The Pennsylvania State University
The Graduate School
Eberly College of Science

COMPUTATIONAL DESIGN AND EXPERIMENTAL CHARACTERIZATION OF PROTEIN DOMAINS

A Thesis in
Chemistry
by
Christina M. Kraemer Pecore

© 2002 Christina M. Kraemer Pecore

Submitted in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

December 2002
We approve the thesis of Christina M. Kraemer Pecore.

Date of Signature

Juliette T. J. Lecomte
Associate Professor of Chemistry
Thesis Co-Advisor
Chair of Committee

B. Tracy Nixon
Associate Professor of Biochemistry and Molecular Biology

Peter C. Jurs
Professor of Chemistry

Christopher J. Falzone
Lecturer in Chemistry

John R. Desjarlais
Director, Computational Biology
Xencor, Inc., Monrovia, California
Thesis Co-Advisor
Special Member

Andrew G. Ewing
Professor of Chemistry
Head of the Department of Chemistry
ABSTRACT

This thesis describes the first successful computational redesign of a β-sheet protein. The high-resolution backbone structure of the WW domain from human peptidyl-prolyl cis-trans isomerase (hPin1) was used as input into an algorithm developed previously for the generation of amino acid sequences compatible with a given backbone geometry (Sequence Prediction Algorithm, or SPA). Various weights of protein folding parameters such as hydrophobic burial, hydrophilic exposure, electrostatic interactions, Lennard-Jones potential and side chain torsional energy were applied for the selection of a sequence of amino acids which, when placed on the target experimental backbone in the predicted conformations, resulted in a native-like protein model in terms of packing, electrostatic interactions, hydrophobic and hydrophilic patterning, and other necessary criteria. One protein sequence was chosen for experimental characterization to assess the usefulness of SPA in selecting a primary structure compatible with the wild-type WW fold. A synthetic gene for this protein was cloned into a previously developed plasmid using standard molecular biology techniques. Although expression was achieved to a high level as a fusion system with the N-terminus of calmodulin, difficulties in purifying the protein product using a variety of modern techniques prevented the structural analysis of the material, and a different approach to the design was adopted.

To avoid the summary rejection of sequences on the basis of a few minor steric clashes during the process of computational design, backbone flexibility was implemented to relieve potential strains. Flexibility was mimicked through the
application of a Monte Carlo algorithm, which allowed for random movements in the 
$(\phi, \psi)$ space of the wild-type backbone. These manipulations produced an ensemble of 
backbones, all of which had a root mean squared deviation <0.3 Å to the wild-type 
backbone. Each of these backbones in the ensemble was used as input into SPA; the 
result was a free energy matrix of each allowed amino acid and its rotamers, which was 
used to determine the probability of occurrence of an amino acid and rotamer 
combination in the lowest-energy sequence. This new method was referred to as SPANS 
for Sequence Prediction Algorithm on Numerous States. Three promising sequences of 
36 amino acids were chosen to test the power of the improved algorithm. The 
corresponding proteins were prepared with standard molecular biology methods, again as 
fusion systems with the N-terminus of calmodulin. Circular dichroism (CD) data 
indicated the presence of a WW-like target fold in one of the designed proteins (referred 
to as SPANS-WW2) via positive ellipticity centered on ~230 nm. Others have identified 
this feature as a characteristic of the CD spectrum of wild-type WW domains, although in 
the instance of the WW domain from hPin1, the wild-type protein utilized in this study, 
the signal is stronger. The positive ellipticity of SPANS-WW2 increased in intensity 
two-fold after the sample was heated to 95 °C for five minutes at pH 7 and subsequently 
cooled – a process referred to as annealing. Additionally, a single mutation in the wild-
type WW domain (W29A) yielded a protein with a positive ellipticity comparable to that 
of the SPANS-WW2 designed protein. These observations suggested that the reduction 
in CD signal was due to interactions between the solvent-exposed tryptophan and the 
aromatic amino acids located near this residue, rather than backbone conformation. The
specificity of the SPANS-WW2 fold was confirmed by initial 1D and 2D proton NMR spectroscopy, which indicated the presence of the target WW-like fold. However, the thermal stability of the designed protein was decreased compared to that of the wild-type WW, as shown by CD thermal denaturation and 1D variable temperature NMR spectroscopy. Efforts towards the characterization of the other two proteins designed by the SPANS method yielded a limited amount of success according to CD data. SPANS highlighted several potentially useful point mutations in these two designed proteins; only one of the mutations increased the presence of a WW-type fold as evidenced by slight changes in the CD signature. The ability to identify interesting or useful point mutations increased the efficacy of the SPANS algorithm in the design of a protein that has high specificity for a target fold.

