here, we test each component of the ORBIT electrostatic model by comparing the ORBIT side chain - backbone hydrogen bond and Coulombic energies to the DelPhi screened Coulombic energies, the ORBIT side chain - side chain hydrogen bond and Coulombic energies to the DelPhi screened Coulombic energies, and the ORBIT polar hydrogen burial penalty and the polar surface area burial penalty to the DelPhi side chain desolvation energies. The results of the DelPhi versus ORBIT comparisons are shown in Figures VI-4 through VI-7 and in Table VI-1.

The current ORBIT force field does not model side chain desolvation effects well. The polar hydrogen burial penalty energies are not well correlated with the DelPhi desolvation energies, as shown in Figure VI-4 and Table VI-1. Buried polar surface area is a better predictor of the DelPhi desolvation energies, as shown in Figure VI-5 and Table VI-1, but the correlation is still weak. Furthermore, the traditional value for both the polar hydrogen burial penalty (2.0 kcal mol^{-1} per hydrogen buried) yields energies that are far larger than the DelPhi desolvation energies.

The sum of the side chain - backbone or side chain - side chain hydrogen bond and Coulombic energies is a poor predictor of the screened Coulombic energies calculated using DelPhi, as shown in Figures VI-6 and VI-7 and Table VI-1. The ORBIT side chain - backbone and side chain - side chain interaction energies are dominated by the hydrogen bond term, as a hydrogen bond with optimal geometry receives an 8 kcal mol^{-1} benefit. Examining

these graphs, many of the data points that are small in magnitude lie along a line. These correspond to interactions where the hydrogen bond energy is zero and the Coulombic component determines the interaction energy. The reasonably linear correlation of these points suggested that using only a Coulombic potential could yield better correlation than using both hydrogen bond and Coulombic terms.

Optimized dielectrics

Since the Coulombic side chain - backbone and side chain - side chain energies provide a better approximation to the DelPhi screened Coulombic energies than the hydrogen bond terms do, it may be sensible to use only the Coulombic terms to calculate side chain - backbone and side chain - side chain interaction energies. By adjusting the dielectric constant or the value of the distance dependent dielectric, it is possible to scale the Coulombic energies so that they are similar in magnitude to the DelPhi screened Coulombic energies. The optimal dielectric constants are found to be 38.1 and 73.5 for side chain - backbone and side chain - side chain interactions, respectively. Alternatively, a distance dependent dielectric of $13.1r$ or $12.8r$ can be used side chain - backbone or side chain - side chain interactions, respectivelty. The agreement between the DelPhi side chain - backbone or side chain - side chain energies and the Coulombic energies is shown in Figures VI-8 through VI-11 and in Table VI-1.

The relationship between side chain - side chain Coulombic energies calculated using a constant dielectric and the screened Coulombic energies is complex, as shown in Figure VI-10. Further analysis reveals that two additional parameters affect the relationship between the screened Coulombic energy calculated using DelPhi and the Coulombic energy: the distance between the side chains and the net charge of the side chains. Interactions between two charged side chains lie along the sigmoidal portion of the curve, shown in

colors, while charge - neutral and neutral - neutral side chain interactions lie along the intersecting straight line shown in gray. Among the charge - charge interactions, the magnitude of the interaction depends on the distance between polar atoms in the side chains. When each region of the curve is fit separately, the agreement between the DelPhi screened Coulombic energies and the Coulombic energies increases significantly.

Optimized atomic solvation parameters

The ORBIT force field contains two terms that can be used to describe the desolvation of polar functional groups; the more accurate term is based on the change in solvent accessible surface area of polar groups upon protein folding. Most of the published sets of atomic solvation parameters use six or seven atom types (typically carbon, uncharged oxygen, charged oxygen, uncharged nitrogen, charged nitrogen, sulfur; the carbon group may be divided into two subgroups)[6] while the current ORBIT polar burial penalty only considers two atom types (hydrophobic and polar). It is therefore likely that the poor correlation between desolvation energies and buried polar surface area is partially due to using only two atom types.

To obtain a set of atomic solvation parameters that would best approximate the desolvation energies predicted by DelPhi, we calculated the change in polar surface area and the desolvation energy of each polar functional group in the set of eight proteins and determined the relationship between these two parameters. We found that significanly more accurate desolvation energies can be obtained using different parameters for each type of polar functional group, given in Table VI-3. The agreement between the DelPhi desolvation energies and the desolvation energies calculated using the optimized atomic solvation parameters is shown in Figure VI-12 and Table VI-1.

The magnitude of the desolvation energies predicted using several sets of published atomic solvation parameters are far larger than the desolvation energies calculated using DelPhi with a probe radius of 0.0[6]. This is not surprising, as these atomic solvation parameters were calculated from octanol-water or vacuum-water transfer studies. In transfer experiments, all water - solute interactions are broken. In contrast, completely buried polar groups in a folded protein often maintain significant interactions with the solvent. By fitting to desolvation energies that result from protein folding rather than transfer, we recover the average effect of interactions with solvent molecules that are beyond the first solvation shell. However, like any surface area based desolvation model, the new parametrization ignores variations in electrostatic environment that can occur at a given degree of solvent accessibility.

Optimized solvent-exclusion model solvation parameters

The accuracy of surface area based models of polar solvation is limited because the effects of solvent beyond the first shell are neglected. An alternative approach, the solvent-exclusion model developed by Lazaridis and Karplus[7], considers the extent to which the solvation energy of an atom is decreased by the proximity of other solute atoms. This model has been used previously in protein design studies conducted by David Baker and coworkers[8]. Briefly, the solvation energy density of an atom is modeled using a Gaussian function. Desolvation of one atom by each other atom can be found by approximating the integral of the solvation energy density within the volume of the other atom, and the effects of additional atoms are additive. The resulting function is pairwise decomposable by atom and requires significantly less computation time than surface area based solvation models.

The solvent-exclusion model was developed to describe solvation of both hydrophobic and polar groups and was parametrized to be compatible with the CHARMM

19 polar hydrogen energy function[9]. The parameter set contains values for the volume, atomic radius, correlation length, free energy of solvation of the isolated atom (ΔG^{free}), and free energy of solvation for the atom in a reference compound (ΔG^{ref}) for each of 17 atom types. For the purposes of this study, we are interested in only the desolvation of polar protein functional groups. Accordingly, it was necessary to adjust the parametrization. We obtained values for ΔG^{free} and ΔG^{ref} for each polar functional group by maximizing the agreement to the DelPhi desolvation energies.

The optimized solvent-exclusion model parameters give slightly better agreement with the DelPhi desolvation energies than the optimized atomic solvation parameters do, as indicated in Table VI-1 and by comparing Figure VI-13 with Figure VI-12. Furthermore, the solvent-exclusion model requires significantly less computation time, as it does not require the generation of a solvent accessible surface. Consequently, it may be advantageous to replace the polar hydrogen burial penalty and the polar area penalty with desolvation energies calculated using the solvent-exclusion model.

Additional Considerations

The optimized dielectric and solvation parameters presented in this paper were derived by maximizing agreement to the energies calculated using DelPhi. Consequently, the manner in which the DelPhi calculations were run will affect the values for the optimized dielectric and solvation parameters. Several of these considerations are presented in Chapter VII. An additional consideration, discussed here, is that the probe radius used in the DelPhi calculations significantly affects the magnitude of the DelPhi desolvation and screened Coulombic energies.

When the probe radius is set to 0.0, as is the case for the calculations presented previously in this chapter, the boundary between the high dielectric solvent and the low

dielectric protein is determined using a van der Waals surface, while a solvent accessible surface is used when the probe radius greater than 0.0. Traditionally, a probe radius of 1.4 has been used, so that the solvent accessible surface is generated using a probe that is the size of a water molecule. Using a probe radius of 0.0 yields smaller desolvation energies and smaller screened Coulombic energies than are obtained using a probe radius of 1.4. Recent results show that using a probe radius of 0.0 better reproduces the experimentally determined strength of salt bridges in barnase[10].

As there is some disagreement about the optimal value for the probe radius in DelPhi calculations, we have determined the optimal dielectric and solvation parameters for both probe radius 0.0, shown in Table VI-1, and probe radius 1.4, shown in Table VI-2. Future experimental work will be required to determine which parameter set is most appropriate for protein design studies.

Conclusions

We have used the FDPB model to develop and test approaches for calculating electrostatic energies in protein design problems. The current parameters used in the ORBIT force field are observed to correlate poorly with the desolvation and interaction energies predicted using DelPhi. We find that using Coulomb's law with a distance dependent dielectric of approximately $10r$ yields significantly more accurate side chain - backbone and side chain - side chain electrostatic energies than using a explicit hydrogen bond term and a Coulombic term. Modeling desolvation effects quickly and accurately has proven somewhat more challenging. A solvent-exclusion model using an optimized parameter set is found to give reasonable desolvation energies in a very short time. While the methods described in this paper do not capture all of the complexity of protein electrostatics, they are extremely fast and significantly more accurate than previously used electrostatic models.

Consequently, the optimized parameters presented here should facilitate the design of stable, well-folded proteins.

Materials and Methods

Test set of proteins. All calculations were performed on eight proteins that are popular targets for design and other biophysical studies: fragment B of protein A, the c-Crk SH3 domain, hen egg white lysozyme, the β1 domain of Streptococcal protein G, ubiquitin, plastocyanin, bovine pancreatic trypsin inhibitor, and rubredoxin. Structural coordinates were obtained from the PDB entries 1cka, 1fc2, 1hel, 1pga, 1ubi, 2pcy, 6pti, and 8rxn, respectively. Explicit hydrogens were added to each structure using BIOGRAF (Molecular Simulations, Inc. San Diego) and the termini were adjusted to carry a net charge of ±1. The resulting structures were minimized for 50 steps using the Dreiding force field; the minimized structures were used in the DelPhi and ORBIT calculations discussed in the following sections.

DelPhi calculations. Finite difference solutions to the linearized Poisson-Boltzmann equation were obtained using the FDPB solver from the computer program DelPhi[11] with a grid spacing of 2.0 grids $Å^{-1}$, an interior dielectric of 4.0, an exterior dielectric of 80.0, and 0.050 M salt. Dielectric boundaries were defined using van der Waals surfaces, which were found to give better agreement with experimental results than solvent accessible surfaces [10], unless otherwise mentioned. The grid size was selected for each protein so that its backbone atoms fill 70 % of the grid. The coordinates of each protein were mapped onto the grid in exactly the same way in each calculation to minimize errors due to differences in grid placement. The PARSE parameter set charges and atomic radii[12] were used in all FDPB calculations. Proline and disulfide bonds were considered part of the backbone in all

calculations. All His, Arg, and Lys residues were modeled with a +1 net charge, all Asp and Glu residues were modeled with a -1 charge, and all other residues were modeled with a net charge of 0. All DelPhi energies were converted to units of kcal mol^{-1} using the relation kT = 0.593 kcal mol^{-1} at 25 °C.

The desolvation energy of a side chain, *i*, is defined as the difference between the electrostatic solvation energy of the side chain in the folded state versus the unfolded state:

$$\Delta\Delta G_{\text{desolv(side chain } i)} = (1/2) \sum_u q_u (\phi^{all} - \phi^{i\ only}) \quad (1)$$

where each *u* is an atom in side chain *i*, q_u is the partial atomic charge of side chain atom *u*, ϕ^{all} is the reaction potential at *u*, generated by the set of partial atomic charge on the side chain, when all of the protein atoms are used to define the dielectric boundary, and $\phi^{i\ only}$ is the reaction potential at *u*, generated by the set of partial atomic charge on side chain *i*, when the atoms on side chain *i* and the local backbone only are used to define the dielectric boundary. The unfolded state was modeled as the side chain and local backbone, mapped to the grid exactly as in the folded state calculations. The local backbone is defined to include the following atoms: CA(*i*-1), C(*i*-1), O(*i*-1), N(*i*), HN(*i*), CA(*i*), C(*i*), O(*i*), N(*i*+1), HN(*i*+1), and CA(*i*+1).

Folded state side chain - backbone screening energies were obtained using the following equation:

$$\Delta G_{\text{screening(sc-bb)}} = \sum_t q_t \phi^{all} \quad (2)$$

where *i* is the side chains of interest, each *t* is an atom in the backbone, q_t is the partial atomic charge of atom *t*, and ϕ^{all} is the reaction potential due to the set of partial atomic

charges on side chain i at t, when all of the protein atoms are used to define the dielectric boundary. The screening energies were then added to the Coulombic energies to obtain screened Coulombic energies:

$$\Delta G_{\text{screened Coulombic(sc-bb)}} = \Delta G_{\text{screening(sc-bb)}} + \Delta G_{\text{Coulombic(sc-bb)}} \quad (3)$$

where the Coulombic energy is calculated using Coulomb's law with the dielectric equal to the dielectric of the protein interior.

Side chain - side chain interactions are obtained using a similar method:

$$\Delta G_{\text{screening(sc-sc)}} = \Sigma\, q_v\, \phi^{all} \quad (4)$$

where i and j are the side chains of interest, each v is an atom in side chain j, q_v is the partial atomic charge of atom v, and ϕ^{all} is the reaction potential due to the set of partial atomic charges on side chain i at v, when all of the protein atoms are used to define the dielectric boundary. The screening energies were then added to the Coulombic energies to obtain screened Coulombic energies:

$$\Delta G_{\text{screened Coulombic(sc-sc)}} = \Delta G_{\text{screening(sc-sc)}} + \Delta G_{\text{Coulombic(sc-sc)}} \quad (5)$$

All of the protein atoms were used to define the dielectric boundary when calculating the screening energies. Side chain - backbone and side chain -side chain interaction energies are assumed to be zero in the unfolded state.

ORBIT calculations. Distance-dependent dielectric Coulombic, hydrogen bond and polar hydrogen burial energies as well as the buried polar surface area of each of the eight proteins were calculated using the ORBIT force field as previously described[3]:

$$E_{\text{Coulombic}} = 322.0637 \left(q_i q_j / \varepsilon R \right) \tag{6}$$

where q_i and q_j are the charges on atoms i and j, respectively, ε is the dielectric constant, R is the distance between atoms i and j,

$$E_{\text{H-bond}} = D_o[\, 5(R_o/R)^{12} - 6(R_o/R)^{10} \,]F(\theta) \tag{7}$$

where R_o is the equilibrium distance, R is the distance between the donor and acceptor heavy atoms, D_o is the well depth, and $F(\theta)$ is defined as follows:

sp^3 donor - sp^3 acceptor	$F=\cos^2\theta\cos^2(\phi\text{-}109.5°)$,	$\theta>90°$, $\phi\text{-}109.5°<90°$	(8)
sp^3 donor - sp^2 acceptor	$F=\cos^2\theta\cos^2\phi$,	$\phi>90°$	(9)
sp^2 donor - sp^3 acceptor	$F=\cos^4\theta$		(10)
sp^2 donor - sp^2 acceptor	$F=\cos^2\theta\cos^2(\max[\phi,\varphi])$		(11)

where θ is the donor-hydrogen-acceptor angle, ϕ is the hydrogen-acceptor-base angle (where the base is the atom covalently attached to the acceptor), and φ is the angle between the normals of the planes defined by the six atoms covalently bonded to the two sp^2 centers.

To facilitate comparisons with the DelPhi energies, the PARSE charge set was also used for the ORBIT calculations. In the surface area calculations, all carbons, sulfurs, and hydrogens bonded to carbons were considered hydrophobic and all oxygens, nitrogens, and

hydrogens bonded to oxygens or nitrogens were considered polar. In all cases, energies and areas were calculated using the minimized crystal structures described in the first methods section.

Atomic solvation parameters. The desolvation energy of each polar functional group was calculated as in equation 1, except that partial atomic charges were considered only for the functional group of interest. The solvent accessible surface area of each polar functional group was calculated using the Lee and Richards definition[13].

Solvent-exclusion model parameters. In the solvent exclusion model, the free energy of solvation of an atom, u, is given by:

$$\Delta G_{solv,\,u} = \sum_{v \neq u} \Delta G_{ref,\,u} - \Sigma\, \Delta G_{free,\,u}\, C$$

$$C = (1/2\, \lambda_u r^2_{uv}\, \pi^{1.5}) \exp[-((r_{uv} - R_u) / \lambda_u)^2]\, V_v$$

where the sum is over all atoms $v \neq u$, $\Delta G_{ref,\,u}$ is the free energy of solvation of atom u in isolation, each v is another solute atom,$\Delta G_{free,\,u}$ is the free energy of solvation of atom u in the context of a reference molecule, λ_u is the correlation length (3.5 Å for neutral groups and 6.0 Å for charged groups), r_{uv} is the distance between atoms u and v, R_u is the van der Waals radius of atom u, and V_v is the volume of atom v. The desolvation energy is defined, as before, as the difference between the solvation energy of a polar group in the context of the folded protein versus in the context of its side chain and local backbone only. To calculate desolvation energies using the above equation, the sum is over all atoms $v \neq u$ where v is not in the same side chain or local backbone as u.

Values for $\Delta G_{ref,\,u}$ and $\Delta G_{free,\,v}$ were calculated for each of the amino acids containing polar atoms by maximizing agreement to the DelPhi desolvation energies. Residues containing the same functional groups were clustered together, so that one set of values was calculated for Asp and Glu, for Asn and Gln, and for Ser and Thr.

Acknowledgements

This work was supported by the Howard Hughes Medical Institute, the Parsons Foundation, an IBM Shared (S. L. M.), a National Institutes of Health training grant, and the Caltech Initiative in Computational Molecular Biology program, awarded by the Burroughs Wellcome Fund (S. A. M.).

References

1. Marshall, S. A., Morgan, C. S. and Mayo, S. L. (2001). Electrostatic interactions significantly affect the stability of designed proteins. accepted, *J. Mol. Biol.*
2. Gilson, M. K., Sharp, K. A. and Honig, B. H. (1987). Calculating the electrostatic potential of molecules in solution: method and error assessment. *J. Comput. Chem.*, **9**, 327-335.
3. Dahiyat, B. I., Gordon, D. B. and Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci.*, **6**, 1333-1337.
4. Gordon, D. B., Marshall, S. A. and Mayo, S. L. (1999). Energy functions for protein design. *Curr. Opin. Struct. Biol.*, **9**, 509-513.
5. Dahiyat, B. I. and Mayo, S. L. (1996). Protein design automation. *Protein Sci.*, **5**, 895-903.
6. Juffer, A. H., Eisenhaber, F., Hubbard, S. J., Walther, D. and Argos, P. (1995). Comparison of atomic solvation parameter sets: Applicability and limitations in protein folding and binding. *Protein Sci.*, **4**, 2499-2509.
7. Lazaridis, T. and Karplus, M. (1999). Effective energy functions for proteins in solution. *Proteins*, **35**, 133-152.
8. Kuhlman, B. and Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci.*, **97**, 10383-10388.
9. Neria, E., Fischer, S. and Karplus, M. (1996). Simulation of activation free energies in molecular systems. *J. Chem. Phys.*, **105**, 1902-1921.
10. Vijayakumar, M. and Zhou, H.-X. (2001). Salt bridges stabilize the folded structure of barnase. *J. Phys. Chem.*, **105**, 7334-7340.
11. Rocchia, W., Alexov, E. and Honig, B. (2001). Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. *J. Phys. Chem.*, **105**, 6507-6514.

12. Sitkoff, D., Sharp, K. and Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.*, **98**, 1978-1988.

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

Table VI-1: Mean energies and errors of electrostatic models, prbrad=0.0

	absolute mean energy (kcal mol^{-1})	RMSD (kcal mol^{-1})	R
1. Side chain desolvation energy			
exact DelPhi	0.271	-	-
polar hydrogen burial = 2.0	2.533	3.826	0.234
polar area burial, $\sigma = 0.1$	7.730	8.057	0.519
polar area burial, σ by group	0.285	0.150	0.817
solvent exclusion model	0.216	0.118	0.864
II. Side chain - backbone screened Coulombic energy			
exact DelPhi	0.413	-	-
H-bond + Coulombic, $\varepsilon = 40\ r$	0.873	1.959	0.529
Coulombic, $\varepsilon = 32.3$	0.333	0.352	0.837
Coulombic, $\varepsilon = 12.3\ r$	0.411	0.144	0.975
III. Side chain - side chain screened Coulombic energy			
exact DelPhi	0.044	-	-
H-bond + Coulombic, $\varepsilon = 40\ r$	0.023	0.230	0.458
Coulombic, $\varepsilon = 73.5$	0.047	0.055	0.850
Coulombic, $\varepsilon = 12.8\ r$	0.029	0.047	0.887

Table VI-2: Mean energies and errors of electrostatic models, prbrad=1.4

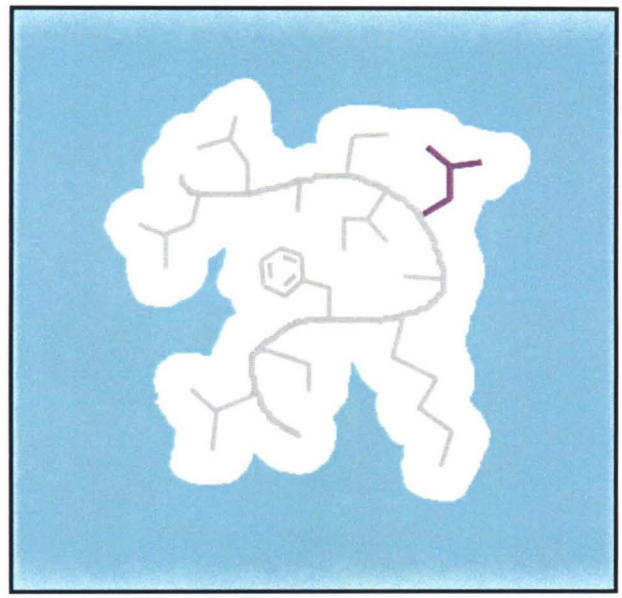
	absolute mean energy (kcal mol^{-1})	RMSD (kcal mol^{-1})	R
I. Side chain desolvation energy			
exact DelPhi	1.371	-	-
polar hydrogen burial = 2.0	2.533	3.335	0.221
polar area burial, $\sigma = 0.1$	7.730	6.942	0.492
polar area burial, σ by group	1.882	0.778	0.895
solvent exclusion model	1.437	0.508	0.931
II. Side chain - backbone screened Coulombic energy			
exact DelPhi	1.010	-	-
H-bond + Coulombic, $\varepsilon = 40\ r$	0.873	1.959	0.568
Coulombic, $\varepsilon = 16.0$	0.793	0.352	0.736
Coulombic, $\varepsilon = 5.0\ r$	1.076	0.144	0.936
III. Side chain - side chain screened Coulombic energy			
exact DelPhi	0.066	-	-
H-bond + Coulombic, $\varepsilon = 40\ r$	0.023	0.230	0.533
Coulombic, $\varepsilon = 48.5$	0.072	0.055	0.599
Coulombic, $\varepsilon = 7.0\ r$	0.054	0.047	0.905

Table VI-3: Optimized solvation parameters for polar functional groups

functional group	ASP (kcal mol^{-1} Å^{-2})	ΔG^{ref} (kcal mol^{-1})	ΔG^{free} (kcal mol^{-1})
1. using van der Waals surface (probe radius = 0.0)			
Arg guanido	5.0	-0.118	0.591
Asn, Gln $CONH_2$	5.0	-0.079	0.403
Asp, Glu COO^-	11.9	-0.186	0.927
His imidazole	2.2	-0.070	0.188
Lys NH_3^+	9.3	-0.116	0.138
Phe aromatic	0.57	-0.018	0.043
Ser, Thr OH	9.1	-0.040	0.978
II. using solvent accessible surface (probe radius = 1.4)			
Arg guanido	37	-0.456	4.07
Asn, Gln $CONH_2$	34	-0.378	2.64
Asp, Glu COO^-	85	-1.446	6.68
His imidazole	26	-0.579	1.85
Lys NH_3^+	48	-0.550	6.65
Phe aromatic	6.1	-0.054	0.412
Ser, Thr OH	49	-0.173	5.21

Figure VI-1. Models used to calculate exact DelPhi side chain desolvation energies. The areas drawn in white were assigned a dielectric constant of 4 (protein interior) and the blue areas have a dielectric constant of 80 (water) and a salt concentration of 50 mM. (a) Folded state side chain solvation. Atoms in side chain *i*, shown in purple, both "generate" and "feel" the electrostatic potential. Side chain and backbone atoms shown in gray are assigned a partial atomic charge of 0. All side chain and backbone atoms are used to define the dielectric boundary. (b) Unfolded state side chain solvation. Atoms in side chain *i*, shown in purple, both "generate" and "feel" the electrostatic potential. Atoms shown in gray are assigned a partial atomic charge of 0. The dielectric boundary is defined using the atoms in side chain *i* and the local backbone only.

