Full Sequence Design of an Alpha-Helical Protein and Investigation of the Importance of Helix Dipole and Capping Effects in Helical Protein Design

Thesis by

Chantal Smith Morgan

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California, USA

2000

(Submitted July 14, 1999)

This thesis is dedicated to the memory of my dear friend Randy.

Acknowledgements

It is hard to begin to acknowledge all of those people who contributed to my research, my success as a graduate student, and to the five years I have enjoyed at Caltech.

First, I wish to thank my advisor, Stephen L. Mayo, for allowing me the freedom to explore many aspects of protein design. I have enjoyed our lab, which is filled with many smart chemists, biochemists, biologists, and physicists, and benefited our multidisciplinary approach to research. In addition, I'd like to thank my committee, consisting of Doug Rees, William Goddard III, and Pamela Bjorkman, for overseeing my research. I am especially indebted to Pamela for counseling me after my first year when I decided to change labs, and for her assistance and the support of her lab during our collaborative project.

I'd like to thank all past and present members of the Mayo Lab for establishing and maintaining a friendly, helpful, and supportive environment. In particular, I'd like to thank Bassil Dahiyat for his support and motivation during the years we overlapped in the lab. I'd like to thank Dirk Bökenkamp and Monica Brekow for assisting me, an organic chemist by training, with molecular biology experiments and for their friendship. I'd like to thank my classmates Arthur Street and D. Ben Gordon, as well as Niles Pierce for help with the computational aspects of my projects and for being great labmates. In addition, I'd like to thank Scott Ross and Cathy Sarisky for help with the NMR structure determination aspect of my project, as well as

for their friendship. I would also like to acknowledge Deepshikha Datta for her true friendship and support as she embarks on her own graduate career.

Gary Hathaway, Dirk Krapf, and Jie Zhou in the mass spectrometry facility have provided me with many spectra and useful advice. I also wish to thank Susanna Horvath and her group for lending me reagents.

As a graduate student at Caltech, I have been fortunate enough to be involved in Caltech's recruiting efforts, the Graduate Student Council, and several student-faculty committees. I sincerely hope that my efforts have contributed to improving the quality of student life for graduate students here at Caltech. I am indebted to the James Irvine Foundation and the Achievement Awards for College Scientists Foundation for support through fellowships.

I came to Caltech as a chemistry student, in part because of the efforts of Dian Buchness, the chemistry option secretary, and the recruiting efforts of the Graduate Office. I am indebted to Danny Howard who helped recruit me to Caltech, for his unwavering friendship and support from my first day here. During my time here, I have made some good friends in addition to those in my lab, including Jim Kempf, my chemistry classmate, and Piet Moeleker, a friend from the GSC. I would like to thank Tracy Johnson, Jan Aura, and Massimo D'Apuzzo for their support during my final years at Caltech. I would like thank Anne Johansen and Tim Melbourne for their friendship and the many *tequilitas* we shared.

I would like to thank my family for their support during the last five years when I've been so far from home. In particular, my grandparents, Rita and George Bissonnette, have been wonderfully supportive of me both at

Princeton and here at Caltech. My father, George Smith, has been unconditionally proud of me.

I would like to thank my non-Caltech friends Charis Simms and Heidi Watson for their support and many trips to visit me during my graduate student career. I would also like to thank my roommate, Michelle Rojas Soto, for her friendship and generous offer of proofreading my thesis.

Caltech will always be a special place to me, in part, because I met my partner, Roberto Zenit, here. *Los ultimos dos años fueron mucho mejores con su amor y apoyo.*

I can not neglect to thank my mother, Alrita Morgan, for her unconditional love and support during my life. Without her, I would not be who I am.

Abstract

Our goal is an objective, quantitative design algorithm based on the physical chemical forces which determine protein structure and stability. To this end, we have developed a cyclical protein design strategy which utilizes theory, computation, and experimentation using a variety of protein systems. We address the inverse folding problem using a protein design algorithm which objectively predicts protein sequences which are compatible with a given fold.

Our protein design methodology was developed using a variety of proteins, and therefore should be generalizable to many folds and motifs. To test the generalizability and expand the size of proteins we have designed, engrailed homeodomain (enh), a 51-residue helix-turn-helix motif, was used as a target motif.

A series of design calculations and experiments on the thirty surface positions of enh were performed to probe the importance of the helix dipole and capping effects in protein design. Rules for which types of residues were allowed at the helix termini were introduced systematically, resulting in progressively more stable proteins. The first design in the series, which had no considerations for the helix dipole or capping effects, was shown to have the same thermal stability as wild-type enh and the protein with the most stringent rules has a T_m of 75 °C, 32° higher than wild-type and the first design. Therefore, helix dipole and capping effects have a large impact on our ability to design stable proteins. The ten core residues of enh were

included in the design calculation. The resulting protein, a 29-fold mutant of wild-type, has a T_m of 81 °C.

The full sequence design of enh was computed stepwise. The eleven boundary residues were designed in the context of the surface-core design. The resulting protein, a 39-fold mutant of wild-type enh, has a melting temperature of 114 °C and is 4.7 kcal/mol more stable than wild-type. The structure of the boundary-surface-core design was solved by NMR techniques and found to be in excellent agreement with the target structure. The top 10 structure have a backbone root-mean-square standard deviation of 0.45 Å and the root-mean-square standard deviation between the model structure and experimental backbones is 1.25 Å.

The side chain selection algorithm was also extended to the design of peptides to bind tightly to MHC class I proteins. A circular dichroism spectrometry assay was developed to determine the peptide dissociation constants. Three designed peptides were bound more tightly to the MHC class I molecule H-2K^d than known peptides. In addition, an investigation of the removal of disulfide bonds from toxin folds is discussed.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction to Protein Function, Structure, and Design

Of the three biopolymers, nucleic acids, oligosaccharides, and proteins, proteins are the most versatile in terms of structure and function. There is a direct relationship between the primary sequence of a protein and its structure, which in turn has an intimate relationship with its function.

I. Protein function

Proteins play extremely diverse roles in biological systems. It is impossible to discuss protein function without mentioning protein structure and vice versa. Perhaps the two are most closely connected in fibrous proteins.

Structural roles of proteins

Fibrous proteins have simple structures which are directly related to their macroscopic shapes[1]. For example, fibroin is a single protein composed of large antiparallel β-sheets. The residues of fibroin are glycine interdigitated with either serine or alanine side chains, which allow the β-strands to stack forming a strong (in the dimension parallel to the sheets) but flexible (in the dimension perpendicular to the sheets) fiber, silk. β-keratin is a similar protein found in feathers, skin, claws, scales, and beaks of reptiles and birds.

α-keratins are α-helical proteins which wind around each other making progressively larger parallel bundles. These structures, arranged similarly to a rope, are the major proteins found in hair and wool. In the case

of collagen, left-handed helices, formed by sequences with every third residue as glycine, and often Gly-Pro-X, wrap around each other in bundles of three in right-handed helices, called tropocollagen molecules. Tropocollagen molecules pack together, forming very strong fibers that are found in tendons and skin.

When elastic fibers are necessary, such as in flexible ligaments and arteries, the keratins or collagen would be undesirable. Instead, a glycine- and alanine-rich protein, elastin, is used. Elastin fibers are cross-linked via lysine side chains, allowing the protein to "snap back" to its original shape when stretched.

Catalysis

Enzymes, like all catalysts, increase reaction rates without being consumed in the reaction themselves. They lower the activation energy needed to transform reactants into products without changing ΔG_{rxn}. Proteins are poised to be very specific catalysts because of the numerous three-dimensional arrangements of the twenty amino acids. In addition, amino acids themselves are chiral, allowing for specific recognition and catalysis of highly stereospecific reactions. The six major classes of enzymes are oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases.

Serine proteases, such as trypsin and chymotrypsin, are examples of hydrolases. Both bind polypeptide chains stereospecifically and use serine side chains to essentially make the peptide backbone carbonyl a good leaving

group, facilitating hydrolysis of the peptide bond. Trypsin features a long, narrow, negatively charged binding pocket in front of the active site. Consequently, it binds long, positively charged side chains in the pocket, and cuts polypeptides specifically following Arg and Lys residues. On the other hand, chymotrypsin has a wide, hydrophobic binding pocket and hydrolyzes polypeptide bonds following the hydrophobic residues His (at high pH), Ile, Leu, Phe, Trp, Tyr, and Val.

Enzymes are not limited to protein substrates. They act on organic molecules *in vivo* and in chemical reactors. They catalyze transfer of carbohydrates to proteins, linking two different biopolymers together. Ligases act on DNA, linking two pieces of DNA together and restriction enzymes specifically cut pieces of DNA.

Recognition

Proteins are adept at recognition because of the variety of tertiary shapes, or structures, and side chains available. Some examples of recognition by proteins are the recognition of DNA to regulate transcription, carbohydrates to regulate cell-cell interactions, and other proteins and small molecules to stimulate immune system responses.

An example of a type of protein which can recognize specific DNA sequences is the homeodomain family. Homeodomains are helix-turn-helix proteins which recognize homeotic genes which control development in eucaryotic cells. These genes feature recurring sequences of 180 base pairs,

called the homeobox, which contains the base pair sequence TAAT. Homeodomains bind as monomers, unlike the procaryotic DNA-binding proteins repressor and Cro, which bind as dimers, and typically have three helices. The structure of engrailed homeodomain bound to DNA has been solved at 2.2 Å resolution and shows that the N-terminal arm of the protein binds in the minor groove and the third helix binds in the major groove, allowing specific binding to DNA with a K_D of 7.9×10^{-11} M as shown in Figure 1-1[2,3]. All homeodomain proteins share very similar helix-turn-helix motifs but recognize different DNA sequences and control different aspects of development.

Two examples of protein-protein recognition are present in the major histocompatability class I immune system proteins. These cell-surface proteins bind small peptides (typically 8-10 residues) for display to T cell receptors[4]. The polymorphic heavy chains of MHC proteins bind peptides which have certain residues (or types of residues) in binding pocket positions but allow a variety of amino acid at non-pocket positions, allowing them to bind an array of peptides from intracellular proteins. The T cell receptor protein samples the top surface of the MHC-peptide complex, initiating a cascade of immune system events if the bound peptide or MHC protein is not recognized as self.

Peptide toxins bind specifically to ion channel proteins and have been isolated from scorpion, spider, and snake venoms. Toxins typically have three conserved disulfide bonds holding together a triple stranded β-sheet

and α-helix. Peptide toxins bind to ion channels with 1:1 stoichiometry at the ion pore entrances, blocking ion flow. Recognition of ion channels by toxins benefits the venomous organism by impeding the nervous system of the stung organism.

Other functions of proteins

The functions of proteins are not limited to those discussed above. Other types of proteins, such as myoglobin, are oxygen transport proteins. Respiratory proteins, including hemoglobin, hemocynanin, hemerythrin, etc., have elaborate three-dimensional structures arranged in such a way to situate a few amino acid side chains to bind porphyrin rings. Metals bound to the porphyrin rings, such as iron and copper, bind oxygen.

II. Protein structure (with an emphasis on helices)

Proteins are polypeptide chains which adopt specific three-dimensional shapes determined by their amino acid sequence. The structure of a protein can be thought of as the result of the optimization of stabilizing interactions between the protein atoms.

Primary structure

The twenty naturally occurring amino acids are chiral (except glycine) monomeric units which are comprised of a central carbon bound to an amino group, a carboxylic acid group, a proton, and a variable side chain. The

carboxylic acid and amino groups of two amino acids are joined in a dehydration reaction to form planar peptide bonds and produce linear polypeptide chains. The characteristics of amino acids are diverse, but can be divided into groups: hydrophobic side chains, which includes Ala, Val, Phe, Pro, Met, Ile, and Leu; charged side chains, which includes Asp, Glu, Lys, and Arg; and polar side chains, which includes Ser, Thr, Tyr, His, Cys, Asn, Gln, and Trp. Glycine and proline residues are special; proline is secondary amine and glycine has a hydrogen in place of a side chain, giving it more flexibility. The order of amino acids in a protein is its primary structure.

Secondary structure

Portions of polypeptides tend to form local, ordered regular three-dimensional shapes, such as α-helices. Since the peptide bonds themselves are planar, the only rotatable backbone bonds are the N-C_α bond, characterized by the bond angle ϕ, and the C_α-C′ bond, characterized by the bond angle ψ. Repeating ϕ and ψ angles lead to ordered secondary structure, such as helices (typically ϕ = -57° and ψ = -47°), parallel β-sheets (ϕ = -119° and ψ = +113°), and antiparallel β-sheets (ϕ = -139° and ψ = +135°). These units of secondary structure can be stabilized independently, in the case of helices, or with other units, in the case of β-strands which form β-sheets together. Turns, which join units of secondary structure, often are stabilized independently via hydrogen bonds between backbone atoms.

Helices. Helices are of special interest for two reasons. First, one fourth of all amino acids of proteins with known structures are in α-helices. Second, the α-helix was first predicted here at Caltech in 1951 by Linus Pauling[5]. Consecutive amino acids with ϕ and ψ angles near -47° and –57° result in a right-handed helical backbone structure with 3.6 residues per turn. The first residue position with helical dihedral angles is called N1, the second, N2, etc. Similarly, the last residue in the helix is called C1, the penultimate, C2, etc. The residues punctuating the helix without helical dihedral angles but which participate in helical i, i + 4 hydrogen bonding are labeled the N-cap and C-cap positions. With this notation, a helix can be described as N-cap, N1, N2, N3, ..., C3, C2, C1, C-cap, as shown in Figure 1-2. Helices are stabilized by hydrogen bonds in between the carboxy oxygen of residue i and the amide proton of residue i + 4, making them independent units of secondary structure.

The helix dipole. The backbone amide N-H bonds of helix backbone hydrogen bonds point towards the N-terminus and the carboxy C-O bonds point towards the C-terminus. Initially, it was thought that the small dipole of each of these bonds which are oriented along the helix axis added along the helix axis and resulted in a large helix dipole, or "helix macrodipole"[6]. On the other hand, *a priori* calculations have predicted that the amide hydrogens at the N-terminus and carbonyl oxygens at the C-terminus which have unsatisfied hydrogen bond donors and acceptors could result in a partial positive character at the helix N-terminal and a partial negative charge

at the C-terminus, and therefore may be responsible for the helix macrodipole[7].

Experimental studies have shown that the effect of the helix dipole at the N-terminus is independent of helix length[8,9]. At the N-terminus of a helix, the first three amide hydrogens have unsatisfied hydrogen bonds because there are no hydrogen bond acceptors preceding them in the helix. As a result, the N-termini of helices have a partial positive charge. Similarly, at the C-terminus of a helix, there are three carbonyls with unsatisfied hydrogen bonds since there are no i + 4 amide hydrogens to make hydrogen bonds to them. As a result, there is a partial negative charge at the C-termini of alpha helices. Therefore, experimental and theoretical results suggest that the helix dipole is an electrostatic effect localized at the termini, rather than a sum of the dipole moment of individual peptide bonds.

Helix propensity. In addition to backbone-backbone hydrogen bonds, helices are stabilized by amino acid side chains. Helices can be stabilized by side chain-side chain interactions, side chain-backbone interactions, as well as the intrinsic helix-forming tendencies, or helix propensities, of the amino acids. Helix propensities have been statistically determined by analysis of the protein data base[10] and by experimental model peptide studies[11,12]. There is good agreement between the propensities obtained from statistical analysis of proteins and model peptide studies. In addition, helix propensities have been calculated as a function of helix position, called normalized positional residue frequencies[13]. From this statistical analysis, it is clear that an amino acid with

high helix propensity, such as lysine, has a much higher probability to be at the C-terminus or middle of a helix (i.e., positions C5-C1, N4, and N5) than at the N-terminus (N1, N2, or N3), presumably because lysine is positively charged and interacts unfavorably with the partial positive charge of the N-terminus. The relationship between side chain charge and propensity to occupy certain regions of the helix was first observed in a statistical analysis of the seven globular proteins of known structure in 1969 by Ptitsyn[14].

In addition to helical residues, propensities have been calculated for the helix N-cap position, the position before N1 which has nonhelical dihedral angles but participates in i, i + 4 hydrogen bonding and for the C-cap position which follows the helix[11,13,15]. Asn, Asp, Ser, and Thr are good N-capping residues because their side chains can act as hydrogen bond acceptors to the open backbone amide protons of the N-terminal residues[11,13,15]. In fact, residues which are good N-capping residues are rare in the middle of helices, perhaps because their side chains can compete for the normal i, i + 4 helix hydrogen bonds[15]. Glycines occupy C-cap positions in 34% of proteins[15], although the effect of the C-capping residue on helix stability is not nearly as dramatic as the effect of the N-capping residue[12]. Similarly, side chain β-sheet and turn propensities have also been determined[16,17].

Tertiary and Quaternary Structure

Elements of secondary structure self-assemble into specific arrangements, called motifs, or folds. Motifs may consist of helices and turns only, such as the helix-turn-helix motif of homeodomains or of β-sheets and turns only, such as the light chain of MHC class I proteins, β_2m. Mixtures of secondary structure elements are common. The arrangement of the motifs of oligomeric proteins with respect to each other is referred to as the quaternary structure.

III. Protein design

Although all of the information determining a protein's fold is contained in the primary sequence, it is nontrivial to determine a fold from sequence alone. This is referred to as the *protein folding problem*. Protein design is the *inverse folding problem*, that is, for a given fold, the selection of side chains which will adopt that target fold.

Early protein design approaches used qualitative, heuristic designs. These proteins had globally correct folds, but often resulted in molten globules[18]. Proteins have also been designed by experimental combinatorial libraries made by random mutagenesis followed by screening for the desired fold[19]. Metal centers[20] and disulfide bonds[21-23] have been engineered into designed proteins. Theoretical models have provided input into protein design as well.

The present work utilizes a cyclical "design cycle" approach, in which iterative cycles of theory, computation, and experiment followed by analysis are used to evaluate and improve the design of peptides. Initial designs utilized a small set of energy terms (i.e., a van der Waals potential) to objectively predict which amino acid side chains were likely to form a given fold. Using rotamer descriptions of the side chains[24,25] and a fixed backbone, a fast scoring algorithm based on the Dead End Elimination Theorem[26-28] was used to find the globally optimal sequence in the optimal rotamer conformation. Backbone atoms are fixed, or included in the template, to limit the degrees of conformational freedom, although the side chain selection algorithm is tolerant to some backbone flexibility[29]. The automated side chain selection algorithm considers specific interactions between the side chains and backbone atoms, and between side chain and side chain atoms.

Initial design efforts in the Mayo Lab included the design of the core residues[30] of the GCN4-p1 coiled coil using a van der Waals potential to account for steric interactions[31]. A quantitative structure activity relation analysis of the experimental results was used to determine that a potential dependent on the hydrophobic and polar buried surface area would greatly improve the scoring ability of the algorithm. In this manner, an analysis of experimental results was fed back to the side chain selection algorithm. Similarly, a subsequent design study of the surface positions examined the effects of inclusion of hydrogen bond optimization, a penalty for buried polar hydrogens, and helix propensities in addition to the standard van der Waals

potential in the design calculation to determine their effectiveness in the force field. Coiled coils designed to have high helix propensity were the most stable, resulting in proteins 10-12 °C more stable than the wild-type coiled coil[32].

A subsequent design study of the core residues of streptococcal protein G β1 domain was used to determine the effect of specific steric constraints and the effect of penalizing exposed hydrophobic surface area[33].

Residue positions that are on the boundary of core and surface positions are designed by a combination of the core and surface energy scoring functions. Design of the core residues and half of the boundary residues of streptococcal protein G β1 domain resulted in the sequence of a hyperthermophilic protein variant[34]. All residues of a small, 28-amino acid ββα motif taken from the zinc finger DNA binding motif of Zif268 were designed, producing a fully-designed protein which folded in the absence of zinc[35].

To date, our protein design efforts have concentrated on the structure, rather than function of proteins. Preliminary work has been done by others to modify existing binding sites, design *de novo* binding sites, and introduce binding sites for reactive metals[18]. Negative design must be introduced into protein design methodology, that is, a method to prevent sequence designs which have lower energies in other conformations than the target fold.

Our future design efforts will include designing enzymes and proteins which bind to each other and to DNA.

IV. Introduction of the present work

The present work describes a diverse study of aspects of protein design. The three molecules used for these studies, engrailed homeodomain, the class I MHC molecule H-2K^d, and agitoxin, are quite different structurally, but all have important recognition properties. Optimization of protein stability was selected as the success criterion, because design of stable proteins is indicative of a good understanding of the physical chemical forces which govern protein structure. In addition, the design of stable proteins is the first step towards the design of proteins with improved or novel functions.

Chapter 2 describes a study of the importance of helix dipole and helix capping effects in our protein design cycle. Using the structure of engrailed homeodomain, a helix-turn-helix motif, rules governing whether or not charged residues are allowed at helix termini were introduced systematically in the design of surface positions. The experimental results show that stricter rules for charged residues at helix termini and N-capping positions result in more stable proteins. Chapter 3 describes the full sequence design of engrailed homeodomain, a hyperthermophilic variant which has been determined by NMR spectrometry to match the target structure.

Chapters 4 and 5 describe work on the design of peptides to bind tightly to the MHC class I molecule H-2K^d. Chapter 4 details a circular dichroism spectrometry assay which was developed to determine peptide

dissociation constants quickly and without the use of radioactivity. Three peptides designed to bind tightly are shown to bind more tightly than endogenous peptides eluted from H-2K^d in Chapter 5.

Chapter 6 described efforts to redesign the disulfide bonds of agitoxin, a small protein held together by three disulfide bonds. A review of subsequent similar work by other groups is presented.

V. References

1. Much of the introduction has been adapted from Mathews, C. & Holde, K.v. *Biochemistry* (The Benjamin/Cummings Publishing Co., Redwood City, CA, 1990).
2. Fraenkel, E., Rould, M., Chambers, K. & Pabo, C. Engrailed homeodomain-DNA complex at 2.2 Å resolution: a detailed view of the interface and comparison with other engrailed structures. *J. Mol. Biol.***284**, 351-361 (1998).
3. Tucker-Kellogg, L., Rould, M., Chambers, K., Ades, S., Sauer, R. & Pabo, C. Engrailed (Gln50->Lys) homeodomain-DNA complex at 1.9 Å resolution: structural basis for enhanced affinity and altered specificity. *Structure* **5**, 1047-1054 (1997).
4. Bjorkman, P.J. & Parham, P. Structure, Function, and Diversity of Class I Major Histocompatibility Complex Molecules. *Annu. Rev. Biochem.* **59**, 253-288 (1990).

5. Pauling, L., Corey, R. & Branson, H. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA* **37**, 205-211 (1951).

6. Wada, A. The alpha-helix as an electric macro-dipole. *Adv. Phiophys.* **9**, 1-63 (1976).

7. Åqvist, J., Luecke, H., Quiocho, F. & Warshel, A. Dipoles localized at helix termini of proteins stabilize charges. *Proc. Natl. Acad. Sci. USA* **88**, 2026-2030 (1991).

8. Lockhart, D. & Kim, P. Internal stark effect measurement of the electric field at the amino terminus of an alpha helix. *Science* **257**, 947-951 (1992).

9. Lockhart, D. & Kim, P. Electrostatic screening of charge and dipole interactions with the helix backbone. *Science* **260**, 198-202 (1993).

10. Muñoz, V. & Serrano, L. Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. *Proteins: Structure, function, and genetics* **20**, 301-311 (1994).

11. Rohl, C., Chakrabartty, A. & Baldwin, R. Helix propagation parameters and N-cap propensities of the amino acids measured in alanine-based peptides in 40 volume percent trifluoroethanol. *Protein Science* **5**, 2623-2637 (1996).

12. Chakrabartty, A., Doig, A. & Baldwin, R. Helix capping propensities in peptides parallel those in proteins. *Proc. Natl. Acad. Sci. USA* **90**, 11332-11336 (1993).

13. Aurora, R. & Rose, G. Helix Capping. *Protein Science* **7**, 21-38 (1998).

14. Ptitsyn, O. Statistical analysis of the distribution of amino acid residues among helical and non-helical regions in globular proteins. *J. Mol. Biol.* **42**, (1969).

15. Richardson, J. & Richardson, D. Amino acids preferences for specific locations at the ends of α-helices. *Science* **240**, 1648-1652 (1988).

16. Smith, C., Withka, J. & Regan, L. A thermodynamic scale for the β-sheet forming propensities of the amino acids. *Biochemistry* **33**, 5510-5517 (1994).

17. Minor, D. & Kim, P. Measurement of the β-sheet-forming propensities of amino acids. *Nature* **367**, 660-663 (1994).

18. Hellinga, H. Rational protein design: combining theory and experiment. *Proc. Natl. Acad. Sci. USA* **94**, 100015-10017 (1997).

19. Kamtekar, S., Schiffer, J., Xiong, H., Babik, J. & Hecht, M. Protein design by binary patterning of polar and nonpolar amino acids. *Science* **262**, 1680-1685 (1993).

20. Hellinga, H. The construction of metal centers in proteins by rational design. *Folding and Design* **3**, R1-R8 (1998).

21. Yan, Y. & Erickson, B. Engineering of betabellin 14D: disulfide-induced folding of a β-sheet protein. *Protein Science* **3**, 1069-1073 (1994).

22. Matsumura, M. & Matthews, B. Stabilization of functional proteins by introduction of multiple disulfide bonds. *Methods in Enzymology* **202**, 336-356 (1991).

23. Pabo, C. & Suchanek, E. Computer-aided model-building strategies for protein design. *Biochemistry* **25**, 5987-5991 (1986).

24. Ponder, J. & Richards, F. Tertiary templates for proteins- use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. of Mol. Biol.* **193**, 775-791 (1987).

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

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

27. Desmet, J., Maeyer, M.D. & Lasters, I. *The dead-end elimination theorem: A new approach to the side-chain packing problem.* 1-307-337 (Birkhauser, Boston, 1994).

28. Goldstein, R. Efficient rotamer elimination applied to protein side-chains and related spin-glasses. *Biophysical Journal* **66**, 1335-1340 (1994).

29. Su, A. & Mayo, S. Coupling backbone flexibility and amino acids sequence selection in protein design. *Protein Science* **6**, 1701-1707 (1997).

30. Classification of residues into core, surface, and boundary groups is discussed in the methods section of Chapter 2.

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

32. Dahiyat, B., Gordon, D. & Mayo, S. Automated design of the surface positions of protein helices. *Protein Science* **6**, 1333-1337 (1997).

33. Dayihat, B. & Mayo, S. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA.* **94**, 10172-10177 (1997).

34. Malakauskas, S. & Mayo, S. Design, structure and stability of a hyperthermophilic protein variant. *Nature Structural Biology* **5**, 470-475 (1998).

35. Dahiyat, B., Sarisky, C. & Mayo, S. *De novo* protein design: towards fully automated sequence selection. *J. of Mol. Biol.* **273**, 789-796 (1997).

Figure 1-1

The engrailed homeodomain-DNA complex[2].

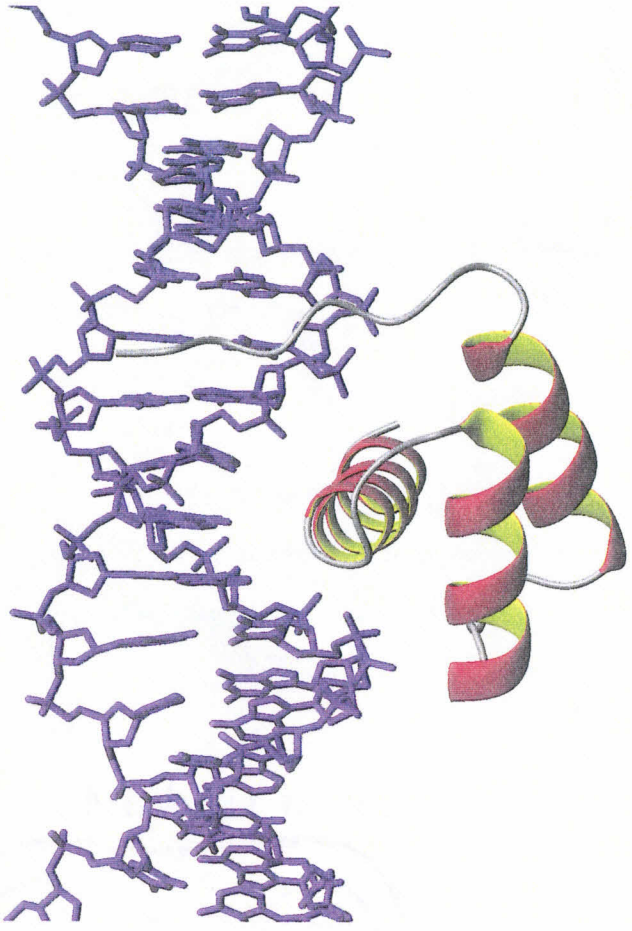
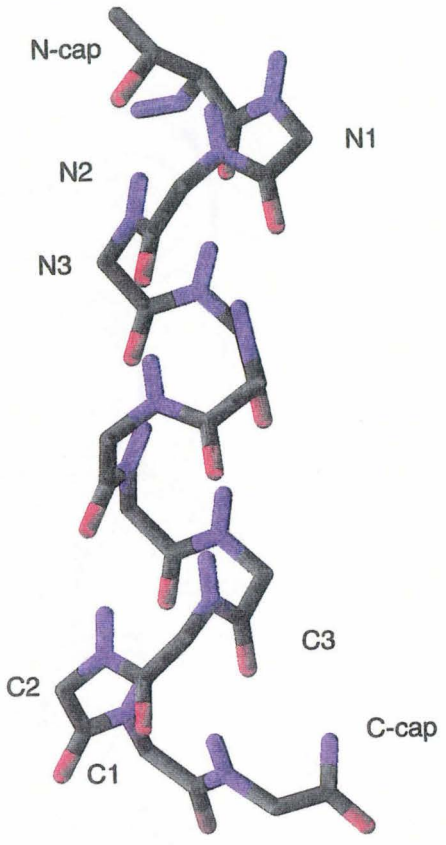


Figure 1-2

Positions of the α-helix.

The side cap of the N-cap residue (Thr) is shown making a hydrogen bond to the i + 2 amide proton.


N-cap
N1
N2
N3
C3
C2
C-cap
C1

Chapter 2

The Effect of the Helix Dipole and N-Capping on the Surface Design of an α-helical Protein

Abstract

The thirty solvent exposed residues of engrailed homeodomain (enh) were designed using an objective side chain selection algorithm. The resulting protein, a twenty-five fold mutant of the fifty-one residue enh, has a T_m of 43 °C, the same as that of enh. To probe the importance of the helix dipole in protein design, helix dipole considerations were systematically introduced into the side chain selection calculation, resulting in protein sequences with dramatically higher stabilities. Introduction of helix capping considerations into the most stable design, 3nc-Ncap, resulted in a twenty-three fold mutant of wild-type with a T_m of 75 °C, well above that of enh. In addition, the calculation of 3nc-Ncap is thirteen times faster than the initial surface design.