This work provides strong experimental evidence of the success of the algorithm in selecting for a sequence that adopted and maintained a β-sheet fold with marginal stability. As most previous attempts at protein design have focused on proteins that were entirely or primarily α-helical, this success is especially noteworthy.
# TABLE OF CONTENTS

## LIST OF FIGURES

x

## LIST OF TABLES

xiv

## LIST OF ABBREVIATIONS

xv

## ACKNOWLEDGEMENTS

xvii

## CHAPTER 1: INTRODUCTION TO THEORETICAL PROTEIN DESIGN

1

1.1 Introduction to Protein Folding Parameters 1

1.2 Definitions of Protein Design Problems 12

1.3 Physical Parameters in Protein Design 14

1.4 History and Importance of Protein Design 19

1.5 Problems of Protein Design 22

1.6 Methods of Protein Design 26

## CHAPTER 2: INTRODUCTION TO EXPERIMENTAL PROTEIN CHEMISTRY

33

2.1 Protein Purification Tools 33

2.2 Protein Fusion Systems 37

2.3 Sodium dodecyl-sulfate Polyacrylamide Gel Electrophoresis 38

2.4 Mass Spectrometry 40

2.5 Ultraviolet/visible (UV/vis) Spectroscopy 42

2.6 Circular Dichroism Spectroscopy 43

2.7 Fluorescence Spectroscopy 45

2.8 Nuclear Magnetic Resonance (NMR) Spectroscopy 46

## CHAPTER 3: INTRODUCTION TO MOLECULAR BIOLOGY

49

3.1 Overview of Molecular Biology Concepts 49

3.2 Codons 52

3.3 Plasmids 53
3.4 Developing and Cloning Insert of Interest into a Cut Vector  59
3.5 Purification of DNA       64
3.6 The Polymerase Chain Reaction (PCR)     68
3.7 DNA Sequencing               70
3.8 Transformations of Bacterial Cells with Plasmid DNA  71

CHAPTER 4: COMPUTATIONAL ALGORITHMS FOR PROTEIN DESIGN:
SEQUENCE PREDICTION ALGORITHM (SPA) AND SEQUENCE
PREDICTION ALGORITHM ON NUMEROUS STATES (SPANS)    74

4.1 Introduction        74
4.2 Amino Acid Baseline Corrections     76
4.3 Secondary Structure Propensity       79
4.4 Hydrophobic/Polar (HP) Patterning     80
4.5 Description of Rotamers     82
4.6 Steric Filtering               84
4.7 Energy Calculations and Parameters Involved  87
4.8 Initialization of Rotamer Roulette Wheel  92
4.9 Sequence Prediction Algorithm (SPA)      93
4.10 Application of Monte Carlo Algorithm in SPANS  95
4.11 Free Energy Matrix for Amino Acid Probability Generation  98

CHAPTER 5: COMPUTATIONAL ALGORITHMS FOR PROTEIN DESIGN:
DE NOVO THREE-HELIX BUNDLES 102

5.1 De Novo Design of Three-Helix Bundles – Introduction  102
5.2 Design of Loops to Connect the Three Helices  112
5.3 Backbone Flexibility               113
5.4 Negative Design                   114
5.5 Structural Specificity and Sequence Diversity  117