(a)



(b)

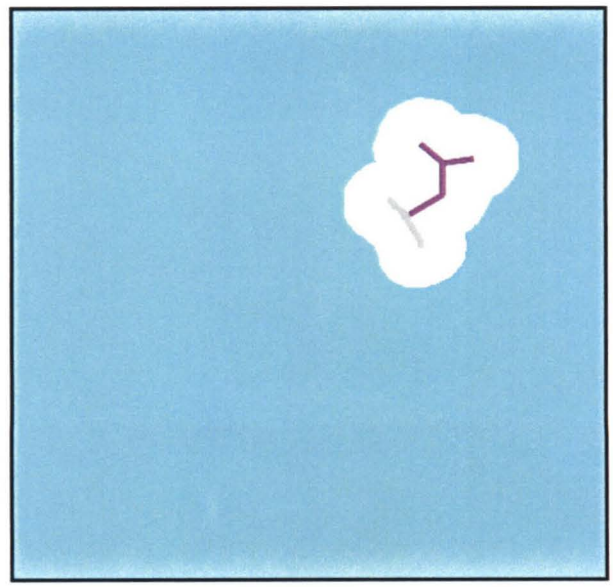


Figure VI-2. Models used to calculate exact DelPhi side chain - backbone screening energies. The areas drawn in white were assigned a dielectric constant of 4 (protein interior) and the blue areas have a dielectric constant of 80 (water) and a salt concentration of 50 mM. Atoms in side chain *i*, shown in orange, "generate" the electrostatic potential and backbone atoms, shown in green, "feel" the electrostatic potential. Side chain atoms shown in gray are assigned a partial atomic charge of 0. All side chain and backbone atoms are used to define the dielectric boundary. Sscreening energies are added to the Coulombic energies to obtain screened Coulombic energies.

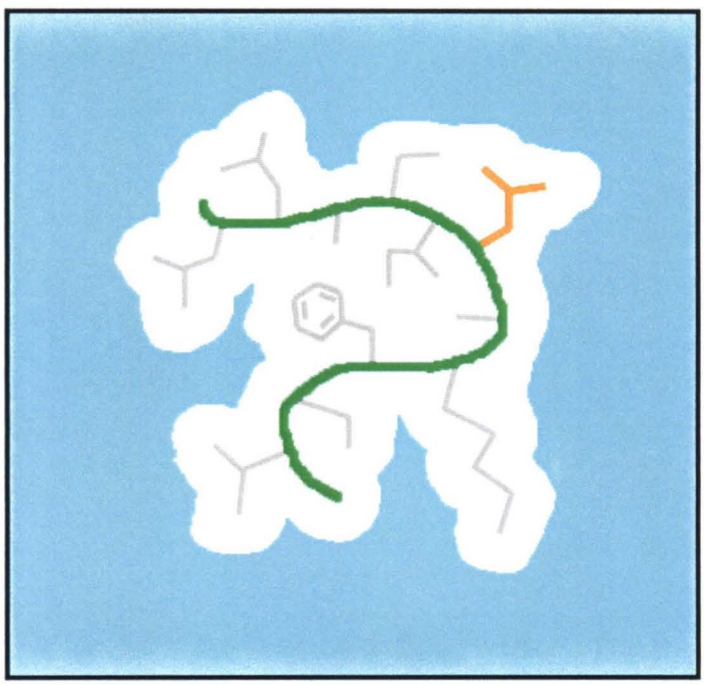


Figure VI-3. Models used to calculate exact DelPhi side chain - side chain screening energies. The areas drawn in white were assigned a dielectric constant of 4 (protein interior) and the blue areas have a dielectric constant of 80 (water) and a salt concentration of 50 mM. Atoms in side chain *i*, shown in orange, "generate" the electrostatic potential and atoms in side chain *j*, shown in green, "feel" the electrostatic potential. Side chain and backbone atoms shown in gray are assigned a partial atomic charge of 0. All side chain and backbone atoms are used to define the dielectric boundary. The screening energies were added to the Coulombic energies to obtain screened Coulombic energies.

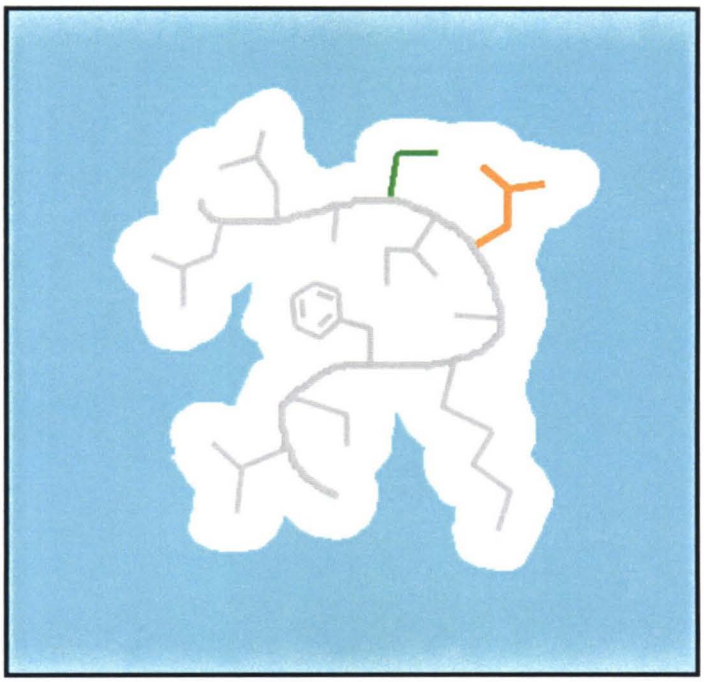


Figure VI-4. Comparison of the side chain desolvation energies calculated using DelPhi *versus* the energies calculated using the ORBIT polar hydrogen burial penalty.

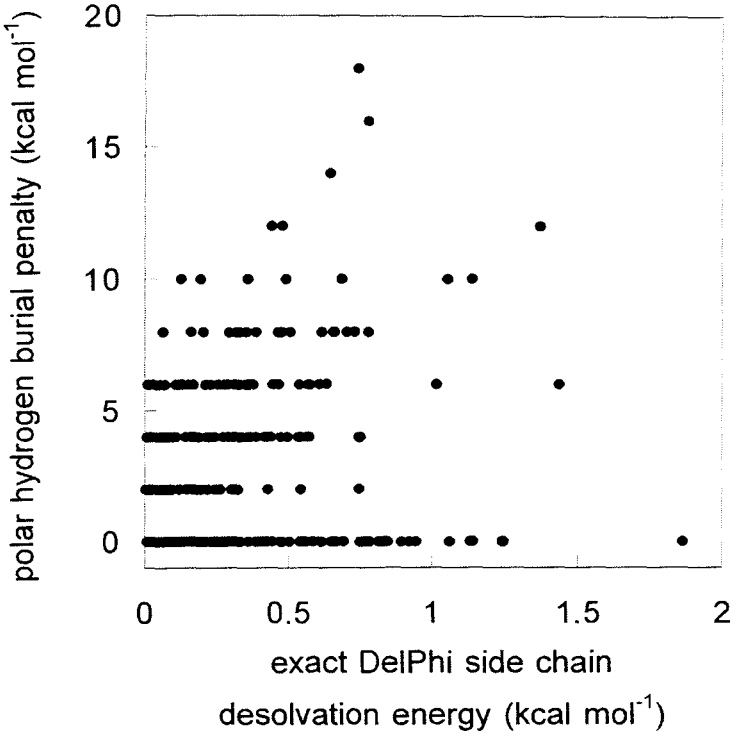

20
15
10
5
0
0
0.5
1
1.5
2
polar hydrogen burial penalty (kcal mol-1)
exact DelPhi side chain
desolvation energy (kcal mol-1)

Figure VI-5. Comparison of the side chain desolvation energies calculated using DelPhi *versus* the energies calculated using the ORBIT polar area penalty.

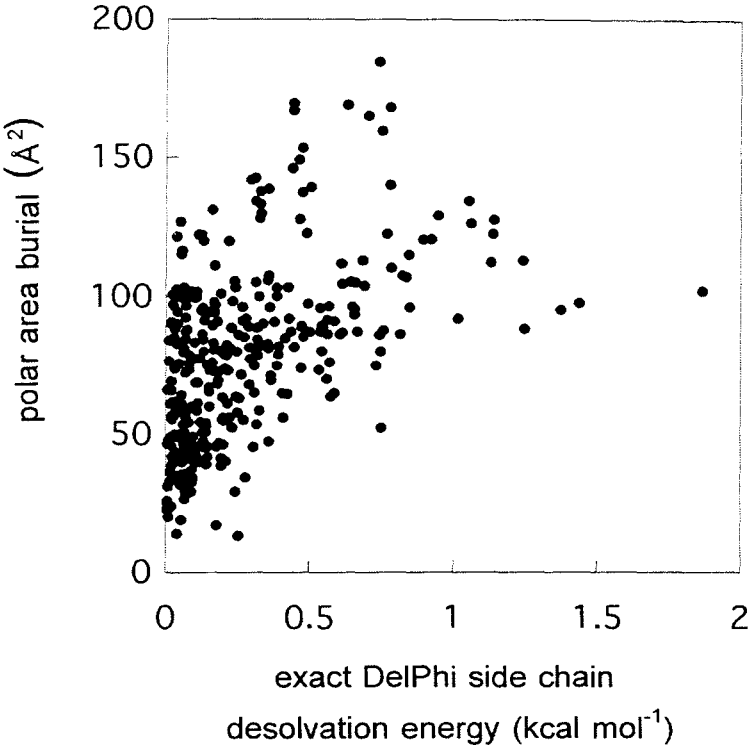

200
150
100
50
0
polar area burial (Å^2)
0
0.5
1
1.5
2
exact DelPhi side chain
desolvation energy (kcal mol^{-1})

Figure VI-6. Comparison of the side chain - backbone screening energies calculated using DelPhi *versus* the energies calculated using the ORBIT hydrogen bond plus Coulombic terms.

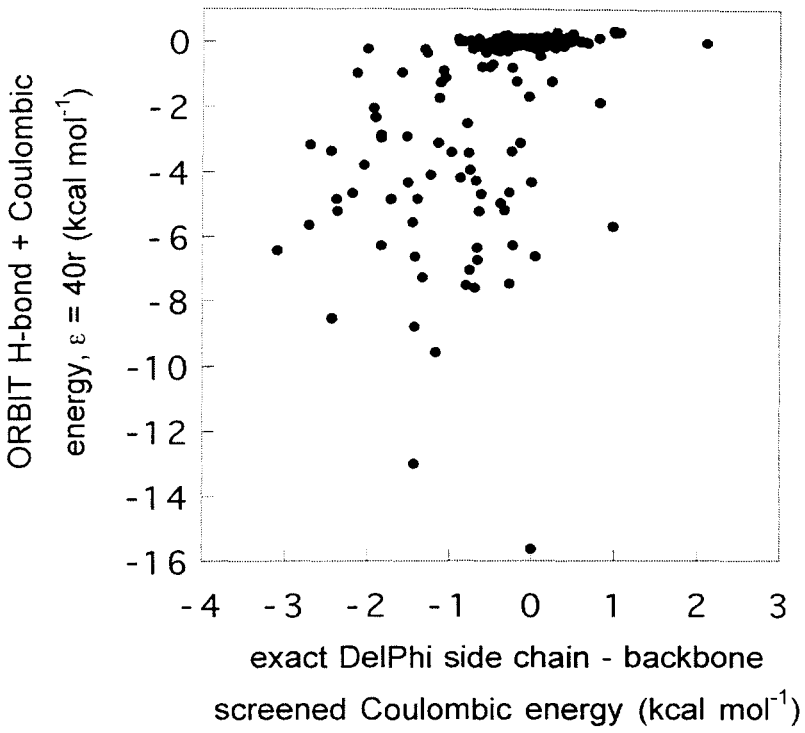

ORBIT H-bond + Coulombic energy, ε = 40r (kcal mol-1)
0
-2
-4
-6
-8
-10
-12
-14
-16
-4
-3
-2
-1
0
1
2
3
exact DelPhi side chain - backbone screened Coulombic energy (kcal mol-1)

Figure VI-7. Comparison of the side chain - side chain screening energies calculated using DelPhi *versus* the energies calculated using the ORBIT hydrogen bond plus Coulombic terms.

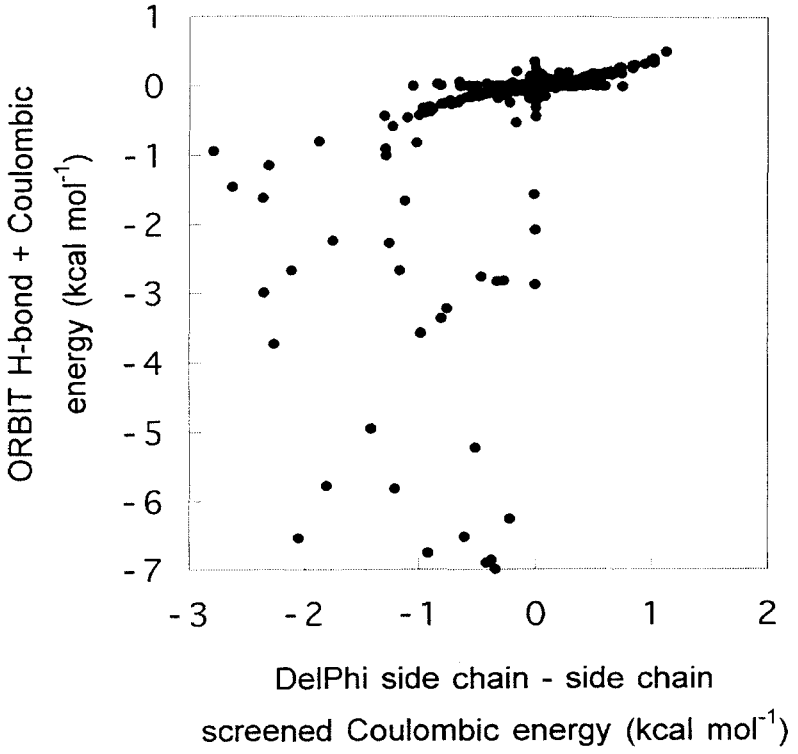

1
0
-1
-2
-3
-4
-5
-6
-7
-3
-2
-1
0
1
2
ORBIT H-bond + Coulombic energy (kcal mol-1)
DelPhi side chain - side chain screened Coulombic energy (kcal mol-1)

Figure VI-8. Comparison of the side chain - backbone screening energies calculated using DelPhi *versus* Coulomb's law using a dielectric of 38.1.

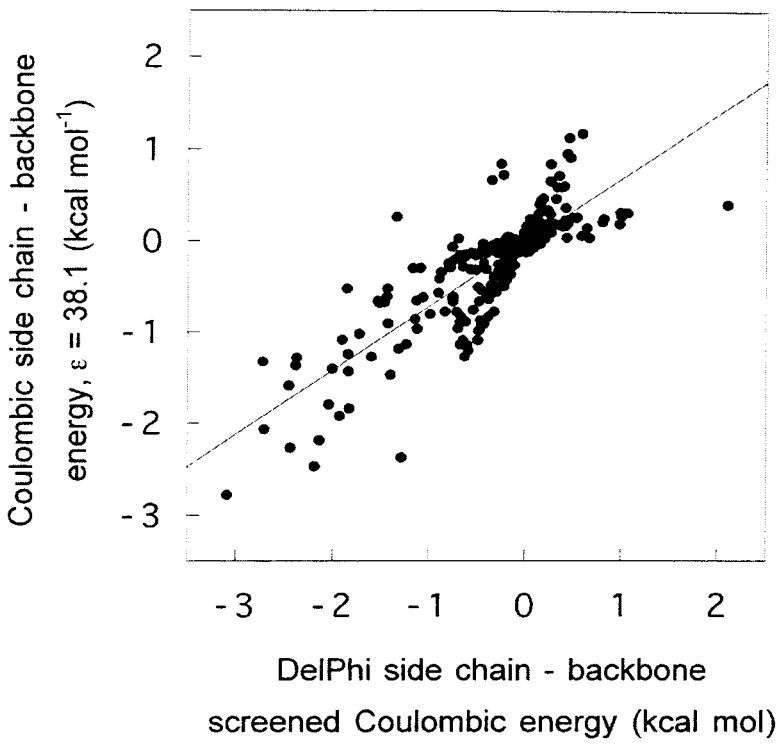

Coulombic side chain - backbone
energy, ε = 38.1 (kcal mol-1)
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
DelPhi side chain - backbone
screened Coulombic energy (kcal mol)

Figure VI-9. Comparison of the side chain - backbone screening energies calculated using DelPhi *versus* Coulomb's law using a dielectric of 13.1*r*.

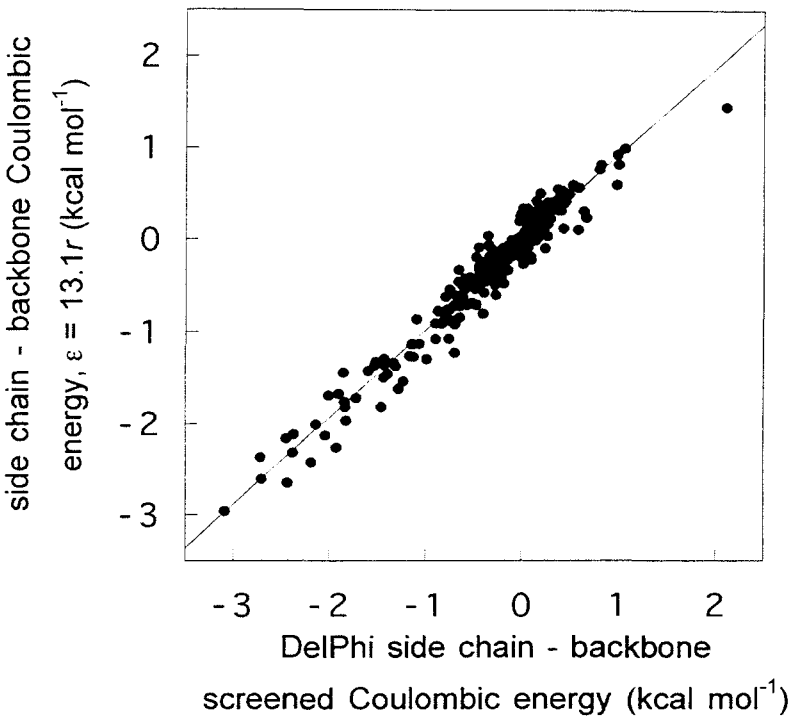

side chain - backbone Coulombic
energy, ε = 13.1r (kcal mol-1)
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
DelPhi side chain - backbone
screened Coulombic energy (kcal mol-1)

Figure VI-10. Comparison of the side chain - side chain screening energies calculated using DelPhi *versus* Coulomb's law using a dielectric of 65. Colored points correspond to charge - charge interactions in which the shortest distance between partial charges in the two side chains is less than 3 Å (red), 3 - 5 Å (orange), 5 - 10 Å (green), or greater than 10 Å (blue). Dark gray points correspond to interactions between a charged amino acid and a polar neutral amino acid, and light gray points correspond to interactions between a pair of polar neutral amino acids.

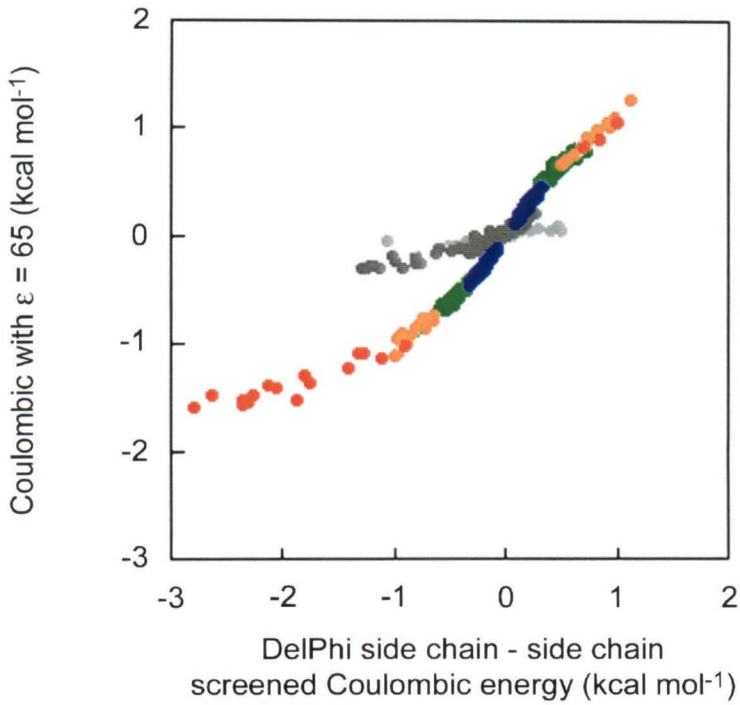

2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
Coulombic with ε = 65 (kcal mol-1)
DelPhi side chain - side chain
screened Coulombic energy (kcal mol-1)

Figure VI-11. Comparison of the side chain - side chain screening energies calculated using DelPhi *versus* Coulomb's law using a dielectric of 12.8*r*.