Introduction

Alpha helices are linked by amide-carboxyl i, i + 4 hydrogen bonds to form ordered helical secondary structure[1]. Helices are autonomous units of secondary structure stabilized by the close packing of backbone atoms, entropy, and intrahelical hydrogen bonds. Helices can be extended to any length, although the average helix length in proteins is twelve residues[2]. Without considering the side chains, these elegantly simple elements of protein secondary structure are polar and can be thought of as a helix dipole.

At the N-terminus of a helix, the first three amide hydrogens have unsatisfied hydrogen bonds since there are no hydrogen bond acceptors in

the helix before them[3]. As a result, the N-termini of helices have a net positive charge. Similarly, at the C-terminus of a helix, there are three carbonyls with unsatisfied hydrogen bonds since there are no i + 4 amide hydrogens to make hydrogen bonds to them. As a result, there is a net negative charge at the C-termini of alpha helices. It is likely that this effect gives rise to the "helix macrodipole." The effect of the helix dipole at the N-terminus is independent of helix length, supporting the hypothesis that it is localized at the termini and not due to an additive effect of aligned polar carbonyl and amide bonds[4,5].

The effect of the helix macrodipole on protein stability has been studied in both model peptides and proteins. Using helical model peptides, Baldwin *et al.* examined the effect of aspartate on helix stability as a function of position and pH (and therefore charge)[6]. Negatively charged Asp increases the helical character when positioned near the helix N-terminus, which has a partial positive charge, and decreases helical character near the C-terminus, which has a partial negative charge. Therefore, negative charges are stabilizing at the N-terminus of proteins and destabilizing at the C-terminus.

The largely electrostatic effect of the helix macrodipole has been observed in proteins. Any of four point mutations which substitute a negatively charged residue for an uncharged residue at the N-termini of two helices of T4 lysozyme have been shown to increase the thermostability of the protein[7,8]. Each of the four mutations (S38D, N144D, T109D, N116D) results in a protein which has a pH-dependent increase in stability, i.e., the mutants

are slightly destabilized at pH 2 when the Asp side chain is neutral and stabilized at pH 5.7-6.9 when the Asp side chain is charged, consistent with the hypothesis that the negative charge of the side chain is stabilizing when positioned at the N-terminal end of helices.

The effect of side chain charge at the C-terminal end of helices was examined in a protein by introduction of two mutations into the ribonuclease T_1 helix, E28Q and D29N[3]. Both conservative mutations resulted in the replacement of a negatively charged side chain by a similar uncharged side chain. Either mutation increased the stability of the protein and introduction of both mutations concurrently produced an additive stabilizing effect with little effect on enzymatic activity.

The identity of the residue in the N-cap position also has an effect on helix stability. This position, which immediately precedes an α-helix, has nonhelical dihedral angles but participates in helical i, i + 4 hydrogen bonding. The side chains of some residues in the N-cap position can hydrogen bond to the otherwise unsatisfied backbone amide hydrogen of subsequent residues. Model studies have been used to determine the N-capping propensities of residues in the N-cap position[9]. The effect of C-capping residues on helix stabilization is not as large as for N-capping residues[10].

Protein stabilization by helix-capping residues was studied in T4 lysozyme by mutation of the N-capping residue T59[11]. Replacement of T59 by Asn, Ser, and Asp resulted in slightly destabilized proteins ($\Delta\Delta G$ = 0.6-0.9 kcal mol^{-1}). Replacement of T59 by the aliphatic side chains Val, Gly, and Ala

which are incapable of making N-cap hydrogen bonds resulted in proteins destabilized by 2.2-2.8 kcal mol^{-1}. Therefore, residues with side chains which are hydrogen bond acceptors, i.e., Asn, Asp, Ser, and Thr, are more stabilizing in the N-cap position of helices than those which can not form hydrogen bonds.

Including the helix macrodipole in protein design

To investigate if including the helix macrodipole in our calculations would improve the design algorithm, a systematic study of the helix macrodipole in the design calculations was undertaken using a helical protein template. Stability was used as the measure of design success, although ultimately these results could be extrapolated to optimize for function, such as catalysis or recognition. Helix propensities of amino acid side chains were not included in the design calculations.

Helix macrodipole considerations were introduced to the design calculations by imposing restrictions at the helix termini which were progressively moved inwards towards the helix center in a stepwise manner as the designs were made more restrictive. Because the protein design algorithm currently does not calculate electrostatic interactions for dipoles, this proved to be a simple method to probe the effect of the helix dipole in protein design. Our design algorithm considers global interactions, so small increases in restrictions at the helix termini can cause a cascade of mutations throughout the protein. Since all surface positions can be designed in each calculation, this approach has an advantage over the introduction of point

mutations because changes at the helix termini are allowed to propagate into the rest of the protein.

The helix-turn-helix motif of *Drosophila* engrailed homeodomain (enh) was selected as the target fold for this design study. The crystal structure of the first fifty-six residues of the mostly helical sixty-one amino acid protein has been solved at a resolution of 2.1 Å[12]. We opted to design the surface residues of a 51 amino acid variant containing residues 6-56 (subsequently renumbered 1-51), shown in Figure 2-1. The renumbered structure is comprised of three helices, residues 5-17, 23-32, and 37-51, all of which are preceded by surface-exposed N-capping residues. The 51-mer can be synthesized by solid phase techniques and appears to have similar secondary structure and slightly higher thermal stability than the 56-residue domain[12].

Thirty of the 51 enh residues were determined to be solvent exposed as described in the Methods. The solvent-exposed residue 34 has a positive ϕ angle and is glycine in the wild-type enh sequence. Glycine lacks a side chain and therefore has a larger range of accessible ϕ angles, so the identity of residue 34 was selected to be glycine in each design. Therefore, surface residue calculations of enh included G34 and the other twenty-nine 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).

In addition to the ten residues allowed at surface positions, three subsets were made for these design calculations, listed in Table 2-1. The first, for N1-N3 positions, contains only neutral and negatively charged side chains to avoid destabilizing electrostatic effects at the N-terminus. Similarly, a

second restricted subset was made for C3-C1 positions containing only neutral and positively charged side chains. A third subset of amino acids which only contains side chains that are capable of N-capping was made for N-capping positions.

Design results

Each surface sequence design is described by its helix positions with restricted allowed amino acids. The first enh design, 0nc, has no computational considerations for the helix macrodipole or N-capping. Therefore, all 10 surface residues were allowed at all 29 designed positions and the remaining core and boundary positions were left as in the wild-type sequence. The 0nc sequence differs from wild-type in twenty-five positions; the predominant difference in the designed sequence is the introduction of seventeen more charged surface residues, shown in Table 2-2. These side chains, mainly Arg, Glu, and Lys, are predicted to form salt bridges and hydrogen bonds as well as interact with solvent in the model structure. There are twelve more putative salt bridges and one fewer hydrogen bonds in the model structure of 0nc than in the wild-type crystal structure (Tables 2-3a, 2-3b). In addition, the N-cap of the second helix is lost by a T22E mutation. The secondary structure is helical and the thermal stability of 0nc is the same as that of wild-type, shown in Figure 2-2. Therefore, design of more than half of the residues of enh results in a helical protein with native stability.

The second sequence design, 1nc, is the first design with helix macrodipole considerations. The N1 (5, 23, 37) and C1 (17, 32) surface positions of the three helices were restricted to the N-terminal and C-terminal

allowable surface amino acids, shown underlined in Table 2-2. Position 51, the C1 residue of helix 3, is a boundary residue and therefore not included in the surface calculation. All other surface positions were selected from the whole set of allowed surface residues. Because Glu and Lys were selected for all three N1 and both C1 positions respectively in the 0nc calculation, application of the restricted side chain groups at the N1 and C1 positions in the 1nc calculation has no net effect and the result of the 1nc calculation is the same as that of 0nc.

In the next enh design, 2nc, the helix macrodipole considerations were extended to the N2 and C2 positions. The N1 and N2 positions (5, 6, 23, 24, 37, 38) and the C1 and C2 surface positions (16, 17, 31, 32, 50), shown underlined in Table 2-2, were restricted N- and C-terminal amino acid types. The other solvent exposed positions in the calculation were allowed to vary among the ten residues allowed at surface positions and could compensate for the N1, N2, C1, and C2 restrictions. The calculated 2nc sequence differed from 0nc/1nc in ten positions, four of which are not at the N1, N2, C2, or C1 positions, demonstrating that local restrictions propagate changes throughout the protein. Many of these mutations involve salt bridge or hydrogen bond networks, tabulated in Table 2-3c. The resulting 2nc molecule has helical secondary structure and a melting temperature of 73 °C and is remarkably more stable than 0nc/1nc, as shown in Figure 2-2.

Similarly, the 3nc calculation included restricted side chain groups at the N1, N2, and N3 N-terminal positions (5, 6, 23, 24, 37, and 38) and C1, C2, C3 C-terminal positions (16, 17, 31, 32, 49, and 50). Of the six N3 and C3 positions, only position 49 is a surface residue; therefore, the 2nc and 3nc

calculations differ only at position 49. The result of this local change propagates throughout helices 2 and 3 and 3nc is a six-fold mutant of 2nc. The sequence changes of 3nc do not perturb the secondary structure or thermal stability of 2nc (Figure 2-2) although there are two more putative hydrogen bonds (Table 2-3d).

To probe the stabilizing effect of the helix macrodipole at just the N-terminus, a design with restrictions at N-terminal positions but not C-terminal positions was computed. Since all three N3 positions are either core or boundary residues, the definition of the 2n and 3n calculations are identical. All ten surface amino acid types were allowed at all surface positions except N1 and N2 (5, 6, 23, 24, 37, and 38) which were restricted to the N-terminal group. The sequence of 2n/3n differs from both 2nc and 3nc in seven positions (Table 2-2). The C1 and C2 positions without C-terminal restrictions revert to those of the 0nc sequence. All other positions have side chains and predicted interactions found in 2nc or 3nc, listed in Table 2-3e. The secondary structure is the same as wild-type, but the thermal stability, measured to be 71 °C, is slightly lower than that of 2nc and 3nc but much higher than 0nc (Figure 2-2). Therefore, N-terminal dipole effects alone can be used to increase stability.

The last avenue of exploration in this series of surface design molecules was the effect of the N-capping residues. All three N-cap positions are surface positions, and therefore have been designed as unrestricted surface positions in all calculations described thus far. In the wild-type sequence all three N-capping residues have good helix-capping propensities[13]. Helix 1 is capped by Ser in the wild-type and all designed

sequences. Helix 2 is capped by Thr in the wild-type and all designed sequences except 0nc/1nc in which it is capped by Glu. In the wild-type sequence, helix 3 is capped by Asn 36, a good capping residue, but was selected as Arg and predicted to form a salt bridge in all designed sequences thus far. To regain the potential 1.0-2.8 kcal mol^{-1} of the stability provided by the helix cap[11,13] at position 36, the 3nc calculation was repeated using the N-cap restricted rotamer group at positions 4, 22, and 36. The resulting sequence, 3nc-Ncap, differs from 3nc in seven positions, listed in Table 2-2. Four of the seven mutations relative to 3nc are not in the N-cap, N1-N3, or C3-C1 positions and all occur in helices 2 and 3. As expected, the N-cap position 36 reverts to Asn. The secondary structure of 3nc-Ncap is the same as wild-type, and the thermal stability of 75 °C is slightly higher than the other surface design sequences with restrictions for the helix dipole (Figure 2-2). Therefore, the helix macrodipole and N-capping effects can be manipulated in concert to optimize stability of a given fold.

Proteins can be stabilized by ion pairs, and in particular, ion pair networks, which allow for more stabilization with a lower entropic cost[14]. Since the designed proteins have large amounts of charged surface residues and the model structures suggest many surface ion pairs, the importance of ion pair stabilization was studied. Chemical denaturation experiments on coiled coils with designed electrostatic interactions to be stabilizing, destabilizing, or neutral, have shown that GdnHCl denaturation can not measure the amount of stabilization or destabilization based on the type of designed electrostatic interactions[15]. GdnHCl is charged, and effectively

screens charge-charge interactions. However, ΔG determined by urea denaturation corresponds to the sum of the stabilizing and destabilizing interactions of the designed ion pairs of the coiled coils. In addition, $[urea]_{1/2}$ measurements correspond linearly to T_ms determined by thermal denaturation experiments.

Chemical denaturation experiments were performed on wild-type engrailed homeodomain and 3nc-Ncap, the most stable surface design, to probe the effect of designed stabilizing salt bridges. The ΔGs of wild-type and 3nc-Ncap were determined to be 4.1 and 3.0 kcal mol^{-1} respectively from GdnHCl denaturation experiments at 1 °C, shown in Figure 2-3. The T_ms of wild-type and 3nc-Ncap are 43 and 75 °C respectively. Because 3nc-Ncap was measured to be less stable than wild-type by GdnHCl denaturation, it is probable that salt bridges are responsible for the increased stabilization of 3nc-Ncap. To insure that the apparent destabilization of 3nc-Ncap was due to the electrostatic interaction screening effect of GdnHCl, chemical denaturation experiments were repeated with urea, a neutral denaturant. The ability of urea to unfold proteins is approximately two-fold lower than GdnHCl, so urea denaturation experiments were performed at a higher temperature, shown in Figure 2-4. The ΔGs at 35 °C were calculated to be 1.2 and 5.0 kcal mol^{-1} for wild-type and 3nc-Ncap respectively. Therefore, the designed protein 3nc-Ncap is stabilized relative to wild-type and that stabilization is due to ion pair interactions. In addition, stabilities may be compared by their $[urea]_{1/2}$ concentration, that is, the concentration of urea

when the proteins are half-unfolded. The $[urea]_{1/2}$ concentrations of wild-type and 3nc-Ncap are 1.75 and 7.25 M respectively. The approximate $\Delta\Delta G$ between wild-type and 3nc-Ncap is 3.8 kcal mol^{-1}.

Conclusions

Introduction of helix macrodipole and N-capping considerations can drastically impart stability to designed proteins. As the design considerations at the helix termini become more restrictive in the Xnc series, the resulting proteins become more stable. Helix capping effects are also important for designing stable proteins. This is demonstrated best by designed sequence 3nc-Ncap, which has a thermal stability 32 °C higher than that of the wild-type sequence.

The increase in stability in the Xnc series as X increases is not due to an increased helix propensity. Helix propensities were not included in the side chain selection algorithm and there is no correlation between the change in helix propensity of the enh variants and stability using the helix propagation parameters and N-cap propensities from model peptide studies [10,13] or α-helical propensities from statistical analysis of the Brookhaven Protein Data Bank (PDB)[16]. As the scatter plot in Figure 2-5 indicates, the two proteins with the lowest sums of helix and capping propensities are wild-type and 3nc-Ncap, the least stable, and most stable variants in this design study.

The increased stability in the Xnc series may not be due to the standard forcefield parameters, i.e., H-bonds, favorable van der Waals interactions, and better designed electrostatic interactions. If increased stability were due to

these factors, one would expect the protein design algorithm, which uses these terms in its forcefield, to choose the most stable protein in the 0nc calculation, when all surface side chains were available in all positions. In addition, there is no correlation between the stabilities of wt and the Xnc series and amounts of buried hydrophobic area, exposed hydrophobic area or buried polar area. Therefore, the increases in stability are most likely due in large part to the effect of the side chains on the helix dipole.

There is no correlation between computationally predicted energies and experimentally determined stabilities in the Xnc series, shown in the scatter plot in Figure 2-6. As it stands now, the protein design algorithm calculates charges as monopoles, not dipoles. Therefore, simple N-term, C-term, and N-cap groups were introduced to compensate for the helix dipole, rather than introducing a more complex method of determining charge-charge interactions.

The calculation speed of sequences in the Xnc series increases as X increases since more residues are limited to subsets of the surface group. The 0nc calculation, with twenty-nine non-glycine designed positions allowed to be any of the ten surface residues, required 87.8 CPU hours[17]. The same twenty-nine surface positions were designed in the 3nc-Ncap calculation, but fifteen were limited to subsets of the ten residues allowed at surface positions, dramatically decreasing the calculation time to 6.68 CPU hours, a 13-fold reduction. The slow step of the DEE side chain selection algorithm scales as the average number of rotamers per position to the fourth power. Therefore, the removal of side chains (and therefore rotamers) from consideration drastically affects the calculation time, as shown in Table 2-4. Two of the side

chains which were removed, arginine and lysine, are long, and have many rotamers to account for the rotations freedom about their four χ angles. Therefore, introduction of the N-capping and N- and C-terminal restrictions greatly reduces the combinatorial complexity of the problem and therefore calculation time.

Materials and methods

Modeling. The engrailed homeodomain structure coordinates were obtained from the PDB entry 1enh[12]. Residues 1-5 were removed from the structure and explicit hydrogens were added using the program Biograf[18] to residues 6-56. The resulting structure was then minimized for fifty steps using the DREIDING force field[19]. The calculations were performed using potential functions, hydrogen bonding, van der Waals interactions, and solvation parameters as described in previous work[20-24].

To determine which positions of engrailed homeodomain are on the solvent-exposed surface, and therefore candidates for surface optimization, a solvent-accessible surface was generated using the Connolly algorithm with a probe radius of 8.0 Å, a dot density of 10 $Å^{-2}$, and a Cα radius of 1.95 Å[22]. Thirty of the fifty-one residues met the criteria of surface positions: the sum of the distances from the Cαatom to the surface along the Cα-Cβ vector and from the Cβ atom to the surface was less than 2.7 Å. The surface positions are 2, 4, 5, 6, 8, 9, 12, 13, 16, 17, 18, 20, 22, 23, 24, 27, 28, 31, 32, 34, 36, 37, 38, 41, 42, 45, 46, 48, 49, and 50. Residues were classified as core positions if the distance

from the Cα along the Cα-Cβ vector to the surface was greater than 5.0 Å and the distance from the Cβ atom to the surface was greater than 2.0 Å. All remaining residues were classified as boundary positions.

The identity of position 34 was fixed as glycine in all designs (as in the wild-type) because it has a positive phi angle of +86.88°, which is generally inaccessible to other side chains. All 10^{29} possible combinations of the hydrophilic amino acids allowed at the surface were considered for the remaining twenty-nine surface positions in the 0nc calculation. Subsequent surface calculations used the surface group as well as the N-term, C-term and N-cap subsets of the surface group, listed in Table 2-1. In addition, a discrete set of rotamers was considered to account for the torsional flexibility of side chains[25]. Helix propensities were not considered in any of the design calculations of this study so N-capping and helix macrodipole effects would be considered independently. Calculations were typically performed with a Silicon Graphics Origin 2000 computer using twenty R10000 processors in parallel.

Protein Synthesis. All proteins (wild-type, 0nc/1nc, 2nc, 3nc, 2n/3n, and 3nc-Ncap) were synthesized with an Applied Biosystems 433A peptide synthesizer. Preloaded resins (0.08 mmol of substituted residue) were used, with subsequent residues coupled via Fmoc chemistry and HTBU/HOBt activation with standard 0.10 mmol scale coupling cycles. Peptides were cleaved from the solid support resin by mixing 200 mg resin with 2 mL trifluroacetic acid (TFA), 100 μL water, 150 mg phenol, 100 μL thioanisole and 50 μL ethanedithiol for ten hours. The proteins were precipitated by addition

of cold methyl *tert*-butyl ether, washed two times with the same solvent, and lyophilized to partially remove the cleavage reaction scavengers. The proteins were further purified by reverse-phase HPLC with a Zorbax C8 column using linear acetonitrile-water gradients (typically 25-30%) 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 masses before use.

Concentration Determination. Protein concentrations were determined by measuring the absorbence of a 10% stock solution in 6M GdnHCl at 280 nm using an extinction coefficient of 7090 for wild-type and 5700 M^{-1} cm^{-1} for all variants. The mean residue ellipticities of the six proteins measured by CD spectrometry were scaled, which was necessary due to difficulty in concentration determination, which has been observed previously for the enh protein[12].

Circular Dichroism Studies. An Aviv 62A DS spectropolarimeter equipped with a thermoelectric cell holder was used to collect CD data. All data have been scaled. Wavelength scan data were obtained from samples containing 20 μM protein in 5.0 mM sodium phosphate, pH 4.5, using a 1.0 mm path length cell. Wavelength scan data were collected from 250-190 nm at even wavelength intervals for two seconds at 1 °C. Scans at 1 °C after the proteins were exposed to high temperature (99 °C) were not significantly different.

Thermal denaturation curves were recorded at 222 nm from 1-99 °C in two degree intervals with a 0.1 second time constant, 10 second averaging time, 2 minute equilibration time, and 1 nm bandwidth. Melting

temperatures were determined by taking the maximum of a plot of d(ellipticity)/dT versus T after smoothing the data. The error in T_m is estimated to be ± 0.5 °C.

Chemical denaturation data were obtained at 222 nm by titrating in a solution of high concentration denaturant into the CD cuvette, and thereby increasing concentrations in a stepwise manner. Guanidine hydrochloride denaturation curves were typically performed from 0-8.0 M guanidine hydrochloride at 1 °C. Urea denaturation curves were typically performed from 0-9.5 M urea at 35 °C.

To calculate ΔG from chemical denaturation data, the molar fraction of folded protein, *f*, was first calculated from the equation $f = ([\theta]-[\theta]_u)/([\theta]_n-[\theta]_u)$, where [θ] is the observed ellipticity at each denaturant concentration and $[\theta]_n$ and $[\theta]_u$ are the ellipticities of the folded and unfolded states, respectively[15]. Assuming a two-state model, the free energy of unfolding at each denaturant concentration was calculated using the equation $\Delta G_u = -RT\ln(f/1-f)$. ΔG_u in the absence of denaturant was calculated by extrapolating a plot of ΔG_u versus denaturant concentration to 0 M denaturant.

References

1. Pauling, L., Corey, R. & Branson, H. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA* **37**, 205-211 (1951).

2. Aurora, R., Creamer, T., Srinivasan, R. & Rose, G. Local interactions in protein folding: lessons from the α-helix. *J. Biol. Chem.* **272**, 1413-1416 (1997).

3. Walter, S., Hubner, B., Hahn, U. & Schmid, F. Destabilization of a protein helix by electrostatic interactions. *J Mol Biol* **252**, 133-143 (1995).

4. Lockhart, D. & Kim, P. Internal stark effect measurement of the electric field at the amino terminus of an α-helix. *Science* **257**, 947-951 (1992).

5. Lockhart, D. & Kim, P. Electrostatic screening of charge and dipole interactions with the helix backbone. *Science* **260**, 198-202 (1993).

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

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

8. Nicholson, H., Anderson, D., Dao-pin, S. & Mathews, B. Analysis of the interaction between charges side chains and the α-helix dipole using designed thermostable mutants of phage T4 lysozyme. *Biochemistry* **30**, 9816-9828 (1991).

9. Doig, A., Chakrabartty, A., Klinger, T. & Baldwin, R. Determination of free energies of N-capping in α-helices by modification of the Lifson-Roig theory to include N- and C-capping. *Biochemistry* **33**, 3396-3403 (1994).

10. Chakrabartty, A., Doig, A. & Baldwin, R. Helix capping propensities in peptides parallel those in proteins. *Proc Natl Acad Sci USA* **90**, 11332-11336 (1993).

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

12. Clarke, N., Kissinger, C., Desjarlais, J., Gilliland, G. & Pabo, C. Structural studies of the engrailed homeodomain. *Protein Science* **3**, 1779-1787 (1994).

13. Rohl, C., Chakrabartty, A. & Baldwin, R. Helix propagation parameters and N-cap propensities of the amino acids measured in alanine-based peptides in 40 volume percent trifluoroethanol. *Protein Science* **5**, 2623-2637 (1996).

14. Muñoz, V. & Serrano, L. Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. *Proteins: Structure, function, and genetics* **20**, 301-311 (1994).

15. This analysis is based on the assumption that calculation time scales linearly with the number of processors used. The error is approximately 10%.

16. *Molecular Simulations Inc. BIOGRAF* (San Diego, California, 1992).

17. Mayo, S., Olafson, B. & Goddard, W. DREIDING: A generic force field for molecular simulations. *Journal of Physical Chemistry* **94**, 8897-8909 (1990).

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

19. Dahiyat, B., Gordon, D. & Mayo, S. Automated design of the surface positions of protein helices. *Protein Science* **6**, 1333-1337 (1997).

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

21. Dahiyat, B., Sarisky, C. & Mayo, S. *De novo* protein design: towards fully automated sequence selection. *J Mol Biol* **273**, 789-796 (1997).

22. Dayihat, B. & Mayo, S. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA.* **94**, 10172-10177 (1997).

23. Ponder, J. & Richards, F. Tertiary templates for proteins- use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* **193**, 775-791 (1987).

Table 2-1
Residue groups.

Residue group	Side chains
Surface	Ala, Arg, Asn, Asp, Glu, His, Lys, Gln, Ser, Thr
N-term	Ala, Asp, Glu, Asn, Gln, Ser, Thr
C-term	Ala, Arg, Asn, Gln, His, Lys, Ser, Thr
N-cap	Asn, Asp, Ser, Thr

Table 2-2
Sequences and stabilities of designed proteins.

S, c, and b indicate surface, core, and boundary residues, respectively, and h indicates a helical position. Vertical lines indicate core and boundary residues which were not designed. The surface residues in the underlined positions were designed using restricted amino acids groups. M is the number of mutations relative to wild-type and T_m is the thermal denaturation temperature in °C. Residues in black retained wild-type identity in each calculation. Residues in blue were the same in each surface calculation and residues in red changed identities.

	----|----1----|----2----|----3----|----4----|----5-	M	T_m
wt	TAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKI	0	43
0nc	|E|SEK|KR||DE||EKD|R|EER||HD||EK|G|R|E||ER||RR|EQE|	25	43
1nc	|E|SEK|KR||DE||EKD|R|EER||HD||EK|G|R|E||ER||RR|EQE|	25	43
2nc	|E|SEE|KR||DE||RRD|R|TEE||RD||QK|G|R|E||ER||RK|EEQ|	24	73
3nc	|E|SEE|KR||DE||RRD|R|TNQ||HD||QK|G|R|E||ED||EK|EQQ|	25	73
2n/3n	|E|SEE|KR||DE||EKD|R|TEQ||RD||EK|G|R|E||ER||RR|EQE|	23	71
3nc-Ncap	|E|SEE|KR||DE||RRD|R|TEE||RD||QK|G|N|E||ER||RR|EQQ|	23	75
	bsbsssscssbcssbcsssbsbsssbbsscbsscscsssccssccssbsssb		
	hhhhhhhhhhhhh hhhhhhhhhh hhhhhhhhhhhhhhh		

Table 2-3
Salt bridges and hydrogen bonds.

Wild type salt bridges and hydrogen bonds are from crystal structure data[12]. All other interactions are from computational model structures.

Table 2-3a

enh-wt		
salt bridges	R10 E14	intrahelical
(3)	R10 E32	helices 1 & 2
	E14 R25	helices 1 & 2
hydrogen bonds	T1 Q39	loop 1 & helix 3
(5)	Q7 L33	helices 1 & 2
	E14 R25	helices 1 & 2
	N18 R25	loop 2 & helix 2
	R19 R48	loop 2 & helix 3
helix capping	S4 Q7	i, i + 3
(3)	S4 Q7	i + 3, i
	T22 R25	i, i + 3
	N36 A38	i, i + 2
3 violations of helix dipole rules		

Table 2-3b

0nc/1nc		
salt bridges	E5 R6	intrahelical
(15)	E5 R9	intrahelical
	L8 D12	intrahelical
	R10 E 14	intrahelical
	E13 R17	intrahelical
	E14 R25	helices 1 & 2
	E16 K17	intrahelical
	D18 R25	loop 2 & helix 2
	R20 E48	loop 2 & helix 3
	E22 R24	loop 2 & helix 2
	R26 E41	helices 2 & 3
	D28 K32	intrahelical
	R36 E41	loop 3 & helix 3
	E41 R45	intrahelical
	R46 E50	intrahelical
hydrogen bonds	T1 Q39	loop 1 & helix 3
(4)	Q7 L33	helices 1 & 2
	H27 E37	helices 2 & 3
	N39 R42	intrahelical
helix capping (1)	S4 Q7	i, i + 3
5 violations of helix dipole rules		

Table 2-3c

2nc		
salt bridges	E5 K9	intrahelical
(13)	E6 R9	intrahelical
	R10 E14	intrahelical
	D12 K47	helices 1 & 3
	E13 R17	intrahelical
	R20 E40	loop 2 & helix 3
	Q21 K32	loop 2 & helix 2
	E24 R27	intrahelical
	R26 E41	helices 2 & 3
	D28 K32	intrahelical
	E28 R36	helix 2 & loop 3
	E41 R45	intrahelical
	K46 E49	intrahelical
hydrogen bonds	D18 R20	loop 2
(3)	S30 L35	helix 2 & loop 3
	Q39 R42	intrahelical
helix capping	S4 Q7	i, i + 3
(2)	T22 R25	i, i + 3
1 violation of helix dipole rules		

Table 2-3d

3nc		
salt bridges	E5 K8	intrahelical
(13)	E6 R9	intrahelical
	R10 E14	intrahelical
	D12 K47	helices 1 & 3
	E13 K17	intrahelical
	E14 R25	helices 1 & 2
	D18 R25	loop 2 & helix 2
	R20 E48	helices 2 & 3
	R26 E41	helices 2 & 3
	D28 K32	intrahelical
	R36 E38	loop 3 & helix 3
	D42 K46	intrahelical
	E45 K46	intrahelical
hydrogen bonds	D18 R20	intrahelical
(5)	D23 Q24	intrahelical
	H27 E37	intrahelical
	S30 L35	helix 2 & loop 3
	Q49 Q50	intrahelical
helix capping	S4 Q7	i, i + 3
(2)	T 22 R25	i, i + 3
0 violations of helix dipole rules		

Table 2-3e

2n (3n)		
salt bridges	E5 K8	intrahelical
(16)	E6 R9	intrahelical
	R10 E14	intrahelical
	D12 K47	helices 1 & 3
	E13 K17	intrahelical
	E14 R25	helices 1 & 2
	E16 K17	intrahelical
	D18 K25	loop 2 & helix 2
	R20 E48	helices 2 & 3
	R26 E37	helices 2 & 3
	R26 E41	helices 2 & 3
	D28 K32	intrahelical
	E31 K32	intrahelical
	R36 E38	loop 3 & helix 3
	E41 R45	intrahelical
	R46 E50	intrahelical
hydrogen bonds	Q24 R27	intrahelical
(3)	S30 L35	helix 2 & loop 3
	Q39 R42	intrahelical
helix capping	S4 Q7	i, i + 3
(1)		
3 violations of helix dipole rules		

Table 2-3f

3nc-Ncap		
salt bridges	E5 K8	intrahelical
(14)	E6 R9	intrahelical
	R10 E14	intrahelical
	D12 K47	helices 1 & 3
	E13 R17	intrahelical
	E14 R25	helices 1 & 2
	D18 R25	loop 2 & helix 2
	R20 E48	helices 2 & 3
	E24 R27	intrahelical
	R26 E37	helices 2 & 3
	R26 E41	helices 2 & 3
	D28 K32	intrahelical
	E38 R42	intrahelical
	E41 R45	intrahelical
hydrogen bonds	Q7 L33	helices 1 & 2
(4)	S30 L35	helix 2 & loop 3
	Q31 K32	intrahelical
	R46 Q50	intrahelical
helix capping	S4 Q7	i, i + 3
(3)	T22 R25	i, i + 3
	N36 E38	i, i + 2
0 violations of helix dipole rules		

Table 2-4
Sequence diversity, number of rotamers, and calculation time for designed molecules.