CHAPTER 6: APPLICATIONS OF SPA/SPANS     121

6.1 Introduction to the WW Domain  121
6.2 The Vector – N-Cam.Y and Nla Fusion System  125
6.3 Obtaining the Gene Coding for the Designed Plasmid  136
6.4 Point Mutations Using PCR Mutagenesis  139
6.5 Expression in BL21or HMS174 Bacterial Cells  142
6.6 Purification and Cleavage of WW Fusion Systems  145
6.7 Experimental Characterization of the WW Proteins  149
6.8 Nla Protease Purification Procedure  155
APPENDIX B: FORTRAN 90 PROGRAMS WRITTEN FOR THE DESIGN OF A THREE-HELIX BUNDLE PROTEIN
LIST OF FIGURES

Figure 1.1: Backbone torsion angles 3
Figure 1.2: Ramachandran plot 3
Figure 1.3: Possible decoy and target energies 25
Figure 1.4: Flowchart of a typical GA used in this work 29
Figure 1.5: HP lattice model conformations 32
Figure 3.1: Plasmid map of pAED4 55
Figure 3.2: Pictorial representation of two possible orientations of an insert cloned into a singly-digested vector 57
Figure 3.3: Flowchart of mutant plasmid construction 62
Figure 4.1: Side chain dihedral angles of valine 83
Figure 4.2: Backbone ensemble used in SPANS 97
Figure 4.3: The sequence of lowest calculated free energy on each of the 30 backbones in the ensemble result in a probability matrix 101
Figure 5.1: High-resolution NMR structure of the Z-domain of protein A from *S. aureus* 104
Figure 5.2: Helix manipulations 107-8
Figure 5.3: The results from an initial MC/GA combination 116
Figure 5.4: The stored sequence on each of the 199 decoys 116
Figure 5.5: Flowchart for sequence diversity/structural specificity results 120
Figure 6.1: High-resolution x-ray crystallography structure of the WW domain found in complex with hPin1 124
Figure 6.2: Representative SDS-PAGE protein gels of the scheme utilized to purify WW designs 127

Figure 6.3: The gene developed for expression of WW proteins as a fusion system with the N-terminal of a calmodulin mutant 128

Figure 7.1: CD spectrum and thermal denaturation of (His)$_6$-tagged SpZ 163

Figure 7.2: SpZ designed sequence and rotamers on the wild-type SpZ backbone 164

Figure 7.3: Results of individual steps towards the de novo design of a three-helix bundle 168

Figure 7.4: De novo-designed three-helix bundle protein 169

Figure 7.5: Ideal contact cutoff for buried positions calculated in the initial MC/GA cycle 172

Figure 7.6: Sequence A and structure A, the optimized sequence/structure pair 172

Figure 7.7: Increasing sequence diversity of the core residues did not correlate with an increase in structural specificity 175

Figure 7.8: Energy of any of the increased core entropy sequences on any of the decoy structures did not correlate with rmsd 176

Figure 8.1: CD spectrum and thermal denaturation profile of wild-type WW 179

Figure 8.2: SPA-WW1 designed sequence and rotamers on the wild-type WW backbone 181

Figure 8.3: SDS-PAGE protein gels of the attempted purification of SPA-WW1 183

Figure 8.4: Glycine-baseline designed sequence and rotamers on the wild-type WW backbone 185

Figure 8.5: SPANS-WW1 designed sequence and rotamers on the wild-type WW backbone 187