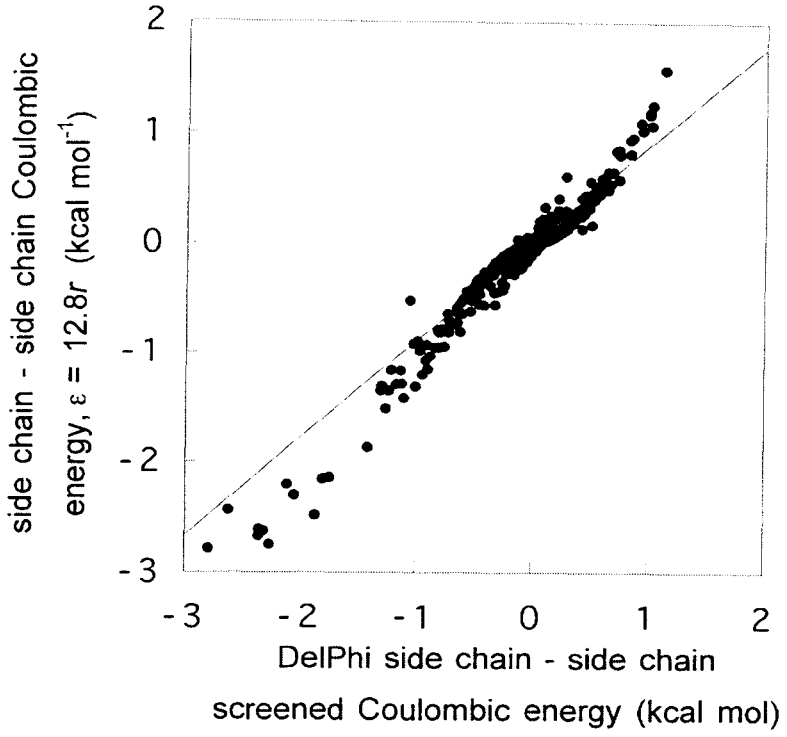

side chain - side chain Coulombic energy, ε = 12.8r (kcal mol-1)
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
DelPhi side chain - side chain screened Coulombic energy (kcal mol)

Figure VI-12. Comparison of the side chain desolvation energies calculated using DelPhi *versus* optimized atomic solvation parameters for each polar functional group, as given in Table VI-3.

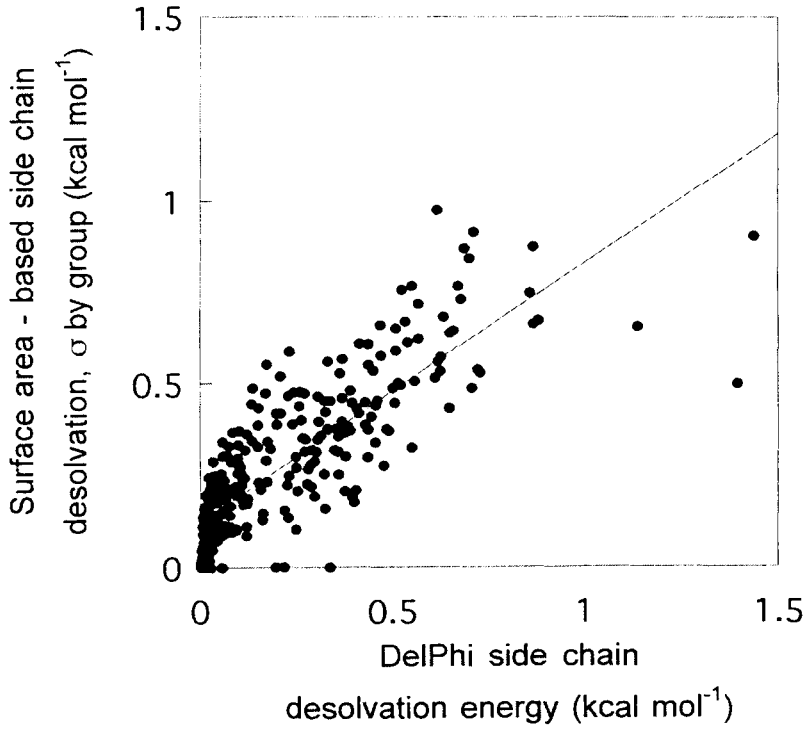
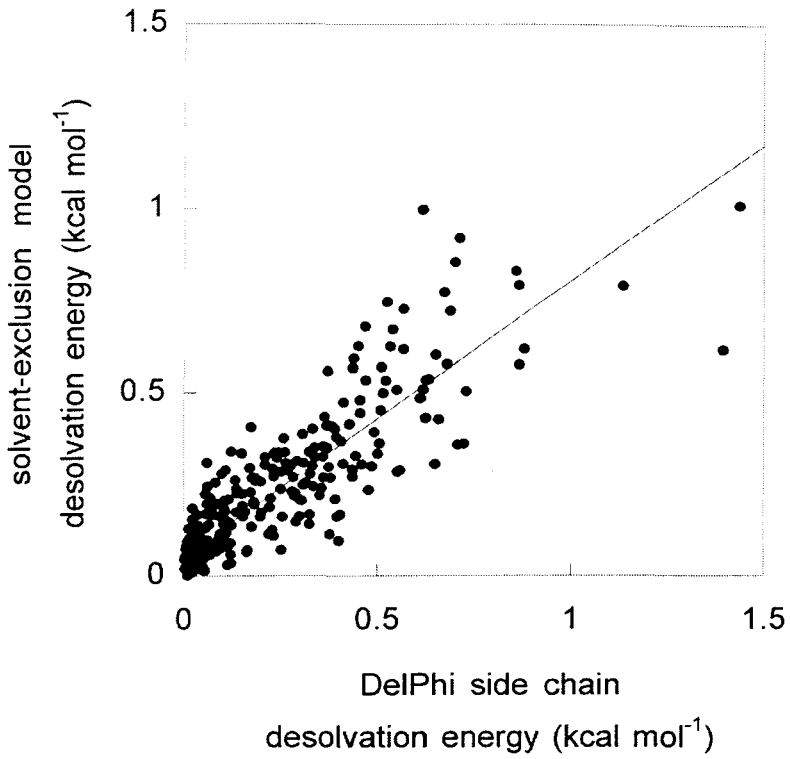

1.5
1
0.5
0
0
0.5
1
1.5
Surface area - based side chain
desolvation, σ by group (kcal mol-1)
DelPhi side chain
desolvation energy (kcal mol-1)

Figure VI-13. Comparison of the side chain desolvation energies calculated using DelPhi *versus* optimized solvent-exclusion model parameters for each polar functional group, given in Table VI-3.


1.5
1
0.5
0
0
0.5
1
1.5
solvent-exclusion model
desolvation energy (kcal mol-1)
DelPhi side chain
desolvation energy (kcal mol-1)

Chapter VII

Electrostatic Models for Protein Design Calculations. II. One and Two Body Decomposable Poisson-Boltzmann Methods

The text of this chapter is adopted from an unpublished manuscript that was coauthored with Emil G. Alexov, Professor Barry Honig, and Professor Stephen L. Mayo.

Abstract

Successfully modeling electrostatic interactions is one of the key factors required for the computational design of proteins with desired physical, chemical, and biological properties. Computational protein design algorithms have typically used fast methods based on Coulomb's law and/or geometry dependent hydrogen bond terms to model electrostatic interactions. These methods fail to accurately account for desolvation of polar protein groups and solvent screening of Coulombic interactions, which strongly attenuate and modulate electrostatic interactions in proteins. Continuum models of electrostatics such as those based on the finite difference Poisson-Boltzmann method (FDPB) have substantially better predictive power, but are intractable for problems with high combinatorial complexity. In this paper, we present one and two body decomposable formulations of the FDPB model. The new methods produce energies that are very similar to the results of traditional FDPB calculations and are compatible with the computational demands of design calculations. These new electrostatic models should significantly aid in efforts to design proteins with desired thermodynamic and functional properties.

Introduction

Electrostatic interactions are often critical determinants of protein structure and function. Proper modeling of electrostatic interactions will likely be crucial for the design of proteins that possess desired physical, chemical and biological properties. However, the currently available electrostatic models that are reasonably accurate are far too slow for protein design calculations. As a result, computational protein design algorithms[1-4] have typically either neglected electrostatic interactions or relied on continuum two-body methods based on Coulomb's law and/or explicit hydrogen bond terms to model electrostatic interactions. These methods fail to accurately account for desolvation of polar protein groups and solvent screening of Coulombic interactions, which both strongly attenuate and modulate electrostatic interactions in proteins. Results of an earlier protein design study indicate that using a simple electrostatic model results in systematic overestimation of the importance of side chain - side chain hydrogen bonds, and underestimation of the contribution of side chain - backbone interactions. Consequently, the stability of proteins designed using a simple electrostatic model is far from optimal[5]. Accurate modeling of electrostatic interactions may prove to be even more critical for the incorporation of functions such as binding and catalysis into designed proteins.

Electrostatic interactions in proteins can be modeled with reasonable accuracy using several methods, including explicit solvent models[6], the protein dipole Langevin dipole (PDLD) method[7], finite difference Poisson-Boltzmann (FDPB) methods[8, 9], and the Born model[10, 11]. In each of these methods, it is necessary to specify the conformation of the protein in order to define the spatial regions that correspond to the protein and the solvent. In addition, explicit solvent models require averaging over a large number of solvent conformations for each protein conformation. In protein design calculations, each possible rotameric sequence (a rotamer is a low energy amino acid conformation) will correspond to

a distinct protein structure and will require an independent PDLD, FDPB, or GB model calculation to determine its electrostatic energy.

The combinatorial complexity of design calculations is typically very high. For example, 10^{67} rotameric sequences were considered in the recently reported design of 29 surface positions in engrailed homeodomain[5]. If each rotameric sequence requires an independent calculation, even if each individual calculation takes only 1 μs, over 10^{53} years would be required to calculate the electrostatic energies of all the rotameric sequences. Furthermore, energies that are not pairwise decomposable are incompatible with deterministic search algorithms such as Dead End Elimination (DEE)[12-14] that are used for sequence selection. To satisfy the computational requirements of protein design calculations, it is necessary to develop force field terms that are two-body decomposable and can be calculated rapidly. Previous studies have used modified versions of the generalized Born model and the Tanford-Kirkwood model to calculate electrostatic energies for large numbers of protein conformations. However, further modifications would be required to enable these models to accurately calculate electrostatic energies for large numbers of different protein sequences.

The GB equation itself is pairwise decomposable. However, to solve the GB equation, Born radii must be calculated for each charged atom. The radii depend on the protein conformation, so the radii calculations are not rigorously pairwise decomposable[15]. In molecular dynamics simulations, Born radii are observed to be reasonably insensitive to slight conformational changes and therefore do not need to be updated at every step[10]. However, an analogous method for minimizing the number of radius calculations in protein design calculations is not obvious. Pairwise decomposable methods for calculating approximate Born radii have been recently developed[15]. While the Born model calculations would be tractable using approximate radii, the error introduced by calculating the Born radii pairwise is not insignificant, especially for short range interactions. Nonetheless, it

may be possible to develop an electrostatic model suitable for protein design calculations based on the Born model.

A modified Tanford-Kirkwood (MTK) model has also been proposed for protein design and protein modeling calculations[16]. Based on the reported timings, the MTK model is computationally demanding but tractable. For example, the homeodomain surface design, comprising about 5500 rotamers and 15,000,000 rotamer pairs, would require about 23 days. A serious limitation of the MTK model is that it neglects variability in the shape of the boundary separating the low dielectric protein from the high dielectric solvent. When considering different side chain conformations of a single amino acid sequence, fixing the dielectric boundary may be reasonable. However, using a single dielectric boundary for different amino acid sequences is likely to result in significant errors.

Changes in protein sequence and conformation will affect the location of the boundary between the high dielectric solvent and the low dielectric protein interior. In the FDPB model, electrostatic energies are sensitive to the structure of the dielectric boundary. If the dielectric boundary is held constant in a design calculation and a small residue is mutated to a larger residue, charged groups in the larger residue will often lie outside of the protein dielectric region and be assigned the solvent dielectric. Similarly, if a large residue is mutated to a smaller residue and the dielectric boundary is not adjusted, the charged groups in the smaller residue will be modeled as being far away from the dielectric boundary even if they are actually in contact with the solvent.

Strategies for incorporating FDPB methods into protein design calculations

Rather than further modifying the GB or MTK models, we have developed new strategies for incorporating the results of a tractable number of FDPB calculations into the energy matrix prior to sequence selection. In each case, the FDPB calculations are conducted

using simplified representations of the protein surface that require knowledge of the identity and conformation of no more than two amino acid side chains at a time. In this study, we have used the FDPB solver from the computer program DelPhi[8] to calculate electrostatic energies for eight proteins that have been targets for design and other biophysical studies: fragment B of Streptococcal protein A, the c-Crk SH3 domain, hen egg white lysozyme, the β1 domain of Streptococcal protein G, ubiquitin, plastocyanin, bovine pancreatic trypsin inhibitor, and rubredoxin. The results of these initial FDPB calculations were compared to the results of FDPB calculations performed using simplified surface representations in order to assess their accuracy. For compatibility with the current design procedure, we have calculated side chain internal energies, side chain - backbone energies, and side chain - side chain energies separately. Both desolvation energies and screened Coulombic energies, described below, were considered.

Polar protein groups can form favorable electrostatic interactions with the solvent; we refer to the resulting energies as electrostatic solvation energies. The difference between the electrostatic solvation energy of a polar group in the folded state versus the unfolded state is the desolvation energy. In design calculations, the backbone conformation is held fixed. The desolvation energy of the backbone can therefore be defined as the difference between the electrostatic solvation energy of the isolated backbone, shown in Figure VII-1b, and the electrostatic solvation energy of the backbone in the presence of all of the side chains, shown in Figure VII-1a. The desolvation energy of a side chain is defined as the difference in the electrostatic solvation energy of the side chain and local backbone alone, shown in Figure VII-2b, versus the electrostatic solvation energy of the side chain in the context of the folded protein, as shown in Figure VII-2a.

Electrostatic interactions between polar protein groups and the solvent also act to screen Coulombic interactions within a protein. The screening energy is always opposite in

sign and weaker in magnitude than the Coulombic energy for a given interaction. The procedures used to calculate side chain - backbone and side chain - side chain screening energies are shown in Figures VII-3 and VII-4, respectively. In all cases, the screening energies and Coulombic energies are added to yield screened Coulombic energies and the screened Coulombic energies predicted by the different electrostatic models are then compared. This is an important point: as solvation energies are strongly correlated with Coulombic energies, a strong correlation between predicted and actual solvation energies does not necessarily indicate that the predicted solvation energies are accurate. In fact, Scarsi and Caflisch have shown that it is possible to observe nearly perfect correlation (r=0.99) between two sets of solvation energies and no correlation (r=0.008) between the corresponding screened Coulombic energies. Based on these observations, they propose that comparison of screened Coulombic energies but not screening energies alone is appropriate for the validation of approximate electrostatic models[17].

A one-body FDPB decomposition

Several physical properties of proteins can be calculated using information about the protein surface. While protein surfaces can not be perfectly represented using pairwise decomposable methods, earlier protein design studies have demonstrated that pairwise or sequence independent approximations can yield satisfactory results for hydrophobic solvation and binary patterning, respectively[18, 19]. Similarly, it may be possible to obtain accurate estimates of the FDPB energies obtained using all the atomic coordinates to define the surface (hereafter referred to as "exact DelPhi energies") from FDPB energies obtained using simplified models for the protein surface that require knowledge of only one or two side chain conformations at a time. The one or two-body FDPB energies could replace some or all of the electrostatic terms currently calculated in protein design force fields.

Since the protein backbone is fixed during design calculations, an approximate surface can be obtained using the atoms from the protein backbone and the side chain of interest only. It is necessary to include the side chain of interest when defining the protein surface to ensure that all protein charges are located in the low dielectric protein region rather than the high dielectric solvent region. The one-body backbone desolvation energy for each side chain is calculated as the difference between the one-body folded state, shown in Figure VII-5a, and the reference state, shown in Figure VII-5b. The total backbone desolvation energy for each protein is the sum of the one-body backbone desolvation energies of each of its side chains. One-body side chain desolvation energies are calculated as the difference in solvation energy between the one-body folded state, shown in Figure VII-6a, and the unfolded state, shown in Figure VII-6b. The one-body side chain - backbone screened Coulombic energy of each side chain is calculated using the model in FigureVII-7.

To test the accuracy of the one body decomposition, we calculated the side chain desolvation energies and the side chain - backbone screened Coulombic energies of all of the polar atoms in the set of eight proteins. Backbone desolvation energies can be calculated reasonably well by summing the desolvation induced by the presence of each side chain, as shown in Figure VII-8. Using the one-body decomposition, the backbone desolvation energy resulting from each side chain can be considered as a component of the internal energy of the side chain in design calculations. The extent to which backbone desolvation energy depends on protein sequence and side chain conformations is not yet known.

The side chain desolvation energies predicted using the one-body method, shown in Figure 8, sometimes match the exact FDPB energies but are often smaller in magnitude. In cases where the one-body energy is underestimated, the side chain is desolvated by other side chains, not just the backbone. The one-body side chain - backbone screened Coulombic

energies correlate well with the exact energies and exhibit small scatter, as shown in Figure VII-7 and Table VII-1.

A two-body FDPB decomposition

In the one-body FDPB method, we calculated side chain and backbone desolvation energies and side chain - backbone screening energies, but not side chain - side chain screening energies. Simply multiplying the one-body potential generated by side chain *i* by the partial atomic charges of side chain *j* is not very accurate (data not shown), especially for charged atoms located at or beyond the dielectric boundary. Side chain - side chain screened Coulombic energies were calculated using a two-body decomposable method that uses only the backbone and two side chains of interest to define the dielectric boundary, as shown in Figure VII-13. The accuracy obtained using a 2-body decomposition of DelPhi is quite good, as shown in Table VII-1 and Figure VII-16.

Two body methods were also used to improve the accuracy of the one-body FDPB calculations. Two-body corrections can be determined from the perturbation in the electrostatic potential generated by one side chain when a second side chain is added to the low dielectric protein region. Alternatively, we can calculate the difference in a given electrostatic energy calculated with and without including a second side chain in the low dielectic protein region. Two-body side chain desolvation energies are calculated using the folded state shown in Figure VII-11a and the unfolded state in Figure VII-11b, and two-body side chain backbone screening energies are calculated using the model shown in Figure VII-12. The effects of each other side chain are summed to obtain the side chain desolvation or side chain - backbone energy for a given side chain. Incorporating the effects of other side chains using the two-body method described above allows accurate calculation of electrostatic energies, as shown in Table VII-1 and Figures VII-14 and VII-15.

The number of pairs in a design calculation is often large, making 2-body FDPB calculations very slow. For instance, the surface design calculation for engrailed homeodomain considers 15,000,000 pairs, which would require over two years of CPU time on a single processor. The time required to complete a 2-body calculation can be significantly reduced by using parallel processing: the homeodomain surface calculation would require about a week on 128 IBM SP3 processors running at 375 MHz. Nonetheless, it would be desirable to reduce the number of pairs calculations that are performed.

Analysis of the side chain desolvation and side chain - backbone screened Coulombic energies indicates that, in most cases, the effect of a second side chain is negligible. The small fraction of 2-body perturbations that are significant involve pairs of residues that are close in space. We conducted additional two-body calculations in which 2-body perturbations were only calculated for pairs that were separated by less than 6 Å or 4 Å. Considering only a limited subset of pairs reduces the total calculation time by over an order of magnitude with only a slight decrease in accuracy, as indicated in Table VII-1 and Figures 17 and 18.

The time required to calculate side chain - side chain screened Coulombic energies can be reduced by using the two-body method only for pairs that are close in space and using Coulomb's law for pairs that are more distant. A dielectric of 47 was found to give the best agreement for pairs separated by over 6 Å and a dielectric of 48 was found to be optimal for pairs separated by over 4 Å. The reduction in accuracy for this hybrid approach is insignificant, as indicated in Table VII-1 and Figure 19: the RMSD increases by only 0.01 kcal mol^{-1} using a 4 Å cutoff.

Additional considerations

Thus far, we have developed and tested new electrostatic models for protein design calculations by maximizing the agreement between the approximate desolvation or screened

Coulombic energies and the "exact" DelPhi energies. The choices we make when calculating the "exact" DelPhi energies will also affect how well the approximate energies will be able to predict experimental results in future design studies. Three such considerations are entropic attenuation of side chain - side chain interactions, unfolded state modeling of side chain - local backbone interactions, and backbone desolvation.

The magnitudes of favorable side chain - side chain interaction energies calculated using DelPhi can be large compared to the experimentally determined contribution of a salt bridge pair[5]. This overestimation likely results because loss of side chain entropy is not considered. Explicitly and accurately modeling the entropy of surface side chains is quite challenging. A simple approach, truncating screened Coulombic interactions that are unreasonably large in magnitude, improved the correlation between predicted and experimentally determined stability in a set of designed homeodomain variants. More sophisticated approaches that model side chains by a conformational ensemble rather than a fixed rotamer may also prove useful.

Interactions between side chains and local backbone are also likely to be smaller in magnitude than the "exact" DelPhi energies calculated in this study because side chain - local backbone interactions are likely present in the unfolded state. While it is difficult to model the unfolded state quickly and accurately, it may make sense to attenuate or truncate side chain - local backbone interactions. Finally, each protein only yields one backbone desolvation energy, so the statistical significance of the backbone desolvation parameters is poor. We plan to use future experimental results to determine the best way to model extremely favorable side chain - side chain interactions, side chain - local backbone interactions, and backbone desolvation.

Conclusions

Accurate electrostatic models, including the FDPB model, require knowledge of the full tertiary structure of the protein in order to define the dielectric boundary between the protein and solvent. As a result, these models can not be applied to protein design calculations, which often consider over 10^{50} possible protein structures. While it is not possible to explicitly consider each possible structure of the dielectric boundary, it is also not prudent to model many protein sequences using a single dielectric boundary. Variation in the dielectric boundary between different sequences threaded along a protein backbone can lead to significant differences in electrostatic energies.

We have found that it is possible to obtain accurate electrostatic energies using simplified surface models that depend on the identity and conformation of the protein backbone and one or two side chains at a time. The simplified surfaces are most accurate in the immediate vicinity of the partial charges that are generating and feeling the electrostatic potential in each calculation. Changes in the dielectric boundary resulting from other nearby side chains are captured in a pairwise fashion. Finally, sequence dependent variation in the dielectric boundary can be neglected if it is reasonably far removed from the partial charges that are generating or feeling the electrostatic potential in a given calculation.