The data for 3nc were not included because the DEE calculation was performed with different parameters.

	number of sequences	number of rotamers	calculation time (CPU hours)
wt	1	--	--
0nc, 1nc	1E29	5.60E67	87.78
2n, 3n	1.18E28	6.73E66	39.48
2nc	2.20E28	4.05E66	7.43
3nc-Ncap	1.97E28	2.85E64	6.68

Figure 2-1

The helix-turn-helix motif of the 51-residue engrailed homeodomain[12].

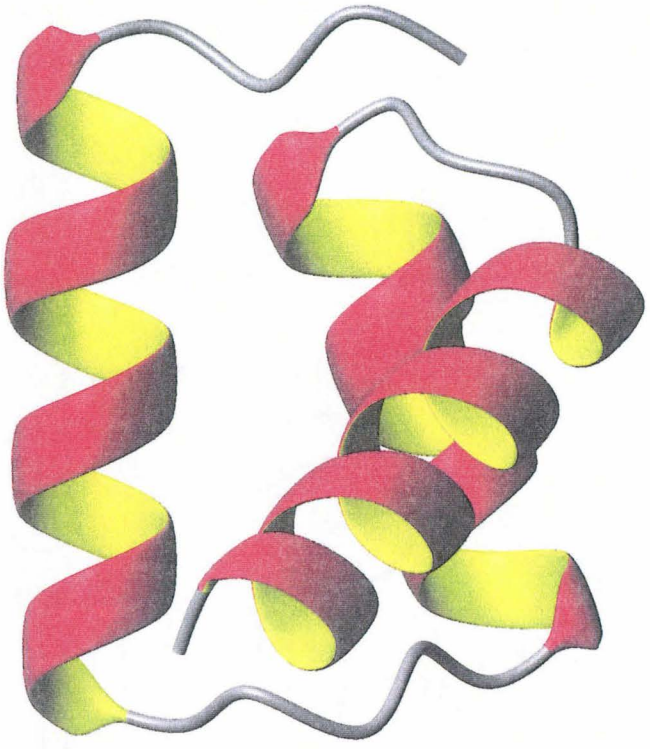
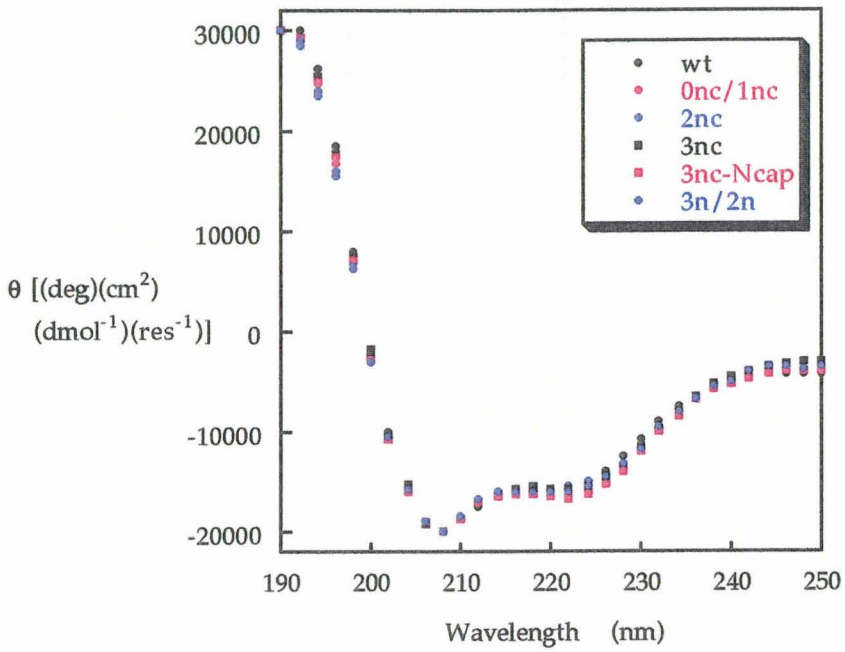


Figure 2-2

CD wavelength scan and thermal denaturation curves.

Wavelength scan data were obtained at 1 °C. Thermal denaturation curves were obtained by monitoring the CD signal at 222 nm.


30000
20000
10000
0
-10000
-20000
θ [(deg)(cm²)
(dmol⁻¹)(res⁻¹)]
190
200
210
220
230
240
250
Wavelength (nm)
wt
0nc/1nc
2nc
3nc
3nc-Ncap
3n/2n

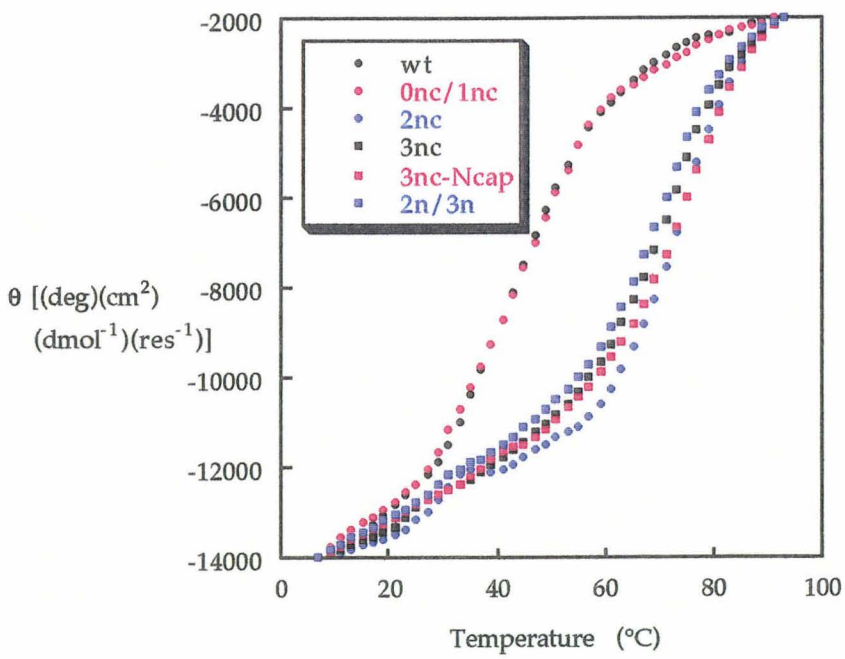

-2000
-4000
-6000
-8000
-10000
-12000
-14000
θ [(deg)(cm²)
(dmol⁻¹)(res⁻¹)]
0
20
40
60
80
100
Temperature (°C)
wt
0nc/1nc
2nc
3nc
3nc-Ncap
2n/3n

Figure 2-3

Guanidine hydrochloride denaturation curves of wild-type engrailed homeodomain and 3nc-Ncap.

The measured ΔG of wild-type and 3nc-Ncap are 4.1 and 3.0 kcal mol^{-1} respectively. These data were obtained at 1 °C at 222 nm.

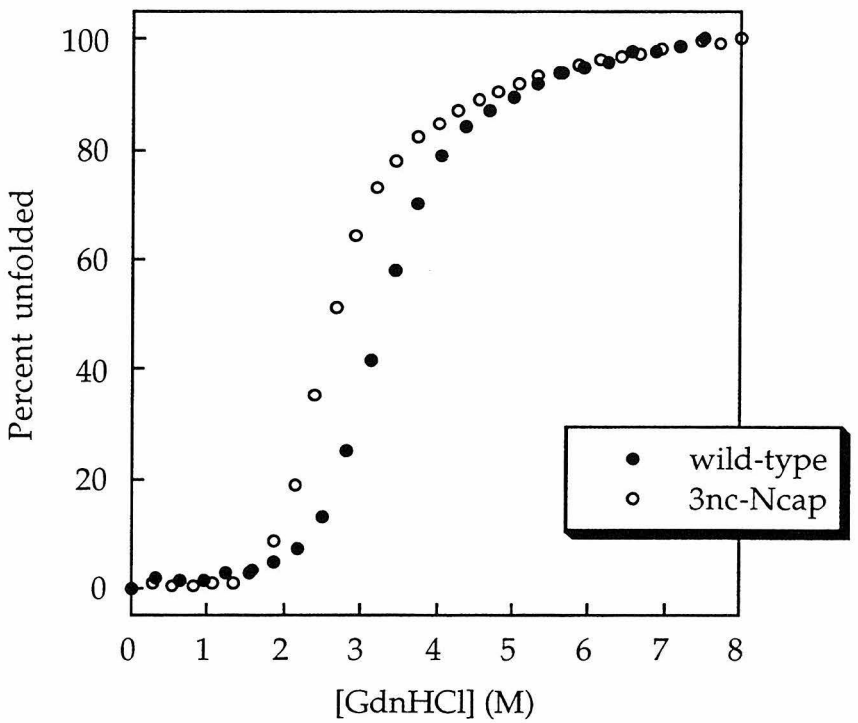

Percent unfolded
100
80
60
40
20
0
0
1
2
3
4
5
6
7
8
[GdnHCl] (M)
wild-type
3nc-Ncap

Figure 2-4

Urea denaturation curves of wild-type engrailed homeodomain and 3nc-Ncap.

The measured ΔG of wild-type and 3nc-Ncap are 1.2 and 5.0 kcal mol^{-1} respectively. These data were obtained at 35 °C at 222 nm.

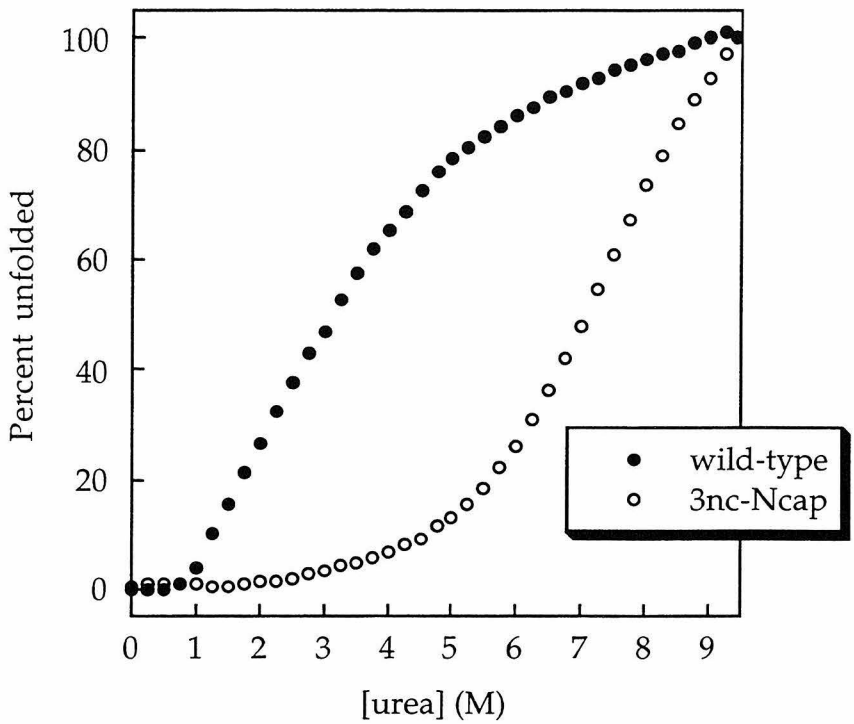

100
80
60
40
20
0
Percent unfolded
0 1 2 3 4 5 6 7 8 9
[urea] (M)
wild-type
3nc-Ncap

Figure 2-5

Correlation between helix and capping propensities and thermal stability.

The difference in helix and N-cap propensities of enh variants do not correlate to stability, demonstrating that stability is not due to an increases in the sum of helix and helix-capping propensities. The difference in propensities were determined by taking the sum of the propensities of the surface residues, since all other residues are the same in each protein. The helix propagation parameters used are from studies of model peptides[13].

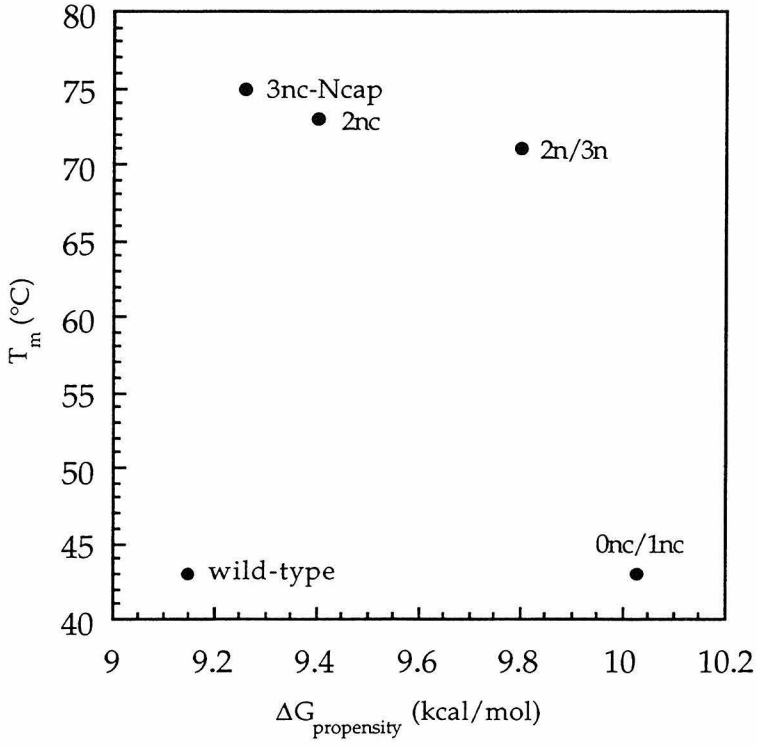
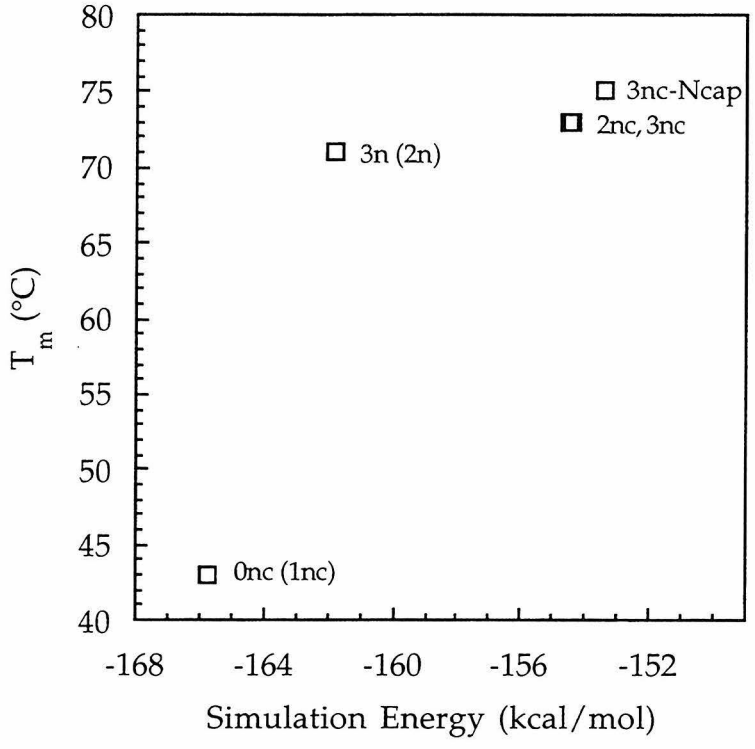

80
75
70
65
60
55
50
45
40
T_m (°C)
3nc-Ncap
2nc
2n/3n
0nc/1nc
wild-type
9
9.2
9.4
9.6
9.8
10
10.2
$\Delta G_{propensity}$ (kcal/mol)

Figure 2-6

Thermal stability versus simulation energy.


80
75
70
65
60
55
50
45
40
T_m (°C)
3nc-Ncap
2nc, 3nc
3n (2n)
0nc (1nc)
-168
-164
-160
-156
-152
Simulation Energy (kcal/mol)

Chapter 3

Full Sequence Design of the Engrailed Homeodomain Fold and Structure Determination of the Resulting Hyperthermophile

Abstract

The protein design cycle was extended to the full sequence design of engrailed homeodomain (enh). The combinatorial complexity of the 51-mer was sufficiently high that it required a two-step calculation. The first design included the forty surface and core residues, and was performed using all helix dipole and capping restrictions (described in Chapter 2). The T_m of the resulting molecule is 81 °C, 38 ° higher than that of wild type. The second calculation included the eleven boundary residues and fixed the identity of the forty surface and core residues to the identity selected in the first calculation. The resulting fully designed molecule, bsc, is a 39-fold mutant of wild type and has a T_m of 114 °C and is 4.7 kcal mol^{-1} more stable than wild-type. The structure of bsc was determined by NMR techniques. The top 10 structures have a backbone root-mean-square standard deviation of 0.45 Å for well-ordered residues 2-48. There is excellent agreement between the target structure and the experimental structure for residues 9-48; the root-mean-square deviation between the mean structure and model backbone is 1.22 Å.

Introduction

Drosophila engrailed homeodomain (enh) was selected as a target fold for protein design because it is almost twice as large as our previous full sequence design target[1], a good quality crystal structure of the wild-type is available[2], and the protein is small enough to be easily synthesized and expressed. The crystal structure of the first fifty-six of the sixty-one amino

acid protein has been solved at a resolution of 2.1 Å[2]. We opted to design the surface residues of a 51 amino acid variant containing residues 6-56 (subsequently renumbered 1-51 and referred to as wild-type). The renumbered structure is comprised of three helices, residues 5-17, 23-32, and 37-51.

The full sequence design of a stable protein with the same fold as enh was the ultimate goal of this protein design project. The surface positions of enh were optimized as described in Chapter 2, so attention was turned to the remaining twenty-one non-surface residues. First, the core residues were designed concurrently with the surface residues. Second, the boundary residues were designed in the context of the surface-core design.

Surface-core design

Once the utility of terminal helix and N-cap restrictions had been established, the surface design was augmented to include the ten core residues (7, 11, 15, 29, 33, 35, 39, 40, 43, 44). The core positions were restricted to the hydrophobic residues Ala, Ile, Leu, Phe, Trp, Tyr, and Val. Of the ten core positions, seven retained the same identity as in the wild-type sequence, position 35 had a conservative mutation from Leu to Ile, and positions 7 and 39 both changed from polar Gln to hydrophobic Leu. The thirty surface positions were designed as in the 3nc-Ncap calculation, utilizing the amino acid groups N-term, C-term, and N-cap at positions N1-N3, C3-C1, and N-cap positions respectively. There were eight surface mutations relative to the 3nc-Ncap sequence to compensate for changes in the core. Of the forty designed positions in the surface-core design (sc1), there are twenty-nine mutations

relative to the wild-type sequence listed in Table 3-1. The model structure of sc1 suggests that it is stabilized by more salt bridges and hydrogen bonds. Sc1 has four more putative salt bridges and five more putative H-bonds than wild-type, indicated in Tables 3-2a and b. Sc1 was designed with residues capable of making N-capping hydrogen bonds at all N-cap positions, but the model structure does not have any predicted helix caps.

The secondary structure of sc1 is the same as wild-type, shown by its CD wavelength scan in Figure 3-1. The thermal denaturation curve of sc1 compared to wild-type is shown in Figure 3-2. The T_m of sc1 is 81 °C, six degrees higher than the optimal surface design, indicating that surface-core design is more successful than surface design alone. The sc1 calculation is more than twice as CPU-intensive as 3nc-Ncap; the number of processor hours increased from 6.68 to 14.48.

Full sequence design

A full sequence design calculation was attempted by including the remaining eleven boundary residues (1, 3, 10, 14, 19, 21, 25, 26, 30, 47, 51) in the surface-core calculation. The surface residues were designed as in the 3nc-Ncap design, the core residues were designed as described in the sc1 calculation, and the remaining eleven boundary residues were selected from residues Ala, Asn, Asp, Gln, Glu, His, Ile, Leu, Lys, Ser, Thr, and Val[3]. In this calculation, the protein design algorithm must search through 4.14×10^{46} unique amino acid sequences[4]. The virtual combinatorial problem is much greater than the number of unique sequences, since discrete sets of allowed

side chain conformations, or rotamers[5-7], are considered at each position raising the number of unique solutions to 3.28×10^{102}. Even with very aggressive parameters, the calculation did not converge to a solution in a reasonable amount of time.

The combinatorial complexity of the calculation was made tractable by computing the full sequence design stepwise. The ten boundary positions were designed in the context of the surface-core design. The identities of surface and core positions were fixed, but the rotamers of the side chains were allowed to vary. The boundary positions were allowed to be any of the twelve boundary residues listed above. This greatly reduced boundary residue calculation resulted in 7.43×10^{11} unique amino acid sequences and 3.12×10^{73} unique rotamer solutions which converged to a solution for the ten boundary positions in 9.0 CPU hours, or 19.2 minutes on 28 processors running in parallel.

The resulting molecule, the boundary-surface-core design (bsc), was designed at all 51 positions and is a 39-fold mutant from the wild type. It is considerably more polar than wild-type, as shown by its sequence in Table 3-1. Bsc has the same secondary structure as wild-type, as shown by its CD spectrum in Figure 3-3.

Stability of the full sequence design

Thermal denaturation CD experiments indicated that bsc was not completely unfolded at 99 °C (Figure 3-4). The T_m was determined to be 114

°C using a -microcalorimeter with a pressurized cell holder. The free energy of unfolding (ΔG) of bsc was determined to be 6.2 kcal mol^{-1} at 1 °C by guanidine hydrochloride (GdnHCl) denaturation experiments, shown in Figure 3-5. Bsc is likely to be more than 1.7 kcal mol^{-1} more stable than wild-type enh since chemical denaturation of proteins by GdnHCl masks salt bridge interactions which may stabilize the protein because of the ionic nature of GdnHCl[8]. Since bsc contains four more putative salt bridges than wild-type, it is more stabilized with respect to wild type than the GdnHCl chemical denaturation results suggest (Table 3-2c). In addition, it was not possible to calculate ΔG from calorimetric experiments because the T_m of bsc is sufficiently high (114 °C) that a post transition baseline of the unfolding transition was unobtainable because of instrument temperature limitations, i.e., 130 °C, shown in Figure 3-6. ΔGs were determined by urea denaturation at 30, 45, and 60 °C to be 5.9, 5.5, and 5.5 kcal mol^{-1} respectively. Figure 3-7 shows a comparison of urea melts of wild type and bsc. The $[urea]_{1/2}$ value of wt at 35 °C is 1.75 M and the $[urea]_{1/2}$ value of scb is 7.75 M at 30 °C. The approximate ΔΔG between wild-type and bsc is 4.7 kcal mol^{-1}.

Assessment of the full sequence design

For NMR studies, bsc was expressed, rather than synthesized because the synthesis of homeodomain variants has not resulted in high yields of protein. In addition, synthesis of ^{15}N labeled protein is prohibitively expensive. All bsc was expressed with an N-terminal methionine, which does

not appear to alter the secondary structure of the protein. Thermal denaturation studies of 3nc-Ncap indicated that the stability of the protein with and without an N-terminal methionine was the same. Both forms of bsc are hyperthermophilic, i.e., the T_ms are greater than 99 °C, and could not be measured by CD thermal denaturation techniques.

To assess the success of the full sequence design, the structure of bsc was determined using two- and three-dimensional NMR techniques for comparison to the target enh crystal structure. The spectra were well dispersed in the amide proton dimension, but not in the alpha proton dimension, typical of helical proteins. Therefore, three-dimensional experiments were performed to make chemical shift assignments, listed in Table 3-3. The fingerprint region of the TOCSY spectrum, Figure 3-8, shows substantial Cα proton overlap. A heteronuclear single quantum coherence spectrum (HSQC)[36] of ^{15}N-labeled bsc was taken to obtain ^{15}N chemical shift data of the spin system roots, shown in Figure 3-9. Three-dimensional TOCSY and NOESY spectra improved chemical shift dispersion.

The NMR solution structure of bsc was determined from 619 experimental restraints (12.1 restraints/residue), including 575 NOE distance restraints, 7 χ1 restraints, 24 ϕ restraints, and 13 hydrogen bonds involving slowing exchanging amide protons, listed in Table 3-4, and shown graphically in Figure 3-10. Sequential and short-range NOEs suggest helical structure for residues 9-17, 23-33, and 36-47, as evidenced by the NOE plot in Figure 3-11a. Structure calculations were performed using X-PLOR[9] with standard protocols for hybrid distance-geometry simulated annealing[10-12].

An ensemble of 40 structures converged with no distance restraint violations greater than 0.3 Å, good covalent geometry, and 97.1% of all residues in allowed regions of the Ramachandran map, listed separately in Table 3-5. The structure of bsc is that of a helix-turn-helix motif indicative of a homeodomain fold, shown in Figure 3-12. The top 10 structures have a backbone root-mean-square (r.m.s.) deviation of the ordered residues 5-48 of 0.45 Å which increases to 0.66 Å for the 40-structure ensemble. The r.m.s. deviation of all heavy atoms in the well-ordered residues 2-48 is 1.68 Å for the ensemble of 40 structures. The termini of the protein are frayed, typical for NMR solution structures; the r.m.s. deviation of backbone atoms of residues 1-51 is 1.15 Å which increases to 2.18 for all heavy atoms, shown in Table 3-5 and Figure 3-13.

Comparison of the backbone of the average restrained minimized structure[12,13] to the wild-type engrailed homeodomain crystal structure reveals excellent agreement for residues 9-48 with an r.m.s. deviation of 1.22 Å. Figure 3-14 shows that the backbones of bsc and wild-type superimpose in the two loops and helices 2 and 3. The ϕ and ψ angles of residues 9-48 in the ensemble of NMR structures and the crystal structure overlap well. Figure 3-15 shows that the majority of the ϕ and ψ angles in the ensemble are clustered around the wild-type values.

The structure of bsc reveals a difference between wild-type and bsc in helix 1. The N-terminal residues of bsc pack against helix 1 rather than adopt a helical conformation as in the model backbone from the crystal structure of enh[14]. The crystal structure of the wild-type protein was determined with an

additional five residues at the N-terminus. Therefore, the target backbone may be incorrect, that is, the structure of 6-56 residue fragment of enh does not necessarily begin with a helix at position 5 and that the initial, non-helical N-terminal residues are responsible for the helix initiation. There are no known homeodomain structures which do not begin with the N-terminal arm that was truncated in our backbone template. In addition, structures of homeodomains typically begin with disordered N-terminal arms which are absent in crystal structures because they are not visible in electron density maps[15,16] or are under-restrained and floppy in NMR structures[17,18]. However, the lack of agreement between residues 5-9 of helix 1 of the target homeodomain backbone and solution structure of bsc could have been circumvented by introduction of negative design into our side chain selection algorithm.

Comparison of the model bsc structure generated by the side chain selection algorithm and NMR solution structure show some differences in side chain packing of core residues. Of the three aromatic residues, W43 and F44 adopt similar conformations to the predicted rotamers and F15 does not, easily shown by the distribution of χ_1 angles in Figure 3-16. The side chain conformation of the model W43 is indistinguishable from those in the ensemble and the χ_1 angles of the model and average restrained minimized structure are almost identical (i.e., they differ by 5°). The ring of F44 is within the standard deviation of our rotamer library. The side chain of F15 is less exposed than in the model structure; it is rotated 98° along χ_1 bringing the

ring into contact with F44 in the core of the experimental structure. The comparison of core residues of the average minimized structure SA and the model structure in Figure 3-17 shows that the aromatic residues cluster closer together than expected. Of the remaining core residues, L33 and I40 have very similar conformations in the predicted and solution structures; I35 does not (Figure 3-18). L29 and L39 do not have well ordered side chains and L7 and L11 can not be compared to the predicted side chain rotamers because the backbones fail to converge at those points.

The design of bsc included restrictions at the solvent-exposed N-cap positions such that only residues capable of making N-capping hydrogen bonds (Asn, Asp, Ser, Thr) were allowed at these positions as described in Chapter 2. The crystal structure of wild-type enh shows that each surface-exposed N-cap position side chain makes a hydrogen bond with the i + 1 residue[2]. In the model structure, none of the three selected rotamers at these positions make hydrogen bonds. However, in the NMR structure, there is weak evidence that D4 and T22 are hydrogen bond acceptors. Although not in a helical conformation, the side chain of D4 makes a hydrogen bond to the amide proton of E5 in 14 of the 40 structures, shown in Figure 3-19a. In the ensemble of 40 structure, 16 have a putative hydrogen bond in between the side chain oxygen of T22 and the amide proton of N23, shown in Figure 3-19b, and 10 have a putative hydrogen bond between the side chain oxygen of T22 and the amide proton of L25, illustrated in Figure 3-19c. The side chain conformation of N-cap residue N36 is not well resolved, as shown by the large distribution of χ_1 angles in Figure 3-16.

The agreement between the calculated model structure and experimentally determined structure of bsc demonstrates the ability of our protein design algorithm to select a sequence which folds into a desired tertiary fold with enhanced stability. Engrailed homeodomain, a 51-residue protein, is nearly twice as large as our previous full sequence design target, fsd1, a 28-residue ββα motif based on Zif268[1]. This is the first application of our design algorithm to a helix-turn-helix motif and demonstrates the generality of our approach which was developed using physical chemical principles and other motifs as test cases.

Factors which contribute to stabilization

The full sequence design of engrailed homeodomain resulted in a hyperthermophilic protein with a melting temperature of 114 °C. The increase from the wild-type stability of 43 °C is not due to an increase in helix propensity (discussed in Chapter 2). Studies of stabilizing factors of hyperthermophilic proteins have identified surface ion pair networks, in particular those connecting different units of secondary structure, as an important difference between hyperthermophiles and their mesophilic counterparts[19] although this is not true for all hyperthermophilic proteins[20]. Typically, hyperthermophilic proteins have an increased number of favorable helix capping, helix dipole, and hydrophobic interactions[21]. The stability of bsc is most likely due to surface salt bridges, helix dipole stabilizing interactions, and perhaps N-capping interactions. The model structure has seven salt bridges and five hydrogen bonds, tabulated in Table 3-2c.

Compared to wild-type, bsc has approximately 20% more exposed hydrophobic area and the same amount of buried hydrophobic area.