Figure 8.6: CD spectrum and thermal denaturation profile of SPANS-WW1 189
Figure 8.7: SPANS-WW2 designed sequence and rotamers on the wild-type WW backbone 192
Figure 8.8: CD spectrum and thermal denaturation profile of SPANS-WW2 194-5
Figure 8.9: Fluorescence spectrum of SPANS-WW2 (folded and unfolded); manual titration of SPANS-WW2 with urea 198
Figure 8.10: Scalar and dipolar connectivities of Trp 6 in SPANS-WW2 199
Figure 8.11: Portions of the SPANS-WW2 structural model used to assign resonances as described in the text 202-3
Figure 8.12: 1-D variable temperature spectra of SPANS-WW2 205
Figure 8.13: Backbone NH-CαH section of the NOESY spectrum of SPANS-WW2 in H₂O 206
Figure 8.14: Section of the NOESY spectrum of SPANS-WW2 in D₂O 207
Figure 8.15: Section of the NOESY spectrum of SPANS-WW2 in D₂O corresponding to contacts between aromatic amino acids and some of the Cα protons 208
Figure 8.16: Section of the NOESY spectrum of SPANS-WW2 in D₂O corresponding to contacts between some of the Cα protons 209
Figure 8.17: SPANS-WW3 designed sequence and rotamers on the wild-type WW backbone 211
Figure 8.18: CD spectrum and thermal denaturation profile of SPANS-WW3 213
Figure 8.19: CD spectrum and thermal denaturation profile of K8P 216
Figure 8.20: CD spectrum and thermal denaturation profile of N17F 218
Figure 8.21: CD spectrum and thermal denaturation profile of W29A 221
Figure 8.22: Overlay of all CD spectra of the WW proteins 222
Figure 9.1: Comparisons of the high-resolution structures of the synthesized WW domain and the WW domain in complex with hPin1 226
Figure A.1: High-resolution x-ray crystallography structure of the cold-shock protein from *Bacillus subtilis* (CspB) 252

Figure A.2: CD spectrum and thermal denaturation profile of CspB11 258

Figure A.3: Chemical denaturation of CspB11, monitored using intrinsic fluorescence 260