The stability of designed proteins has already been demonstrated to be sensitive to the quality of the electrostatic model used in the design calculations. It is likely that electrostatic interactions are at least as important in determining the functional properties of proteins, including binding and catalysis. As a result, the development and testing of accurate electrostatic models is likely to significantly aid in the design of proteins with specific physical, chemical, and biological properties.

Materials and Methods

Test set of proteins. All calculations were performed on eight proteins that are popular targets for design and other biophysical studies: fragment B of protein A, the c-Crk SH3 domain, hen egg white lysozyme, the β1 domain of Streptococcal protein G, ubiquitin, plastocyanin, bovine pancreatic trypsin inhibitor, and rubredoxin. Structural coordinates were obtained from the PDB entries 1cka, 1fc2, 1hel, 1pga, 1ubi, 2pcy, 6pti, and 8rxn, respectively. Explicit hydrogens were added to each structure using BIOGRAF (Molecular Simulations, Inc. San Diego) and the termini were adjusted to carry a net charge of ±1. The resulting structures were minimized for 50 steps using the Dreiding force field; the minimized structures were used in the calculations discussed in the following sections.

Exact DelPhi calculations. Finite difference solutions to the linearized Poisson-Boltzmann equation were obtained using the FDPB solver from the computer program DelPhi[20] with a grid spacing of 2.0 grids $Å^{-1}$, an interior dielectric of 4.0, an exterior dielectric of 80.0, and 0.050 M salt. Dielectric boundaries were defined using van der Waals surfaces, which were found to give better agreement with experimental results than solvent accessible surfaces [21]. The grid size was selected for each protein so that its backbone atoms fill 70 % of the grid. The coordinates of each protein were mapped onto the grid in exactly the same way in each calculation to minimize errors due to differences in grid placement. The PARSE parameter set charges and atomic radii[22] were used in all FDPB calculations. Proline and disulfide bonds were considered part of the backbone in all calculations. All His, Arg, and Lys residues were modeled with a +1 net charge, all Asp and Glu residues were modeled with a -1 charge, and all other residues were modeled with a net charge of 0. All DelPhi

energies were converted to units of kcal mol^{-1} using the relation kT = 0.593 kcal mol^{-1} at 25 °C.

The desolvation energy of the backbone is defined as the difference between the electrostatic solvation energy of the backbone alone and the electrostatic solvation energy of the backbone in the presence of all the protein side chains:

$$\Delta\Delta G_{\text{exact desolv(backbone)}} = (1/2) \sum_t q_t (\phi^{\text{all}} - \phi^{\text{bb only}}) \quad (1)$$

where the sum is over all backbone atoms, each t is a backbone atom, q_t is the partial atomic charge of backbone atom t, ϕ^{all} is the potential at atom t generated by the set partial atomic charges on the backbone when all of the protein atoms are used to define the dielectric boundary, and $\phi^{\text{bb only}}$ is the potential at atom t generated by the set partial atomic charges on the backbone when the backbone atoms only are used to define the dielectric boundary.

The desolvation energy of a side chain, i, is defined as the difference between the electrostatic solvation energy of the side chain in the folded state versus the unfolded state:

$$\Delta\Delta G_{\text{exact desolv(side chain } i)} = (1/2) \sum_u q_u (\phi^{\text{all}} - \phi^{i \text{ only}}) \quad (2)$$

where the sum is over the atoms in side chain i, each u is an atom in side chain i, q_u is the partial atomic charge of side chain atom u, ϕ^{all} is the reaction potential at u, generated by the set of partial atomic charges on the side chain, when all of the protein atoms are used to define the dielectric boundary, and $\phi^{i \text{ only}}$ is the reaction potential at u, generated by the set of partial atomic charges on side chain i, when the atoms on side chain i and the local backbone only are used to define the dielectric boundary. The unfolded state was modeled as the side chain and local backbone, mapped to the grid exactly as in the folded state

calculations. The local backbone is defined to include the following atoms: CA(i-1), C(i-1), O(i-1), N(i), HN(i), CA(i), C(i), O(i), N(i+1), HN(i+1), and CA(i+1).

Folded state side chain - backbone screening energies were obtained using the following equation:

$$\Delta G_{\text{exact screening(sc-bb)}} = \sum_t q_t \, \phi^{\text{all}} \tag{3}$$

where the sum is over the backbone atoms, i is the side chains of interest, each t is an atom in the backbone, q_t is the partial atomic charge of atom t, and ϕ^{all} is the reaction potential due to the set of partial atomic charges on side chain i at t, when all of the protein atoms are used to define the dielectric boundary. The screening energies were then added to the Coulombic energies to obtain screened Coulombic energies:

$$\Delta G_{\text{screened Coulombic(sc-bb)}} = \Delta G_{\text{screening(sc-bb)}} + \Delta G_{\text{Coulombic(sc-bb)}} \tag{4}$$

where the Coulombic energy is calculated using Coulomb's law with the dielectric equal to the dielectric of the protein interior.

Side chain - side chain interactions are obtained using a similar method:

$$\Delta G_{\text{exact screening(sc-sc)}} = \sum_v q_v \, \phi^{\text{all}} \tag{5}$$

where the sum is over atoms in side chain j, i and j are the side chains of interest, each v is an atom in side chain j, q_v is the partial atomic charge of atom v, and ϕ^{all} is the reaction potential due to the set of partial atomic charges on side chain i at v, when all of the protein

atoms are used to define the dielectric boundary. The screening energies were then added to the Coulombic energies to obtain screened Coulombic energies:

$$\Delta G_{\text{screened Coulombic(sc-sc)}} = \Delta G_{\text{screening(sc-sc}} + \Delta G_{\text{Coulombic(sc-sc)}} \quad (6)$$

All of the protein atoms were used to define the dielectric boundary when calculating the screening energies. Side chain - backbone and side chain -side chain interaction energies are assumed to be zero in the unfolded state.

One-body FDPB calculations. One-body FDPB energies were calculated for backbone desolvation energies, side chain desolvation energies, and side chain - backbone screened Coulombic interaction energies. Folded state solvation energies for the protein backbone were calculated as in the exact DelPhi calculations, except that side chains other than the side chain of interest were not included:

$$\Delta\Delta G_{\text{1-body desolv(backbone)}} = (1/2) \sum_t q_t \, (\phi^{\text{bb},\, i} - \phi^{\text{bb only}}) \quad (7)$$

where each t is a backbone atom, q_t is the partial atomic charge of backbone atom t, $\phi^{\text{bb},\, i}$ is the potential at atom t generated by the set partial atomic charges on the backbone when side chain i and the backbone atoms only are used to define the dielectric boundary, and $\phi^{\text{bb only}}$ is the potential at atom t generated by the set partial atomic charges on the backbone when the backbone atoms only are used to define the dielectric boundary.

Similarly, side chain desolvation and side chain - backbone screened Coulombic interactions were calculated as in the exact DelPhi calculations, except only the side chain of interest and the backbone were used to construct the dielectric boundary:

$$\Delta\Delta G_{\text{1-body desolv(side chain } i)} = (1/2) \sum_u q_u \, (\phi^{\text{bb}, i} - \phi^i) \tag{8}$$

$$\Delta G_{\text{1-body screening(sc-bb)}} = \sum_t q_t \, \phi^{\text{bb}, i} \tag{9}$$

where i is the side chain of interest, each u is an atom in side chain i, q_u is the partial atomic charge of atom u, $\phi^{\text{bb}, i}$ is the potential at atom u generated by the set partial atomic charges on side chain i when side chain i and the backbone atoms only are used to define the dielectric boundary, and ϕ^i is the potential at atom u generated by the set partial atomic charges on side chain i when side chain i atoms only are used to define the dielectric boundary.

Two-body FDPB calculations. Two-body FDPB side chain desolvation energies, side chain - backbone screened Coulombic interaction energies, and side chain - side chain screened Coulombic energies were calculated. The two-body side chain - side chain calculation is performed using the same method as was used to calculate the exact side chain - side chain screening energies, except that the dielectric boundary is defined using only the backbone and two side chains of interest:

$$\Delta G_{\text{2-body screening}(i,j)} = \sum_v q_v \, \phi^{\text{bb}, i, j} \tag{10}$$

where i and j are the two side chains of interest, each v is an atom in side chain j, q_v is the partial atomic charge of atom v, and $\phi^{\text{bb}, i, j}$ is the reaction potential due to the set of partial atomic charges on side chain i at v, when the backbone and side chains i and j only are used to define the dielectric boundary.

The side chain desolvation and side chain - backbone screened Coulombic energies were calculated as the sum of a one body energy and a two body correction energy:

$$\Delta\Delta G_{\text{2-body desolv(side chain } i)} = \Delta\Delta G_{\text{1-body dslv}(i)} + \sum_{j \neq i} [(1/2) q_u \, \phi^{\text{bb}, i, j} - \Delta\Delta G_{\text{1-body dslv}(i)}] \quad (11)$$

$$\Delta G_{\text{2-body screening(sc-bb)}} = \Delta G_{\text{1-body scrn(sc-bb)}} + \sum_{j \neq i} [(1/2) q_t \, \phi^{\text{bb}, i, j} - \Delta G_{\text{1-body scrn(sc-bb)}}] \quad (12)$$

where the sums are over all side chains $j \neq i$, i is the side chain of interest, each u is an atom in side chain i, q_u is the partial atomic charge of u, and ϕ is the potential at the location of u. First, the one-body energies were calculated as described previously. Next, two-body corrections were calculated using the atoms for the backbone, the side chain of interest, and one "perturbing" side chain, j, to define the dielectric boundary. Two body energies are calculated using each residue other than the side chain of interest as the perturbing residue. Pairwise contributions were calculated by adding the one-body energy to the sum of the two-body correction terms. For two-body calculations using only pairs that are close in space, the distance between side chains i and j is defined as the minimum distance between an atom with non-zero partial atomic charge on side chain i and any atom on side chain j.

Acknowledgements

This work was supported by the Howard Hughes Medical Institute (S. L. M. and B. H.), the Parsons Foundation, an IBM Shared Universities Research Grant (S. L. M.), a National Institutes of Health training grant, and the Caltech Initiative in Computational Molecular Biology program, awarded by the Burroughs Wellcome Fund (S. A. M.).

References

1. Dahiyat, B. I., Gordon, D. B. and Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci.*, **6**, 1333-1337.
2. 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.
3. Koehl, P. and Levitt, M. (1999). De novo protein design. I. In search of stability and specificity. *J. Mol. Biol.*, **293**, 1161-1181.
4. Werniach, L., Hery, S. and Wodak, S. J. (2000). Automatic protein design with all atom force-fields by exact and heuristic optimization. *J. Mol. Biol.*, **301**, 713-736.
5. Marshall, S. A., Morgan, C. S. and Mayo, S. L. (2001). Electrostatic interactions significantly affect the stability of designed proteins. *J. Mol. Biol.*
6. Levy, R. M. and Gallicchio, E. (1998). Computer simulations with explicit solvent: Recent progress in the thermodynamic decomposition of free energies and in modeling electrostatic effects. *Annu. Rev. Phys. Chem.*, **49**, 531-567.
7. Papazyan, A. and Warshel, A. (1997). Continuum and dipole-lattice models of solvation. *J. Phys. Chem.*, *101*, 11254-11264.
8. Gilson, M. K., Sharp, K. A. and Honig, B. H. (1987). Calculating the electrostatic potential of molecules in solution: method and error assessment. *J. Comput. Chem.*, **9**, 327-335.
9. Honig, B. and Nicholls, A. (1995). Classical electrostatics in biology and chemistry. *Science*, **268**, 1144-1149.
10. Still, W. C., Tempczyk, A., Hawley, R. C. and Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, **112**, 6127-6129.
11. Dominy, B. N. and Brooks, C. L., III. (1999). Development of a generalized Born model parametrization for proteins and nucleic acids. *J. Phys. Chem. B*, **103**, 3675-3773.

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

13. Desmet, J., De Maeyer, M., Hazes, B. and Lasters, I. (1994) The dead-end elimination theorem: a new approach to the side-chain packing problem. In *The protein folding problem and tertiary structure prediction* (K. Merz, Jr and S. Le Grand, ed) 307-337, Birkhauser, Boston.

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

15. Qiu, D., Shenkin, P. S., Hollinger, F. P. and Still, W. C. (1997). The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate born radii. *J. Phys. Chem. A*, **101**, 3005-3014.

16. Havranek, J. J. and Harbury, P. B. (1999). Tanford-Kirkwood electrostatics for protein modeling. *Proc. Natl. Acad. Sci.*, **96**, 11145-11150.

17. Scarsi, M. and Caflisch, A. (1999). Comment on the validation of continuum electrostatics models. *J. Comput. Chem.*, **20**, 1533-1536.

18. Street, A. G. and Mayo, S. L. (1998). Pairwise calculation of protein solvent accessible surface areas. *Fold. Des.*, **3**, 253-258.

19. Marshall, S. A. and Mayo, S. L. (2001). Achieving stability and conformational specificity in designed proteins via binary patterning. *J. Mol. Biol.*, **305**, 619-631.

20. Rocchia, W., Alexov, E. and Honig, B. (2001). Extending the applicability of the nonlinear Poisson-Boltzmann equation: Multiple dielectric constants and multivalent ions. *J. Phys. Chem.*, **105**, 6507-6514.

21. Vijayakumar, M. and Zhou, H.-X. (2001). Salt bridges stabilize the folded structure of barnase. *J. Phys. Chem.*, **105**, 7334-7340.

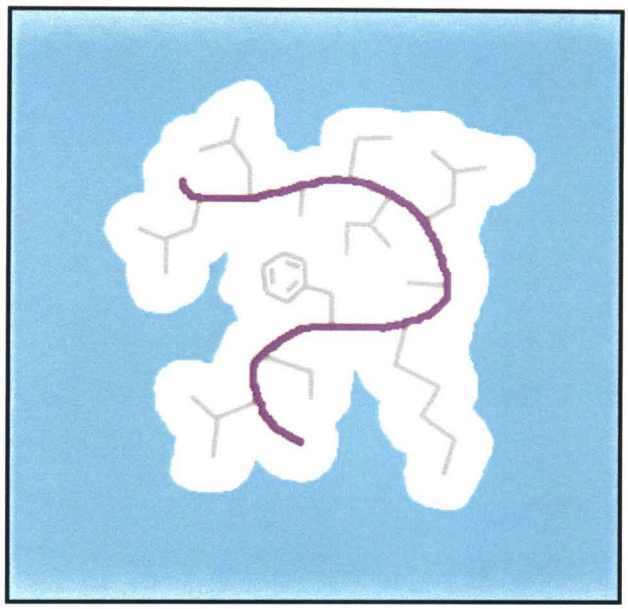
22. Sitkoff, D., Sharp, K. and Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.*, **98**, 1978-1988.

Table VII-1: Mean energies and errors of the electrostatic models

	absolute mean energy (kcal mol^{-1})	RMSD (kcal mol^{-1})	R^2
I. Backbone desolvation energy			
exact DelPhi	9.656	-	-
1-body	8.763	1.012	0.998
II. Side chain desolvation energy			
exact DelPhi	0.205	-	-
1-body	0.079	0.203	0.516
2-body, all pairs	0.197	0.018	0.997
2-body, 6 Å cutoff	0.195	0.022	0.996
2-body, 4 Å cutoff	0.186	0.035	0.989
III. Side chain - backbone screened Coulombic energy			
exact DelPhi	0.413	-	-
1-body	0.360	0.113	0.989
2-body, all pairs	0.404	0.026	0999
2-body, 6 Å cutoff	0.401	0.033	0.998
2-body, 4 Å cutoff	0.393	0.046	0.997
IV. Side chain - side chain screened Coulombic energy			
exact DelPhi	0.044	-	-
2-body, all pairs	0.045	0.012	0.993
2-body, 6 Å cutoff	0.044	0.012	0.994
2-body, 4 Å cutoff	0.044	0.013	0.992

Figure VII-1. Models used to calculate exact DelPhi backbone desolvation energies. The areas drawn in white were assigned a dielectric constant of 4 (protein interior) and the blue areas have a dielectric constant of 80 (water) and a salt concentration of 50 mM. (a) Folded state backbone solvation. The backbone atoms, shown in purple, both "generate" and "feel" the electrostatic potential. The side chain atoms, shown in gray, are assigned partial atomic charges of 0. All side chain and backbone atoms are used to define the dielectric boundary. (b) Reference state backbone solvation. The backbone atoms, shown in purple, both "generate" and "feel" the electrostatic potential. The dielectric boundary is defined using the backbone atoms only.

(a)



(b)

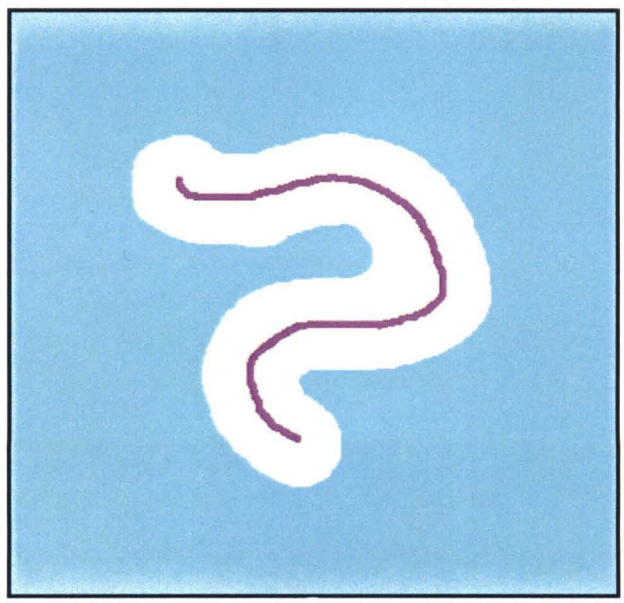
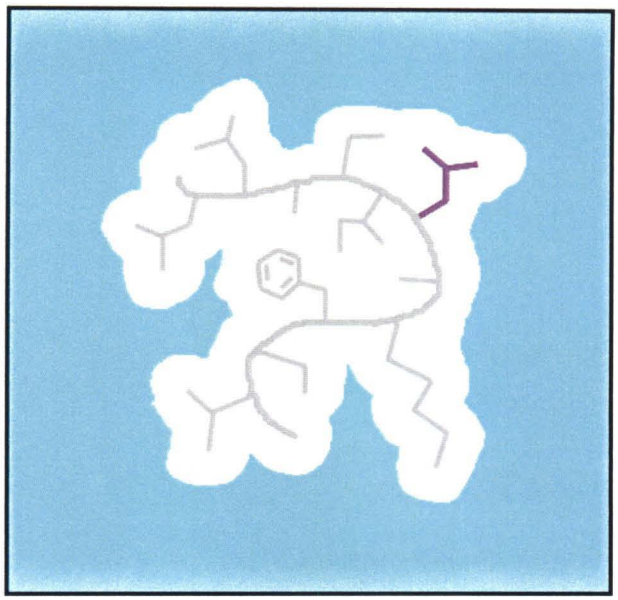


Figure VII-2. Models used to calculate exact DelPhi side chain desolvation energies. The areas drawn in white were assigned a dielectric constant of 4 (protein interior) and the blue areas have a dielectric constant of 80 (water) and a salt concentration of 50 mM. (a) Folded state side chain solvation. Atoms in side chain *i*, shown in purple, both "generate" and "feel" the electrostatic potential. Side chain and backbone atoms shown in gray are assigned a partial atomic charge of 0. All side chain and backbone atoms are used to define the dielectric boundary. (b) Unfolded state side chain solvation. Atoms in side chain *i*, shown in purple, both "generate" and "feel" the electrostatic potential. Atoms shown in gray are assigned a partial atomic charge of 0. The dielectric boundary is defined using the atoms in side chain *i* and the local backbone only.

(a)



(b)

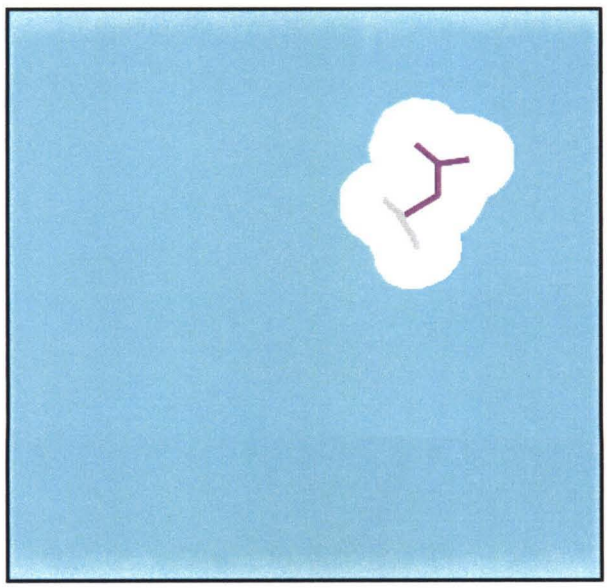


Figure VII-3. Models used to calculate exact DelPhi side chain - backbone screening energies. The areas drawn in white were assigned a dielectric constant of 4 (protein interior) and the blue areas have a dielectric constant of 80 (water) and a salt concentration of 50 mM. Atoms in side chain *i*, shown in orange, "generate" the electrostatic potential and backbone atoms, shown in green, "feel" the electrostatic potential. Side chain atoms shown in gray are assigned a partial atomic charge of 0. All side chain and backbone atoms are used to define the dielectric boundary. Sscreening energies are added to the Coulombic energies to obtain screened Coulombic energies.

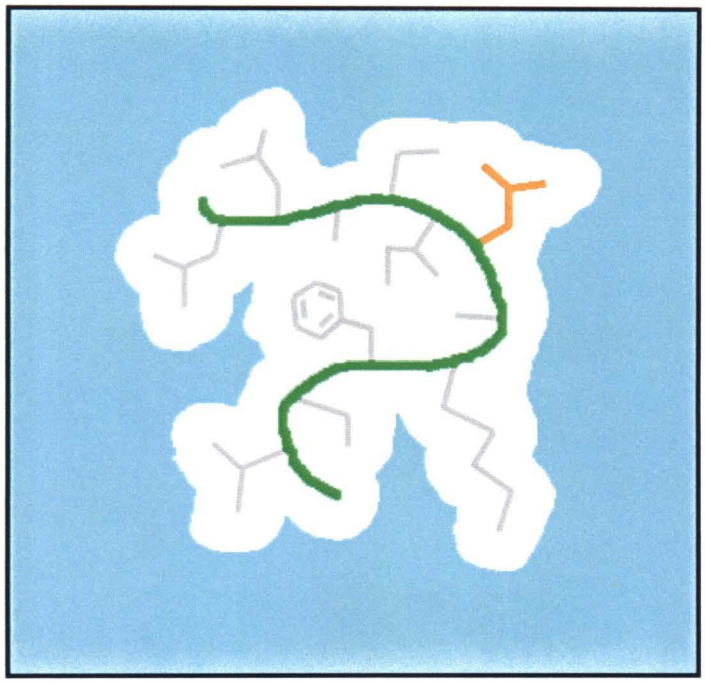


Figure VII-4. Models used to calculate exact DelPhi side chain - side chain screening energies. The areas drawn in white were assigned a dielectric constant of 4 (protein interior) and the blue areas have a dielectric constant of 80 (water) and a salt concentration of 50 mM. Atoms in side chain *i*, shown in orange, "generate" the electrostatic potential and atoms in side chain *j*, shown in green, "feel" the electrostatic potential. Side chain and backbone atoms shown in gray are assigned a partial atomic charge of 0. All side chain and backbone atoms are used to define the dielectric boundary. The screening energies were added to the Coulombic energies to obtain screened Coulombic energies.