Using our side chain selection algorithm based on the DEE theorem in concert with experimental validation has allowed us to design a hyperthermostable protein which adopts a target fold without the use of disulfide bridges, metal binding sites, or oligamerization. Wild-type enh (residues 1-51) has a melting temperature of 43 °C. Optimal surface design resulted in the 3nc-Ncap sequence, which has an experimentally determined T_m of 75 °C. We determined that including N-capping and helix dipole effects in our sequence selection rules greatly enhances our ability to design stable proteins (discussed in Chapter 2). Concurrent surface and core design results in a sequence with even greater thermostability; sc1 has a melting temperature of 81 °C. Finally, addition of the boundary residues results in a protein which is a 39-fold mutant of wild type and has a melting temperature of 114 °C. Since the surface, surface-core, and boundary-surface-core designs of enh are progressively more stable, the stability of bsc must arise from the design of all positions of the protein, and not just the core positions, for example. Using a quantitative design method based on physical chemical principles we are able to successfully design proteins using the power of computational techniques.

Future directions

The logical extension of the present work is the design of a functional mutant of enh with increased thermal stability. Engrailed homeodomain is a transcription factor, which binds to DNA as shown in Figure 1-1. The

residues which recognize DNA are in the N-terminal arm and helix 3, the recognition helix.

In this design study, the protein was optimized for stability, not function. The N-terminal arm, which binds in the minor groove of DNA, was truncated, and the residues of helix 3 which contact the DNA were redesigned. The simplest approach to design a more stable DNA-binding variant of enh is to design only helices 1 and 2 of the protein with the N-terminal arm. For a slightly more sophisticated approach, the residues which contact DNA could maintain their identity while the rest of the protein is redesigned. These residues include R5, I47, Q50, and N51, which make direct contact with DNA base pairs, and T3, A4, A5, K46 Q50, A54 which make water-mediated contacts with DNA bases, and T6, Y25, R31, R53, K46, W48, and R53, which contact the DNA backbone in the major groove[22,23].

If it is possible to design a more thermostable DNA-binding enh variant as described above, we will include the residues which contact DNA into the design algorithm with the DNA. The protein can be redesigned to optimize DNA-binding contacts, and potentially increase binding affinity. If this is successful, the enh fold could be redesigned to bind to another sequence of DNA.

Materials and Methods

Computational modeling. The calculations were performed as described in previous work[1,24-27].

Protein Synthesis. Sc1 and bsc were synthesized with an Applied Biosystems 433A peptide synthesizer. Preloaded resins (0.08 mmol of

substituted residue) were used, with subsequent residues coupled via Fmoc chemistry and HTBU/HOBt activation with standard 0.10 mmol scale coupling cycles. Peptides were cleaved from the solid support resin by mixing 200 mg resin with 2 ml trifluroacetic acid (TFA), 100 μL water, 150 mg phenol, 100 μL thioanisole and 50 μL ethanedithiol for ten hours. The proteins were precipitated by addition of cold methyl *tert*-butyl ether, washed two times with the same solvent, and lyophilized to partially remove the cleavage reaction scavengers. The proteins were further purified by reverse-phase HPLC with a Zorbax C8 column using linear acetonitrile-water gradients (typically 25-30%) 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 masses before use.

Protein Expression. The bsc gene was synthesized from four primers by recursive PCR[28] and cloned into pET-11a (Novagen). Unlabeled bsc was expressed in BL21(DE3) *E. coli* cells (Strategene) by IPTG induction[29] and isolated using freeze-thaw methods[30]. The bsc gene was designed with codons used by *E. coli* in highly expressed genes to encourage high expression levels. All of the bsc protein was expressed with an N-terminal methionine and purified by HPLC as described above. ^{15}N labeled bsc was expressed in cells in M9 minimal media using $^{15}NH_4Cl$ as the sole nitrogen source. Protein masses were determined by MALDI-TOF or electrospray mass spectrometry and were found to be within one mass unit of the expected masses before use.

CD Wavelength Scans and Denaturation Curves. An Aviv 62A DS spectropolarimeter equipped with a thermoelectric cell holder was used to collect CD data. Wavelength and temperature scan data were obtained from samples containing 20 μM protein in 5.0 mM sodium phosphate, pH 4.5, using a 1.0 mm path length cell. Wavelength scan data were collected from 250-190 nm at even wavelength intervals for two seconds at 1 °C. Scans at 1 °C after the proteins were exposed to high temperature (99 °C) were not significantly different. Thermal denaturation curves were recorded at 222 nm from 1-99 °C in two degree intervals with a 0.1 second time constant, 10 second averaging time, 2 minute equilibration time, and 1 nm bandwidth. Melting temperatures were determined by taking the maximum of a plot of d(ellipticity)/dT versus T after smoothing the data. The error in T_m is estimated to be ± 0.5 °C. All data have been scaled.

Chemical denaturation data were obtained at 222 nm by titrating in a solution of high concentration denaturant into the CD cuvette, thereby increasing concentrations in a stepwise manner. Samples were equilibrated for 240 seconds and data were collected for 300 seconds. Guanidine hydrochloride denaturation curves were typically performed from 0-8.0 M guanidine hydrochloride at 1 °C. Urea denaturation curves were typically performed from 0-9.5 M urea at 35 °C. To calculate ΔG from chemical denaturation data, the molar fraction of folded protein, *f*, was first calculated from the equation $f = ([\theta]-[\theta]_u)/([\theta]_n-[\theta]_u)$, where [θ] is the observed ellipticity at each denaturant concentration and $[\theta]_n$ and $[\theta]_u$ are the ellipticities of the

folded and unfolded states, respectively[15]. Assuming a two-state model, the free energy of unfolding at each denaturant concentration was calculated using the equation $\Delta G_u = -RT\ln(f/1-f)$. ΔG_u in the absence of denaturant was calculated by extrapolating a plot of ΔG_u versus denaturant concentration to 0 M denaturant.

Calorimetry. Differential scanning calorimetry experiments of bsc were done with a Calorimetry Sciences Corporation N-DSC-II calorimeter. A solution of bsc was dialyzed overnight against pH 4.5 mM sodium phosphate and degassed before use. The experiment was performed with 2.2 mg/ml protein under 3 atm pressure. The temperature cycled between 25 °C and 130 °C at a scan rate of 1 °C/min. The melting temperature of bsc is too high to obtain a post transition baseline and therefore precludes a complete thermodynamic analysis by calorimetry.

Protein Preparation for the HX Experiment. A homonuclear sample of lyophilized bsc was dissolved to 1 mM in 700 µl 5 mM sodium phosphate in H_2O and adjusted to pH 4.5. The sample was relyophilized and brought up in 700 µl D_2O, the pH was quickly readjusted to pH 4.5 and the experiment was started soon afterwards to avoid missing quickly exchanging peaks.

NMR Structure Determination. NMR samples were prepared in 5 mM sodium phosphate solution at pH 4.5 (uncorrected) in H_2O:D_2O (90:10) or 99.9% D_2O with 1.0-1.3 mM concentrations. Self-association of bsc was

observed at concentrations higher than 1.3 mM. NMR spectra were collected at 25 °C using a Varian UnityPlus 600 MHz spectrometer equipped with a Nalorac inverse probe with a self-shielded z-gradient.

A series of NMR experiments were performed to determine the structure of bsc. Spin system root assignments were made with homonuclear dqf-COSY[31], TOCSY[32], and NOESY[33] experiments. Spectra were processed with VNMR[34] and crosspeaks were assigned using the program ANSIG[35] (Table 3-3).

150 ms mixing time NOESY experiments were used for crosspeak assignment. In addition, 100 ms mixing time heteronuclear NOESY data was collected to use for peak area integration since the ratio of peak intensities to atomic distances are more constant than in the analogous 150 ms mixing time NOESY experiment. The 150 ms NOESY assignments were transcribed into the 100 ms NOESY data.

A heteronuclear single quantum coherence spectrum (HSQC)[36] of ^{15}N-labeled bsc was taken to obtain ^{15}N chemical shift data of the spin system roots, shown in Figure 3-9. Three-dimensional HSQC-TOCSY and HSQC-NOESY experiments[37] were obtained to reduce overlap and facilitate chemical shift determination, listed in Table 3-3. The three-dimensional experiments simplify 2D data, allowing for increased NOE assignments. With 3D data, more NOE data was made available, illustrated simply in the secondary structure tables Figures 3-11a, 3-11b. Three-dimensional NOE assignments were used to assign ambiguous 150 ms NOESY crosspeaks.

To begin the calculations of possible structures, the homonuclear 100 ms NOESY and heteronuclear three-dimensional NOESY spectra were integrated using ANSIG. Stereospecific assignments were removed from β, γ, and δ protons and a restraints file was generated. NOE crosspeaks were binned into categories of strong, medium, and weak (Table 3-4). Artificially high intensities due to peak overlap were adjusted[38].

Hydrogen bond restraints were obtained by performing a hydrogen exchange (HX) G-COSY experiment. Peak intensities were monitored for 48 hours. The resulting intensity decays were fitted with exponential curves to obtain the protection factors listed in Table 3-6. Hydrogen bonds were only assigned if the protons had protection factors greater than 100, were involved in backbone-backbone interactions, were in units of regular secondary structure (i.e., helices), and there were supporting NOEs such as $\alpha N_{i,i+4}$ crosspeaks. There were thirteen hydrogen bonds that met these criteria, mostly in helices 2 and 3. Each hydrogen bond allowed for two restraints: an N-O distance of 2.4-3.5 Å and H_N-O distance of 1.5-2.8 Å. Hydrogen bond restraints are tabulated in Table 3-6 and shown in Figure 3-10.

Two experiments were performed to obtain ϕ restraints. The first, a 2D-NOESY experiment, did not yield useful data since the all-helical bsc does not have good chemical dispersion and the measurement of coupling constants is limited by peak overlap. In addition, the small coupling constants typical for helices are not easily determined by measuring linewidths. To circumvent these problems, a 3D HNHA quantitative J

correlation experiment was performed[39]. In this experiment, the ratio of intensities of the H_α-N-H_N and H_N-N-H_N crosspeaks were used to calculate coupling constants. For ratios less than 0.20, ϕ angles were restrained to α-space. Similarly, ratios greater than 0.64, typical for extended conformations, were retrained to β–space. The HNHA experiment is well suited to bsc because it overcomes the lack of dispersion by separation in the ^{15}N dimension and because it relies on measurement of peak intensities, rather than linewidths. Twenty-four ϕ restraints were obtained, listed in Table 3-7 and shown in Figure 3-10. Twenty-three are helical ϕ angles.

χ_1 restraints were obtained from an ECOSY experiment[40]. Linear combinations of multiple quantum filtered COSY spectra were used to simplify in-phase multiplet components from passive couplings which distort coupling constants after which $\alpha\beta$ coupling constants can be measured directly. This technique requires resolved $\alpha\beta$ crosspeaks, nondegenerate Hβ1 and Hβ2 chemical shifts, and coupling constants large enough to measure. Seven χ_1 restraints were obtained, shown in Table 3-8 and Figure 3-10.

A series of possible structures corresponding to the list of restraints were generated by the program X-PLOR[9] using standard protocols for hybrid distance-geometry simulated annealing[10-12]. An ensemble of regularized structures was made by substructure embedding followed by 18 ps of high temperature (2000 K) dynamics followed by 50 ps of cooling to low

temperature (100 K). This ensemble was refined by 2 cycles of 75 ps cooling to 100 K, beginning at 1000 K (first 2 cycles) and one cycle of 125 ps at 500 K (subsequent cycle) followed by 500 steps of conjugate gradient minimization. Nonbonded contacts were optimized by a quartic repulsive potential and the final REPEL radius was 0.8 Å. NOE-derived distance restraints were utilized with a force constant of 50 kcal mol^{-1} $Å^{-2}$; a force constant of 200 kcal mol^{-1} rad^{-2} was used for dihedral restraints. Peak assignments and X-PLOR runs were done iteratively to assign previously unassigned NOE crosspeaks. 100 distance geometry structures were calculated which resulted in an ensemble, <SA>, of 41 structures with no restraint violations greater than 0.3 Å, r.m.s. deviations from idealized bond angles and impropers less than 1°, or r.m.s. deviations from ideal bond lengths less than 0.01 Å after regularization and refinement.

Acknowledgements

I would like to thank Scott Ross for assistance with the NMR data collection and interpretation. In addition, I would like to thank Catherine Sarisky for generous help with ANSIG and interpretation.

References and Notes

1. Dahiyat, B. & Mayo, S. De novo protein design: fully automated sequence selection. *Science* **278**, 82-87 (1997).
2. Clarke, N., Kissinger, C., Desjarlais, J., Gilliland, G. & Pabo, C. Structural studies of the engrailed homeodomain. *Protein Science* **3**, 1779-1787 (1994).

3. Initially, the boundary calculation allowed Phe, Trp, and Tyr residues at boundary positions in addition to those listed. The side chain selection algorithm selected five aromatic residues at boundary positions, possibly because the aromatic side chains are over-represented in the rotamer library.

4. Ten core positions (7 possible amino acids), 11 boundary positions (12 possible amino acids), 6 N-term surface positions (7 possible amino acids), 6 C-term positions (8 possible amino acids), 3 N-cap positions (4 possible amino acids),and 14 unrestricted surface positions (10 possible amino acids) result in $7^{10} \times 12^{11} \times 7^{6} \times 8^{6} \times 4^{3} \times 10^{14} = 4.14 \times 10^{46}$ possible amino acid sequences.

5. Janin, J., Wodak, S., Levitt, M. & Maigret, B. Conformation of amino acids side chains in proteins. *J. Mol. Biol.* **125**, 357-386 (1978).

6. Ponder, J. & Richards, F. Tertiary templates for proteins- use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* **193**, 775-791 (1987).

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

8. Monera, O., Kay, C. & Hodges, R. Protein denaturation with guanidine hydrochloride or urea provides a different estimate of stability depending on the contributions of electrostatic interactions. *Protein Science* **3**, 1984-1991 (1994).

9. Brünger, A. *X-PLOR, version 3.1, A system for X-ray crystallography and NMR* (Yale Univ. Press, New Haven, CT, 1992).

10. Kuszewski, J., Nilges, M. & Brünger, A. Sampling and efficiency of metric matrix geometry - a novel partial metrization algorithm. *J. Biomol. NMR* **2**, 33-56 (1992).

11. Nilges, M., Clore, C. & Gronenborn, A. Determination of three-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. *FEBS Lett.* **229**, 317-324 (1998).

12. Nilges, M., Kuszewski, J. & Brünger, A. in *Computational Aspects of the Study of Biological macromolecules by NMR* (eds. Hoch, J., Poulson, F. & Redfield, C.) 451-457 (Plenum Press, New York, 1991).

13. An average structure was generated by superimposing the ensemble of structures, lining up the equivalent side chain atoms, and then averaging the coordinates of the ensemble, followed by 500 steps of conjugate gradient (Powel) minimization.

14. This packing arrangement is supported by two NOEs between residues 2 and 5, two NOEs between 2 and 8, three NOEs between 2 and 11, one NOE between 3 and 7, and three NOEs between 2 and 11.

15. Klemm, J., Rould, M., Aurora, R., Herr, W. & Pabo, C. Crystal structure of the Oct-1 Pou domain bound to an octamer sire: DNA recognition with tethered DNA-binding molecules. *Cell* **77**, 21-32 (1994).

16. Ceska, T., Lamers, M., Monaci, P., Nicosia, A. & Cortese, R. The X-ray structure of an atypical homeodomain present in the rat liver

transcription factor LFB1/HNF1 and implications for DNA binding. *EMBO Journal* **12**, 1805-1810 (1993).

17. Qian, Y., Furukubo-Tokunaga, K., Resendez-Perez, D., Müller, M., Gehring, W. & Wüthrich, K. Nuclear magnetic resonance solution structure of the fushi tarazu homeodomain from Drosophila and comparison with the Antenapedia homeodomain. *J. Mol. Biol.* **238**, 333-345 (1994).

18. Schott, O., Billeter, M., Leiting, B., Wider, G. & Wüthrich, K. The NMR solution structure of the non-classical homeodomain from rat liver LFB1/HNF1 transcription factor. *J. Mol. Biol.* **267**, 673-683 (1997).

19. Aguilar, C., Sanderson, I., Moracci, M., Ciaramella, M., Nucci, R., Rossi, M. & Pearl, L. Crystal structure of the b-glycosidase from the hyperthermophilic archeon Sulfolobus solfataricus: resilience as a key factor in thermostability. *J. Mol. Biol.* **271**, 789-802 (1997).

20. Henning, M., Darimont, B., Sterner, R., Kirschner, K. & Jansonius, J. 2.0 Å structure of indole-3-glycerol phosphate synthase from the hyperthermophile Sulfolobus solfataricus: possible determinants of protein stability. *Structure* **3**, 1295-1306 (1995).

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

22. Dahiyat, B., Gordon, D. & Mayo, S. Automated design of the surface positions of protein helices. *Protein Science* **6**, 1333-1337 (1997).

23. Dahiyat, B., Sarisky, C. & Mayo, S. *De novo* protein design: towards fully automated sequence selection. *J. Mol. Biol.* **273**, 789-796 (1997).

24. Dayihat, B. & Mayo, S. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA.* **94**, 10172-10177 (1997).

25. Prodromou, C. & Pearl, L. Recursive PCR: a novel technique for total gene synthesis. *Protein Engineering* **5**, 827-829 (1992).

26. Alexander, P., Fahnestock, S., Lee, T., Orban, J. & Bryan, P. 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 (1992).

27. Johnson, B. & Hecht, M. Recombinant proteins can be isolated from *E. coli* cells by repeated cycles of freezing and thawing. *Bio/technology* **12**, 1357-1360 (1994).

28. Piantini, U., Sorenson, O. & Ernst, R. Multiple quantum filters for elucidating NMR coupling networks. *JACS* **104**, 6800-6801 (1982).

29. Bax, A. & Davis, D. MLEV-17-based two-dimensional homonuclear magnetization transfer spectroscopy. *J. Magnetic Reson.* **65**, 355-360 (1985).

30. Janeer, J., Meier, B., Bachmann, R. & Ernst, R. Investigation of exchange processes by two-dimensional NMR spectroscopy. *J. Chem. Phys.* **71**, 4546 (1979).

31. *Varian Associates* (Palo Alto, CA).

32. Kraulis, P. ANSIG - A program for the assignment of protein H-1 2D-NMR spectra by interactive computer-graphics. *J. Magnetic Reson.* **84**, 627-633 (1989).

33. Kay, L., Keifer, P. & Saarinen, T. Pure absorption gradient enhanced heteronuclear single quantum coherence spectroscopy with improved sensitivity. *JACS* **114**, 10663-10665 (1992).

34. Zhang, O., Kay, L., Olivier, J. & Forman-Kay, J. Backbone H-1 and N-15 resonance assignments of the N-terminal SH3 domain of DRK in folded and unfolded states using enhanced-sensitivity pulsed-field gradient NMR techniques. *J. Biomolecular NMR* **4**, 845-858 (1994).

35. Dahiyat, B. *newovfix.tcl* (California Institute of Technology, Pasadena, 1997).

36. Kubinowa, H., Grzesiek, S., Delaglio, F. & Bax, A. Measurement of H-N-N-alpha J-couplings in calcium-free calmodulin using new 2D and 3D water-flip-back methods. *J. Biomolecular NMR* **4**, (1994).

37. Griesinger, C., Sorenson, O. & Ernst, R. Two-dimensional correlation of connected NMR transitions. *JACS* **107**, 6394-6396 (1985).

Table 3-1

Sequences and thermal denaturation temperatures (T_m) of wild-type enh, the surface-core mutant (sc1), and the full sequence design (bsc). S denotes surface positions, c denotes core positions, and b denotes boundary positions. Helical positions are denoted by h. Vertical lines represent boundary residues, which were not designed in sc1. M is the number of mutations relative to wild-type. Residues in blue maintained the wild-type identity and residues in red are different from those in the wild-type sequence.

	----|----1----|----2----|----3----|----4----|----5-	M	T_m
wt	TAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKI	0	43
sc1	|K|DEQLKR|LEE|FKRD|R|TNQ||HDL|QKLGINEELIEDWFRR|EQQ|	29	81
bsc	KKIDEQLKRKLEEIFKRDKRLTNQLLHDLAQKLGINEELIEDWFRRKEQQK	39	114
	bsbssscssbcssbcsssbsbsssbbsscbsscscsssccssccssbsssb		
	hhhhhhhhhhhhh hhhhhhhhhh hhhhhhhhhhhhhhh		

Table 3-2
Salt bridges and hydrogen bonds in wild-type enh (2a), scl (2b), and bsc (2c). Wild type salt bridges and hydrogen bonds are from Clarke *et al.* All other interactions are from computational model structures.

Table 3-2a

enh-wt		
salt bridges	R10 E14	intrahelical
(3)	R10 E32	helices 1 & 2
	E14 R25	helices 1 & 2
H-bonds	T1 Q39	loop 1 & helix 3
(5)	Q7 L33	helices 1 & 2
	E14 R25	helices 1 & 2
	N18 R25	loop 2 & helix 2
	R19 R48	loop 2 & helix 3
helix capping	S4 Q7	i, i + 3
(3)	T22 R25	i, i + 3
	N36 A38	i, i + 2

Table 3-2b

scl		
salt bridges	E5 K9	intrahelical
(7)	R10 E14	intrahelical
	E13 R17	intrahelical
	R20 E48	loop2 & helix 3
	R26 E41	helices 2 & 3
	E41 R45	intrahelical
	D42 R46	intrahelical
H- bonds	D4 Q6	loop 1 & helix 1
(10)	D18 R20	loop 2 & helix 2
	T22 R25	loop 2 & helix 2
	N23 Q24	intrahelical
	H27 E37	helices 2 & 3
	D28 K32	intrahelical
	S30 I35	helix 2 & loop 3
	Q31 K32	intrahelical
	N36 E38	loop 3 & helix 3
	Q49 Q50	intrahelical
helix capping (0)		

Table 3-2c

bsc		
salt bridges	R9 Q13	intrahelical
(7)	R10 Q13	intrahelical
	E12 K47	helices 1 & 3
	R20 E48	loop 2 & helix 3
	D28 K32	intrahelical
	E38 R45	intrahelical
	D42 R46	intrahelical
H-bonds	Q6 R9	intrahelical
(5)	T22 L25	loop 2 & helix 2
	N23 Q24	intrahelical
	H27 E37	helices 2 & 3
	Q31 K32	intrahelical
helix capping (0)		

Table 3-3
Chemical shift assignments of bsc.

	N	HN	HA	others
0 Met			4.14	HB 2.16, HG 2.59
1 Lys	124.7	7.99	4.28	HG 1.46, HD 1.75
2 Lys	119.5	8.32	3.39	HB1 1.91, HB2 1.79, HG1 1.48, HG2 1.28, HD 1.73, HE 2.94
3 Ile	121.3	8.33	3.99	HD1 0.78
4 Asp	126.6	7.88	4.62	HB1 3.01, HB2 2.70
5 Glu	120.1	8.61	3.93	HB 2.07, HG 2.37
6 Gln	117.6	8.31	4.02	HB1 2.16, HB2 2.07, HG1 2.45, HG2 2.37
7 Leu	121.4	7.98	4.04	HB1 1.75, HB2 1.47, HG 1.61, HD 0.90
8 Lys	121.5	8.20	4.36	
9 Arg	117.0	7.64	4.02	HE 7.47, HD 3.21
10 Lys	119.2	7.42	4.05	HB1 1.96, HB2 1.81, HG1 1.57, HG2 1.45, HD1 2.89, HD2 3.21
11 Leu	119.8	8.12	3.71	HB1 0.84, HB2 0.52, HG 1.31, HD1 0.00, HD2 -0.11
12 Glu	118.1	8.35	4.17	HB1 2.26, HB2 2.12, HG1 2.63, HG2 2.54
13 Glu	118.4	7.61	4.08	HB1 2.19, HG1 2.44, HG2 2.33
14 Ile	120.3	7.83	3.73	HB 2.00, HG11 1.73, HG12 1.23, HG2 0.90, HD1 0.82
15 Phe	119.9	8.62	4.44	HB1 3.36, HB2 3.22, HZ 7.11, HD 7.26, HE 7.23
16 Lys	116.0	7.67	4.02	HE1 3.21
17 Arg	117.0	7.66	4.26	HE 7.37
18 Asp	118.6	8.32	4.65	HB1 2.73, HB2 2.64
19 Lys	119.5	8.05	4.10	HB1 1.75, HB2 1.55, HG1 1.27, HG2 1.15, HE 2.89
20 Arg	119.2	8.05	4.43	HE 7.40
21 Leu	123.7	8.28	4.23	HB1 1.29, HB2 1.07, HG 0.91, HD 0.53
22 Thr	112.4	6.82	4.43	HB 4.57, HG2 1.22
23 Asn	119.9	8.99	4.30	HB1 2.79, HB2 2.72, HD21 7.66, HD22 6.98, HD2 7.68
24 Gln	119.0	8.48	4.00	HB1 2.12, HB2 1.92, HE21 7.76, HE22 6.86, HG 2.40, HE2 7.27
25 Leu	120.8	7.64	4.19	HB1 1.67, HB2 1.59, HD1 0.99, HD2 0.92
26 Leu	118.7	8.27	3.87	HG 1.52, HD1 0.90, HD2 0.78, HB 1.66
27 His	115.1	7.88	4.33	HB 3.36
28 Asp	120.6	8.39	4.36	HB1 2.87, HB2 2.72
29 Leu	120.4	8.84	4.02	
30 Ala	121.2	8.14	4.02	HB 1.55
31 Gln	115.5	7.58	4.02	
32 Lys	119.3	8.16	4.03	
33 Leu	115.5	8.48	4.28	HB1 1.72, HB2 1.44, HG 1.79, HD1 0.76, HD2 0.69
34 Gly	108.1	7.73	3.88	
35 Ile	113.8	7.40	4.58	HB 1.68, HG11 1.33, HG12 0.69, HG2 0.83, HD1 0.93
36 Asn	119.6	8.17	4.49	HD21 7.79, HD22 7.05, HB 2.85, HD2 7.45
37 Glu	123.5	9.31	3.71	HB 2.00, HG 2.29

38 Glu	117.8	8.94	4.05	HB1 2.08, HB2 2.01, HG 2.41
39 Leu	118.9	7.53	4.22	HB1 1.88, HB2 1.56, HG 1.71, HD1 0.99, HD2 0.91
40 Ile	118.7	7.22	3.71	HB 1.98, HG11 1.68, HG12 0.89, HG2 1.00, HD1 0.78
41 Glu	119.2	8.43	3.98	HG1 2.46, HG2 2.34, HB 2.14
42 Asp	118.3	8.17	4.40	HB1 2.87, HB2 2.73
43 Trp	120.1	8.00	3.94	HB1 3.47, HB2 3.27, HD1 7.09, HE1 9.90, HE3 7.02, HZ2 7.22, HZ3 6.22, HH2 6.65
44 Phe	118.1	8.63	3.68	HZ 7.19, HB 3.16, HD 7.37, HE 7.26
45 Arg	118.7	8.32	4.02	HE 7.37, HD 3.23
46 Arg	117.5	7.72	3.95	HD1 3.07, HD2 3.00, HE 7.46, HB 1.64
47 Lys	120.1	7.55	3.68	HB1 1.06, HB2 0.83, HG1 0.34, HE 2.35
48 Glu	117.9	7.87	4.12	HB1 2.09, HB2 2.00, HG1 2.42, HG2 2.34
49 Gln	117.6	7.63	4.22	HB1 2.11, HB2 1.99, HG1 2.42, HG2 2.35
50 Gln	120.7	7.86	4.22	
51 Lys	127.8	7.95	4.13	

Table 3-4
Restraints used in the NMR structure determination of bsc.

Experimental restraints	
Distance restraints from NOEs	
Intraresidue restraints $\|i-j\| = 0$	220
Sequential restraints $\|i-j\| = 1$	102
Medium range restraints $\|i-j\| = 2$	38
Medium range restraints $\|i-j\| = 3$	56
Long range restraints $\|i-j\| = 4$	35
Long range restraints $\|i-j\| > 4$	64
Total NOE distance restraints	575
Hydrogen bond restraints	13
χ_1 restraints	7
ϕ restraints	24
Total number of experimental restraints	619
Total number of experimental restraints/residue	12.1

Table 3-5
Structural statistics, atomic root-mean-square deviations, and Lennard-Jones potential energies of the ensemble of NMR structures, <SA>. SA is the average minimized structure.

Structural statistics	
R.m.s. deviations from ideal stereochemistry	
Bonds	0.0042Å
Angles	0.6454°
Dihedrals	0.2631°
Impropers	0.4014°
NOEs	0.0513 Å
PROCHECK Ramachandran Map Analysis	
Most favored regions	71.4 %
Additionally allowed regions	22.6 %
Generously allowed regions	3.2 %
Disallowed regions	2.9 %
Atomic r.m.s. deviations	
Backbone atoms <SA>	
Residues 2-48	0.661 Å
Residues 1-51	1.152 Å
All heavy atoms <SA>	
Residues 2-48	1.677 Å
Residues 1-51	2.184 Å
Comparison of SA and target backbone atoms	
Residues 9-48	1.252 Å
Residues 2-46	2.033 Å
Residues 1-51	2.396 Å
Lennard-Jones potential energies	
<SA>	257 kcal mol^{-1}
$<SA>_{min}$	183 kcal mol^{-1}

Table 3-6
Hydrogen bond protection factors and hydrogen bond restraints used in the structure determination of bsc. N-O distances were restrained to 2.4-3.5 Å and H_N-O distances were restrained to 1.5-2.8 Å.

Protection factors	Restraints
23	11O 15N, 11O 15H_N
15	12O 16N, 12O 16H_N
5	22O 26N, 22O 26H_N
10	23O 27N, 23O 27H_N
124	27O 31N, 27O 31H_N
80	28O 32N, 28O 32H_N
35	29O 33N, 29O 33H_N
84	30O 34N, 30O 34H_N
34	31O 35N, 31O 35H_N
19	32O 36N, 32O 36H_N
35	36O 40N, 36O 40H_N
146	37O 41N, 37O 41H_N
36	38O 42N, 38O 42H_N
52	40O 44N, 40O 44H_N
68	41O 45N, 41O 45H_N

Table 3-7
Phi restraints used in the structure determination of bsc.

	J(Hz)	restraint
K3	3.3	-60° ± 20°
E6	3.9	-60° ± 20°
Q7	5.1	-60° ± 20°
L12	4.7	-60° ± 20°
E13	3.4	-60° ± 20°
E14	5.2	-60° ± 20°
F16	4.7	-60° ± 20°
K17	4.5	-60° ± 20°
N24	3.6	-60° ± 20°
Q25	4.5	-60° ± 20°
L27	4.0	-60° ± 20°
H28	4.5	-60° ± 20°
D29	4.9	-60° ± 20°
L30	4.5	-60° ± 20°
A31	3.6	-60° ± 20°
K33	5.2	-60° ± 20°
I36	9.9	-120° ± 20°
N37	4.1	-60° ± 20°
E38	2.6	-60° ± 20°
E39	5.1	-60° ± 20°
E42	3.8	-60° ± 20°
D43	4.2	-60° ± 20°
W44	3.6	-60° ± 20°
F45	4.7	-60° ± 20°

Table 3-8
Chi1 restraints used in the structure determination of bsc.