Figure A.4: Schematic of the PCR-assisted mutational method used in the CspB project 262
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>Helical angle and translation parameters</td>
<td>111</td>
</tr>
<tr>
<td>7.1</td>
<td><em>De novo</em>-designed three-helix bundle parameters</td>
<td>166</td>
</tr>
<tr>
<td>7.2</td>
<td>Results of placing sequences of increasing core diversity on the target and decoy structures</td>
<td>174</td>
</tr>
<tr>
<td>7.3</td>
<td>Energy gap and z-score of sequences of different core diversity on structure A</td>
<td>174</td>
</tr>
<tr>
<td>9.1</td>
<td>PSIPRED results of WW designs and wild-type WW</td>
<td>224</td>
</tr>
<tr>
<td>9.2</td>
<td>Solubilities of WW proteins</td>
<td>229</td>
</tr>
</tbody>
</table>
LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Å</td>
<td>Angstrom</td>
</tr>
<tr>
<td>Amp</td>
<td>ampicillin</td>
</tr>
<tr>
<td>ANS</td>
<td>anilinonaphthalene-8-sulfonic acid</td>
</tr>
<tr>
<td>APS</td>
<td>ammonium persulfate</td>
</tr>
<tr>
<td>asp</td>
<td>atomic solvation parameters</td>
</tr>
<tr>
<td>ATP</td>
<td>adenosine triphosphate</td>
</tr>
<tr>
<td>CD</td>
<td>circular dichroism spectroscopy</td>
</tr>
<tr>
<td>CIAP</td>
<td>calf intestinal alkaline phosphatase</td>
</tr>
<tr>
<td>CspB</td>
<td>cold shock protein from <em>Bacillus subtilis</em></td>
</tr>
<tr>
<td>Da</td>
<td>Daltons</td>
</tr>
<tr>
<td>DEE</td>
<td>dead-end elimination</td>
</tr>
<tr>
<td>dNTP</td>
<td>deoxynucleoside triphosphate</td>
</tr>
<tr>
<td>ddNTP</td>
<td>dideoxynucleoside triphosphate</td>
</tr>
<tr>
<td>DQF-COSY</td>
<td>double quantum filtered-correlated spectroscopy</td>
</tr>
<tr>
<td>DTT</td>
<td>dithiothreitol</td>
</tr>
<tr>
<td>E</td>
<td>energy</td>
</tr>
<tr>
<td>EDTA</td>
<td>ethylenediaminetetraacetic acid</td>
</tr>
<tr>
<td>ESI</td>
<td>electrospray ionization</td>
</tr>
<tr>
<td>EtBr</td>
<td>ethidium bromide</td>
</tr>
<tr>
<td>ΔG°</td>
<td>the change in Gibbs free energy at standard state</td>
</tr>
<tr>
<td>GA</td>
<td>genetic algorithm</td>
</tr>
<tr>
<td>hPin1</td>
<td>peptidyl-prolyl <em>cis-trans</em> isomerase from <em>Homo sapiens</em></td>
</tr>
<tr>
<td>hYap</td>
<td>Yes-associated protein from <em>Homo sapiens</em></td>
</tr>
<tr>
<td>HP</td>
<td>hydrophobic/polar</td>
</tr>
<tr>
<td>IgG</td>
<td>immunoglobulin G</td>
</tr>
<tr>
<td>IMAC</td>
<td>immobilized metal affinity column</td>
</tr>
<tr>
<td>IPTG</td>
<td>isopropyl-1-thio-β-D-galactopyranoside</td>
</tr>
<tr>
<td>Kd</td>
<td>dissociation constant</td>
</tr>
<tr>
<td>LB</td>
<td>Luria-Bertani broth</td>
</tr>
<tr>
<td>Lk</td>
<td>link</td>
</tr>
<tr>
<td>MALDI</td>
<td>matrix-assisted laser desorption ionization</td>
</tr>
<tr>
<td>MC</td>
<td>Monte Carlo</td>
</tr>
<tr>
<td>MOPS</td>
<td>3-(N-morpholino)propanesulfonic acid</td>
</tr>
<tr>
<td>N</td>
<td>the native state</td>
</tr>
<tr>
<td>N-Cam</td>
<td>N-terminus of calmodulin from <em>Homo sapiens</em></td>
</tr>
<tr>
<td>NOESY</td>
<td>nuclear Overhauser-effect spectroscopy</td>
</tr>
<tr>
<td>NTA</td>
<td>nitrilotriacetic acid</td>
</tr>
<tr>
<td>NTP</td>
<td>nucleoside triphosphate</td>
</tr>
<tr>
<td>NTCB</td>
<td>2-nitro-5-thiocyanobenzoic acid</td>
</tr>
</tbody>
</table>
NMR: nuclear magnetic resonance
OPLS: optimized potentials for liquid simulations
PAGE: polyacrylamide gel electrophoresis
PCR: polymerase chain reaction
P<sub>i</sub>: inorganic phosphate (PO<sub>4</sub><sup>3-</sup>)
PNK: phosphonucleotide kinase
pSer: phosphoserine
PSIPRED: protein secondary structure prediction program
pThr: phosphothreonine
QAE: quaternary aminoethyl
rmsd: root mean squared deviation
RPM: revolutions per minute
sasa: solvent-accessible surface area
SDS: sodium dodecyl sulfate
SPA: sequence prediction algorithm
SPANS: sequence prediction algorithm on numerous states
SpZ: protein A, Z domain from Staphylococcus aureus
T: temperature
TE: 10 mM Tris, 1 mM EDTA, pH 8
TEMED: N,N,N',N'-tetramethylethlenediamine
TOCSY: total correlation spectroscopy
TMAO: trimethyl amine N-oxide
TPPI: time-proportional phase-incrementation frequency discrimination
Tw: twist
U: the unfolded state
UV: ultraviolet
Wr: writhe
WT: wild-type
ACKNOWLEDGEMENTS

I would like to express my sincere thanks to John Desjarlais, who sparked my interest in protein design. John gave me a good start and the freedom to follow my interests. He was generous with his assistance as I was beginning my graduate career, and never doubted my abilities. Additionally, I would like to thank the past members of the Desjarlais lab for their support and criticism. It was always appreciated.

I am deeply indebted to Juliette Lecomte for financial assistance as well as personal and scientific guidance. Her generosity and kindness carried me through this past year. Moreover, her patience and knowledge made this thesis possible. I truly believe I am a better person and scientist because of her tutelage. I especially enjoyed being a part of her group – a better collection of people I have never met. Thanks all!

I thank my committee members for their time and participation in my thesis. I express my appreciation for their efforts, questions, comments, and suggestions, as well as their unending patience. I also would like to thank the Chemistry department for the partial grant-in-aid for the summer of 2002.

To my friends, most especially Sue the Splitter, Christie, Janell, and Jane, as well as Janet, Mike and others who have since left, I express my deep gratitude. I couldn’t have done it without all of you. So long, and thanks for all the kheer!