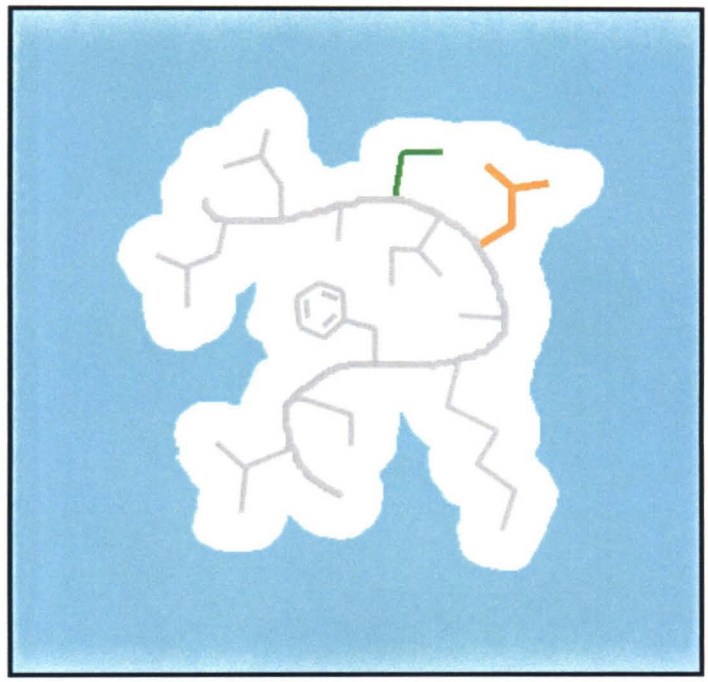


Figure VII-5. Models used to calculate one-body FDPB backbone desolvation energies. The areas drawn in white were assigned a dielectric constant of 4 (protein interior) and the blue areas have a dielectric constant of 80 (water) and a salt concentration of 50 mM. (a) Folded state backbone solvation. The backbone atoms, shown in purple, both "generate" and "feel" the electrostatic potential. Atoms in side chain i, shown in gray, are assigned a partial atomic charge of 0. Backbone and side chain i atoms only are used to define the dielectric boundary. (b) Reference state backbone solvation. The backbone atoms, shown in purple, both "generate" and "feel" the electrostatic potential. The dielectric boundary is defined using the backbone atoms only.

(a)

(b)

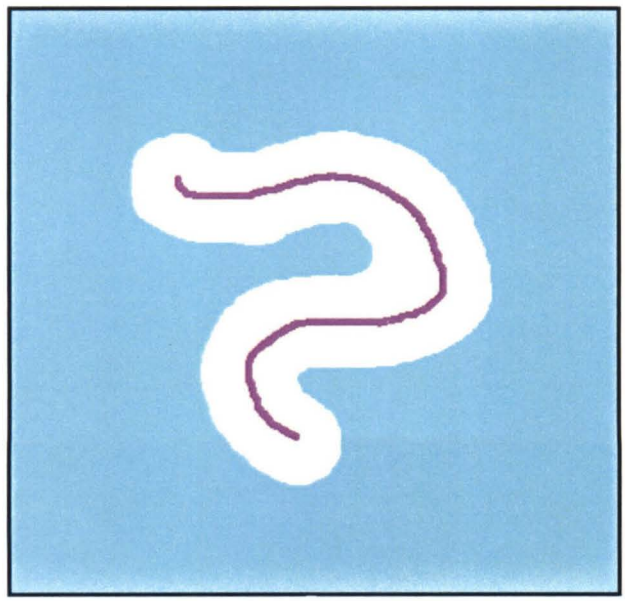
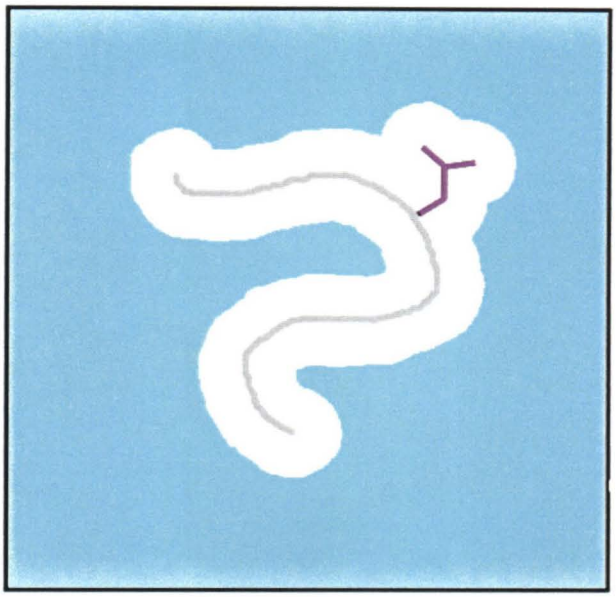


Figure VII-6. Models used to calculate one-body FDPB side chain desolvation energies. The areas drawn in white were assigned a dielectric constant of 4 (protein interior) and the blue areas have a dielectric constant of 80 (water) and a salt concentration of 50 mM. (a) Folded state side chain solvation. Atoms in side chain *i*, shown in purple, both "generate" and "feel" the electrostatic potential. Backbone atoms shown in gray are assigned a partial atomic charge of 0. Backbone and side chain *i* atoms only are used to define the dielectric boundary. (b) Unfolded state side chain solvation. Atoms in side chain *i*, shown in purple, both "generate" and "feel" the electrostatic potential. Atoms shown in gray are assigned a partial atomic charge of 0. The dielectric boundary is defined using the atoms in side chain i and the local backbone only.

(a)



(b)

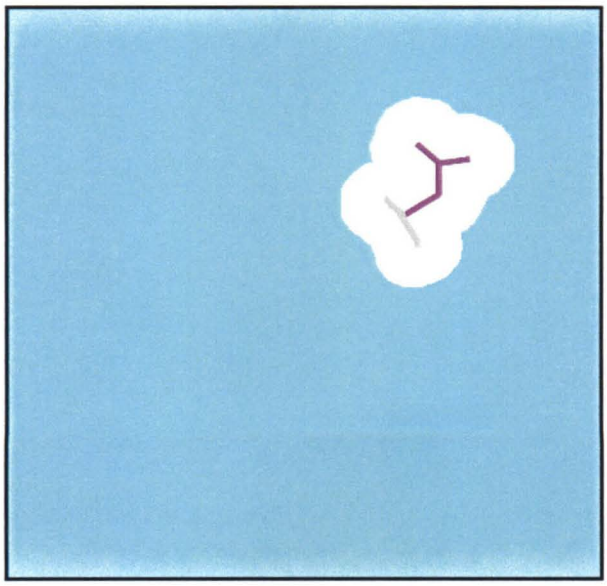


Figure VII-7. Models used to calculate one-body FDPB side chain - backbone screening energies. The areas drawn in white were assigned a dielectric constant of 4 (protein interior) and the blue areas have a dielectric constant of 80 (water) and a salt concentration of 50 mM. Atoms in side chain *i*, shown in orange, "generate" the electrostatic potential and backbone atoms, shown in green, "feel" the electrostatic potential. Backbone and side chain *i* atoms only are used to define the dielectric boundary. The screening energies were added to the Coulombic energies to obtain screened Coulombic energies.

Figure VII-8. Comparison of the backbone desolvation energies calculated using "exact" DelPhi *versus* the one-body decomposition of DelPhi.

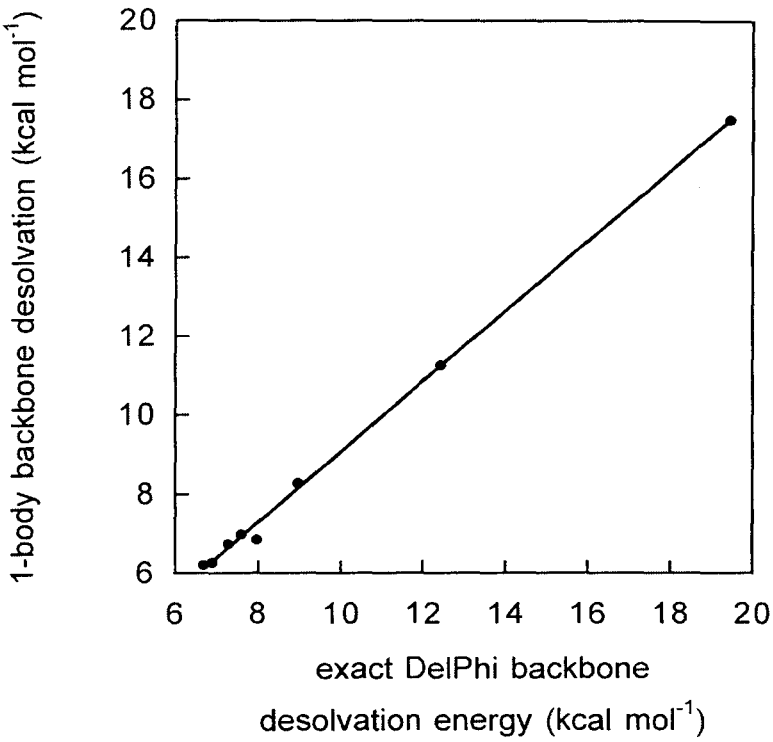

1-body backbone desolvation (kcal mol-1)
20
18
16
14
12
10
8
6
6
8
10
12
14
16
18
20
exact DelPhi backbone
desolvation energy (kcal mol-1)

Figure VII-9. Comparison of the side chain desolvation energies calculated using "exact" DelPhi *versus* the one-body decomposition of DelPhi.

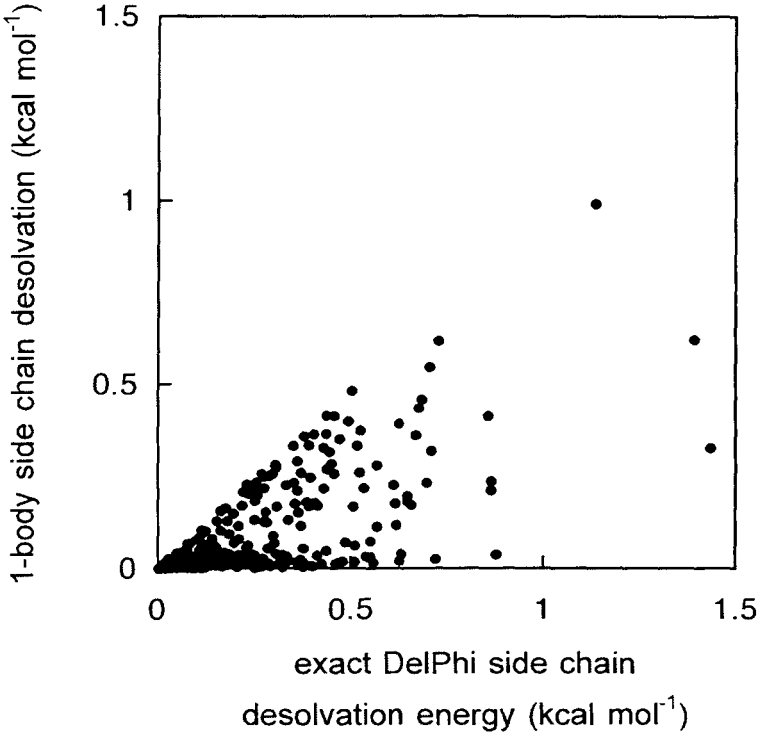

1.5
1
0.5
0
0
0.5
1
1.5
1-body side chain desolvation (kcal mol-1)
exact DelPhi side chain
desolvation energy (kcal mol-1)

Figure VII-10. Comparison of the side chain - backbone screened Coulombic energies calculated using "exact" DelPhi *versus* the one-body decomposition of DelPhi.

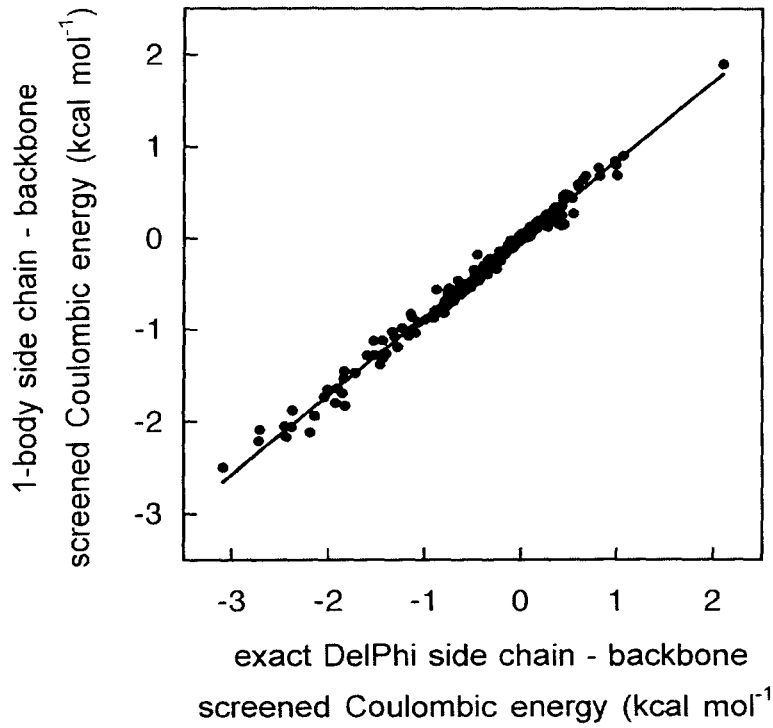
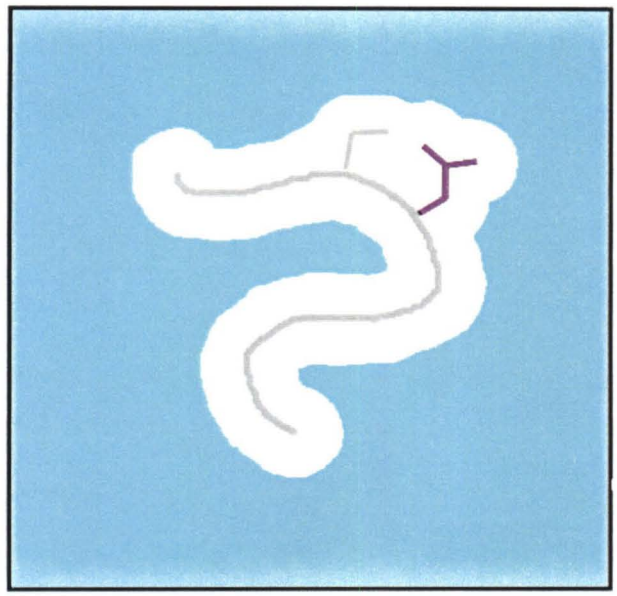

1-body side chain - backbone
screened Coulombic energy (kcal mol-1)
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
exact DelPhi side chain - backbone
screened Coulombic energy (kcal mol-1)

Figure VII-11. Models used to calculate two-body FDPB side chain desolvation energies. The areas drawn in white were assigned a dielectric constant of 4 (protein interior) and the blue areas have a dielectric constant of 80 (water) and a salt concentration of 50 mM. (a) Folded state side chain solvation. Atoms in side chain *i*, shown in purple, both "generate" and "feel" the electrostatic potential. Side chain and backbone atoms shown in gray are assigned a partial atomic charge of 0. Backbone, side chain *i*, and side chain *j* atoms only are used to define the dielectric boundary. (b) Unfolded state side chain solvation. Atoms in side chain *i*, shown in purple, both "generate" and "feel" the electrostatic potential. Atoms shown in gray are assigned a partial atomic charge of 0. The dielectric boundary is defined using the atoms in side chain *i* and the local backbone only.

(a)



(b)

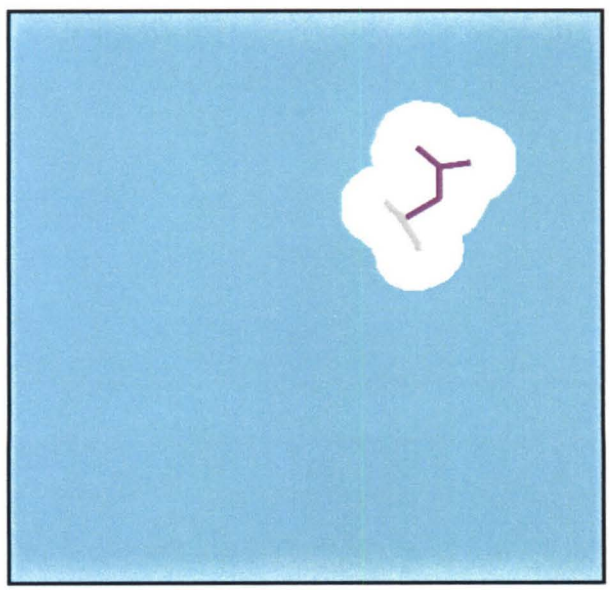


Figure VII-12. Models used to calculate two-body FDPB side chain - backbone screening energies. The areas drawn in white were assigned a dielectric constant of 4 (protein interior) and the blue areas have a dielectric constant of 80 (water) and a salt concentration of 50 mM. Atoms in side chain *i*, shown in orange, "generate" the electrostatic potential and backbone atoms, shown in green, "feel" the electrostatic potential. Side chain atoms shown in gray are assigned a partial atomic charge of 0. Backbone, side chain *i*, and side chain *j* atoms only are used to define the dielectric boundary.

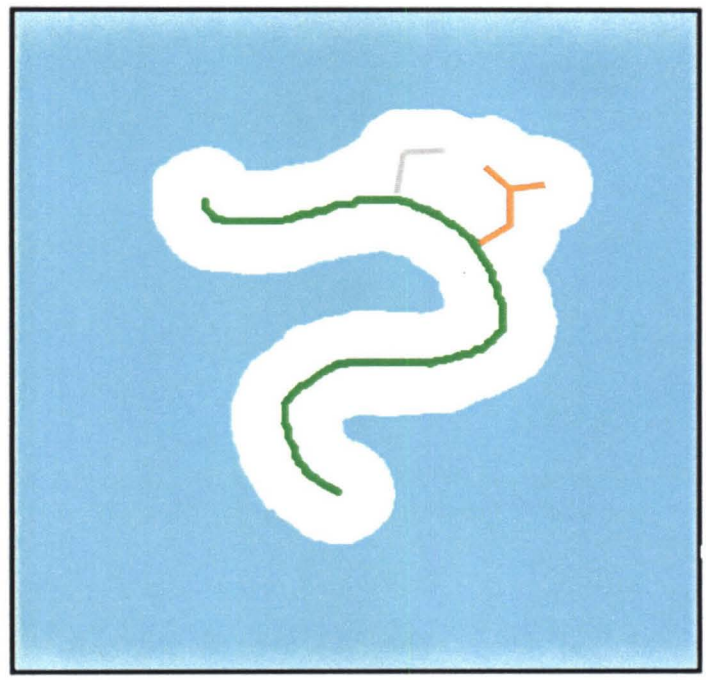


Figure VII-13. Models used to calculate two-body FDPB side chain - side chain screening energies. The areas drawn in white were assigned a dielectric constant of 4 (protein interior) and the blue areas have a dielectric constant of 80 (water) and a salt concentration of 50 mM. Atoms in side chain *i*, shown in orange, "generate" the electrostatic potential and atoms in side chain *j*, shown in green, "feel" the electrostatic potential. Backbone atoms shown in gray are assigned a partial atomic charge of 0. Backbone, side chain *i*, and side chain *j* atoms only are used to define the dielectric boundary. The screening energies were added to the Coulombic energies to obtain screened Coulombic energies. The screening energies were added to the Coulombic energies to obtain screened Coulombic energies.

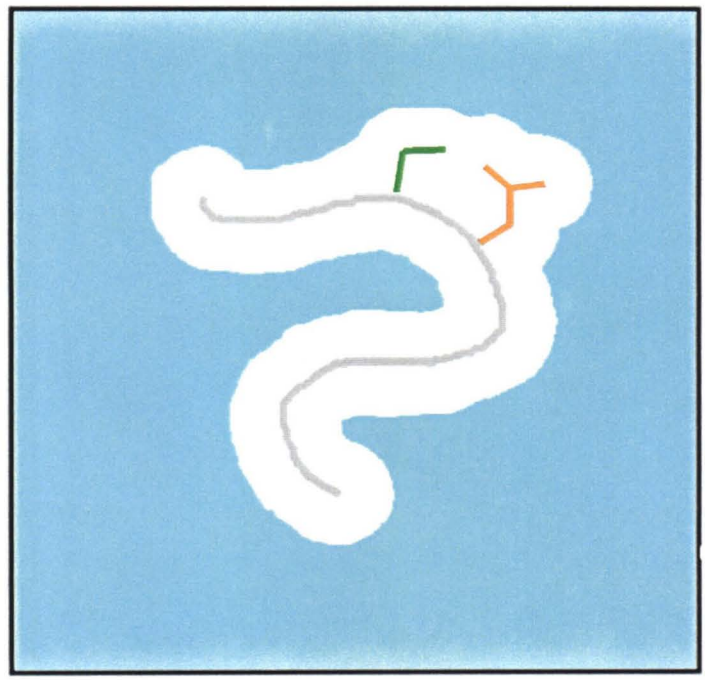


Figure VII-14. Comparison of the side chain desolvation energies calculated using "exact" DelPhi *versus* the two-body decomposition of DelPhi.