Residue	restraint
16	-60° ± 60°
22	180 ± 60°
24	-60 ± 60°
26	180 ± 60°
29	-60 ± 60°
34	-60 ± 60°
44	180 ± 60°

Figure 3-1

CD wavelength scans of wild-type and sc1 at 1 °C.

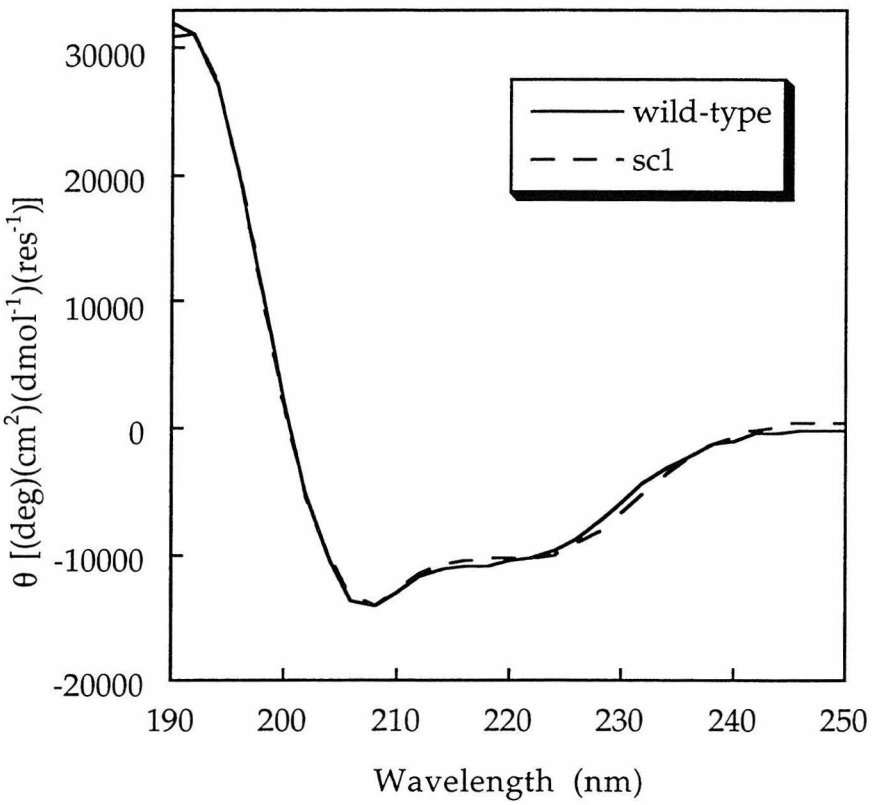

30000
20000
10000
0
-10000
-20000
190
200
210
220
230
240
250
θ [(deg)(cm²)(dmol⁻¹)(res⁻¹)]
Wavelength (nm)
wild-type
sc1

Figure 3-2

Thermal denaturation curves of wild-type and sc1 at 222 nm.

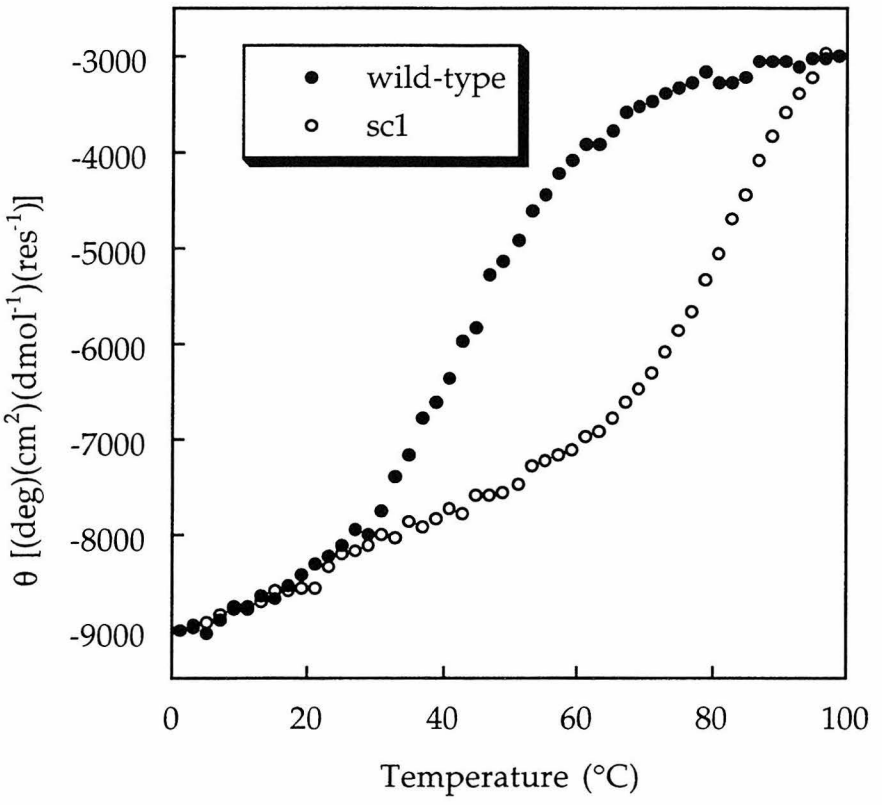

wild-type
sc1
-3000
-4000
-5000
-6000
-7000
-8000
-9000
θ [(deg)(cm²)(dmol⁻¹)(res⁻¹)]
0
20
40
60
80
100
Temperature (°C)

Figure 3-3

CD wavelength scans of wild-type and bsc at 1 °C.

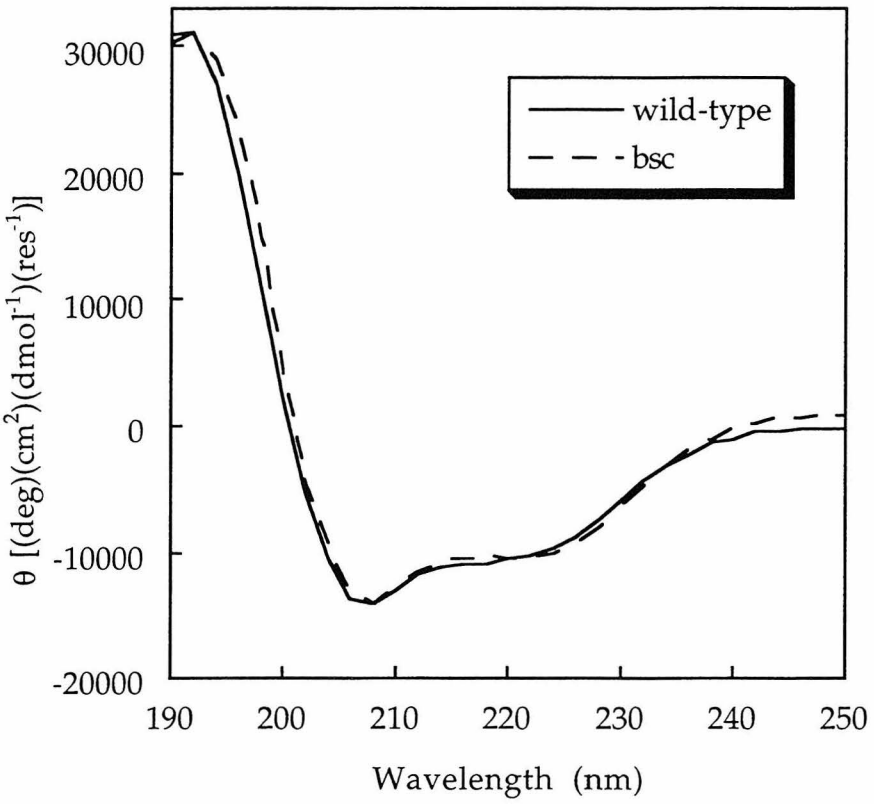

30000
20000
10000
0
-10000
-20000
190
200
210
220
230
240
250
θ [(deg)(cm²)(dmol⁻¹)(res⁻¹)]
Wavelength (nm)
wild-type
bsc

Figure 3-4

Thermal denaturation curves of wild-type and bsc at 222 nm.

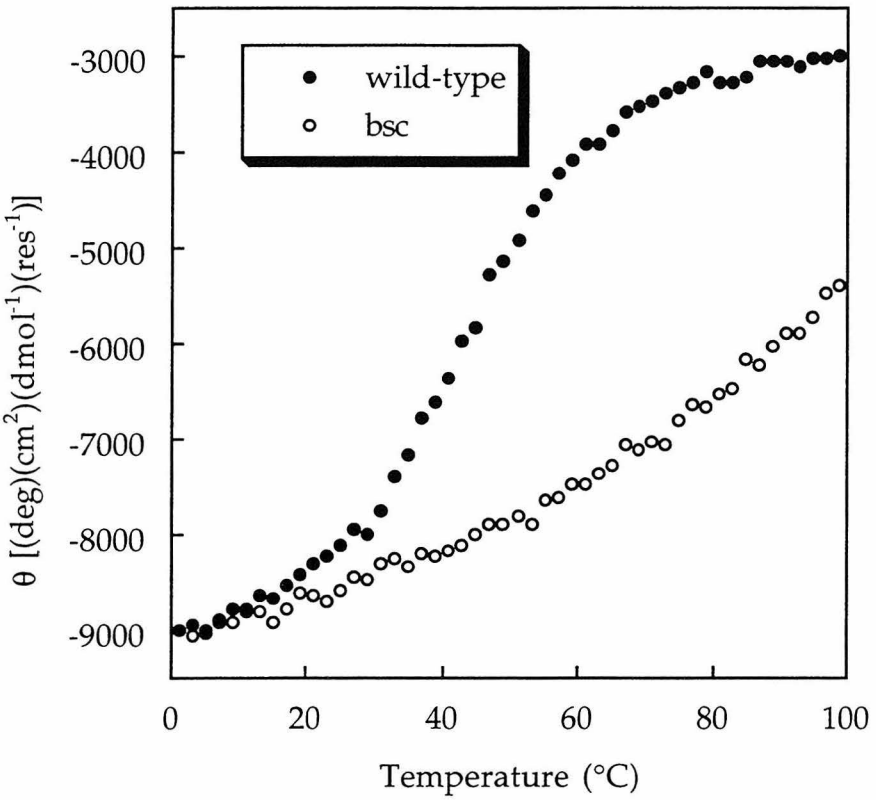

wild-type
bsc
-3000
-4000
-5000
-6000
-7000
-8000
-9000
θ [(deg)(cm²)(dmol⁻¹)(res⁻¹)]
0
20
40
60
80
100
Temperature (°C)

Figure 3-5

Guanidine hydrochloride denaturation curves of wild-type and bsc. Data were obtained at 1 °C and 222 nm.

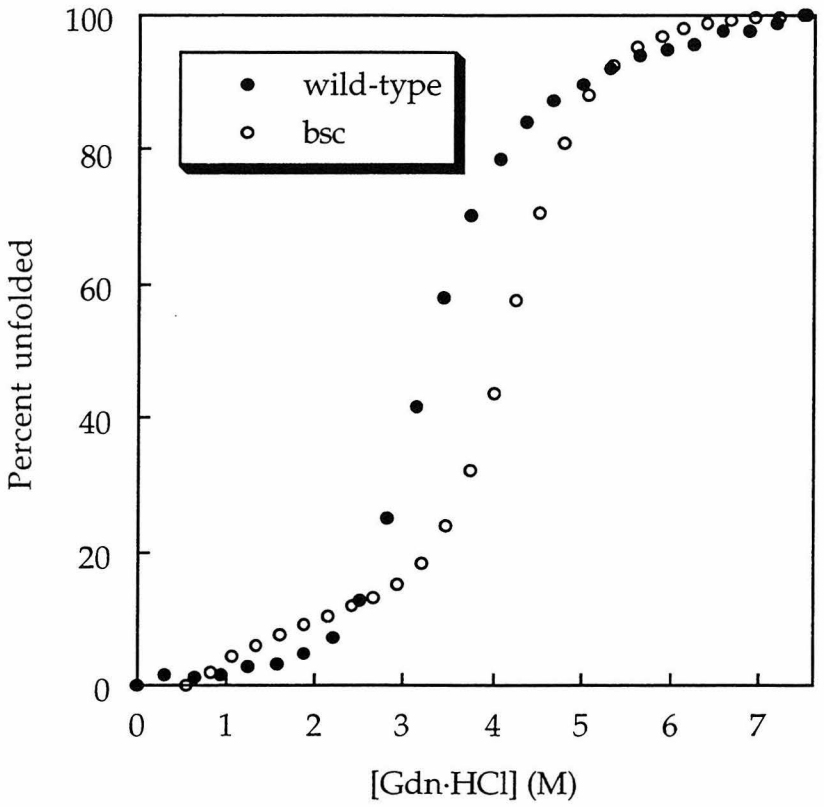

100
80
60
40
20
0
Percent unfolded
wild-type
bsc
0
1
2
3
4
5
6
7
[Gdn·HCl] (M)

Figure 3-6

Differential scanning calorimetry profile of bsc. The T_m of bsc is 114 °C.

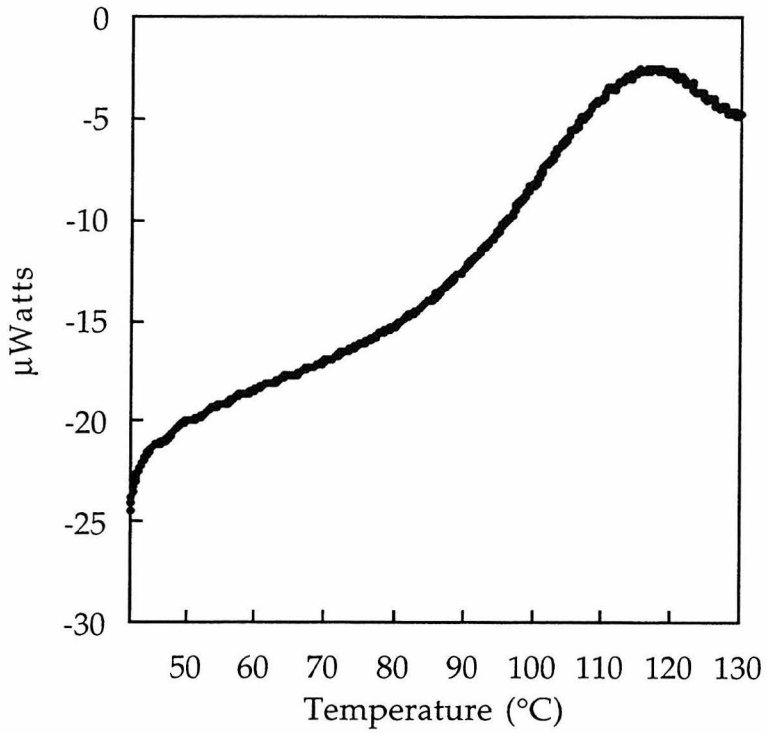

0
-5
-10
-15
-20
-25
-30
µWatts
50
60
70
80
90
100
110
120
130
Temperature (°C)

Figure 3-7

Urea denaturation curves of wild-type and bsc. Data were obtained at 222 nm and 35 °C (wild-type) and 30 °C (bsc). At 60 °C, the urea denaturation curve of bsc indicates that it does not completely unfold at 9.5 M urea.

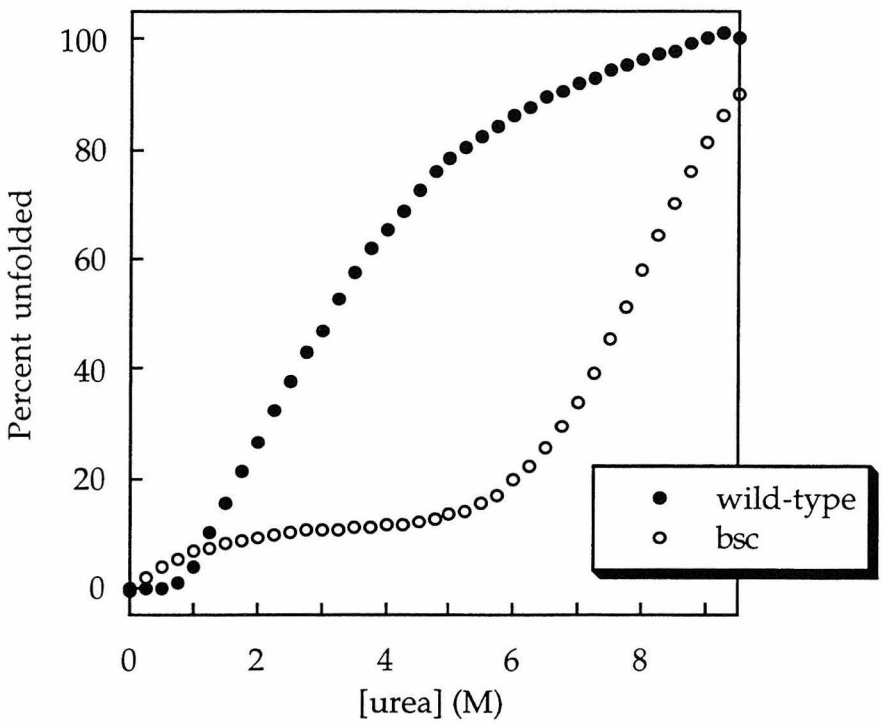

100
80
60
40
20
0
Percent unfolded
0
2
4
6
8
[urea] (M)
wild-type
bsc

Figure 3-8

The fingerprint region of the TOCSY spectrum. The peaks are well resolved, indicative of a folded protein. The absence of low field peaks is consistent with an α-helical structure and there is good amide proton dispersion.

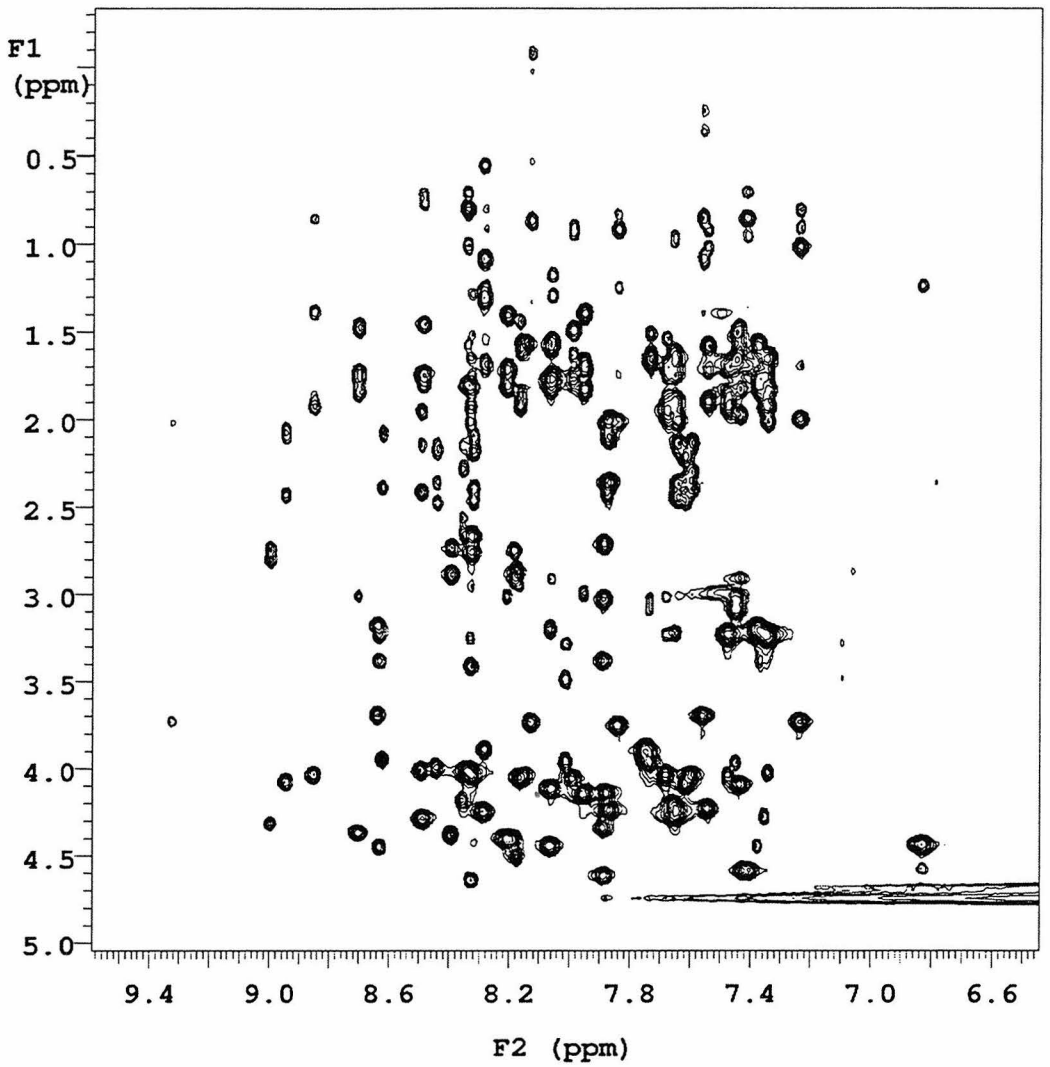

F1
(ppm)
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
9.4
9.0
8.6
8.2
7.8
7.4
7.0
6.6
F2 (ppm)

Figure 3-9

The heteronuclear single quantum coherence spectrum (HSQC) of ^{15}N-labeled bsc. Chemical shifts of the spin system roots were obtained, facilitating assignment of the 3D NOESY and TOCSY data.

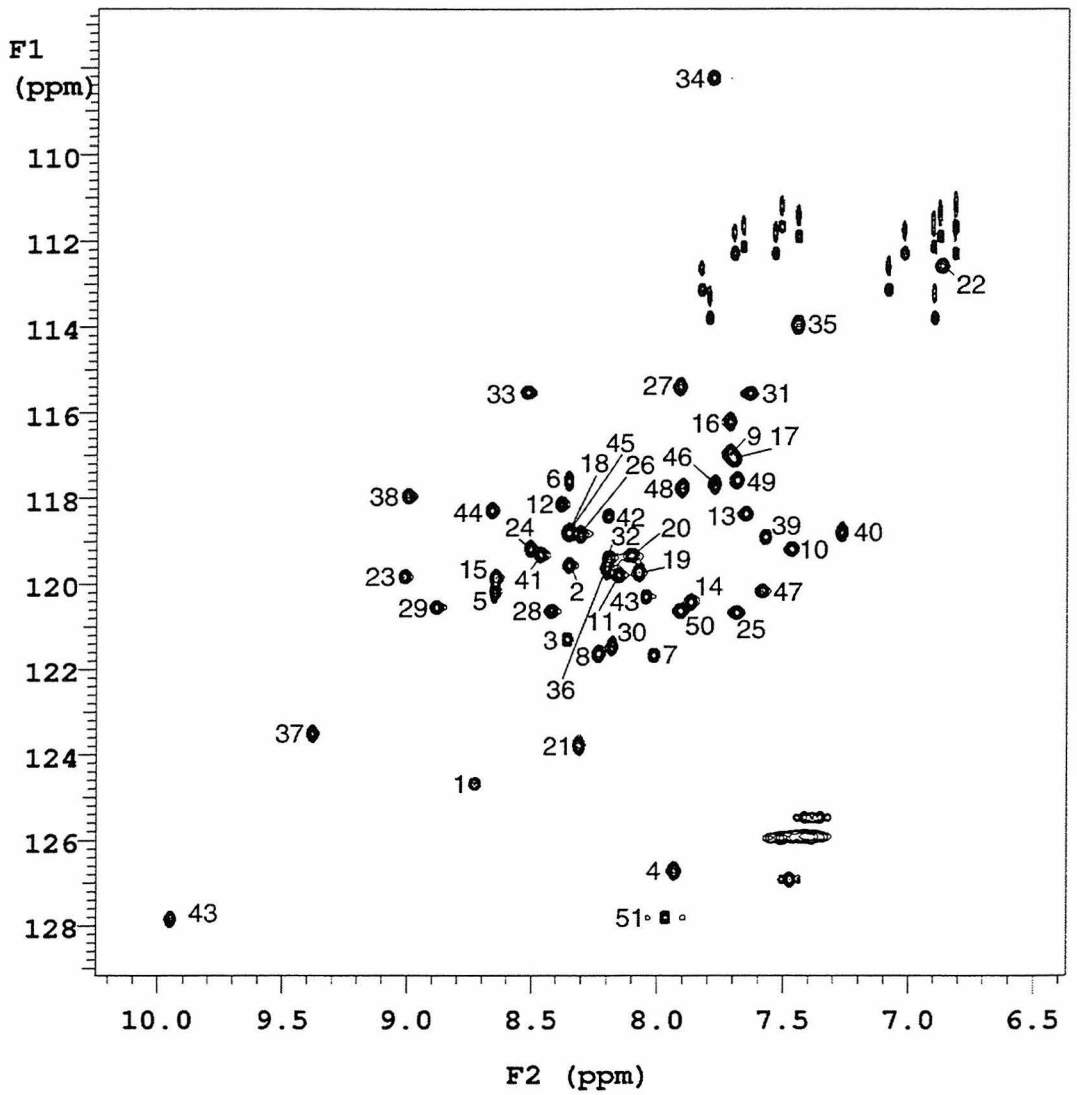
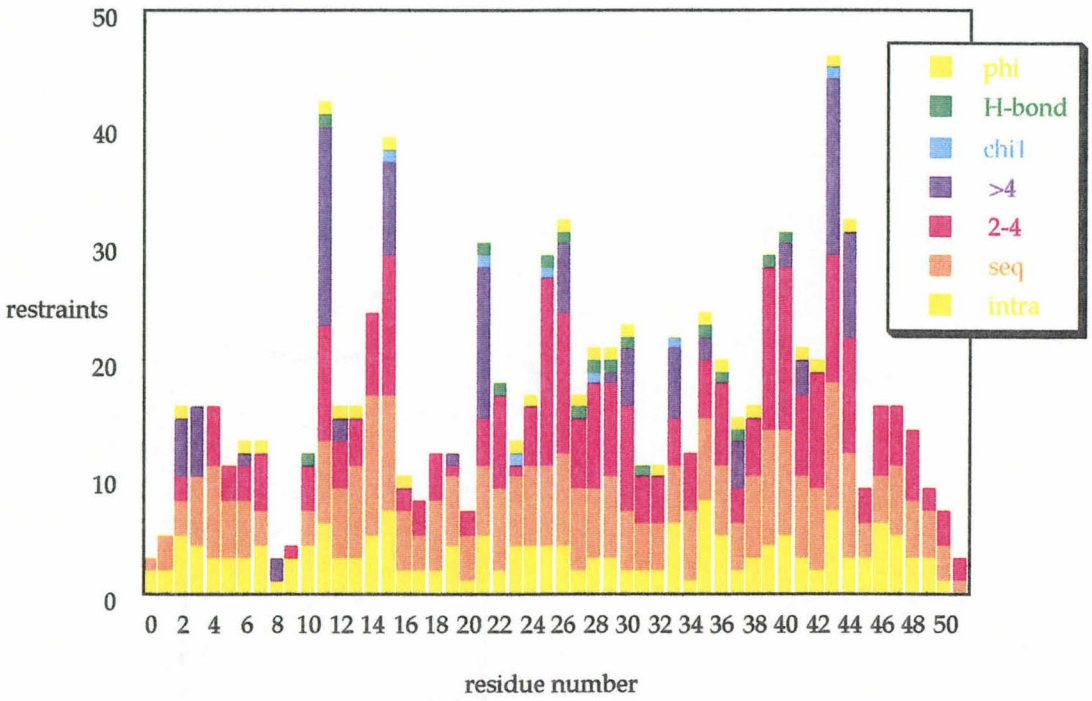

F1
(ppm)
110
112
114
116
118
120
122
124
126
128
10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
F2 (ppm)
34
22
35
33
27
31
16
9
17
45
18
26
46
6
48
49
38
44
12
42
13
39
40
24
32
20
10
19
23
15
41
2
14
47
29
5
28
43
50
25
11
30
3
8
7
36
37
21
1
4
43
51

Figure 3-10

Experimental restraints/residue.

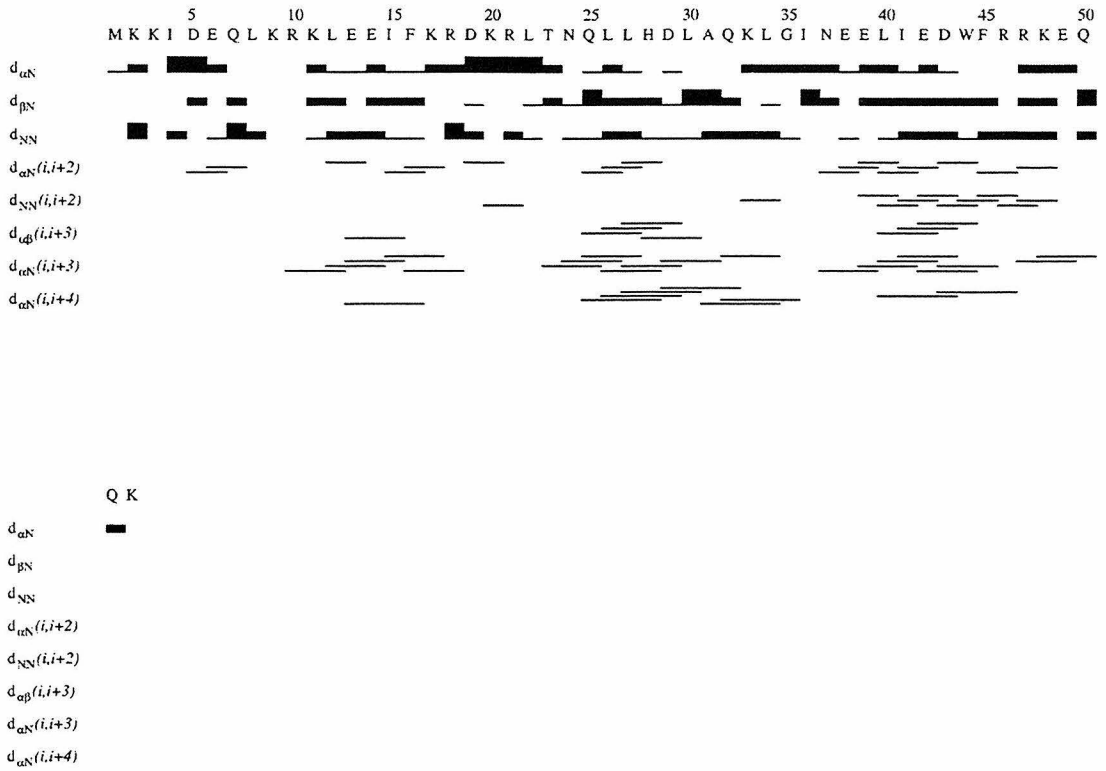

50
40
30
20
10
0
restraints
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50
residue number
phi
H-bond
chi1
>4
2-4
seq
intra

Figures 3-11

Sequential and short-range NOE connectivities of bsc.

Figure 3-11a shows data from all NOESY experiments and Figure 11b shows data from 2D experiments only, illustrating how much less data was obtainable in the absence of 3D data. In helical protein regions, strong connectivities are expected for αN, NN and αN(*i*,*i*+3) and weaker connectivities are expected for βN and αβ(*i*, *i*+3).