Finally, I would like to dedicate this thesis to my family, especially my mother, Sara, Alex and his bright, beautiful future, and particularly Paul. Throughout it all, you were there.

This is also dedicated to the memory of Paul Kraemer. I still miss you, Dad.
1.1 Introduction to Protein Folding Parameters

Proteins are linear heteropolymers comprised of individual amino acids linked by peptide bonds. There are twenty naturally occurring amino acids in biological systems; the general formula for 18 of these amino acids is $\text{H}_3\text{N}-\text{C}_\alpha\text{HR}-\text{COO}^-$, where $R$, the side chain, is endowed with different properties. Cysteine, for example, has the capability of forming disulfide bonds or bridges to another cysteine through the oxidation of its $–\text{SH}$ group ($R = \text{CH}_2\text{SH}$). Other amino acids are polar or non-polar; all polar amino acids are amphiphilic, in that they have both polar and non-polar portions. Some have side chain $\text{pK}_a$s that influence the protein’s stability at different pH conditions. Still other side chains have an aromatic substituent. The last two amino acids have $R = \text{H}$ (glycine) and $\text{H}_2\text{N} – \text{CH} – \text{COO}^-$ (proline). In the large majority of naturally occurring peptides and proteins, the stereochemistry at the $\alpha$ carbon is $L$, and further stereochemistry in $R$ is unique (both threonine and isoleucine have an additional center of asymmetry at $C\beta$; only one isomer of each occurs naturally). The order in which the amino acids occur is termed the primary structure or amino acid sequence.
The backbone moieties undergo a condensation reaction to form an amide linkage or a peptide bond; this bond has sp\(^2\) character, and therefore rotation about the linkage bond is hindered. The configuration about this bond is either \textit{trans} or \textit{cis}; \textit{trans} is preferred over \textit{cis} by \(~1000:1\), except in the case of Xxx-Pro linkages, where the \textit{trans} is only favored over the \textit{cis} by 4:1 (because of steric interactions between side chain and backbone atoms, as will be discussed). The \textit{cis-trans} configuration is represented by the backbone dihedral angle \(\Omega\) (Figure 1.1); other backbone dihedral angles are labeled \(\phi\) and \(\psi\) (Figure 1.1). The peptide bond has polar character, which becomes important in the folding of a protein containing helices. The backbone atoms (C=O and NH) are also capable of hydrogen-bonding with other backbone moieties; each residue except Pro can accept and donate at least one hydrogen bond (for a minimum of two per residue).

The side chain identity, hindered rotation around the peptide bond and the hydrogen-bonding character of the backbone atoms are the main parameters dictating the secondary and tertiary structures of native proteins. The structure of the side chain and the rigidity of the peptide bond provide for certain preferred conformations of the amino acid and the backbone. If \(\Omega\) is assumed to be approximately \(\pm 178^\circ\) (or the chain is all in \textit{trans}), excluded volume effects limit the number of \((\phi, \psi)\) combinations that the chain can adopt. These allowed values are mapped out by a Ramachandran plot (Figure 1.2); glycine, with two \(\alpha\)-hydrogens and no side chain, has the largest number of possible backbone conformations, whereas proline has the smallest number, as it is a cyclic imino acid. Typically, the bulkier the amino acid, the fewer allowed regions are indicated on the Ramachandran plot. Using only steric constraints due to the amino acids and the
Figure 1.1. The thick green bond is the peptide bond; the configuration of the backbone atoms involved in this bond describes the $\Omega$ angle. The other two backbone angles are depicted. (Courtesy of the Computer Graphics Lab, University of California at San Diego.)