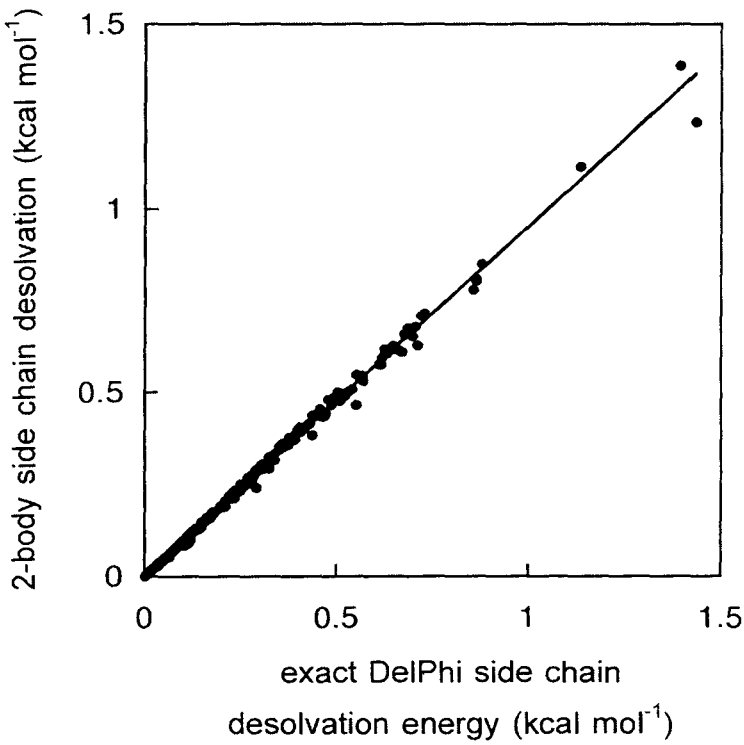

1.5
1
0.5
0
0
0.5
1
1.5
2-body side chain desolvation (kcal mol-1)
exact DelPhi side chain
desolvation energy (kcal mol-1)

Figure VII-15. Comparison of the side chain - backbone screened Coulombic energies calculated using "exact" DelPhi *versus* the two-body decomposition of DelPhi.

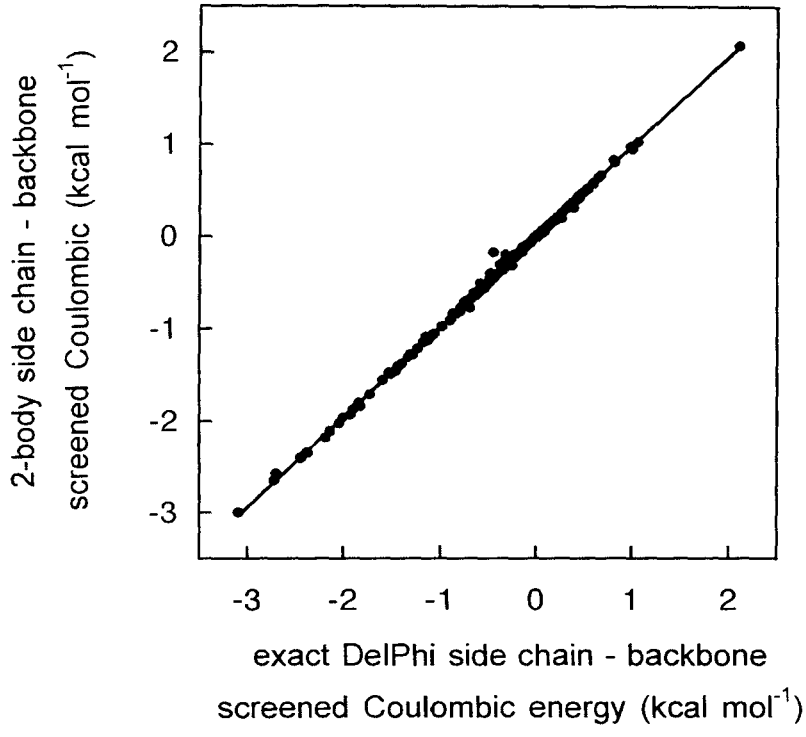

2-body side chain - backbone
screened Coulombic (kcal mol-1)
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
exact DelPhi side chain - backbone
screened Coulombic energy (kcal mol-1)

Figure VII-16. Comparison of the side chain - side chain screened Coulombic energies calculated using "exact" DelPhi *versus* the two-body decomposition of DelPhi.

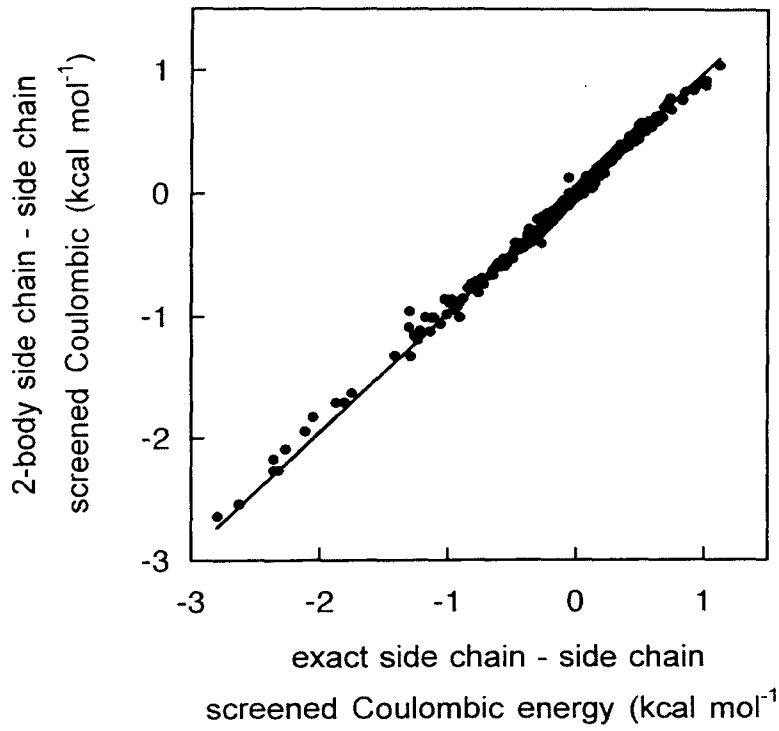
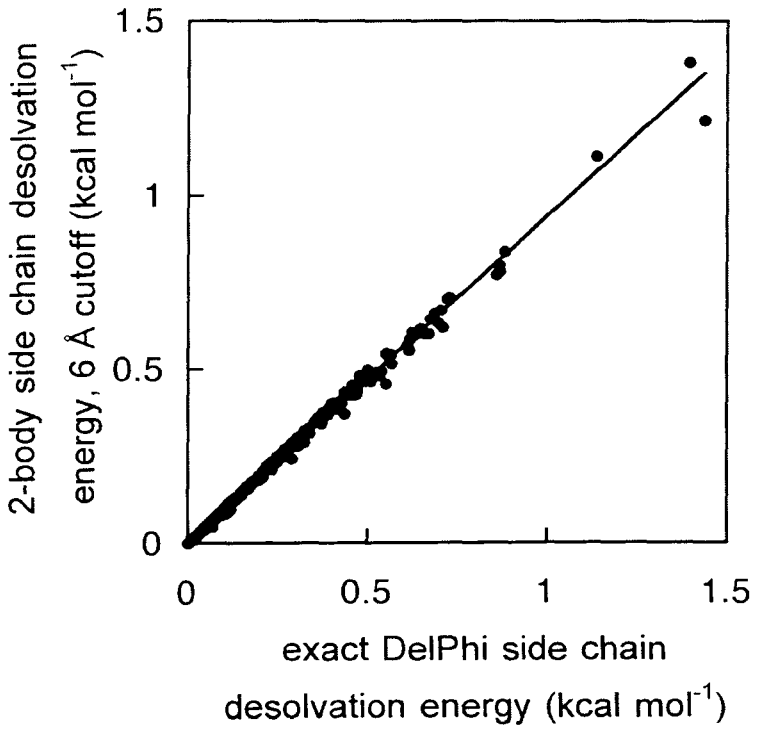

2-body side chain - side chain
screened Coulombic (kcal mol-1)
1
0
-1
-2
-3
-3
-2
-1
0
1
exact side chain - side chain
screened Coulombic energy (kcal mol-1)

Figure VII-17. Accuracy of the two-body method for calculating side chain desolvation energies using (a) only pairs separated by less than 6 Å, and (b) only pairs separated by less than 4 Å, determined by comparing approximate energies to exact DelPhi energies.

(a)


2-body side chain desolvation energy, 6 Å cutoff (kcal mol-1)
1.5
1
0.5
0
0
0.5
1
1.5
exact DelPhi side chain desolvation energy (kcal mol-1)

(b)

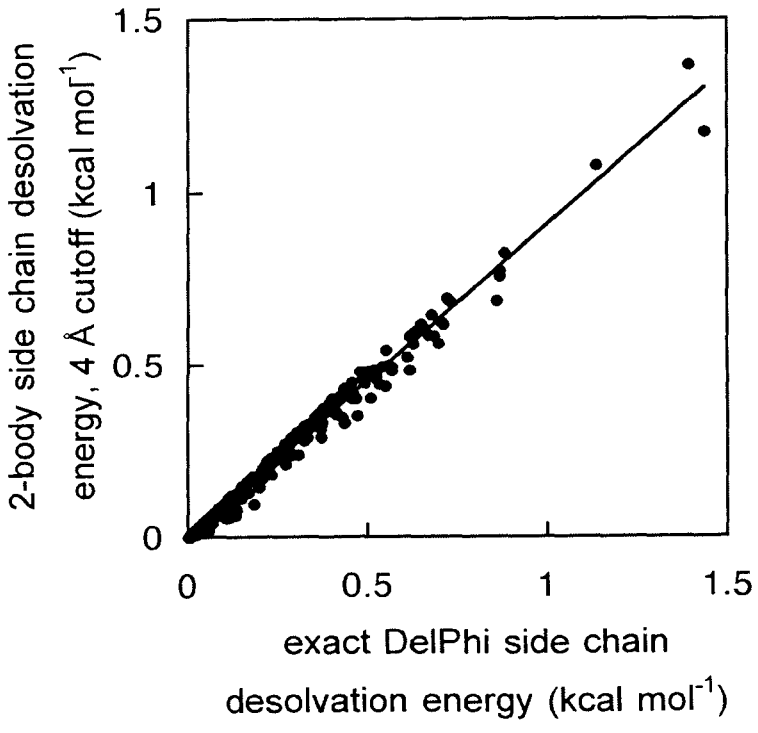
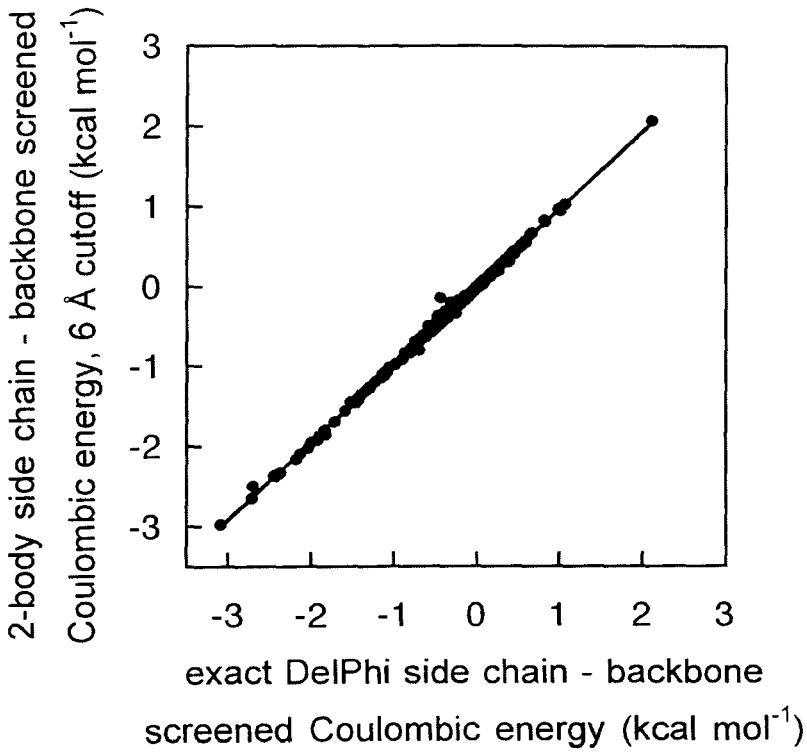

2-body side chain desolvation energy, 4 Å cutoff (kcal mol-1)
1.5
1
0.5
0
0
0.5
1
1.5
exact DelPhi side chain desolvation energy (kcal mol-1)

Figure VII-18. Accuracy of the two-body method for calculating side chain - backbone screened Coulombic energies using (a) only pairs separated by less than 6 Å, and (b) only pairs separated by less than 4 Å, determined by comparing approximate energies to exact DelPhi energies.

(a)


2-body side chain - backbone screened Coulombic energy, 6 Å cutoff (kcal mol-1)
3
2
1
0
-1
-2
-3
-3 -2 -1 0 1 2 3
exact DelPhi side chain - backbone screened Coulombic energy (kcal mol-1)

(b)

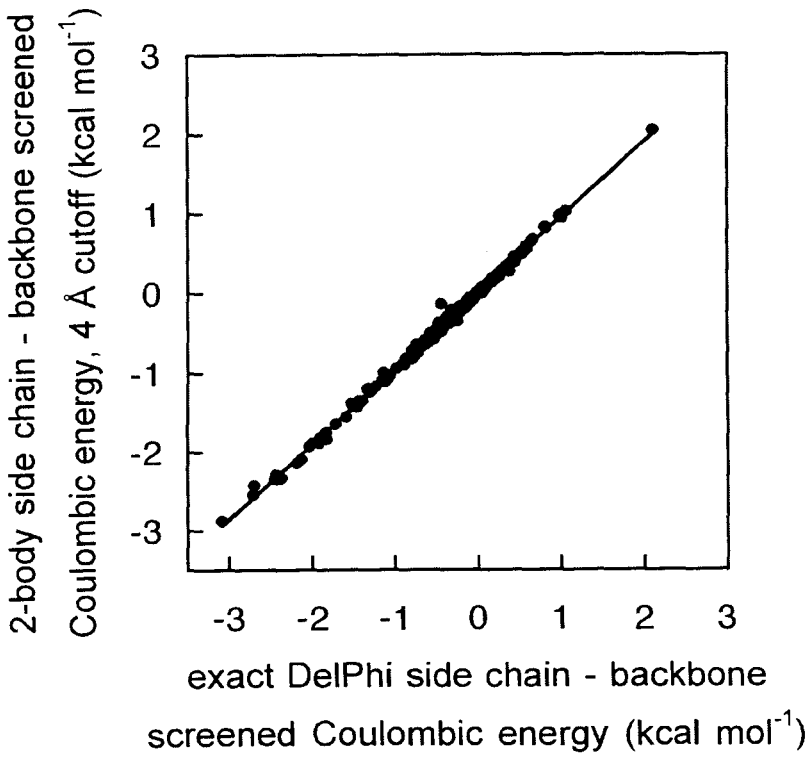
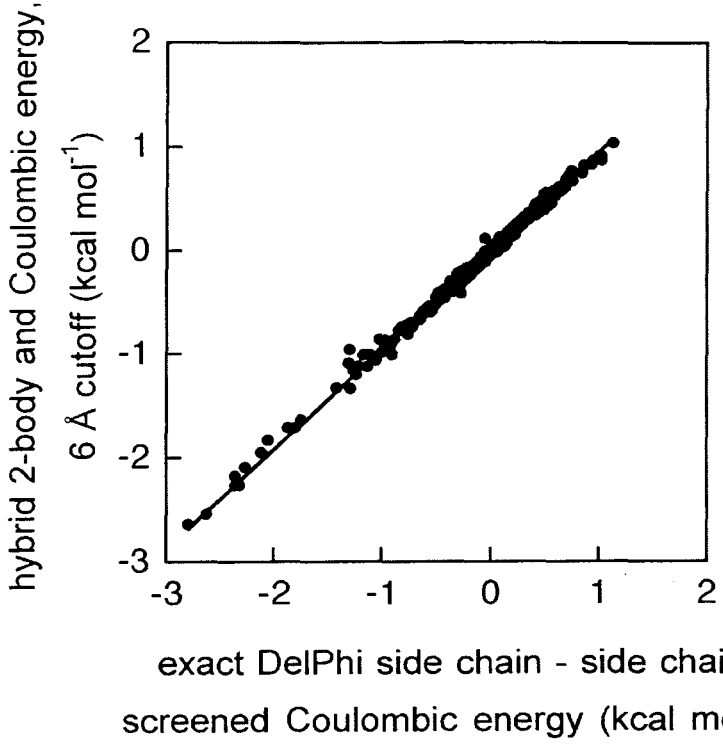
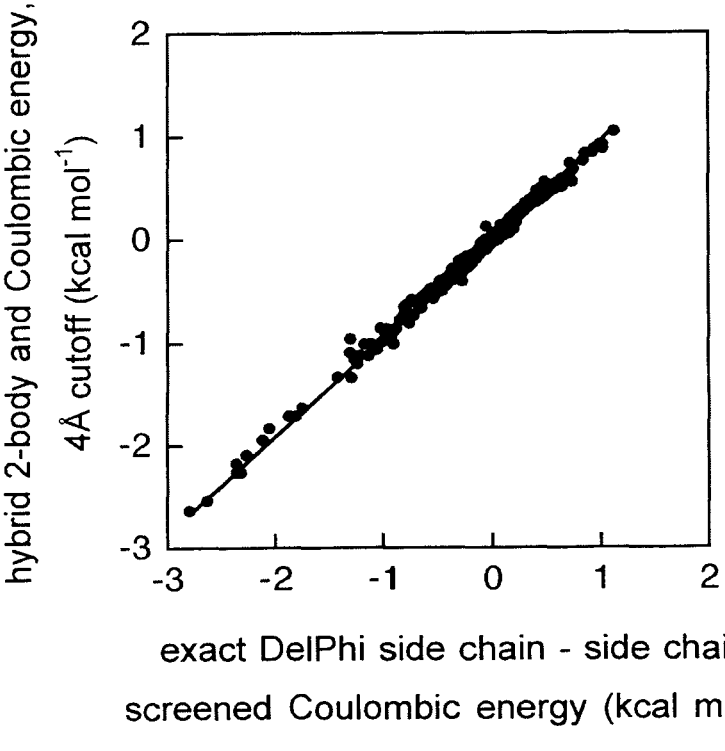

2-body side chain - backbone screened Coulombic energy, 4 Å cutoff (kcal mol-1)
3
2
1
0
-1
-2
-3
-3 -2 -1 0 1 2 3
exact DelPhi side chain - backbone screened Coulombic energy (kcal mol-1)

Figure VII-19. Accuracy of the two-body method for calculating side chain - side chain screened Coulombic energies using (a) only pairs separated by less than 6 Å, and (b) only pairs separated by less than 4 Å, determined by comparing approximate energies to exact DelPhi energies.

(a)


hybrid 2-body and Coulombic energy, 6 Å cutoff (kcal mol-1)
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
exact DelPhi side chain - side chain screened Coulombic energy (kcal mol-1)

(b)


hybrid 2-body and Coulombic energy, 4Å cutoff (kcal mol-1)
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
exact DelPhi side chain - side chain screened Coulombic energy (kcal mol-1)

Appendix A:

Core and boundary design of a SH3 domain

Introduction

Computational protein design methods have typically been applied to proteins with mixed or helical secondary structure, as design of beta sheet proteins has proved more challenging. Since protein design methods continually advance, we wanted to determine whether design of beta sheet proteins had become feasible yet. A SH3 domain was selected as the target fold, as it is one of the smallest all-beta protein structures. SH3 domains function as adapter domains, typically mediating interactions between tyrosine kinases and their substrates in a variety of signal transduction pathways[1]. SH3 domains are composed of two antiparallel beta sheets that are aligned perpendicular to each other, three loop regions and a small 3_{10} helix. Residues in the RT and n-Src loops and the 3_{10} helix form a binding site for proline-rich helices[2].

Core Design

The c-crk SH3 domain contains 15 core positions, including one aspartate and two glycines, as shown in Figures A-1 and A-2. The core residue Asp 17 appears to stabilize a turn by forming multiple hydrogen bonds to backbone amides. Glycine and buried polar residues have not yet been incorporated into the protein design algorithm, so these three positions were held fixed in the design calculations. The protein design algorithm selected a point mutation, W37V, as the lowest energy sequence for the core of 1cka. Circular dichroism wavelength scans of the wild type SH3 domain, wt, and the core redesign, cr, differ significantly, as shown in Figure A-3. CD signals from small all-beta proteins are significantly more variable than CD signals from helical proteins. In addition, aromatic residues can contribute significantly to CD signal[3], so the discrepancy may result from

contributions of Trp 37 versus Val 37. Variant cr is destabilized by 13 °C relative to the wild type protein, as shown in Figure A-4.

Boundary Design

The c-crk SH3 domain contains 14 boundary residues, as shown in Figures A-1 and A-2. The boundary sequence selected using ORBIT was a 12-fold mutant from wild type (E2Q, L7I, D24W, R27E, R29L, E33Q, E40K, K45Q, M48Y, V51E, Y53E, Y57E). The CD wavelength scan of variant b1, shown in Figure A-5, suggests that the protein is not fully folded at 1 °C. Variant, br1 is significantly destabilized relative to wild type, as shown in Figures A-6 and A-7. Its thermal unfolding transition is only weakly cooperative and lacks a well defined pretransition. The thermal denaturation temperature of br1 is approximately 14 °C, which is 35 °C less than wild type.

The predicted structures were examined manually to identify the mutations that caused this decrease in stability. Some of the mutated residues appeared to made potentially destabilizing interactions with the rest of the protein while other mutations seemed benign. To better understand the source of the decreased stability of br1, a second protein containing five seemingly harmless mutations (L7I, R29L, K45Q, M48Y, Y57E) from br1 in a wild type background was characterized. Variant br2 has a well-behaved thermal unfolding transition, shown in Figure A-6 but is still significantly less stable than wild type: its melting temperature is only 24°C.

One possible explanation for the low thermal stability of br1 and br2 is the choice of solvation potential parameters. Arthur Street, another graduate student in the Mayo group, developed a new pairwise method to calculate surface areas and two new sets of atomic solvation parameters[4]. The atomic solvation parameters differ depending on whether a polar area burial penalty or a polar hydrogen burial penalty is used to model the desolvation

of polar groups. The initial boundary calculations were performed using the polar hydrogen burial penalty; I repeated the br2 calculation using the parameter set optimized for inclusion of the polar area burial penalty. The predicted sequence, br3, is a four-fold mutation from wild type (L7I, R29L, K45R, Y57K). CD thermal denaturation experiments, shown in Figure A-6, demonstrate that the thermal denaturation temperature of br3 is 55 °C, which is 7 °C higher than the wild type thermal denaturation temperature.

Two additional boundary variants were then generated using the parameter set that produced br3. The boundary residues form two groups that interact minimally, so sequences were calculated separately for each group. The first protein, br4a, is a five-fold mutant (E2L, R27L, R29L, E40K, K45R) from wild type and the second, br4b, is a four-fold mutant (L7I, D24K, E33Q, Y53R). Thermal denaturation experiments show that br4a undergoes an irreversible transition, likely aggregation, at elevated temperatures, as shown in Figure A-8. Variant br4b lacks a cooperative unfolding transition, as shown in Figure A-7. CD wavelength scans, shown in Figure A-5, indicate that br4b has some secondary structure at 1°C, but it is likely that the protein is never properly folded.