3-11a


5 10 15 20 25 30 35 40 45 50
M K K I D E Q L K R K L E E I F K R D K R L T N Q L L H D L A Q K L G I N E E L I E D W F R R K E Q
$d_{\alpha N}$
$d_{\beta N}$
d_{NN}
$d_{\alpha N}(i,i+2)$
$d_{NN}(i,i+2)$
$d_{\alpha\beta}(i,i+3)$
$d_{\alpha N}(i,i+3)$
$d_{\alpha N}(i,i+4)$
Q K
$d_{\alpha N}$
$d_{\beta N}$
d_{NN}
$d_{\alpha N}(i,i+2)$
$d_{NN}(i,i+2)$
$d_{\alpha\beta}(i,i+3)$
$d_{\alpha N}(i,i+3)$
$d_{\alpha N}(i,i+4)$

3-11b

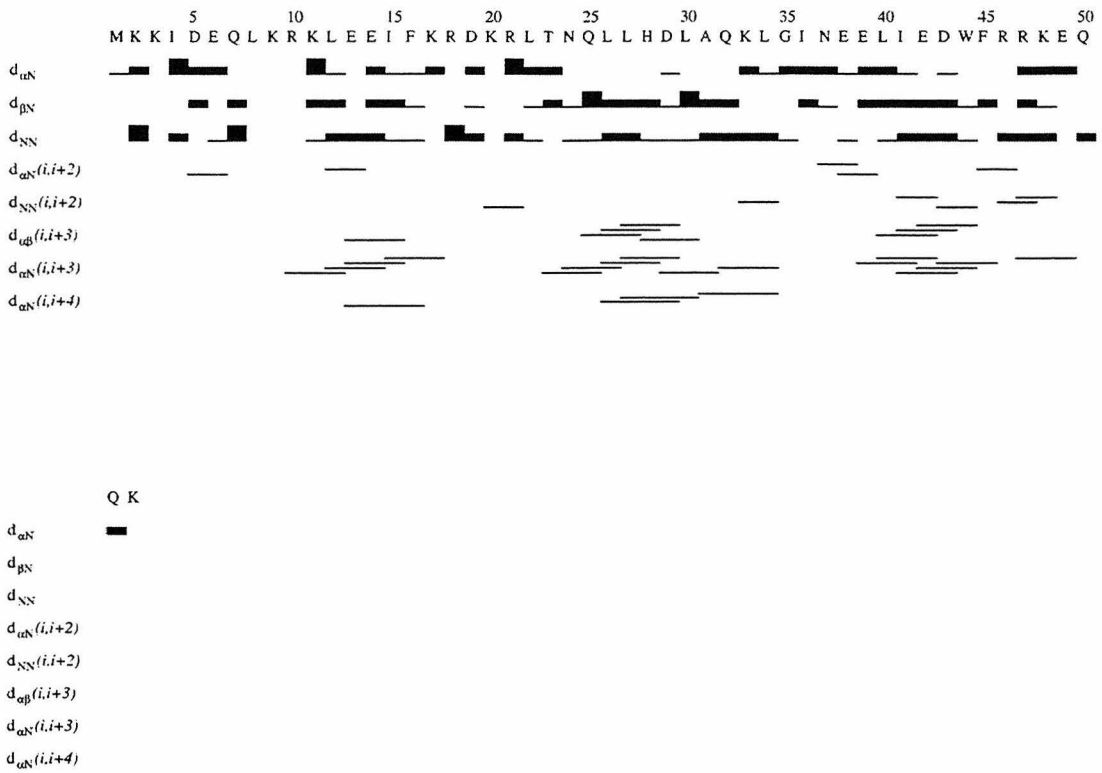

5 10 15 20 25 30 35 40 45 50
M K K I D E Q L K R K L E E I F K R D K R L T N Q L L H D L A Q K L G I N E E L I E D W F R R K E Q
$d_{\alpha N}$
$d_{\beta N}$
d_{NN}
$d_{\alpha N}(i,i+2)$
$d_{NN}(i,i+2)$
$d_{\alpha\beta}(i,i+3)$
$d_{\alpha N}(i,i+3)$
$d_{\alpha N}(i,i+4)$
Q K
$d_{\alpha N}$
$d_{\beta N}$
d_{NN}
$d_{\alpha N}(i,i+2)$
$d_{NN}(i,i+2)$
$d_{\alpha\beta}(i,i+3)$
$d_{\alpha N}(i,i+3)$
$d_{\alpha N}(i,i+4)$

Figure 3-12

The structure of bsc is a helix-turn-helix motif.

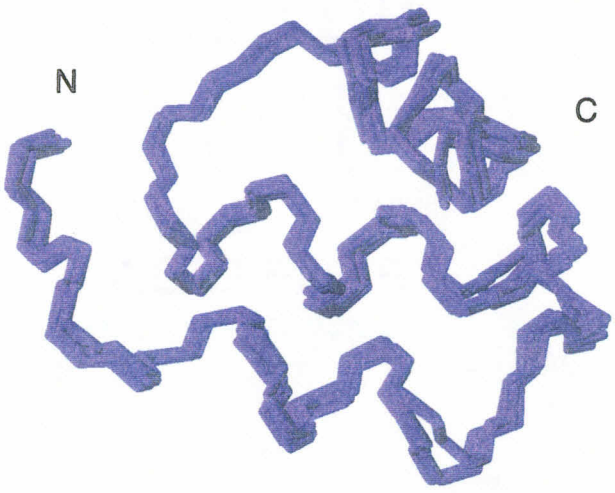

N
C

Figure 3-13

The structure of bsc.

Residues 1-51 of the ensemble of 40 structures are shown.

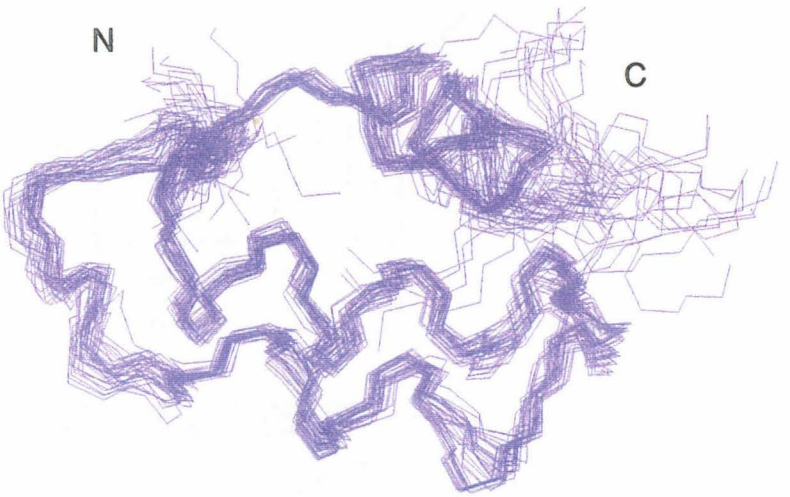

N
C

Figure 3-14

Superposition of the backbones of wild-type (red) and bsc (blue).

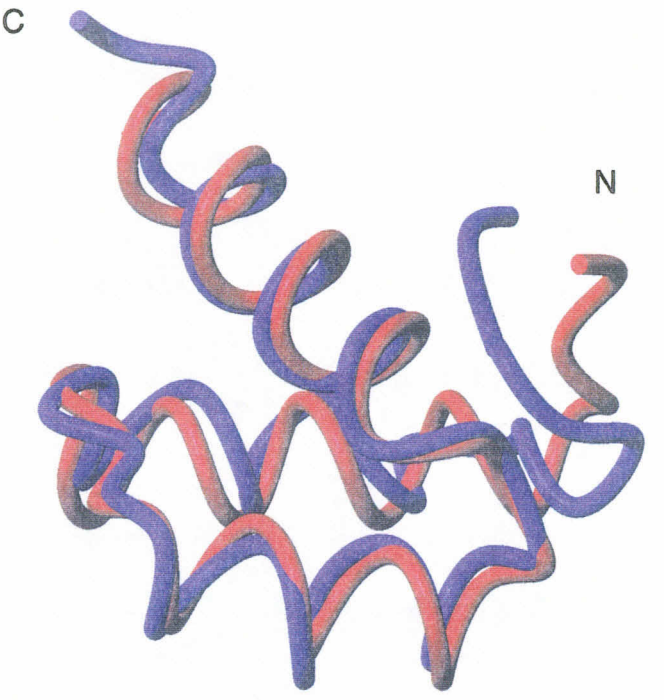

C
N

Figure 3-15

ϕ and ψ angle plots of wild-type and the ensemble of 40 bsc structures. Wild-type is shown in red.

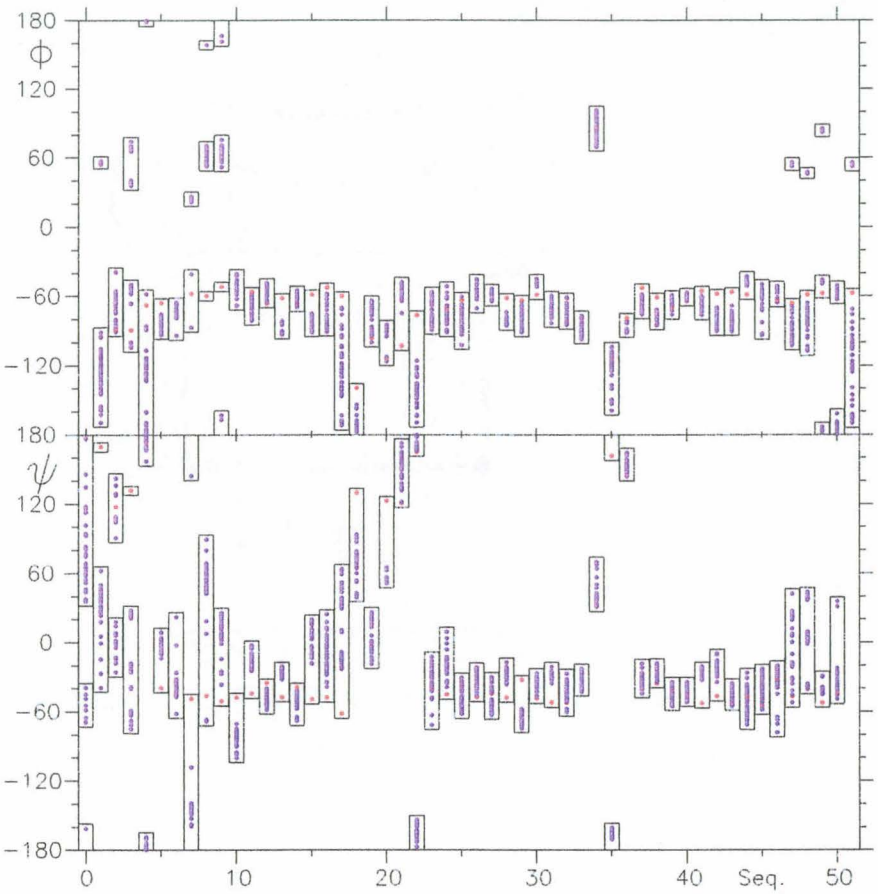

180
Φ
120
60
0
−60
−120
180
Ψ
120
60
0
−60
−120
−180
0
10
20
30
40
Seq.
50

Figure 3-16

χ_1 angle plot of wild-type and the ensemble of 40 bsc structures. Wild-type is shown in red.

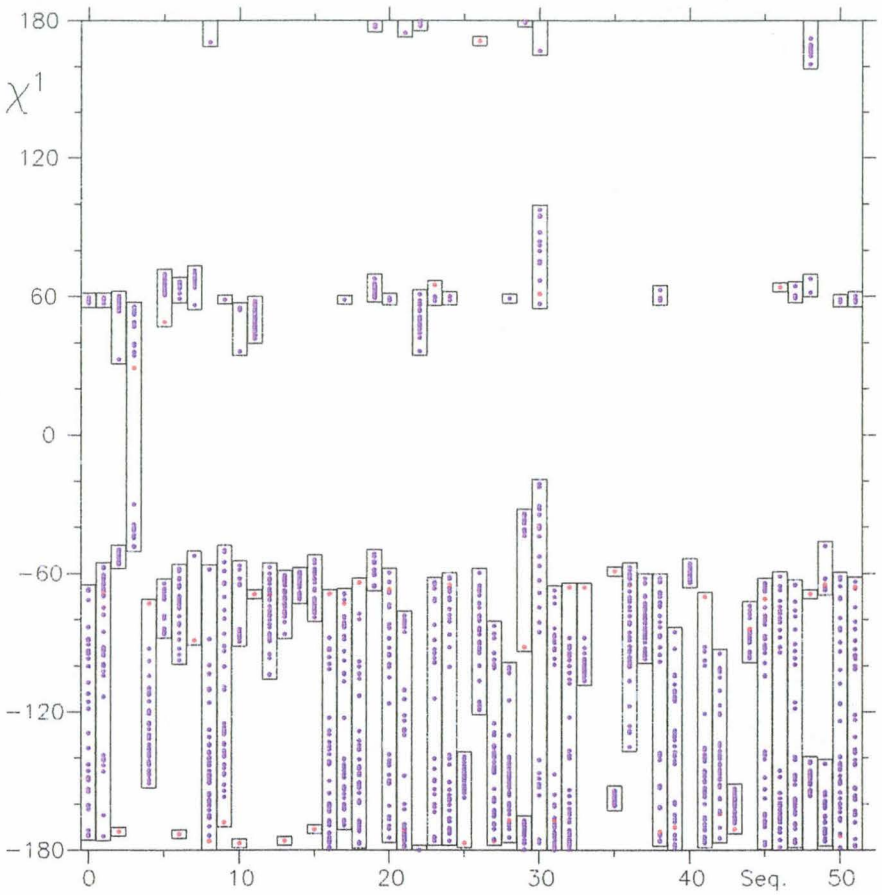

180
χ^1
120
60
0
−60
−120
−180
0
10
20
30
40
Seq.
50

Figure 3-17

Comparison of aromatic core residues of bsc and wild-type. Wild-type is shown in red and bsc is shown in blue. W43 and F44 adopt predicted rotamers (within standard deviations). F15 is rotated along χ_1 with respect to the model structure and is further in the core.

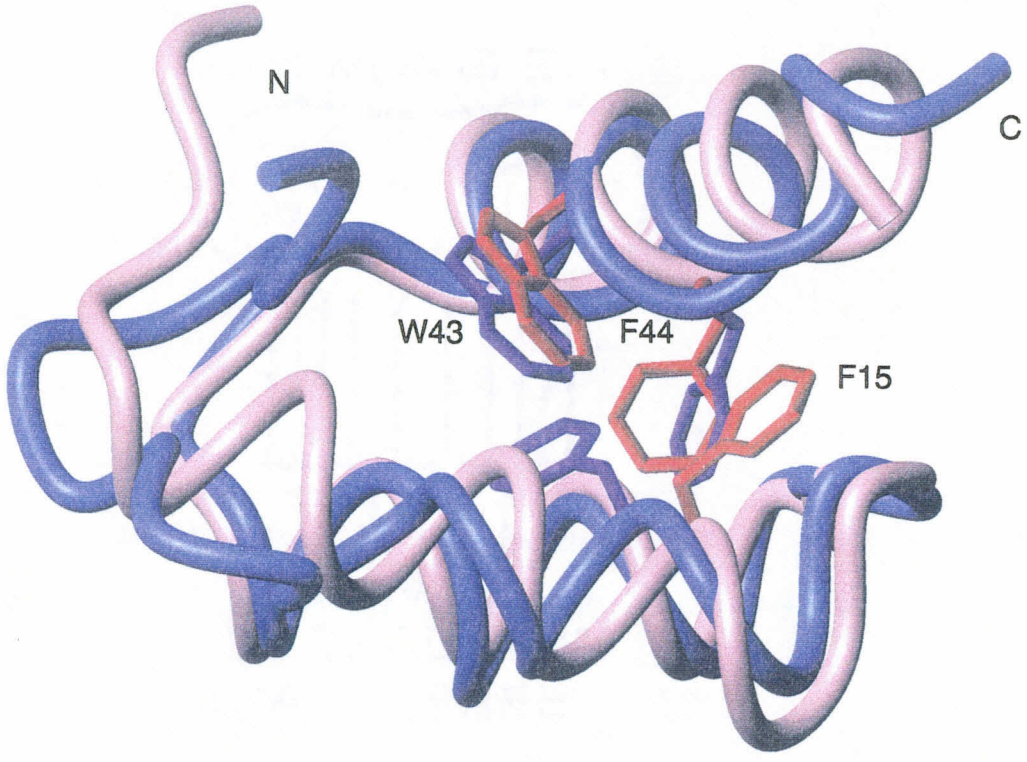

N
C
W43
F44
F15

Figure 3-18

Comparison of three aliphatic core residues of bsc and wild-type. Bsc is shown in blue and wild-type is shown in red.

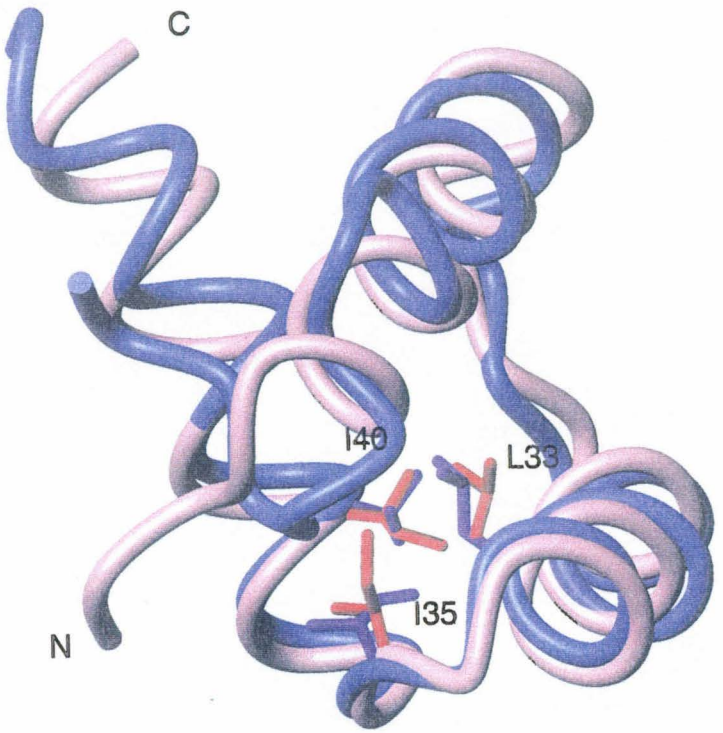

C
I40
L33
I35
N

Figure 3-19

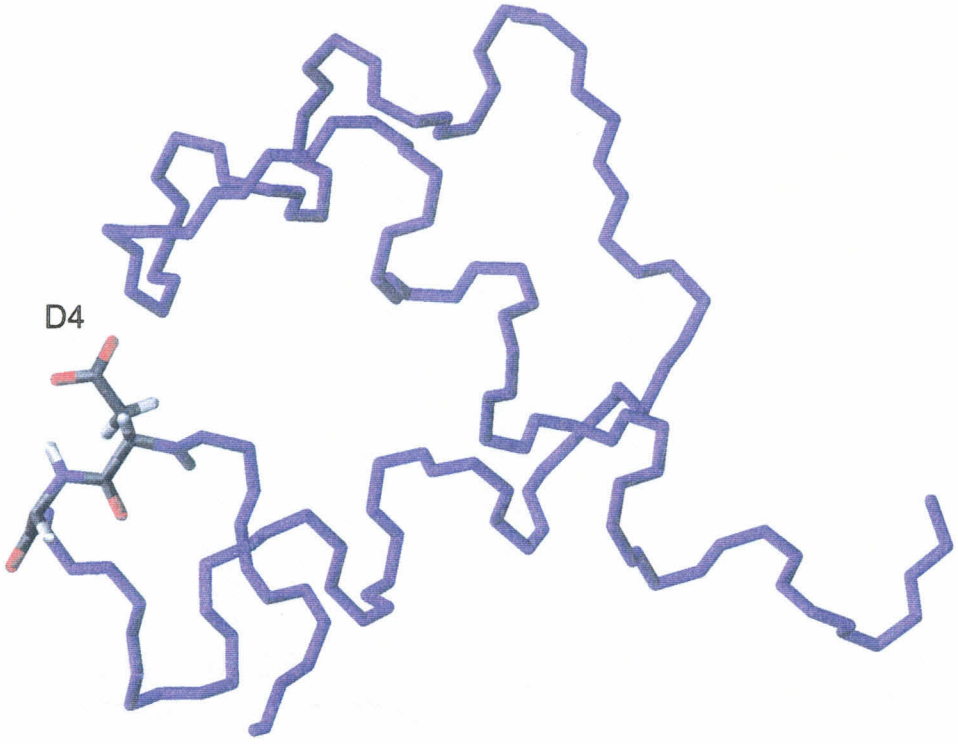
Helix capping interactions.

A: The capping H-bond between the side chain of D4 and amide proton of E5.

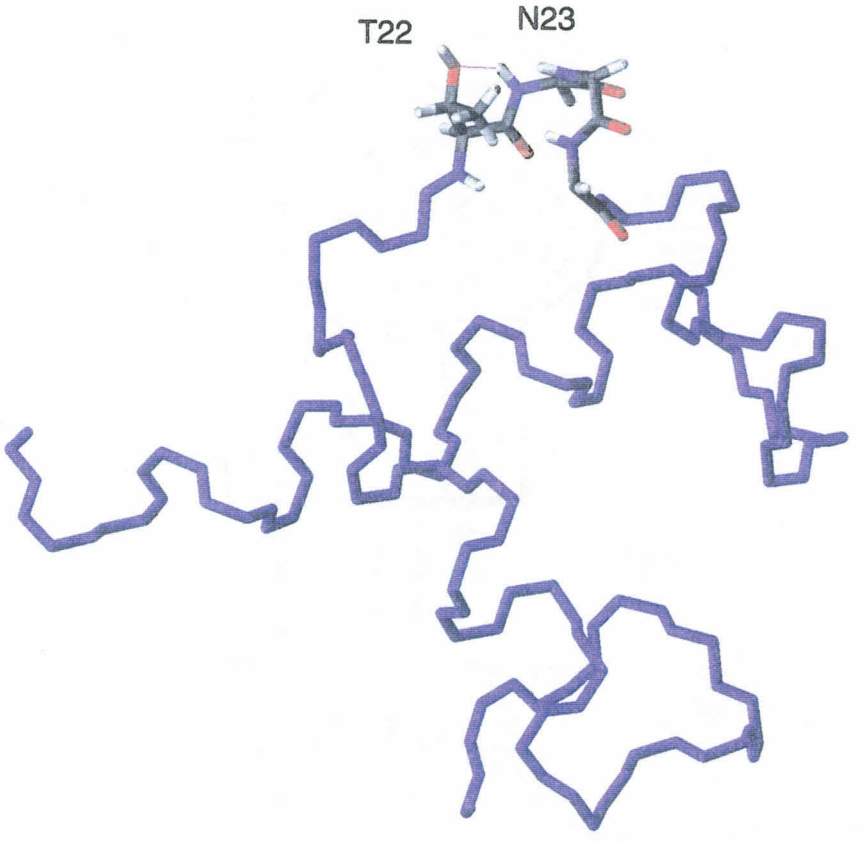
B: The capping H-bond between the side chain of T22 and amide proton of N23.

C: The capping H-bond between the side chain of T22 and amide proton of L25.

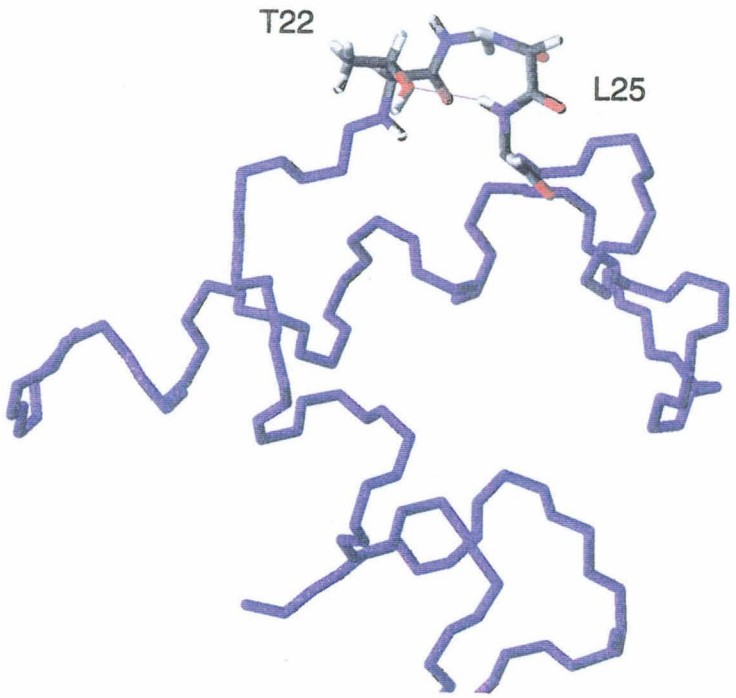
3-19a



3-19b



3-19c



Chapter 4

Circular Dichroism Determination of Class I MHC-Peptide Equilibrium Dissociation Constants

The text of this chapter has been adapted from a published manuscript that was co-authored with Professors Pamela J. Bjorkman and Stephen L. Mayo as well as James M. Holton and Barry D. Olafson.

Chantal S. Morgan, James M. Holton, Barry D. Olafson, Pamela J. Bjorkman, and Stephen L. Mayo, *Protein Science*, 6, 1771 (1997).

Abstract

Class I major histocompatibility complex (MHC) molecules bind peptides derived from degraded proteins for display to T cells of the immune system. Peptides bind to MHC proteins with varying affinities, depending upon their sequence and length. We demonstrate that the thermal stability of the MHC-peptide complex depends directly on peptide binding affinity. We use this correlation to develop a convenient method to determine peptide dissociation constants by measuring MHC-peptide complex stability using thermal denaturation profiles monitored by circular dichroism. The equation relating K_D and T_m is $K_D = 10^{\wedge} [(43.3 - T_m)/2.47]$, $R = 0.990$.

Introduction

Class I major histocompatibility complex (MHC) proteins bind fragments of cytosolic proteins for presentation to the immune system. These cell-surface proteins are heterodimers, composed of a heavy chain (~45 kD), a noncovalently associated light chain, β_2-microglobulin (β_2m, 12 kD), and a noncovalently bound peptide (typically 8-10 residues), illustrated in Figure 4-1. The heavy chain is composed of three domains, α_1, α_2, and α_3, and contains a peptide binding groove formed by two α-helices and a β-sheet of domains α_1 and α_2. (reviewed in Bjorkman & Parham, 1990).

MHC proteins derived from different alleles have preferences for particular amino acids, called anchor residues, at specific positions in the peptide sequences. For example, the murine class I MHC protein H-2K^d has a preference for binding peptides containing tyrosine at position two (P2), and a hydrophobic amino acid at position nine (P9), as in the peptide NP1 (TYQRTRALV, see 4-1). The dissociation constant for the K^d/NP1 complex is two orders of magnitude lower than the complex formed with NP10, an analogous peptide with phenylalanine at P2. A peptide variant with threonine at P2 has no detectable binding to K^d. Similarly, NP1 binds to K^d three orders of magnitude more tightly than NP15, a peptide with the P9 hydrophobic residue valine of NP1 replaced by serine. (Fahnestock *et al.*, 1994). Therefore, in K^d, as in other MHC molecules, the identities of the anchor residue are responsible for the majority of the binding specificity.

Empty class I molecules lacking peptides are considerably less stable than their filled counterparts (Ljunggren *et al.*, 1990; Schumacher *et al.*, 1990; Townsend *et al.*, 1990). Circular dichroism (CD) thermal melting curves of empty and peptide filled K^d were used to demonstrate that the binding of a peptide confers about 4.4 kcal/mol of stability to the empty heterodimer (Fahnestock *et al.*, 1992). We have shown that this stability is not pH dependent in the range of pH 5.5-7.0 (Reich *et al.*, 1997). Different peptides stabilize MHC molecules by different amounts, as evidenced by the dissociation rate of β_2m, which is reflective of MHC-peptide complex

stability (Parker *et al.*, 1992). However, the amount of stability the peptide confers to the MHC-peptide complex and how tightly the peptide binds have not been correlated by this method. In the case of the murine class I MHC molecule H-2K^d, peptide equilibrium dissociation constants have been measured by equilibrium dialysis assays (Fahnestock *et al.*, 1994). This paper establishes a correlation between the thermal denaturation temperatures and the dissociation constants of K^d-peptide complexes.

Results and discussion

A soluble version of the murine class I MHC protein, H-2K^d, was loaded with peptides by incubation with a ten-fold molar excess of peptide at room temperature for 12 hours. Approximately 30% of the K^d molecules as purified from transfected cell supernatants were occupied with endogenous peptides (Fahnestock *et al.*, 1992). This initial semi-filled population is expected to make only a small contribution to the apparent melting temperature (T_m).

Thermal denaturation curves of K^d-peptide complexes were measured by monitoring the CD ellipticity at 223 nm (Figure 4-2). Melting temperatures were determined for the K^d-peptide complexes listed in Table 4-1. The dissociation constants (K_Ds) of these peptides, which range over four orders of magnitude, were previously measured by equilibrium dialysis (Fahnestock *et al.*, 1994). Despite the presence of endogenous peptides in a fraction of the MHC proteins, we found a logarithmic

relationship between the T_ms of the K^d-peptide complexes and the peptide dissociation constants for nonameric peptides (Figure 4-3). These results demonstrate that the amount of thermal stability the peptide imparts to the K^d-peptide complex is a direct result of how tightly the peptide binds.

The equation relating K_D and T_m for those peptides with known dissociation constants, $K_D = 10^{\wedge} [(43.3 - T_m)/2.47]$, can be used to calculate dissociation constants for nonameric peptides (Table 4-1). The error in the T_m measurement is estimated to be ±0.5 °C, which results in an error in the calculated K_D of approximately half an order of magnitude, which is comparable to the error in equilibrium dialysis determination of K_D.

Given a set of peptides with known dissociation constants that can be used to generate a standard curve, it should be possible to determine K_D's by thermal denaturation for other class I MHC proteins. This methodology may extend to class II MHC proteins and other proteins which bind peptides, such as the viral MHC homolog UL18 (Fahnestock *et al.*, 1995). Recently, shifts in thermal stability have been used to determine how well peptides with unnatural amino acids bound to the MHC molecule HLA-B*2705 (Krebs *et al.*, 1998).