Figure 1.2. Ramachandran plot of alanine (courtesy of the Department of Crystallography, Birkbeck College, University of London). The white areas correspond to conformations where atoms in the polypeptide chain come closer than the sum of their van der Waals radii. These regions are sterically disallowed for all amino acids except glycine. The red regions correspond to conformations where there are no steric clashes. The yellow areas show the allowed regions if the atoms are permitted to come closer than their van der Waals radii.
rigidity of the peptide bond, the chain conformation is limited to only a few secondary structures (such as $\alpha$-helices and $\beta$-sheets). Near-optimal hydrogen bonds form between backbone moieties in these secondary structures, and other noncovalent forces such as van der Waals forces interactions are also present (as will be discussed in Section 3). Steric hindrance of the side chain atoms also affects the side chain orientations, and will be discussed in Chapter 4. These steric constraints result in unique secondary structure propensities inherent to each amino acid. If the primary structure contains a string of amino acids with high propensity for one type of secondary structure as determined by statistical analyses of natural proteins, it is likely that this secondary structure element will be present in that area of the protein. However, one or two amino acids with low propensity for that secondary structure can sometimes serve to break it; for example, the relatively rigid ring of Pro limits the preferred Ramachandran ($\phi$) angle to approximately $-60^\circ$, and therefore is typically found in turns. The occurrence of Pro in a region meant to be helical can sometimes disrupt the $2^\circ$ structure. It can thus be said that primary structure guides secondary structure through steric constraints and hydrogen bonding possibilities.\(^1\)

The tertiary structure of the protein is the overall fold of the secondary structure elements. It can be a variety of topologies, as evidenced by the folds adopted by native and designed proteins. It is directed in part by the secondary structure and therefore is largely dependent on the characteristics of the sequence as described above; Anfinsen’s hypothesis, formulated in the mid-1950s, poses that the information determining the $3^\circ$
structure of the protein resides in the chemistry of the $1^\circ$ structure. It is this primary-tertiary dependence that is the focus of this study.

It is observed that many small proteins exist in either of two states depending upon the solution conditions: the native (folded) state, which is typically a narrow ensemble of highly related conformations, or in the unfolded (denatured) state. The unfolded state is considered a broad ensemble of structures that are about 1.3-2-fold greater in volume than the native state, and is therefore still quite compact. For these proteins, the acquisition of secondary and tertiary structure occurs cooperatively and the transition from unfolded to folded conformations is “two-state.” If collapse of the unfolded ensemble into the native structure involves detectable population of any intermediate states, the folding pathway is multi-stated. This situation is often encountered in large, multidomain proteins.

Two important properties of proteins are their thermodynamic stability and the specificity of their fold. Stability refers to what extent the molecules populate the native state as opposed to the unfolded state, or the standard free energy difference between the native and unfolded states. Well-defined structure can be lost by raising or lowering the temperature, increasing the concentration of certain molecules such as guanidine or urea, and so forth; this is known as denaturation or unfolding of the protein chain. Generally, the higher the temperature or concentration of molecule needed to unfold at least half of the molecules, the more stable the protein. Specificity describes to what extent the protein molecules are populating one fold under native conditions; if every folded protein molecule has the same three-dimensional structure, specificity is high. Natural proteins
tend to have high specificity for a particular fold. However, in the past few years, improvements in molecular biology and biophysical tools have led to the discovery of a new class of proteins that are “natively disordered” – some of the proteins in this class adopt a more structured fold upon binding of a ligand or cofactor.\textsuperscript{5} In this case, the disorder is sometimes necessary for functionality; a rigid apoprotein might not be able to bind its cofactor and thus will not remain functional. Most native proteins are stable in a wide variety of conditions with a vast array of characteristics dictating their stability and functionality. Other proteins can exhibit stability at extreme conditions, such as high heat, high or low pH, or high pressure. Still other proteins have a high affinity for various functional groups or moieties, such as heme molecules or DNA strands, or are stable in hydrophobic matrices, as is the case for membrane-bound proteins and porins. More is known about well-folded proteins because they are more readily studied using standard biophysical techniques. This work will focus on the computational design of proteins intended to be monomeric and well folded, and the experimental characterization of these designed proteins.