The Importance of Rotamer Libraries

The instability of the designed core variant is likely due to limitations in the rotamer library used in the design calculations. Valine, the residue selected in the design calculation, is much smaller than the wild type tryptophan, so the core is probably underpacked in cr. The wild type residue has an unusual (-12°, or almost eclipsed) $\chi 2$ angle. Strained rotamers are not included in the rotamer libraries used for design calculations, so the wild type rotamer was not considered. Placing a side chain with a low energy $\chi 2$ conformation at position 37 results in van der Waals clashes with other core residues. As a result, valine was selected, as it is the largest side chain with no $\chi 2$ angle. The results of the core redesign indicate that a

well-packed c-crk SH3 domain core can not be made using a fixed backbone structure and canonical low energy rotamers. Future study could incorporate backbone flexibility or consider rotamers with suboptimal covalent geometry in conjunction with a self energy term for each rotamer.

The Importance of Binary Patterning

The choice of solvation parameters strongly influenced the number of polar and nonpolar residues that were selected at boundary positions. Proteins with too many nonpolar residues exposed on their surface, such as br4b, are prone to aggregation. Proteins that bury large amounts of polar surface area, such as br1, are very unstable. At the end of this project, I could not identify a set of solvation parameters that would reliably select sequences with a proper balance of hydrophobic and polar interactions. The binary patterning project, described in Chapter 2, was initiated to develop a method to select a proper balance of hydrophobic and polar residues prior to sequence selection.

Materials and Methods

Modeling. Coordinates for the wild type c-crk SH3 domain were obtained from PDB entry 1cka, a crystal structure solved to 1.5 Å resolution[2]. Explicit hydrogens were added using the program Biograf (Molecular Simulations, Inc., San Diego) and the resulting structure was minimized for 50 steps using the Dreiding force field[5]. The program Resclass was used to classify positions as core, boundary, or surface[6].

Sequence Selection: Core Design. Amino acid identities and conformations were optimized at the following core positions: 4, 6, 10, 18, 20, 26, 37, 39, 49, and 54. Positions 12 and 47, although classified as core, were fixed to wild type Gly. Position 17 was excluded, as it

makes multiple hydrogen bonds to the backbone. Ala, Val, Leu, Ile, Phe, Tyr, and Trp were considered at each variable core position. Variable side chains were represented as discrete rotamers from the Dunbrak and Karplus backbone dependent rotamer library[7]. Rotamers were also included at ±1 standard deviation about $\chi 1$ for aliphatic residues and at ±1 standard deviation about $\chi 1$ and $\chi 2$ aromatic residues. Side chain - backbone and side chain - side chain energies were calculated using a force field containing van der Waals solvation terms in conjunction with the following parameters: van der Waals scale factor[8]: 0.90, hydrophobic burial benefit[9]: 0.0232 kcal mol^{-1} $Å^{-2}$, nonpolar exposure factor: 1.0, polar hydrogen burial penalty[10]: 2.0 kcal per buried polar hydrogen. The optimal rotameric sequence was determined using the dead-end elimination (DEE) theorem[11-13].

Sequence Selection: Boundary Design. In the br1 calculation, amino acid identities and conformations were optimized at the following positions: 2, 7, 17, 24, 27, 28, 29, 33, 40, 41, 45, 48, 51, 53, and 57. All residues other than Cys, Pro, Gly, and Met were considered at each variable position. Variable side chains were represented as discrete rotamers from the Dunbrak and Karplus backbone dependent rotamer library[7]. Rotamers were also included at ±1 standard deviation about $\chi 1$ for aliphatic residues and at ±1 standard deviation about $\chi 1$ and $\chi 2$ aromatic residues. Side chain - backbone and side chain - side chain energies were calculated using a force field containing van der Waals, solvation, hydrogen bond, and electrostatic terms in conjunction with the following parameters: van der Waals scale factor[8]: 0.90, dielectric[10]: $40r$, solvation parameters[4] including hydrophobic burial benefit: 0.048 kcal mol-1 $Å^{-2}$, nonpolar exposure factor: 1.6, and polar hydrogen burial penalty[10]: 2.0 kcal mol^{-1} per buried polar hydrogen (including template hydrogens). The optimal rotameric sequence was determined as for the core variant.

Variant br2 contained five mutations predicted in the br1 calculation that appeared, by manual inspection, to be innocuous: L7I, R29L, K45Q, M48Y, and Y57E.

In the br3 calculation, amino acid identities and conformations were optimized at positions 7, 29, 45, 48, and 57. The conformations of the wild type amino acids were optimized at the remaining boundary positions. All residues other than Cys, Pro, Met and Gly were considered at variable positions 7, 29, 45, and 57; all residues other than Cys, Pro and Gly were considered at variable position 48 which is Met in the wild type protein. Variable side chains were represented as discrete rotamers from the Dunbrak and Karplus backbone dependent rotamer library[7]. Rotamers were also included at ±1 standard deviation about $\chi 1$ and $\chi 2$ for all residues. The force field parameters were as for the b1 calculation, with the following modifications: hydrophobic burial benefit: 0.026 kcal mol-1 $Å^{-2}$, polar area burial penalty: 0.1 kcal mol^{-1} $Å^{-2}$, and the polar hydrogen burial penalty was not used[4]. The optimal rotameric sequence was determined as for the core variant.

The br4a calculation optimized amino acid identities and conformations at positions 2, 27, 28, 29, 40, 41, 45, and 57. The conformations of the wild type amino acids were optimized at the remaining positions. All residues other than Cys, Pro, Met, and Gly were considered at each variable position. The rotamer library, force field, and optimization algorithm were as for the br3 calculation.

The br4b calculation optimized amino acid identities and conformations at positions 7, 24, 33, 48, 51, and 53. The conformations of the wild type amino acids were optimized at the remaining positions. All residues other than Cys, Pro, Met, and Gly were considered at variable positions 7, 24, 33, 51, and 53; all residues other than Cys, Pro, and Gly were considered at residue 48, which is Met in the wild type protein. The rotamer library, force field, and optimization algorithm were as for the br3 calculation.

Peptide Synthesis and Purification. 57 residue SH3 domains were synthesized on an Applied Biosystems Model 433A peptide synthesizer using FMOC chemistry. Peptides were cleaved from the resin using a cocktail containing phenol, ethanedithiol, thioanisole, water, and trifluoroacetic acid, following the manufacturer's protocol. All peptides were purified by HPLC using a reverse-phase C8 prep column (Zorbax) and linear acetonitrile-water gradients containing 0.1% trifluoroacetic acid. Peptide masses were determined using MALDI-TOF or electrospray mass spectrometry and were found to be within one mass unit of expected values.

Circular Dichroism Studies. Circular dichroism data were obtained using an Aviv 62A DS spectropolarimeter equipped with a thermoelectric cell holder. Wavelength scans and thermal denaturation data were obtained from samples containing approximately 50 mM protein, 20 mM sodium chloride, and 20 mM sodium phosphate adjusted to pH 6.0 or 100 mM sodium phosphate adjusted to pH 6.0. Thermal denaturation data were acquired every degree from 2 °C to 99 °C using a 1.5 min equilibration time. Melting temperatures for well-behaved thermal denaturation transitions were determined by fitting to a two state transition as previously described[14]. An approximate thermal denaturation temperature for variant b1 was obtained using the method of John and Weeks[15]. All nonlinear regression calculations were performed using KaleidaGraph (Synergy Software). Thermal denaturation was followed by monitoring CD ellipticity at 222 nm.

References

1. Musacchio, A., Saraste, M. and Wilmanns, M. (1994). High resolution crystal structures of tyrosine kinase SH3 domains complexed with proline-rich peptides. *Nat. Struct. Biol.*, **1**, 546-551.
2. Wu, X., Knudsen, B., Feller, S. M., Zheng, J., Sali, A., Cowburn, D., Hanafusa, H. and Kuriyan, J. (1995). Structural basis for the specific interaction of the lysine-containing proline-rich peptides with the N-terminal SH3 domain of c-Crk. *Structure*, **3**, 215-226.
3. Woody, R. W. and Dunker, A. K. (1996) Aromatic and cystine side-chain circular dichroism in proteins. In *Circular dichroism and the conformational analysis of biomolecules* (G. D. Fasman, ed) Plenum Press, New York.
4. Street, A. G. and Mayo, S. L. (1998). Pairwise calculation of protein solvent accessible surface areas. *Fold. Des.*, **3**, 253-258.
5. Mayo, S. L., Olafson, B. D. and Goddard, W. A., III. (1990). Dreiding - a generic force-field for molecular simulations. *J. Phys. Chem.*, **94**, 8897-8909.
6. Dahiyat, B. I. and Mayo, S. L. (1997). *De novo* protein design: fully automated sequence selection. *Science*, **278**, 82-87.
7. Dunbrack, R. L. and Karplus, M. (1993). Backbone dependent rotamer library for proteins - an application to side-chain prediction. *J. Mol. Biol.*, **230**, 543-574.
8. Dahiyat, B. I. and Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci., USA*, **94**, 10172-10177.
9. Dahiyat, B. I. and Mayo, S. L. (1996). Protein design automation. *Protein Sci.*, **5**, 895-903.
10. Dahiyat, B. I., Gordon, D. B. and Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci.*, **6**, 1333-1337.

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

12. Desmet, J., De Maeyer, M., Hazes, B. and Lasters, I. (1994) The dead-end elimination theorem: a new approach to the side-chain packing problem. In *The protein folding problem and tertiary structure prediction* (K. Merz, Jr and S. Le Grand, ed) 307-337, Birkhauser, Boston.

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

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

15. John, D. M. and Weeks, K. M. (2000). van't Hoff enthalpies without baselines. *Protein Sci.*, **9**, 1416-1419.

Figure A-1. Tertiary structure of the c-crk SH3 domain. Core residues are colored red, boundary residues are colored green, and surface residues are colored navy.

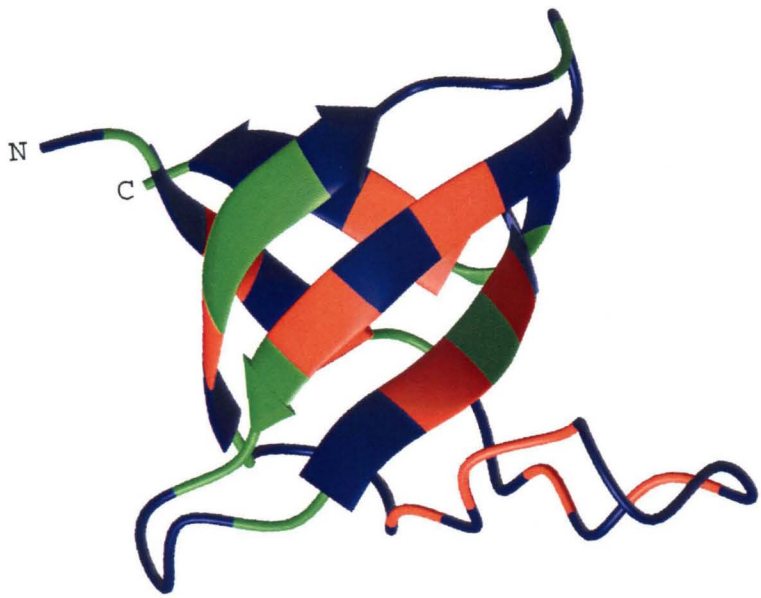

N
C

Figure A-2. Sequences of the wild type and designed variants. Thermal denaturation temperatures, where appropriate, are given to the left of the sequence. Classification of residues as core (c), boundary (b), or surface (s) is denoted below the sequence alignment. In the ribbon diagram, core residues are colored red, boundary residues are colored green, and surface residues are colored navy. Residues held fixed in a given calculation are marked "|" in the sequence alignment.

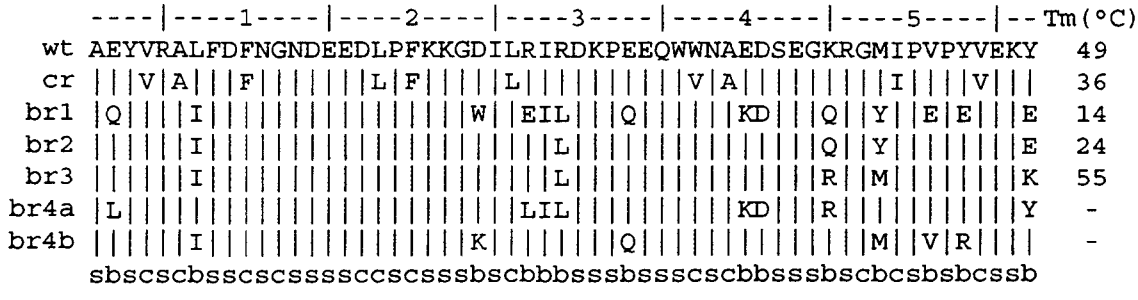

----|----1----|----2----|----3----|----4----|----5----|--Tm(°C)
wt AEYVRALFDFNGNDEEDLPFKKGDILRIRDKPEEQWWNAEDSEGKRGMIPVPYVEKY 49
cr |||V|A|||F|||||||L|F|||||L||||||||||V|A||||||||||I||||V||| 36
br1 |Q||||I||||||||||||||||W||EIL|||Q||||||KD|||Q||Y||E|E|||E 14
br2 ||||||I|||||||||||||||||||||L||||||||||||||Q||Y||||||||E 24
br3 ||||||I|||||||||||||||||||||L||||||||||||||R||M||||||||K 55
br4a |L|||||||||||||||||||||||||LIL||||||||||KD|||R||||||||||Y -
br4b ||||||I||||||||||||||||K|||||||Q||||||||||||||M||V|R|||| -
sbscscbsscscsssscccscsssbscbbbsssbsssscscbbsssbscbcsbsbcssb

Figure A-3. Wavelength scans of the wild type c-crk SH3 domain (blue) and the core design variant (red) at 1 °C.

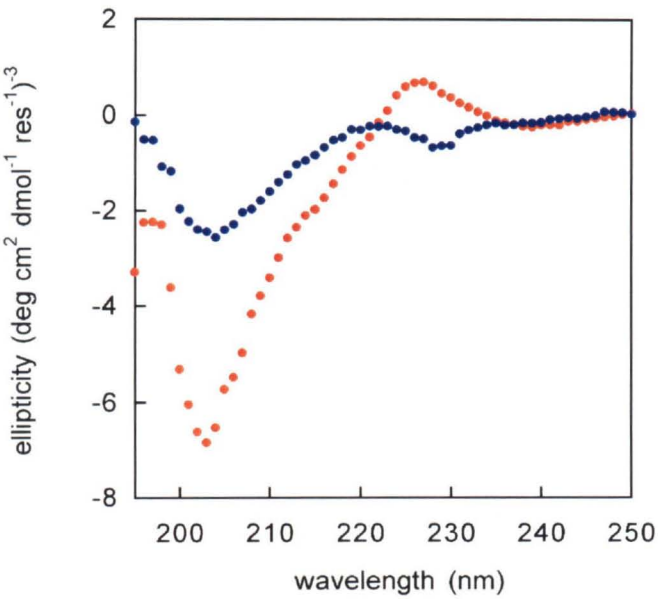

2
0
-2
-4
-6
-8
200
210
220
230
240
250
wavelength (nm)
ellipticity (deg cm^2 dmol^{-1} res^{-1})$^{-3}$

Figure A-4. Thermal denaturation of the wild type c-crk SH3 domain (blue) and the core design variant (red) monitored at 222 nm.

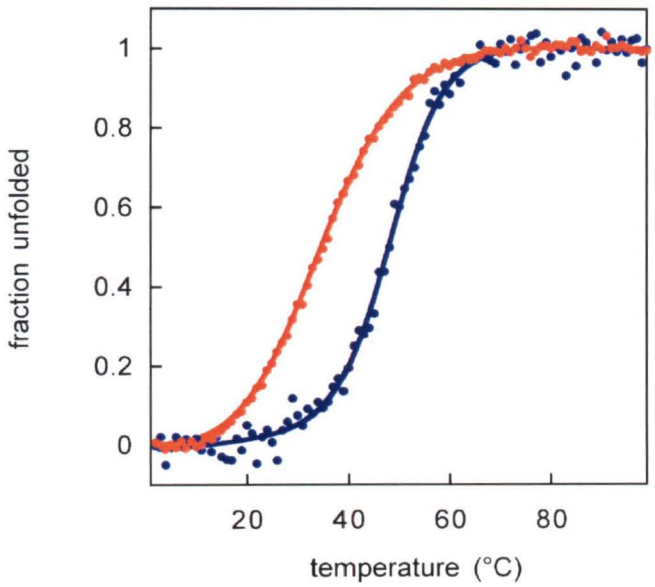

fraction unfolded
1
0.8
0.6
0.4
0.2
0
20
40
60
80
temperature (°C)

Figure A-5. Wavelength scans of the wild type c-crk SH3 domain (blue) and the boundary design variants br1 (orange), br4a (green), and b4rb (red).

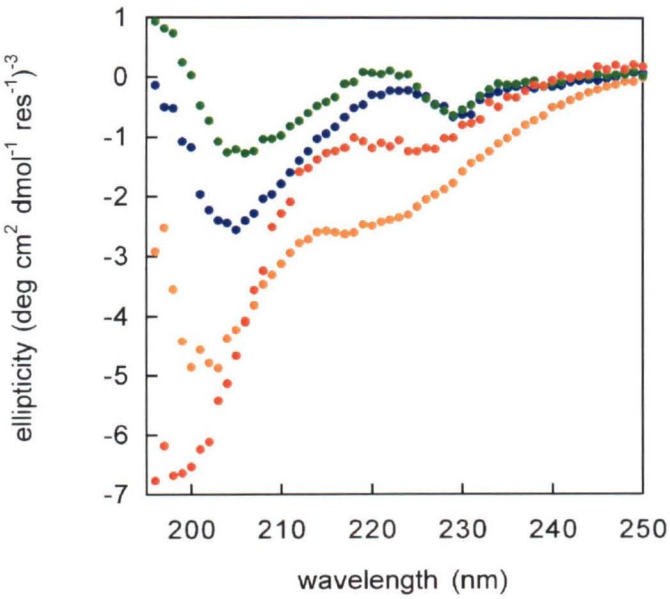

1
0
-1
-2
-3
-4
-5
-6
-7
200
210
220
230
240
250
wavelength (nm)
ellipticity (deg cm^2 dmol^{-1} res^{-1})$^{-3}$

Figure A-6. Thermal denaturation of the wild type c-crk SH3 domain (blue) and designed variants br2 (violet) and br3 (turquoise).

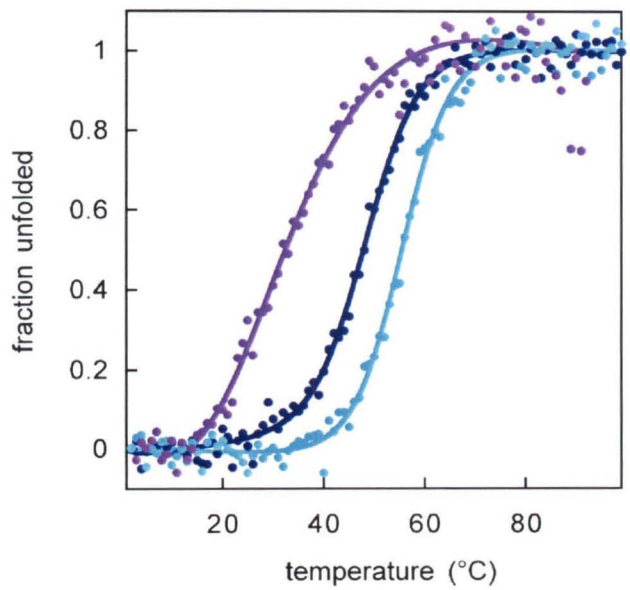

fraction unfolded
1
0.8
0.6
0.4
0.2
0
20
40
60
80
temperature (°C)

Figure A-7. Uncooperative thermal denaturation transitions of designed variants br1 (orange) and br4b (red).

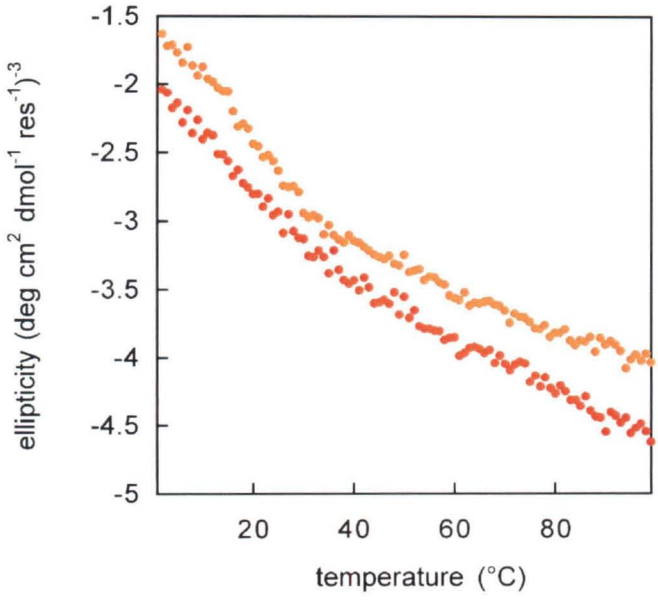
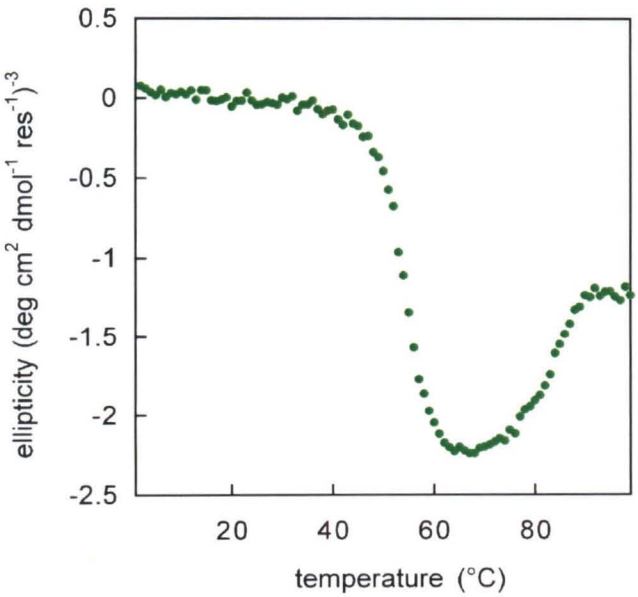

-1.5
-2
-2.5
-3
-3.5
-4
-4.5
-5
20
40
60
80
ellipticity (deg cm^2 dmol^{-1} res^{-1})$^{-3}$
temperature (°C)

Figure A-8. The irreversible thermal denaturation transition of designed variant br4a (green).