One issue with this methodology concerns the octameric peptide KD3, YIPSAEKI. Its T_m was determined to be 57 °C, which results in a calculated K_D that is significantly larger than the K_D measured by equilibrium dialysis. Although the crystal structure of an octamer bound to MHC Class I molecule H2-K^b reveals the octamer in an extended

conformation, this may not be the case for KD3 bound to K^d (Fremont *et al.*, 1992). T_m measurements are comparable only for complexes with peptides bound with the same backbone conformation. The first residue of KD3 is tyrosine, which presumably binds as an anchor residue in the P2 pocket which would result in the loss of all N-terminal contacts. The loss of N-terminal contacts between the peptide and MHC protein has been previously shown to destabilize an HLA-A2 MHC-peptide complex by 4.6 kcal/mol (Bouvier *et al.*, 1994). This may explain why KD3 has lower thermal stability than expected judging from its binding constant.

Materials and methods:

Expression and Purification of Soluble H-2K^d in CHO Cells: A stable CHO cell line expressing a secreted form of K^d was used as previously described (Fahnestock *et al.*, 1992). The protein was purified using immunoaffinity columns constructed with immobilized monoclonal antibodies, either 34-1-2 (anti-H-2K^d) (Ozato *et al.*, 1982) or BBM.1 (anti-β_2m) (Parham *et al.*, 1983). Approximately 70% of the purified protein product does not contain endogenous peptides (Fahnestock *et al.*, 1992).

Peptide Synthesis: Peptides were synthesized on an Applied Biosystems 433A peptide synthesizer. Preloaded resins were used, with subsequent residues coupled via Fmoc chemistry and HTBU/HOBt activation with standard 0.25 mmol scale coupling cycles. Peptides were cleaved from the solid support resin by mixing 200 mg resin with 2 ml

trifluroacetic acid (TFA), 100 μL water, 150 mg phenol, 100 μL thioanisole and 50 μL ethanedithiol for two hours. The peptides were precipitated by addition of cold methyl *tert*-butyl ether, washed four times with the same solvent, and lyophilized to partially remove the cleavage reaction scavengers. The peptides were further purified by reverse-phase HPLC with a Vydac C8 column using linear acetonitrile-water gradients (typically 20-30%) containing 0.1% TFA. Peptide masses were determined by MALDI-TOF mass spectrometry and were found to be within one mass unit of the expected masses.

CD Thermal Denaturation Curves: An Aviv 62A DS spectropolarimeter equipped with a thermoelectric cell holder was used to collect CD data. Data were obtained from samples containing 10 μM MHC and 100 μM peptide in 5.0 mM sodium phosphate, pH 7.0 using a 1.0 m m path length cell. Thermal denaturation curves were recorded at 223 n m over an appropriate temperature range with a 0.1 second time constant, 10 second averaging time, 3 minute equilibration time, and 1 nm bandwidth. Melting temperatures were determined by taking the maximum of a plot of d(ellipticity)/dT versus T after averaging the data with a moving window of 3 degrees. The error in T_m is estimated to be ± 0.5 °C.

Acknowledgments

We thank B. Dahiyat, G. Hathaway, J. Johnson, and D. Penny for technical help, C. White for insightful discussions, and M. Ary for

comments on the manuscript. C.S.M. is supported by the James Irvine Foundation and ARCS Foundation. P.J.B. acknowledges support from the Arthritis Foundation. S.L.M. acknowledges support from the Rita Allen Foundation, the David and Lucile Packard Foundation, and the Searle Scholars Program/The Chicago Community Trust.

References

Bjorkman PJ, Parham P. 1990. Structure, Function, and Diversity of Class I Major Histocompatibility Complex Molecules. *Annu Rev Biochem 59*:253-288.

Fahnestock ML, Johnson JL, Feldman RMR, Neveu JM, Lane WS, Bjorkman PJ. 1995. The MHC Class I Homolog Encoded by Human Cytomegalovirus Binds Endogenous Peptides. *Immunity 3*:583-590.

Fahnestock ML, Johnson JL, Feldman RMR, Tsomides TJ, Mayer J, Narhi LO, Bjorkman PJ. 1994. Effects of Peptide Length and Composition on Binding to an Empty Class I MHC Heterodimer. *Biochemistry 33*:8149-8158.

Fahnestock ML, Tamir I, Narhi L, Bjorkman PJ. 1992. Thermal Stability Comparison of Purified Empty and Peptide-filled Forms of a Class I MHC Molecule. *Science 258*:1658-1662.

Fremont DH, Masazumi M, Stura EA, Peterson, PA, Wilson, IA. 1992. Crystal structures of two viral peptides in complex with murine MHC class I H-2K^b.

Krebs S, Folkers G, Rognan D. 1998. Binding of rationally designed non-natural peptides to the human leukocyte antigen HLA-B*2705. *J Peptide Sci 4*:378-388.

Ljunggren H, Stam N, Ohlen C, Neefjes J, Hoglund P, Heemels M, Bastin J, Schumacher T, Townsend A, Karre K, Ploegh H. 1990. Empty MHC Class I Molecules Come Out in the Cold. *Nature 346*:476-480.

Ozato K, Mayer NM, Sachs DH. 1982. Monoclonal Antibodies to Mouse Major Histocompatibility Complex Antigens: A Series of Hybridoma Clones Producing Anti-H-2^d Antibodies and an Examination of Expression of H-2^d Antigens on the Surface of These Cells. *Transplantation 34*:113-120.

Parham P, Androlewicz M, Holmes N, Rothenberg B. 1983. Arginine-45 is a Major Part of the Antigenic Determinant of Human β2-Microglobulin Recognized by Mouse Monoclonal-Antibody BBM.1. *J Biol Chem 258*:6179-6186.

Parker KC, DiBrino M, Hull L, Coligan JE. 1992. The β2-Microglobulin Dissociation Rate is an Accurate Measure of the Stability of MHC Class I Heterotrimers and Depends on Which Peptide is Bound. *J Immunol 149*:1896-1904.

Reich Z, Altman JD, Boniface JJ, Lyons DS, Kozono H, Ogg G, Morgan CS, Davis MM. 1997. Stability of empty and peptide-loaded class II major histocompatability complex molecules at neutral and endosomal pH: comparison to class I proteins. *Proc Natl Acad Sci USA 94*:2495-2500.

Schumacher TMM, Heemels M, Neefjes J, Kast W, Melief C, Ploegh H. 1990. Direct Binding of Peptide to Empty MHC Class I Molecules on Intact Cells and In Vitro. *Cell 62*:563-567.

Townsend A, Elliot T, Cerundolo V, Foster L, Barber B, Tse A. 1990. Assembly of MHC Class I Molecules Analyzed In Vitro. *Cell 62*:285-295.

Table 4-1

Sequences, thermal denaturation temperatures, and dissociation constants for nonameric peptides complexed to H-2K^d.

Peptide	Sequence	T_m (°C)	K_D(M)[a] determined by equilibrium dialysis	K_D (M)[b] determined by CD thermal denaturation
KD1	SYFPEITHI	61.0	4.7×10^{-8}	6.6×10^{-8}
NP1[c]	TYQRTRALV	61.0	6.9×10^{-8}	6.6×10^{-8}
NP22	TYQRTCALV	58.0	9.5×10^{-7}	1.1×10^{-6}
NP10	TFQRTRALV	57.0	5.6×10^{-6}	2.8×10^{-6}
NP15	TYQRTRALS	55.0	3.0×10^{-5}	1.8×10^{-5}
NP17	TYQRTRALK	52.0	2.7×10^{-4}	3.0×10^{-4}
NP25	TYGGGGGLV	52.0	1.5×10^{-4}	3.0×10^{-4}
no added peptide[d]	–	45.0	–	–

[a]Fahnestock et al., 1994.

[b]This study.

[c]The previously reported T_m for the NP1/K^d complex is 57 °C (Fahnestock et al., 1992). In this study, a higher T_m is observed for the same complex, possibly because the protein used in the previous study was denatured and reassembled prior to the addition of peptide.

[d]Approximately 30% of this protein contains a mixture of endogenous peptides, as discussed in the text.

Figure 4-1

Structure of MHC class I H-2K^b complexed to a nonameric peptide (Fremont *et al.*, 1992). The heavy chain is shown in purple, the light chain is shown in red, and the bonds of the peptide are shown.

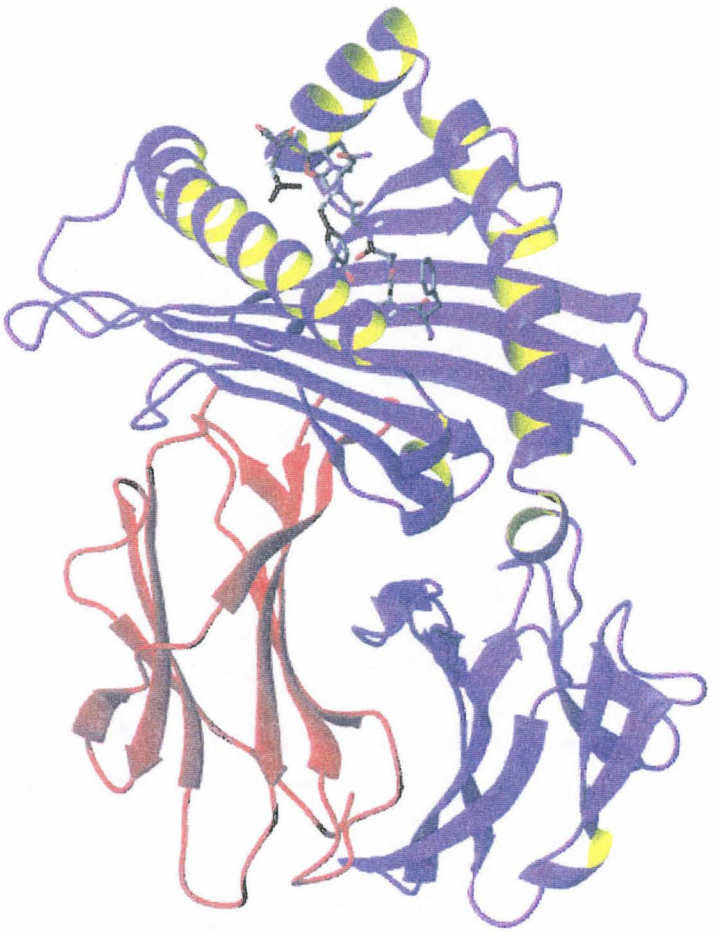


Figure 4-2

Representative thermal denaturation curves of K^d without exogenous peptide (purple squares), complexed to NP25 (black triangles), and complexed to NP22 (red circles). The curves have been smoothed and normalized.

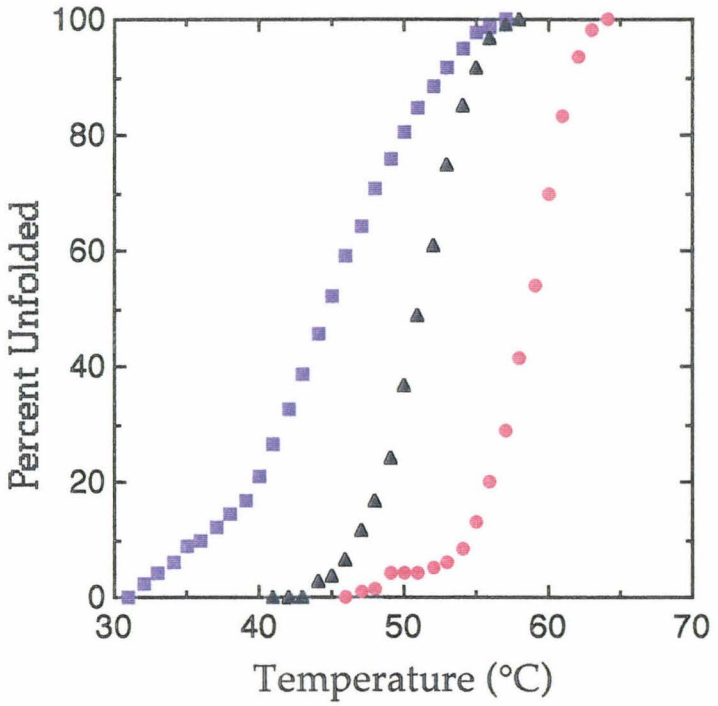
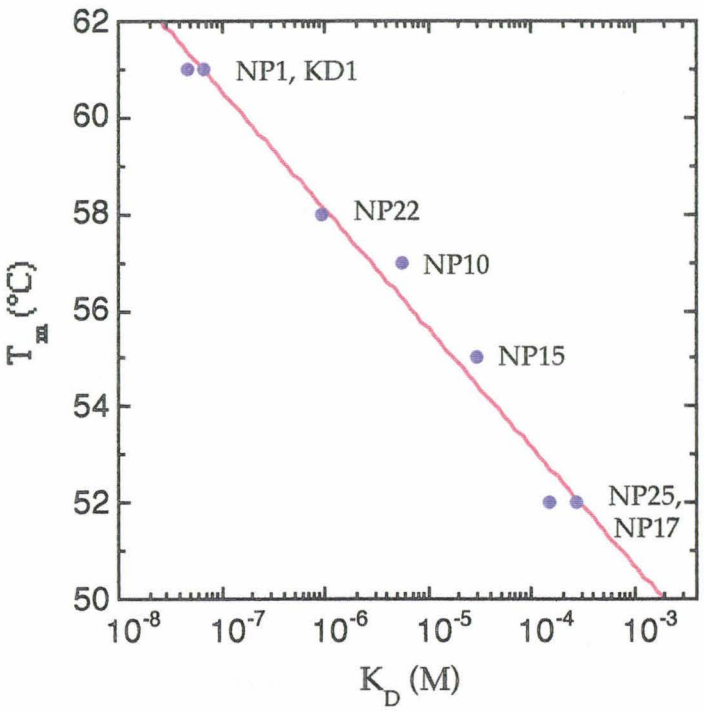

100
80
60
40
20
0
Percent Unfolded
30
40
50
60
70
Temperature (°C)

Figure 4-3

K_D versus T_m for the peptides listed in Table 4-1. The equation relating K_D and T_m is $K_D = 10^{\wedge} [(43.3 - T_m)/2.47]$, R = 0.990.


62
60
58
56
54
52
50
T_m (°C)
NP1, KD1
NP22
NP10
NP15
NP25,
NP17
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
K_D (M)

Chapter 5

The Design of Peptides to Bind to MHC Class I Proteins

Abstract

In this study, nonameric peptides were computationally designed to bind to the major histocompatability class I molecule H-2K^d. Several computational approaches are described. Design success was assessed by experimental determination of the dissociation constants of the designed peptides. Experimental feedback was used to redesign the peptides using an iterative approach between computational design and experimental data. Ultimately, two designed peptides were shown to bind with dissociation constants of 1.6×10^{-9} and 1.1×10^{-8} M, more tightly than peptides which have been eluted from H-2K^d which have K_Ds in the range of $4.7 \times 10^{-8} - 2.7 \times 10^{-4}$ M.

Introduction

Protein binding interactions play a role in cell adhesion, cell signaling, electron transfer, recognition, signal transduction, transcription, and many other biological processes. The extension of protein design to protein-protein interactions is crucial for the ability to design many protein functions.

One particular protein-protein binding interaction is the noncovalent binding of a peptide by major histocompatability (MHC) class I molecules. Before traversing the cell membrane, MHC proteins bind peptide fragments of intracellular proteins for display on cell surfaces. These peptide-MHC complexes are displayed to cytotoxic T-cells of the immune system. If the complex is not recognized as self either because of the peptide or MHC

protein, a cascade of immune system events takes place, resulting in the destruction of the cell displaying the MHC-peptide complex.

Many autoimmune diseases, such as ankylosing spondylitis and insulin-dependent diabetes mellitus, are caused by autoreactive $CD4^+$ or $CD8^+$ T lymphocytes which recognize self peptides bound to MHC molecules. Expression of MHC class I-restricted allele specific blocking peptides in cells has been shown to lower recognition of the host MHC-peptide complex by T cells, thereby preventing virus-induced autoimmune diabetes[1].

Binding of peptides by MHC molecules is both specific and promiscuous. Different alleles specifically bind some peptides but not others, but each allele can bind many peptides, making the design of a novel peptide to bind to a specific allele feasible. Tightly binding peptides could potentially be blocking peptides, taking the place of intracellular peptide fragments which would be recognized by T cells.

Computational design and experimental results

Peptide design of the MHC-peptide complex must involve both the MHC molecule and the peptide. MHC class I molecules are large cell-surface heterodimeric proteins. The extracellular portion is composed of a heavy chain (~45 kD) and a noncovalently associated light chain, β_2-microglobulin (β_2m, 12 kD). The heavy chain is composed of three domains, α_1, α_2, and α_3, and contains a peptide binding groove formed by two α-helices and a β-sheet

results[5]. The energy of the van der Waals interactions and hydrogen bonds between the peptide *in vacuo* and in the context of the protein were calculated and subtracted to provide an energy associated with peptide binding. Surface area corrections were also performed to include the energy associated with burial of hydrophobic and hydrophilic surfaces upon binding. Together, these energies were used to generate an energy of peptide binding, which could be correlated to the peptide dissociation constant and the melting temperature of the K^d-peptide complex.

Initially, two nonameric sequences were predicted to bind tightly to K^d using different computational approaches[6]. The first, CP1, was a full DEE result, which was calculated by allowing all nine peptide positions to vary simultaneously among the twenty amino acids. The sequence is listed in Table 5-1. The protein design algorithm selected the anchor residues correctly, selecting a tyrosine at position 2 (P2) and a hydrophobic residue at P9. Previously, we have shown that peptide binding constants can be determined by monitoring thermal denaturation by circular dichroism spectrometry[7]. Experimental determination of the T_m of CP1 complexed to K^d demonstrated that CP1 did not bind very tightly; the dissociation constants was calculated to be 1.9×10^{-3} M (Table 5-1).

The individual side chains of the second peptide, CP2, were calculated independently. Bound peptides have extended conformations and later computational analysis by Schueler-Furman *et al.* has shown that peptide side chains do not interact with each other, but rather with the peptide backbone and the MHC binding groove[5]. Nine calculations were performed

designing one position in each calculation and setting the other eight to alanine, for example, XAAAAAAAA, AXAAAAAAA, etc. The CP2 sequence is the result of the nine calculations; the lowest energy side chain is selected at each position (Table 5-1). The CP2 sequence is very similar to CP1; the sequences only vary at P3 and P5, shown in Table 5-1. The T_m of CP2 complexed to K^d, 47 °C, was used to obtain a calculated dissociation constant of 3.2×10^{-2} M, approximately an order of magnitude worse than CP1.

CP1 and CP2 did not bind as well as hoped, so the calculations were re-evaluated. The most probable cause was that the protein design algorithm did not have any enthalpic parameters and did not handle charges well. The charged side chains of CP1 and CP2 make the peptides very soluble and the algorithm did not contain adequate terms for polar burial of hydrophilic surface and desolvation.

A new sequence was computed as in the first calculation but without any charged residues or methionines yielding sequence CP3. CP3 was synthesized, complexed to K^d, and found to raise the T_m of the MHC-peptide complex to 65 °C. The calculated dissociation constant of CP3 is 1.6×10^{-9} M. This nanomolar dissociation constant is lower than those of the viral and self peptides NP1 and KD1 (which had been eluted from K^d) which are 6.9×10^{-8} M and 4.7×10^{-8} M respectively[3].

Another tightly binding peptide was predicted by using the energy list of peptides in positions P1-P9 from the CP2 calculation. For each position P1-P9, a rank-ordered energy was calculated for all twenty residues. The residue

with the lowest energy which appears in the sequence of tightly binding peptides NP1, KD1 or KD2 was selected to comprise the sequence of CP4. Like CP3, this chimeric sequence also has a measured T_m of 65 °C and a nanomolar dissociation constant.

Since charged residues may have been responsible for the high dissociation constants of CP1 and CP2, the CP1 sequence was used to generate CP5, which has a glutamate to arginine mutation at position 6 (E6R) to test our ability to handle charged side chains. This change from a negative to positive charge resulted in a peptide which bound an order of magnitude more tightly than CP1. The T_m of the K^d-CP5 complex is 52 °C and the calculated dissociation constant is 3.0×10^{-4} M.

The next design, CP6, is an R6F mutant of viral peptide NP1. Substitution of the charged residue with a hydrophobic residue at this exposed position causes the loss of two orders of magnitude of binding affinity as compared to NP1, which may be due to solvation effects.

Peptide CP7 was an accidental I9E mutation of CP4. It does not bind well to K^d because K^d requires the anchor residue at P9 to be hydrophobic. The dissociation constant of CP7 is 4.7×10^{-5} M, ~30,000 times worse than CP4.

Since predicted peptide CP4 has a low dissociation constant, its sequence was used as a starting place for refinement of the sequence prediction. Charged residues and methionine were reintroduced into the calculation and the peptide CP8 was calculated. CP8 is a T6K and E7M mutant of CP4 and binds three orders of magnitude less tightly. At this

point, single mutations were made to the CP4 sequence. CP9 is a T1D mutant which binds three orders of magnitude less tightly also. CP10 is an E7M mutant and binds one order of magnitude less tightly. Since CP5, CP6, CP8, and CP9 all bind K^d less tightly than CP4, this confirmed that the algorithm did not handle charges, solvation, and entropic effects very well at that time, which has since been improved[8].

Conclusions

The computational design of peptides CP3 and CP4 to bind tightly to K^d was successful. This design methodology could be used to design peptides to bind to specific MHC alleles, perhaps blocking display of immunogenic self peptides to the immune system, and thereby preventing autoimmune diseases. However, many of the peptide designs were not successful, and there are several approaches to improving the computational design methodology.

To improve the binding constants of the predicted peptides, solvation effects must be included. For example, KD1, a peptide which has been eluted from K^d and found to bind tightly, contains two charged side chains, demonstrating the necessity of repeating the experiments with surface area terms.

Rather than consider the MHC binding groove as part of the template, the side chains which contact the peptide should be included in the calculation maintaining their identities but allowing their rotamers to vary.

This would include any conformational changes the side chains make upon peptide binding in the calculation.

Once the calculations are optimized using K^d, it would be interesting to use another MHC class I protein and repeat the computational and experimental cycles. Since our methodology of measuring dissociation constants by CD spectrometry does not extend to other proteins without a standard curve[9], new standard curves for each MHC allele or a different way of evaluating the computed peptides or would be necessary.

Methods

Expression and purification of soluble H-2K^d in CHO cells: A stable CHO cell line expressing a secreted form of K^d using a glutamine synthetase amplification system was used[10]. The protein was purified using immunoaffinity columns constructed with immobilized monoclonal antibodies, either 34-1-2 (anti-H-2K^d)[11] or BBM.1 (anti-β_2m)[12]. Approximately 70% of the purified protein product does not contain endogenous peptides[10].

Peptide synthesis: Peptides were synthesized with an Applied Biosystems 433A peptide synthesizer. Preloaded resins were used, with subsequent residues coupled via Fmoc chemistry and HTBU/HOBt activation with standard 0.25 mmol scale coupling cycles. Peptides were cleaved from the solid support resin by mixing 200 mg resin with 2 ml trifluroacetic acid (TFA), 100 µl water, 150 mg phenol, 100 µl thioanisole and 50 µl ethanedithiol for two hours. The peptides were precipitated by addition of cold methyl *tert*-butyl ether, washed four times with the same solvent, and

lyophilized to remove some cleavage reaction scavengers. The peptides were further purified by reverse-phase HPLC with a Vydac C8 column using linear acetonitrile-water gradients (typically 20-30%) containing 0.1% TFA. Peptide masses were determined by MALDI-TOF mass spectrometry and were found to be within one mass unit of the expected masses.

CD thermal denaturation curves: An Aviv 62A DS spectropolarimeter equipped with a thermoelectric cell holder was used to collect CD data. All data were obtained from samples containing 10 μM MHC protein and 100 μM peptide in 5.0 mM sodium phosphate, pH 7.0 with a 3 minute equilibration time. Thermal denaturation curves were recorded using a 1.0 mm path length cell and were monitored at 223 nm over an appropriate temperature range with a 0.1 second time constant, 10 second averaging time, and 1 nm bandwidth. Melting temperatures were determined by taking the maximum of a plot of d(ellipticity)/dT versus T after averaging the data with a moving window of 3 degrees. The error in T_m is estimated to be ± 0.5 °C.

Acknowledgements

We thank the Bjorkman Lab for their gift of the CHO cell line which expresses K^d. In addition, we thank D. Penny and J. Johnson for their help with tissue culture.

References

1. Herrath, M.v., Coon, B., Lewicki, H., Mazarguil, H., Gairin, J. & Oldstone, M. *In vivo* treatment with a MHC class I-restricted blocking

peptide can prevent virus-induced autoimmune diabetes. *J. Immunology* **161**, 5087-5096 (1998).

2. Bjorkman, P.J. & Parham, P. Structure, function, and diversity of class i major histocompatibility complex molecules. *Annu. Rev. Biochem.* **59**, 253-288 (1990).

3. Fahnestock, M.L., Johnson, J.L., Feldman, R.M.R., Tsomides, T.J., Mayer, J., Narhi, L.O. & Bjorkman, P.J. Effects of peptide length and composition on binding to an empty class I MHC heterodimer. *Biochemistry* **33**, 8149-8158 (1994).

4. Fremont, D., Matsumura, M., Stura, E., Peterson, P. & Wilson, I. Crystal structures of two viral peptides in complex with murine MHC class I H-2K^b. *Science* **257**, 919-927 (1992).

5. Schueler-Furman, O., Elber, R. & Margalit, H. Knowledge-based structure prediction of MHC class I bound peptides; a study of 23 complexes. *Folding & Design* **3**, 549-564 (1998).

6. The calculations described here were performed by Barry D. Olafson.

7. Morgan, C., Holton, J., Olafson, B., Bjorkman, P. & Mayo, S. Circular dichroism determination of class I MHC-peptide equilibrium dissociation constants. *Protein Science* **6**, 1771-1773 (1997).

8. Street, A. & Mayo, S. Pairwise calculation of protein solvent-accessible surface areas. *Folding & Design* **3**, 235-258 (1998).

9. Lakey, J. & Raggett, E. Measuring protein-protein interactions. *Curr. Opin. Struc. Biol.* **8**, 119-123 (1998).

10. Fahnestock, M.L., Tamir, I., Narhi, L. & Bjorkman, P.J. Thermal stability comparison of purified empty and peptide-filled forms of a class I MHC molecule. *Science* **258**, 1658-1662 (1992).

11. Ozato, K., Mayer, N.M. & Sachs, D.H. Monoclonal antibodies to mouse major histocompatibility complex antigens: a series of hybridoma clones producing anti-H-2D antibodies and an examination of expression of H-2D antigens on the surface of these cells. *Transplantation* **34**, 113-120 (1982).

12. Parham, P., Androlewicz, M., Holmes, N. & Rothenberg, B. Arginine-45 is a major part of the antigenic determinant of human beta2-microglobulin recognized by mouse monoclonal-antibody BBM.1. *J. Biol. Chem.* **258**, 6179-6186 (1983).

Table 5-1

Sequences, thermal denaturation temperatures, and dissociation constants for predicted nonameric peptides complexed to K^d. The dissociation constants are calculated by CD thermal denaturation experiments[7].

Peptide	Sequence	T_m (°C)	K_D (M)
CP1	DYFRKEMHI	50.0	1.9×10^{-3}
CP2	DYHRREMHI	47.0	3.2×10^{-2}
CP3	TYFHTLHFI	65.0	1.6×10^{-9}
CP4	TYFRTREHI	65.0	1.6×10^{-9}
CP5	DYFRKRMHI	52.0	3.0×10^{-4}
CP6	TYQRTFALV	56.0	7.2×10^{-6}
CP7	TYFRTREHE	54.0	4.7×10^{-5}
CP8	TYFRKRMHI	58.0	1.1×10^{-6}
CP9	DYFRTREHI	58.0	1.1×10^{-6}
CP10	TYFRTRMHI	63.0	1.1×10^{-8}
NP1	TYQRTRALV	61	6.6×10^{-8}
KD1	SYFPEITHI	61	6.6×10^{-8}

Chapter 6

Protein Design of Toxin Folds

Abstract

Agitoxin 2, a 38-residue polypeptide neurotoxin isolated from scorpion venom, was made synthetically in the reduced form. Agitoxin 2 contains three conserved disulfide bonds, C8-C28, C14-C33, and C18-C35, which were oxidized. Circular dichroism studies of reduced and oxidized agitoxin demonstrate that agitoxin is folded correctly after oxidation, and therefore the disulfide bridges were correctly paired. Design calculations to remove the disulfide bridges are described. Subsequent work by others towards replacing disulfide bonds with the isosteric unnatural amino acid L-α-aminobutyric acid in other toxin folds suggest that the C8-C28 and C14-C33 disulfide bonds of agitoxin are not critical for structure, but redesign of the C18-C35 disulfide bond may result in an unfolded molecule if the other two disulfide bonds are present. Other approaches to protein design of toxin folds are discussed.

Introduction

Rational protein design seeks to design well-folded stable proteins. Our protein design strategy has done so without the use of metal binding sites or disulfide bonds. Along that vein, we attempted to remove conserved disulfide bonds from a native fold, while restricting the design to natural amino acids. Agitoxin, a small scorpion toxin, was selected as the target fold.

Polypeptide toxins have been isolated from scorpions, spiders, and plants based on their ability to block ion channels. Toxins bind with varying affinities to both voltage- and Ca^{2+}-dependent K^+ channels. The mechanism

of action of some peptide toxins is presumed to be due to a single molecule binding to a receptor on the outside surface of an ion channel, thereby occluding the ion pore entrance and blocking ion flow[1]. The ability of toxins to bind to ion channels has led to their use as tools to identify ion pore-forming regions of K^+ channels[2]. In addition, toxin folds have been used as calipers to measure the diameter of ion channel pores[3].

There are two general types of scorpion toxin folds: large toxins (60-66 residues) which block Na^+ channels, and small toxins (31-38 residues) which recognize K^+ channels, the most diverse family of ion channels[4]. Scorpion toxins usually have a triple stranded βsheet, an α-helix, an extended fragment, and three disulfide bonds. Larger toxins have more elements of secondary structure and up to five disulfide bonds. Six half-cystines and one glycine are the only conserved residues among the family of scorpion toxin folds. These residues lie in the consensus sequence $X_nCX_nCX_3CX_nGXCX_nCXCX_n$. Some toxins from other organisms share the same consensus sequence, such as insect defensins[4], while others have very similar sequences, such as the K^+ channel inhibitor BgK from a sea anemone[5].

Agitoxins 1, 2, and 3 are small polypeptide neurotoxins isolated from the venom of the scorpion *Leiurus quinquestriatus hebraeus* based on their ability to block Shaker K^+ channels[6]. They contain 38 amino acids including 6 cysteines in the pattern $X_7CX_5CX_3CX_7GXCX_4CXCX_3$, characteristic of scorpion folds. Agitoxins 1, 2, and 3 differ only at positions 7, 29, and 41. Agitoxin 2 has been shown to fold into a βαββ motif held together with the three disulfide bonds C8-C28, C14-C33, and C18-C35, shown in Figure 6-1[7].

Agitoxin 2 (subsequently referred to as agitoxin) was chosen as the target fold for protein redesign of disulfide bonds because it has three disulfide bonds, a good quality NMR structure is available, and it is small enough to synthesize easily.