0.5
0
-0.5
-1
-1.5
-2
-2.5
20
40
60
80
ellipticity (deg cm^2 dmol^{-1} res^{-1})$^{-3}$
temperature (°C)

Appendix B:

Double mutant cycle analysis of cation-π interactions

Abstract

Protein design models and force fields aim to describe the covalent and noncovalent interactions that contribute to protein folding and stability. Interactions between the quadrupole moment on aromatic residues and other charged or polar functional groups in the protein have thus far not been included in protein design force fields, although theoretical studies and analysis of known protein structures suggest that cation-π interactions can make a significant contribution to protein stability. Here, we have used double mutant cycle analysis to determine the contribution of cation-π interactions that have been introduced into protein G and engrailed homdeodomain.

Introduction

The results of a recent study by Gallivan and Dougherty suggest that cation-π interactions, or the electrostatic interactions between positive charges and π-electron systems such as aromatic amino acids, can contribute more to protein stability than salt bridges[1]. They demonstrate that salt bridge formation is strongly opposed by desolvation effects and attenuated by solvent screening, while cation-π interactions are only diminished slightly by interactions with water. Gallivan and Dougherty find that most cation-π interactions in proteins are located so that the aromatic group is mostly buried and the cation maintains interactions with solvent.

While cation-π interactions have been observed in many protein structures[2, 3], the contribution of cation-π interactions to protein stability had not been determined experimentally. In this study, we used double mutant cycle analysis to determine the stability conferred by two engineered cation-π interactions. These interactions were incorporated

into the surfaces of protein G and engrailed homeodomain. In both cases, solvent exposed helical (i,i+4) positions were used, as such sites allow for cation-π interactions with favorable geometry and minimal interactions with the rest of the protein. Positions in the middle of the helix were selected to minimize the effect of the helix dipole. Double mutant cycle analysis[4, 5] was then used to measure the energetic contribution of the engineered cation-π interactions.

Results

Initial study focused on incorporating a cation-π interaction into positions 28 and 32 of protein G. These positions, shown in Figure B-1, are exposed to solvent and located in the middle of the one helix in protein G. After considering all possible cation-π pairs in both orientations, the R28-W32 interaction was selected for experimental analysis. The R28-W32 cation-π interaction is predicted to be among the most favorable of the interactions considered and does not clash with surrounding residues. Furthermore, analysis of cation-π interactions in known protein structures indicates that Trp is far more likely than Phe and Tyr to form a cation-π interaction, and Arg is somewhat more likely than Lys to form a cation-π interaction[3].

Urea denaturation was selected as the best experimental method for determining the stability of the R28-W32 interaction. Like other chemical denaturation methods, urea denaturation experiments can be used to determine the free energy of unfolding, ΔG_u, of a protein. Urea was selected rather than guanidinium chloride, as guanidinium can compete with Arg to form cation-π interactions and changes in ionic strength can mask electrostatic contributions to stability[6]. However, the urea denaturation experiment conducted on the double alanine mutant demonstrated that protein G variants can be nearly fully folded even at 9 M urea, as shown in Figure B-2.

As the protein G variants do not denature in urea under otherwise moderate solution conditions, thermal denaturation experiments were used instead. Although accurate free energies can not be obtained directly thermal denaturation, trends in thermal stability correlate with trends in ΔG_u for closely related proteins[7]. Approximate free energies can be calculated using a constant value for ΔC_p obtained using another method. The results of the thermal denaturation experiments, shown in Table B-1 and Figure B-3, indicate that the engineered cation-π interaction makes a very small contribution to the stability of protein G: the $\Delta\Delta T_m$ is 1.1 °C and $\Delta\Delta G_u$ is 0.5 kcal mol^{-1}, which is within experimental error.

A second engineered cation-π interaction, shown in Figure B-4, was introduced to the surface of engrailed homdeodomain. As before, the interaction occurs between Arg and Trp and is located at (i,i+4) positions on a solvent exposed helical face. The homeodomain variants lack a well-defined posttransition in thermal denaturation experiments (data not shown). However, unlike protein G, homdeodomain unfolds in the presence of moderate urea concentrations. Therefore, urea denaturation experiments were used for the double mutant cycle analysis. The chemical denaturation transitions of the homdeodomain variants are all fairly similar, as shown in Table B-2 and Figure B-5, and the interaction energy predicted using double mutant cycle analysis indicates that the cation-π interaction is slightly unfavorable. However, the cooperativity of the transistions, given by the m-value, ranges from 0.73 to 0.91 kcal mol^{-1} M^{-1}, suggesting that the transitions may not all be two-state[8]. Accordingly, the free energies calculated assuming a two-state transition and the interaction energy calculated from the double mutant cycle may not be accurate.

Conclusions

The cation-π interactions that were introduced to protein G and homeodomain do not contribute significantly to protein stability. However, this is likely due to the experimental

systems selected for this study rather than the fundamental ability of cation-π interactions to stabilize proteins. While Arg and Trp residues were introduced in positions where they could form a favorable interaction, it is possible that it is more energetically favorable for either or both of the cation-p partners to interact with other protein functional groups instead. Although it is easy to introduce helical (*i*,*i*+4) interactions, more favorable cation-π interactions would be expected in areas where the aromatic residue is more buried. However, double mutant cycle analysis would not be valid for such an interaction, since removal of the tryptophan probably results in structural rearrangement.

As more sophisticated electrostatic models are developed for protein design, it is likely that design algorithms will begin selecting cation-π interactions. Many of the newer charge sets, including the PARSE charge set[9] that was used for the Poisson-Boltzmann calculations in Chapters V - VII, include net quadrupole moments for the aromatic residues. So long as the quadrupole moment is included, the Coulombic interaction between an aromatic side chain and a cationic side chain can be significant. Therefore, it may prove unnecessary to add a specialized cation-π term to the force field. Once cation-π interactions are selected in the context of the entire protein, we may obtain more conclusive experimental data on the effect of cation-π interactions on protein stability.

Materials and Methods

Modeling. Structural coordinates for protein G were obtained from PDB entries 1pga (protein G). Hydrogens were added to the remaining residues using BIOGRAF (Molecular Simulations, Inc., San Diego). The resulting strucutre was minimized for 50 steps using the Dreiding force field[10]. A designed homeodomain variant with optimized surface and core sequence[11] was used as the template for subsequent computational and experimental studies on homdeodomain. The side chains forming the cation-π interactions were modeled using

the backbone dependent rotamer library developed by Dunbrak and Karplus[12]. Rotamers were also included at ±1 standard deviation about $\chi 1$ and $\chi 2$.

Selecting sites for introduced cation-π interactions. Protein G residues 28 and 32 were considered as the site for the first cation-π interaction. These residues are located in the center of the helix and are solvent exposed. Arg and Lys were considered for the cation group and Phe, Tyr, and Trp were considered for the π group. All possible cation-π combinations were considered. Two sites were considered for the homeodomain cation-π interaction: 9 and 13, and 42 and 46. Only Arg-Trp cation-π interactions were considered at these positions. Two calculations were performed to test each candidate site. First, a cation-π pair was placed at the site and the surrounding residues (*i*-4, *i*-1, *i*+1, *i*+3, *i*+5, and *i*+8) were mutated to Ala. The conformations of the cation-π pair were optimized using the force field described below. Next, the interaction energy between the cation-π pair and the surrounding residues was calculated for each cation-π pair using the force field described below. Conformations for the residues involved in the cation-π interaction were held fixed, while the conformations of the surrounding residues were allowed to vary. In all calculations, the optimal rotameric conformation was determined using the dead-end elimination theorem[13-15] Pairs were selected based on their geometry, cation-π interaction energy, and interactions with the remainder of the protein.

Force fields. The geometry of the cation-π pairs was optimized using van der Waals interactions scaled by 0.9[16] and Coulomb's law calculated using a distance dependent dielectric of $2r$ in conjunction with partial atomic charges from the OPLS force field[17]. The OPLS charge set includes a net quadrupole moment for aromatic groups. The interaction

energy between the cation-π pairs and the rest of the protein was calculated using the standard ORBIT parameters and charge set[18].

Protein expression. The following constructs were used for the double mutant cycle analysis on protein G: 28A/32A, 28A/32W, 28R/32A, and 28R/32W. Double mutant cycle analysis on homeodomain used the constructs 9A/13A, 9A/13W, 9R/13W, and 9R/13W. All constructs were generated by site-directed mutagenesis using inverse PCR and confirmed using DNA sequencing. Recombinant proteins were expressed in BL21 (DE3) *Escherichia coli* cells (Stratagene) and isolated using the freeze-thaw method[19]. The proteins were purified by reversed-phase HPLC using a C8 prep column (Zorbax) and linear water-acetonitrile gradients with 0.1 % trifluoroacetic acid. Protein masses were checked using MALDI-TOF or electrospray mass spectrometry; all masses were within one unit of the expected weight.

Circular dichroism studies. CD data were collected using an Aviv 62A DS spectropolarimeter equipped with a thermoelectric cell holder and an autotitrator. Samples for thermal denaturation contained 50 μM protein and 50 mM sodium phosphate adjusted to pH 4.5 and samples for urea denaturation contained 5 μM protein and 50 mM sodium phosphate adjusted to pH 4.5. To maintain constant pH, the urea stock solution also was adjusted to pH 4.5. Thermal denaturation data were acquired every 1 °C from 1 °C to 99 °C with an equilibration time of 90 seconds and an averaging time of 30 seconds. Reversibility of the thermal unfolding transitions was confirmed. Thermal denaturation temperatures were determined by fitting to a two-state transition as previously described[20]. Urea denaturation data was acquired every 0.2 M from 0.0 M to 9.0 M with a 9 minute mixing time and 100 second averaging time. ΔG_u was calculated assuming a two-state transition and using the linear extrapolation model[21]. In the case of protein G, 0.621 kcal mol^{-1}, a

value obtained from calorimetric studies conducted on the wild type protein G^{22}, was used for ΔC_p. This value for ΔC_p and values for ΔH and T_m obtained from the thermal denaturation data were used to calculate ΔG_u for the protein G variants. Denaturation experiments were monitored at 218 nm for protein G variants and at 222 nm for homeodomain variants.

Double mutant cycle analysis. The strength of the cation-π interactions was calculated using the following equation:

$$\Delta G^{\text{cation-}\pi} = (\Delta G^{RW} - \Delta G^{AA}) - [(\Delta G^{RA} - \Delta G^{AA}) + (\Delta G^{AW} - \Delta G^{AA})] \qquad (1)$$

where ΔG^{RW} is the free energy of unfolding of the RW mutant, ΔG^{RA} is the free energy of unfolding of the RA mutant, ΔG^{AW} is the free energy of unfolding of the AW mutant, ΔG^{AA} is the free energy of unfolding of the AA mutant. The contribution of the cation-π interaction to the thermal stability of protein G was calculated similarly, with all of the free energies in the preceeding equation replaced by thermal denaturation temperatures.

References

1. Gallivan, J. P. and Dougherty, D. A. (2000). A computational study of cation-π interactions vs salt bridges in aqueous media: implications for protein engineering. *J. Am. Chem. Soc.*, **122**, 870-874.
2. Ma, J. C. and Dougherty, D. A. (1997). The cation-π interaction. *Chem.Rev.*, **97**, 1303-1324.
3. Gallivan, J. P. and Dougherty, D. A. (1999). Cation-π interactions in structural biology. *Proc. Natl. Acad. Sci., USA*, **96**, 9459-9464.
4. Horovitz, A., Serrano, L., Avron, B., Bycroft, M. and Fersht, A. R. (1990). Strength an co-operativity of contributions of surface salt bridges to protein stability. *J. Mol. Biol.*, **216**, 1031-1044.
5. Serrano, L., Horovitz, A., Avron, B., Bycroft, M. and Fersht, A. R. (1990). Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. *Biochemistry*, **29**, 9343-9352.
6. Monera, O. D., Kay, C. M. and Hodges, R. S. (1994). Protein denaturation with guanidine-hydrochloride or urea provides a different estimate of stability depending on the contributions of electrostatic interactions. *Protein Sci.*, **3**, 1984-1991.
7. Becktel, W. J. and Schellman, J. A. (1987). Protein stability curves. *Biopolymers*, **26**, 1859-1877.
8. Soulages, J. L. (1998). Chemical denaturation: protential impact of undetected intermediates in the free energy of unfolding and *m*-values obtained from a two-state assumption. *Biophys. J.*, **75**, 484-492.
9. Sitkoff, D., Sharp, K. and Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.*, **98**, 1978-1988.

10. Mayo, S. L., Olafson, B. D. and Goddard, W. A., III. (1990). Dreiding - a generic force-field for molecular simulations. *J. Phys. Chem.*, **94**, 8897-8909.

11. Morgan, C. S. (2000) Ph. D. Thesis. California Institute of Technology, Pasadena, CA.

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

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

14. Desmet, J., De Maeyer, M., Hazes, B. and Lasters, I. (1994) The dead-end elimination theorem: a new approach to the side-chain packing problem. In *The protein folding problem and tertiary structure prediction* (K. Merz, Jr and S. Le Grand, ed) 307-337, Birkhauser, Boston.

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

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

17. Jorgensen, W. L. and Tirado-Rives, J. (1988). The OPLS potential functions for proteins. Energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.*, **110**, 1657-1666.

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

19. Johnson, B. H. and Hecht, M. H. (1994). Recombinant proteins can be isolated from E. coli cells by repeated cycles of freezing and thawing. *Biotechnology*, **12**, 1357-1360.

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

21. Santoro, M. M. and Bolen, D. W. (1988). Unfolding free-energy changes determined by the linear extrapolation method .1. unfolding of phenylmethanesulfonyl α-chymotrypsin using different denaturants. *Biochemistry*, **27**, 8063-8068.

22. Alexander, P., Fahnestock, S., Lee, T., Orban, J. and Bryan, P. (1992). Thermodynamic analysis of the folding of the streptococcal protein G IgG-binding domains B1 and B2: why small proteins tend to have high denaturation temperatures. *Biochemistry*, **31**, 3597-3603.

Table B-1. Thermal denaturation data: protein G variants

	T_m[1] (°C)	ΔH_u[2] (kcal mol^{-1})	ΔG_u[3] (kcal mol^{-1})
AA	85.7	51.5	5.35
AW	85.2	54.5	5.89
RA	81.1	45.3	4.27
RW	81.7	45.5	4.31

[1]Midpoint of the thermal denaturation transition

[2]Enthalpy of unfolding, calculated assuming ΔC_p = 0.621 kcal K mol^{-1}

[3]Free energy of unfolding at 25 °C, calculated assuming ΔC_p = 0.621 kcal K mol^{-1}

Table B-2. Urea denaturation data: homeodomain variants

	ΔG_u[1] (°C)	C_m[2] (M)	m[3] (kcal $mol^{-1}M^{-1}$)
AA	4.82	6.6	0.73
AW	5.99	6.6	0.91
RA	5.58	6.6	0.85
RW	5.36	6.4	0.84

[1]Free energy of unfolding at 25 °C

[2]Midpoint of the unfolding transition

[3]Slope of ΔG_u versus denaturant concentration

Figure 1. Modeled structure of the cation-π interaction introduced to protein G. The side chains forming the cation-π interaction are shown in green and the surrounding residues are shown in red.

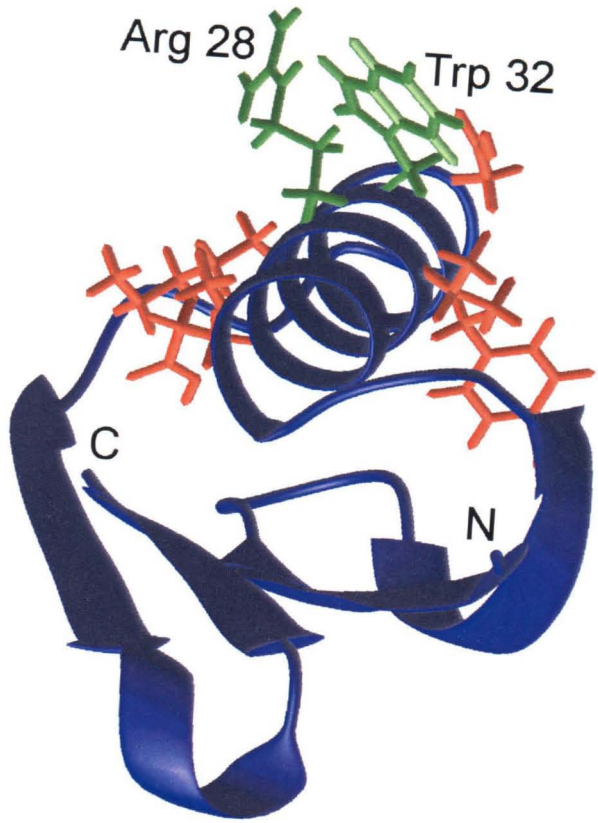

Arg 28
Trp 32
C
N

Figure 2. Urea denaturation of the protein G 28A/32A variant. Even at 9 M urea, the protein appears to be folded.

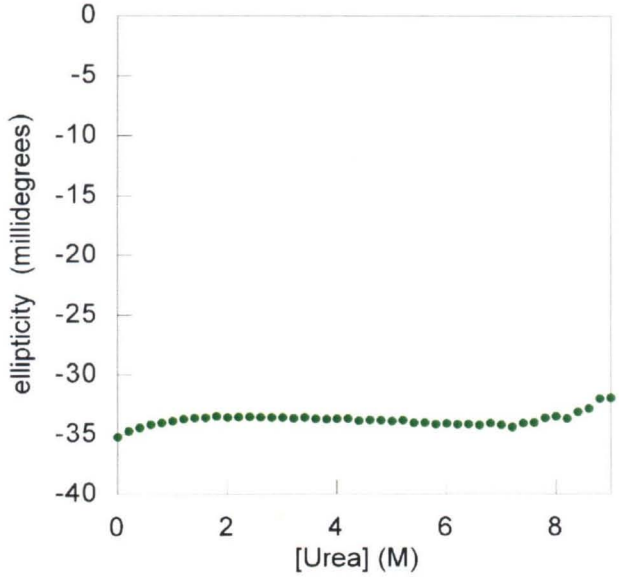

0
-5
-10
-15
-20
-25
-30
-35
-40
ellipticity (millidegrees)
0
2
4
6
8
[Urea] (M)

Figure 3. Thermal denaturation of protein G variants 28A/32A, shown in green, 28A/32W, shown in blue, 28R/32A, shown in in red, and 28R/32W, shown in orange.

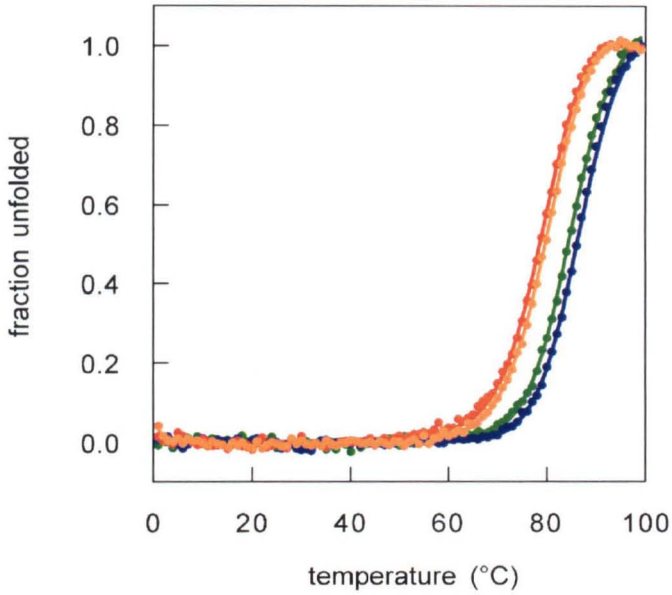

fraction unfolded
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
temperature (°C)

Figure 4. Modeled structure of the cation-π interaction introduced to engrailed homeodomain. The side chains forming the cation-π interaction are shown in green and the surrounding residues are shown in red.

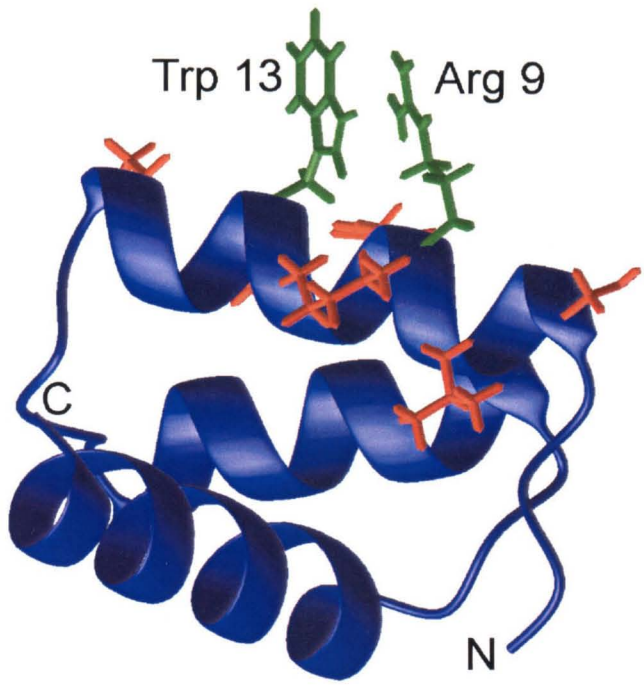
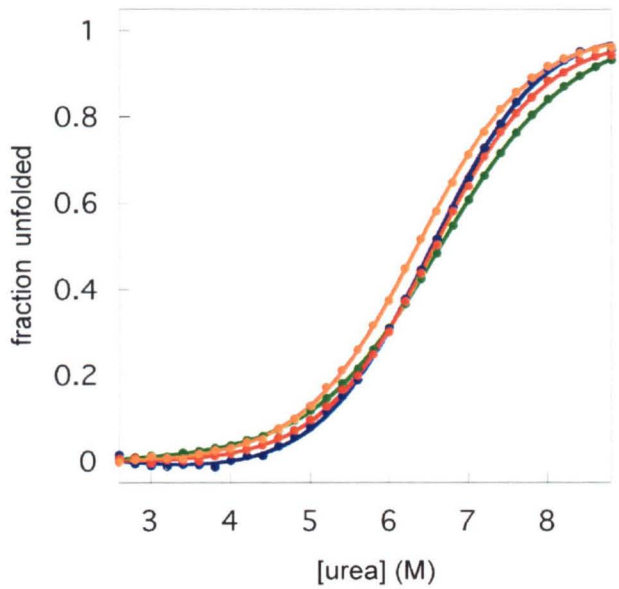

Trp 13
Arg 9
C
N

Figure 5. Urea denaturation of the homeodomain variants 9A/13A, shown in green, 9A/13W, shown in blue, 9R/13A, shown in in red, and 9R/13W, shown in orange.


1
0.8
0.6
0.4
0.2
0
fraction unfolded
3
4
5
6
7
8
[urea] (M)