Synthesis of agitoxin

Wild type agitoxin was synthesized in the reduced form by standard solid-phase peptide synthesis techniques. Reduced agitoxin is unfolded at 25 and 98 °C, as shown by the circular dichroism (CD) wavelength scans in Figure 6-2. The cysteine residues were oxidized by treatment with a 10:1 mixture of reduced and oxidized glutathione. After oxidation, the CD wavelength scans are quite different; the peptide appeared to be folded at moderate and high temperatures, shown in Figure 6-3. The disulfide bridges were paired in a single orientation, evidenced by a single HPLC peak. The pairings were assumed to be correct since the molecule has the expected CD wavelength signature of a folded βαββ toxin. The melting temperature of agitoxin is above 90 °C.

Protein design of agitoxin

The goal of our protein design work on agitoxin was to redesign the half-cystine residues of agitoxin thereby eliminating the disulfide bonds. The approach included computing new residues to replace the half-cystines systematically as well as a full core design. No metal binding sites, new

disulfide bonds, or unnatural amino acids would be introduced, restricting the design to naturally occurring amino acids.

The NMR solution structure of agitoxin structures was reported in 1995[7]. We chose to use the second of the ensemble of seventeen structures as our template, since it has the lowest root-mean-square deviation to the mean structure. Of the thirty-eight agitoxin residues, G26, M23, as well as the half-cystines C8, C14, C18, C28, C33, and C35 were calculated to be core residues, as described in the Methods of Chapter 2.

Design of the core residues resulted in mutations of the half-cystines to alanine residues, whether the calculations were performed by designing one disulfide bridge at a time or all core residues at once. Radical scaling of the van der Waals radii resulted in selection of slightly larger hydrophobic residues. We were not confident that these mutations would produce correctly folded molecules and the design aspect of the project was postponed.

Removal of disulfide bonds in other toxin folds

Subsequent work on the disulfide bonds of peptide toxins has been done by other groups. In addition, other approaches to protein design of toxin folds have been reported. Since the work on scorpion toxin folds is limited, many types of toxin folds are discussed.

To probe the importance of the disulfide bonds in folding and function, several groups have systematically replaced two half-cystines with the unnatural amino acids L-α-aminobutyric acid (Aba) or alanine residues.

Others have designed new disulfide bonds into existing toxin folds and redesigned the non-cystine residues of folds.

Leiurotoxin I, also called scyllatoxin, is neurotoxin from the Israeli scorpion *Leiurus quinquestriatus hebraeus* and blocks Ca^{2+}-activated K^+ channels[8]. Leiurotoxin is small, containing 31 residues and three disulfide bonds in an αββ motif, analogous to the βαββ motif of agitoxin without the first β-strand, shown in Figure 6-4. Two of the three disulfide bonds were removed systematically by replacing the two cysteine residues with Aba, which is isosteric with cysteine but incapable of forming disulfide bonds. Removal of the C3-C21 disulfide bond (analogous to C8-C28 in agitoxin) resulted in a fully active molecule with the correct disulfide C8-C26 and C12-C28 pairings and an identical structure as leiurotoxin[9,10]. However, removal of the C12-C28 bond (analogous to C18-C35 in agitoxin) results in an inactive molecule with a disordered structure. In addition, enzymatic cleavage of this peptide indicated that the two remaining disulfide bridges are mispaired, such that C3-C8 and C21-C26 disulfide bridges are formed. Therefore, the C3-C21 disulfide bond stabilizes the molecule whereas the C12-C28 bond is responsible for structure.

Similar studies have been performed on charybdotoxin, a 37 residue polypeptide isolated from the venom of same scorpion *Leiurus quinquestriatus hebraeus*[11]. Charybdotoxin is very similar to the agitoxins, with thirty-seven residues and three disulfide bonds maintaining an αββ motif, shown in Figure 6-5[4,12]. Replacement of the C7-C28 bridge connecting the loop and

first β-strand (analagous to the C8-C28 disulfide bridge of agitoxin) results in a molecule which folds correctly with the correct pairing of the other disulfide bonds in high yields, though with decreased nativelike characteristics and 180-fold lower affinity for K^+ channels[13]. When the cysteines of the C13-C33 bridge between the helix and second β-strand (analagous to the C14-C33 disulfide bridge of agitoxin) are replaced by Aba, the resulting polypeptide has correct disulfide bridges, is folded, and has a 9-fold reduction of biological activity. However, the other disulfide bonds are more likely to be mispaired. Removal of the third disulfide bridge also between the helix and last strand, C17-C35, (analagous to the C18-C35 disulfide bridge of agitoxin) results in a molecule with low native-like characteristics, mispaired remaining disulfide bonds, and 580-fold reduced biological activity. Therefore, some disulfide bonds may be present in toxin folds to induce correct cystine pairings between others which may be responsible for structure and stability.

κ-Bungarotoxin is a polypeptide neurotoxin isolated from *Bungarus multicinctus*, the banded Krait snake, and blocks nicotinic acetylcholine receptors. Unlike the scorpion toxins above, bungarotoxin is a sixty-six residue homedimeric β-sheet protein with five disulfide bonds, shown in Figure 6-6[14]. Systematic replacement of the half-cystines with alanine residues indicated different roles of the disulfide bridges in folding and function. Removal of the C46-C58 or C59-C64 disulfide bridges results in molecules which do not fold, indicative of a structural role for these disulfide

bonds[15]. Polypeptides lacking either disulfide bonds C3-C21 or C14-C42 fold and show native biological activity, although removal of both bonds concurrently prevents the protein from folding at all. Removal of the C27-C31 disulfide bond, which has no analogous bond in shorter toxins, results in a folded protein with slightly reduced activity, indicating that this disulfide bond may be responsible for specificity, not structure.

Although disulfide bonds are often essential for correct folding of toxin, an exact structure is not necessarily essential for toxicity. Two polypeptide neurotoxins, Oh-6A and Oh-6B, isolated from venom of the king cobra *Ophiophagus hannah*, share the same seventy amino acid sequence but not the same conformation[16]. One of the five disulfide bonds, C26-C30, is present in *cis-trans* isomers. Both Oh-6A and Oh-6B have the same affinity for nicotinic acetylcholine receptors.

Protein design of toxin folds

Toxin folds have not been extensively used as a tool for protein design and little progress towards toxin redesign has been made to date. There are still no toxin folds which have been engineered to fold without disulfide bonds or metal sites and few which fold without unnatural amino acids.

Another approach towards toxin design is the introduction of a metal binding site. Using the Zn^{2+} binding site of carbonic anhydrase as a model, a Cu^{2+} binding site was engineered into the charybdotoxin fold by mutating eight residues, including introduction of three histidine residues to serve as metal-binding ligands[17]. The resulting protein, which still contains the three

conserved toxin disulfide bonds, binds Cu^{2+} with a K_D of 4.2×10^{-8} M and folds similarly to native charybdotoxin.

The four disulfide bonds of cardiotoxin, a 60-residue polypeptide isolated from cobra venom, were removed and the resulting molecule was stabilized by introduction of two Ca^{2+} binding sites[18]. All eight half-cystine residues were replaced by glycine. The Ca^{2+} binding sites were designed by a 5-fold mutation, L1E, L26E, S28E, L48E, and S55E, to utilize glutamate carboxyl groups as Ca^{2+} ligands. In the presence of calcium, the structure of the designed cardiotoxin is similar to that of the native toxin and biological activity was reduced by only 35%.

A new disulfide bond was engineered into heat-labile enterotoxin LT-I to probe conformational restriction and increase stability[19]. LT-I, which is a large heterohexameric protein produced by enterotoxigenic *Escherichia coli*, does not share structural features with scorpion toxins. The designed protein is a N40C and G166C double mutant of LT-I, shares the same structure as the wild-type, and has a 6 °C increase in thermal stability.

Conclusions

Agitoxin was successfully synthesized in the reduced form. The peptide was found to make the correct disulfide bonds upon oxidation. The relative experimental ease of obtaining the protein makes agitoxin a good target fold for protein design, which necessitates the ability to make designed proteins quickly. However, our preliminary calculations and experiments towards removing disulfide bonds were unsuccessful.

Later research which involved replacing disulfide bonds from similar neurotoxins, charybdotoxin and leiurotoxin, with unnatural amino acids suggest that the agitoxin disulfide bond C8-C28 would be the most logical bond to redesign first[9,10,13]. Removal of the analogous disulfide bridge in charybdotoxin and leiurotoxin resulted in folded functional mutants.

Based on the results of other toxin disulfide bond studies, it is likely that the C8-C28 agitoxin disulfide bond could be redesigned without adversely affecting the structure or function of agitoxin. Similarly, the C14-C33 agitoxin disulfide bridge could be redesigned resulting in a functional molecule, but with a high incidence of incorrect pairing of the other two disulfide bonds. It is likely that removal of the C18-C35 disulfide bridge of agitoxin would result in an unfolded, unfunctional protein with the other disulfide bonds mispaired. However, if the redesign of the other two disulfide bonds is successful, it may be possible to remove the C18-C35 disulfide bond, since its role may be to prevent mispairing of the other two bonds, rather than an inherent structural role.

It is possible that small toxin folds may not be amenable to protein design. They may be constrained by the three conserved disulfide bonds in such a way that is not reproducible with only the natural twenty amino acids, without covalent bonds, or metal binding sites.

Materials and methods

Synthesis of agitoxin: Agitoxin was synthesized on an Applied Biosystems 433A peptide synthesizer. A preloaded lysine resin was used with subsequent residues coupled via Fmoc chemistry and HTBU/HOBt

activation with standard 0.10 mmol scale coupling cycles. The peptide was cleaved from the solid support resin by mixing 200 mg resin with 2 ml trifluroacetic acid (TFA), 100 μL water, 150 mg phenol, 100 μL thioanisole and 50 μL ethanedithiol for two hours. The peptide was precipitated by addition of cold methyl *tert*-butyl ether, washed four times with the same solvent, and lyophilized to partially remove the cleavage reaction scavengers. The peptide was further purified by reverse-phase HPLC with a Vydac C8 column using linear acetonitrile-water gradients (typically 15-25%) containing 0.1% TFA. The reduced peptide mass was determined by MALDI-TOF mass spectrometry to be 4097.3 (expected 4098.0). The concentration of agitoxin was determined by amino acid analysis.

Formation of disulfide bonds: The three disulfide bonds of agitoxin were formed by dissolving the reduced peptide (either before or after an initial HPLC purification) in a solution of 20 mM sodium phosphate pH 7.8 containing 5.0 mM reduced glutathione, 0.50 mM oxidized glutathione, and 0.20 M NaCl. The solution was stirred at room temperature for three hours and then purified by HPLC as described above. Analytical HPLC analysis using a 0-45% gradient over 45 minutes showed a shift in retention time from 26 minutes for the reduced form to 22 minutes for the oxidized form. The mass of the oxidized agitoxin was determined to be 4092.9 (expected 4092.0).

CD Studies: An Aviv 62A DS spectropolarimeter equipped with a thermoelectric cell holder was used to collect CD data. Data were obtained from samples containing 13.8 μM agitoxin in 5.0 mM sodium phosphate at pH 4.5 and 7.0 using a 1.0 mm path length cell. Wavelength scan data were collected from 250-190 nm at even wavelength intervals for two seconds at 25

°C. Scans at 25 °C after the proteins were exposed to high temperature (99 °C) were not significantly different. Thermal denaturation curves were recorded at 222 nm from 24-98 °C in two degree intervals with a 0.1 second time constant, 10 second averaging time, 2 minute equilibration time, and 1 nm bandwidth. The melting temperature is sufficiently high (> 90 °C) that it could not be determined by CD spectrometry.

References

1. Park, C. & Miller, C. Interaction of charybdotoxin with permeant ions inside the pore of a K^+ channel. *Neuron* **9**, 307-313 (1992).
2. MacKinnon, R., Heginbotham, L. & Abramson, T. Mapping the receptor-site for charybdotoxin, a pore-blocking potassium channel inhibitor. *Neuron* **5**, 767-771 (1990).
3. Goldstein, S., Pheasant, D. & Miller, C. The charybdotoxin receptor of a shaker K^+ channel - peptide and channel residues mediating molecular recognition. *Neuron* **12**, 1377-1388 (1994).
4. Bontems, F., Roumestand, C., Gilquin, B., Ménez, A. & Toma, F. Refined structure of charybdotoxin: common motifs in scorpion toxins and insect defensins. *Science* **254**, 1521-1523 (1991).
5. Cotton, J., Crest, M., Bouet, F., Alessandri, N., Gola, M., Forest, E., Karlsson, E., Castañeda, O., Harvey, A., Vita, C. & Ménez, A. A potassium-channel toxin from the sea anemone *Bunodosoma granulifera*, an inhibitor for Kv1 channels. *Eur. J. Biochem.* **244**, 192-202 (1997).
6. Garcia, M., Garcia-Calvo, M., Hidalgo, P., Lee, A. & MacKinnon, R. Purification and characterization of three inhibitors of voltage-

dependent K^+ channels from *Leiurus quinquestriatus* var. *hebraeus* venom. *Biochemistry* **33**, 6834-6839 (1994).

7. Krezel, A., Kasibhatla, C., Hildalgo, P., MacKinnon, R. & Wagner, G. Solution structure of the potassium channel inhibitor agitoxin 2: caliper for probing channel geometry. *Protein Science* **4**, 1478-1489 (1995).

8. Chicchi, G., Gimenez-Gallego, G., Ber, E., Garcia, M., Winquist, R. & Cascieri, M. Purification and characterization of a unique, potent inhibitor of apamin binding from *leiurus-quinquestriatus-hebraeus* venom. *J. Biol. Chem.* **263**, 10192-10197 (1988).

9. Calabro, V., Sabatier, J., Blanc, E., Lecomte, C., Rietschoten, J.V. & Darbon, H. Differential involvement of disulfide bridges on the folding of a scorpion toxin. *J. Peptide Res.* **50**, 39-47 (1997).

10. Sabatier, J., Lecomte, C., Mabrouk, K., Darbon, H., Oughideni, R., Canarelli, S., Rochat, H., Martin-Eauclaire, M. & VanRietschoten, J. Synthesis and characterization of leiurotoxin I analogs lacking one disulfide bridge: evidence that disulfide pairing 3-21 is not required for full toxin activity. *Biochemistry* **35**, 10641-10647 (1996).

11. Gimenez-Gallo, G., Navia, M., Reuben, J., Katz, G., Kaczorowski, G. & Garcia, M. Purification, sequence, and model structure of charybdotoxin, a potent selective inhibitor of calcium-activated potassium channels. *Proc. Natl. Acad. Sci. USA* **85**, 3329-3333 (1988).

12. Bontems, F., Roumestand, C., Gilquin, B., Ménez, A. & Toma, F. Refined structure of charybdotoxin: common motifs in scorpion toxins and insect defensins. *Science* **254**, (1991).

13. Drakopoulou, E., Vizzavona, J., Neyton, J., Aniort, V., Bouet, F., Virelizier, H., Ménez, A. & Vita, C. Consequence of the removal of evolutionary conserved disulfide bridges on the structure and function of charybdotoxin and evidence that particular cysteine spacings govern specific disulfide bond formation. *Biochemistry* **37**, 1292-1301 (1998).

14. Sutcliffe, M., Dobson, C. & Oswald, R. Solution structure of neuronal bungarotoxin determined by two-dimensional NMR spectroscopy: calculation of tertiary structure using systematic homologous model building, dynamical simulated annealing, and restrained molecular dynamics. *Biochemistry* **31**, 2962-2970 (1992).

15. Grant, G., Luetje, C., Summers, R. & Xu, X. Differential roles for disulfide bonds in the structural integrity and biological activity of κ-bungarotoxin, a neuronal nicotinic acetylcholine receptor antagonist. *Biochemistry* **37**, 12166-12171 (1998).

16. Lin, S.-r., Chang, L.-s. & Chang, C.-c. Disulfide isomers of α-neurotoxins from kind cobra (*Ophiophagus hannah*) venom. *Biochem. and Biophys. Res. Comm.* **254**, 104-108 (1999).

17. Vita, C., Roumestand, C., Toma, F. & Ménez, A. Scorpion toxins as natural scaffolds for protein engineering. *Proc. Natl. Acad. Sci. USA* **92**, 6404-6408 (1995).

18. Chen, S.-t., Yang, M.-t., Wu, S.-y. & Wang, K.-t. Protein engineering III: computer aided design of a disulfide-bond free cobra venom

cardiotoxin with a novel confirmation and biological activity from synthetic peptides. *J. Chinese Chem. Soc.* **44**, 331-335 (1997).

19. Akker, F.v.d., Feil, I., Roach, C., Platas, A., Merritt, E. & Hol, W. Crystal structure of heat-labile enterotoxin from *Escherichia coli* with increased thermostability introduced by an engineered disulfide bond in the A subunit. *Prot. Sci.* **6**, 2644-2649 (1999).

20. Martins, J., Zhang, W., Tartar, A., Lazdunski, M. & Borremans, F. Solution conformation of leiurotoxin I (scyllatoxin) by ^{1}H nuclear magnetic resonance assignment and secondary structure. *FEBS Lett.* **260**, (1990).

21. Bontems, F., Gilquin, B., Roumestand, C., Ménez, A. & Toma, F. Analysis of side-chain organization on a refined model of charybdotoxin: structural and functional implications. *Biochemistry* **31**, 7756-7764 (1992).

Figure 6-1

NMR solution structure of agitoxin 2[7]. The conserved disulfide bonds, C8-C28, C14-C33, and C18-C35 are indicated.

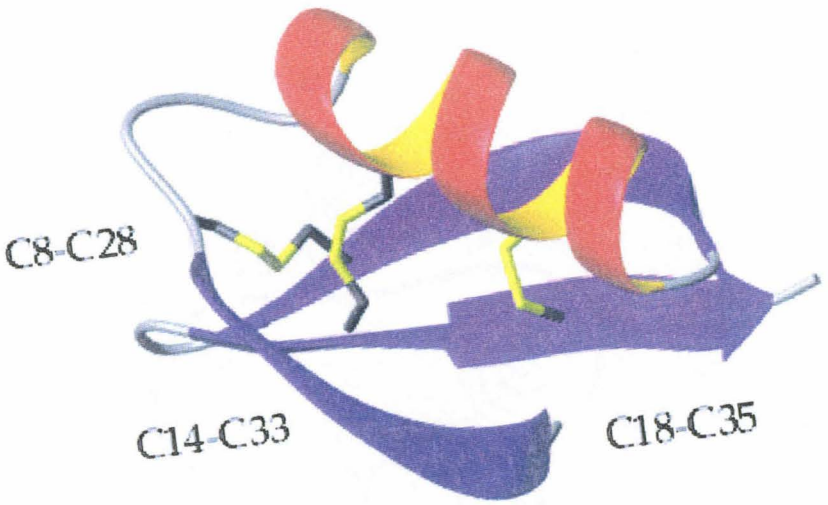

C8-C28
C14-C33
C18-C35

Figure 6-2

The CD wavelength scans of 13.8 μM reduced agitoxin, pH 7.0, in 5mM sodium phosphate at 25 °C (blue) and 98 °C (red) are indicative of an unfolded protein.

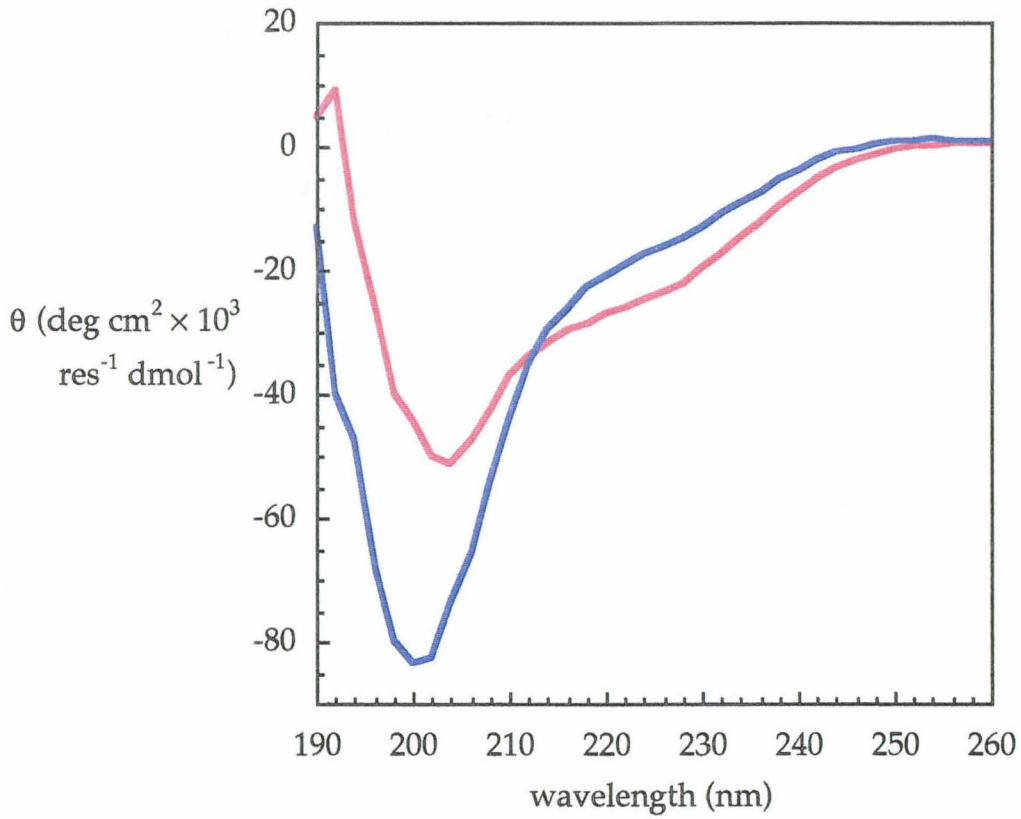

20
0
-20
-40
-60
-80
θ (deg cm^2 × 10^3 res^{-1} dmol^{-1})
190
200
210
220
230
240
250
260
wavelength (nm)

Figure 6-3

The CD wavelength scans of 13.8 μM oxidized agitoxin, pH 7.0, in 5mM sodium phosphate at 28 °C (blue) and 98 °C (red) are indicative of a folded structure.

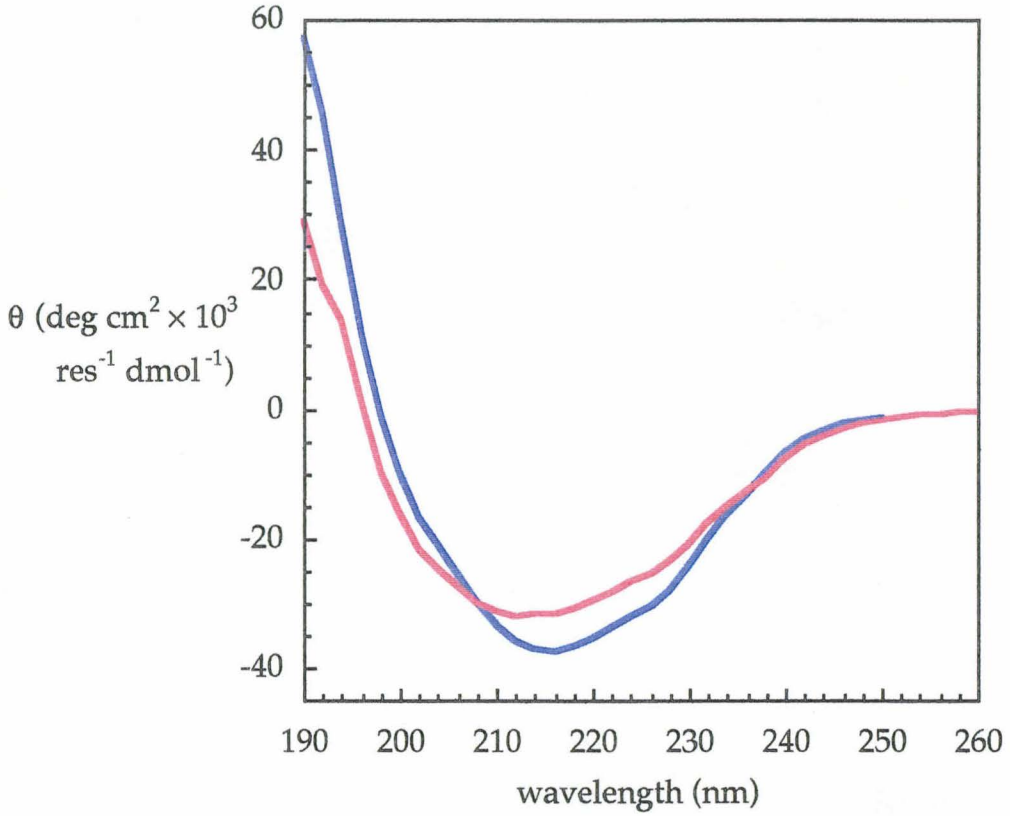

60
40
20
0
-20
-40
θ (deg cm² × 10³ res⁻¹ dmol⁻¹)
190
200
210
220
230
240
250
260
wavelength (nm)

Figure 6-4

NMR solution structure of leiurotoxin I^{20}. This 31-residue protein has an analogous structure to agitoxin without the first β-strand. Conserved disulfide bridges are shown.

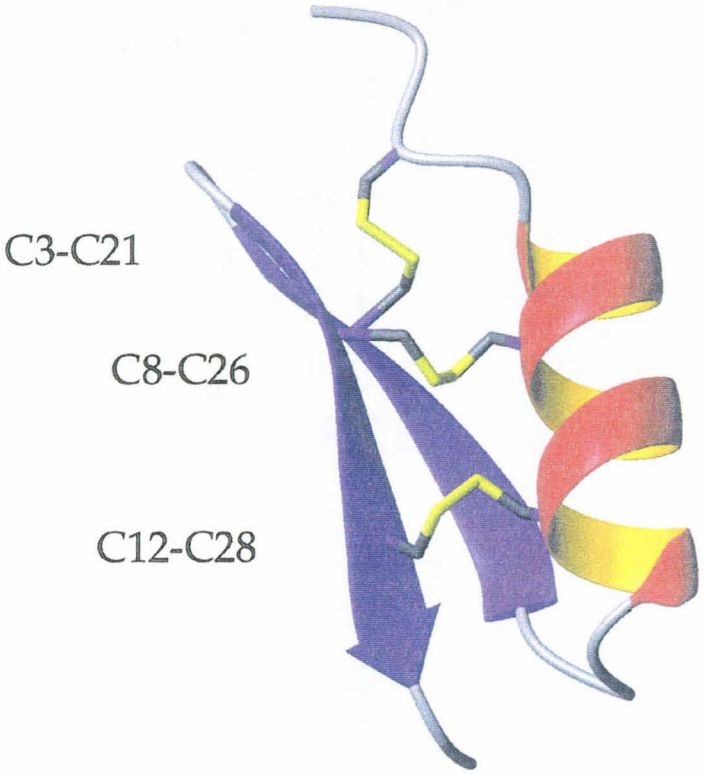

C3-C21
C8-C26
C12-C28

Figure 6-5

NMR solution structure of charybdotoxin[12,21]. The conserved disulfide bridges are indicated.

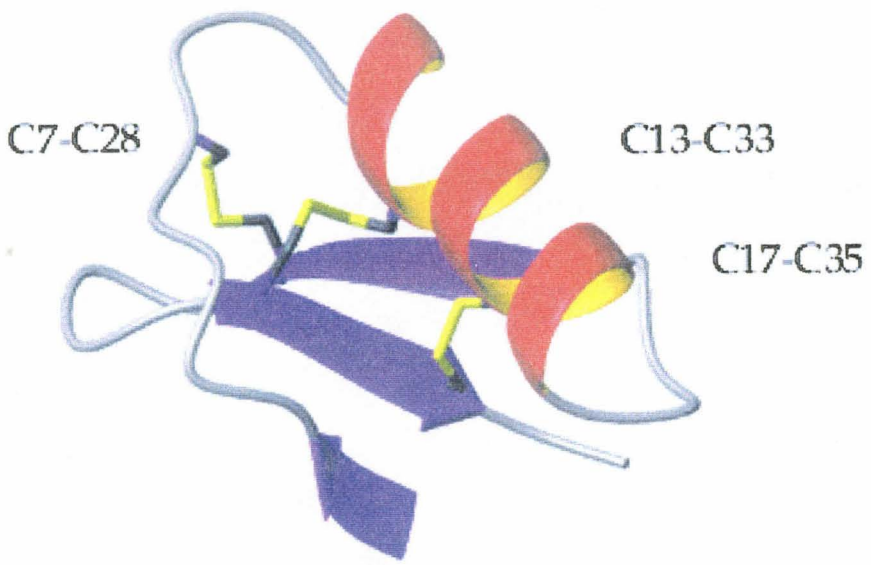
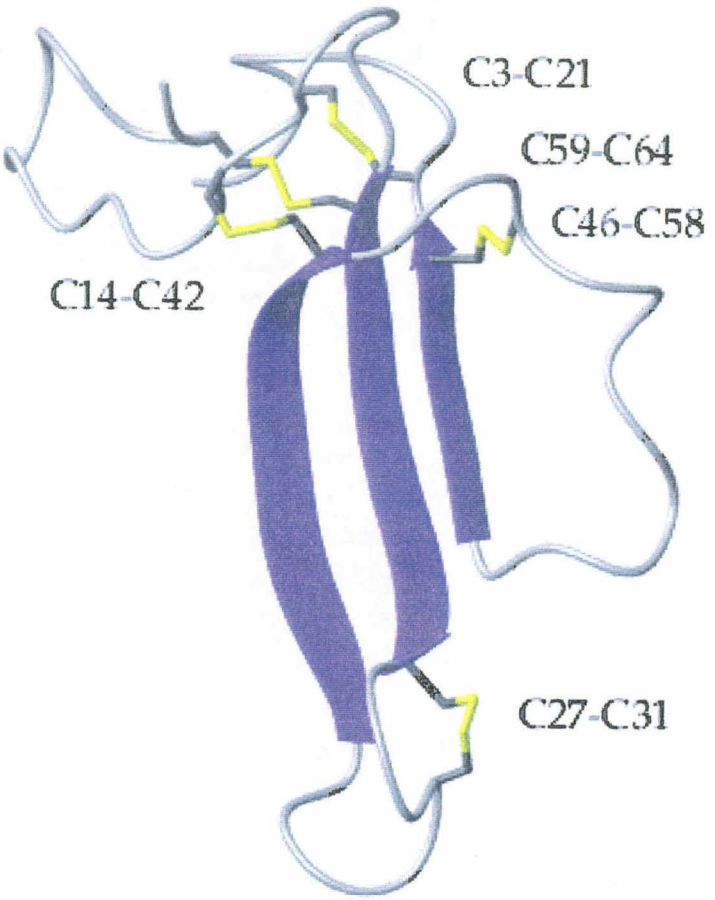

C7-C28
C13-C33
C17-C35

Figure 6-6

NMR solution structure of κ-bungarotoxin[14]. This structure is a larger toxin fold held together by five disulfide bridges.


C3-C21
C59-C64
C46-C58
C14-C42
C27-C31