Various parameters drive the folding of a water-soluble protein into a cooperatively folded structure or multimer of structures. Most water-soluble proteins fold so as to bury their hydrophobic residues within the interior of the protein. This interior is typically well packed – the extent of packing is sometimes compared to the packing in an organic crystal. This packing restricts solvent molecules from entering the hydrophobic core and maximizes the attractive forces (or van der Waals interactions) between the atoms.\textsuperscript{6,7} The hydrophobic portions are usually sequestered from solvent,
whereas the polar side chains are mostly exposed to water or, if buried, hydrogen-bonded to or participating in an electrostatic interaction with other buried polar side chains. Additionally, the protein backbone forms internal hydrogen bonds to itself, the side chains or water molecules. The polar side chains and backbone atoms can also interact with water molecules, either through electrostatic interactions, hydrogen bonds or, to a lesser extent, van der Waals forces. Because water forms hydrogen bonds with itself, the dominant driving force behind protein folding is postulated to be the hydrophobic interaction, or the burial of hydrophobic portions of the protein. This is supported by the fact that non-polar solvents denature proteins; these solvents are thought to reduce the free energy of the unfolded state by lessening the strength of the hydrophobic effect. However, non-polar solvents also interfere with electrostatic interactions of the protein to solvent, as well as solvent structure.\textsuperscript{1,8} Further studies by Sauer et al have shown that proteins can tolerate a significant amount of surface position mutations from polar to hydrophobic character, indicating that burying all hydrophobic residues is not as imperative as it may seem.\textsuperscript{9} This may possibly be due to the reverse hydrophobic effect, where a hydrophobic residue that is somewhat exposed to solvent in the folded protein becomes hyperexposed in the unfolded state.\textsuperscript{10} In thermodynamic terms, it appears that the hydrophobic effect is either or both entropy- and enthalpy-driven, depending on the temperature, and is due to the large increase in heat capacity ($\Delta C_p$) upon transfer of non-polar groups into water\textsuperscript{11} as well as the affinity of water molecules for each other.\textsuperscript{1} The weakening of the hydrophobic interaction accounts for the observation of denaturation at low temperature, termed “cold denaturation”.\textsuperscript{2}
The contribution of salt bridges to the force driving protein folding is difficult to quantify. The strength of these interactions depends upon pH and ion concentration; the pH determines the total charge of the protein, whereas the ions shield the salt bridge interactions. Certain studies report that ion pairing stabilizes proteins when it is present between charged pairs at surface positions rather than buried in the core. It has been shown that on average approximately 5 ion pairs are formed in a 150-residue protein; of these, only 1 ion pair is buried. Furthermore, ion pairs are not highly conserved within protein families, which indicates that ion pairing is not the dominant force driving protein folding. However, other studies mention that anisotropic interactions, such as buried salt bridges, might be useful in specifying a certain native fold.

The strength of hydrogen-bonding to the overall enthalpy of the folded protein is also difficult to calculate. The dielectric constant is lower in the interior of a well-folded protein, making the electrostatic interactions that exist in the protein core stronger; this implies that hydrogen bonds that exist inside the folded protein should be stronger than those that occur between the unfolded protein and the solvent. Mutational analyses show that the extent to which a hydrogen bond increases the stability of a protein is entirely dependent on the site and nature of the hydrogen bond. Helix-coil transition appears largely driven by hydrogen bonding, and a recent study by Pace et al, on the basis of a body of work on 89 mutants of six different proteins, suggested that hydrogen-bonding can be as important as the hydrophobic effect for the stabilization of globular proteins. Additionally, Schiffer et al have shown that surface hydrogen bonds can significantly increase the stability of a protein domain. Kauzmann and Dill, however, along with
several theoretical studies, argue that although hydrogen bonds are important, they are not the driving force behind protein folding.\textsuperscript{8,17-19} However, hydrogen bonding does favor internal organization within the folded protein, thus increasing the activation energy between the native state and the other competing states; additionally, hydrogen-bonding between backbone polar groups is the fundamental feature of repetitive and non-repetitive secondary structure elements. It is clear from these contrasting results that the true contribution of hydrogen bonding is difficult to assess. Based on these studies, it appears that buried (and, probably to a lesser extent, exposed) electrostatic interactions such as salt bridges and hydrogen bonds, while possibly destabilizing the protein because of desolvation of the polar moieties, may enhance specificity of the native fold by disfavoring other, competing folds that do not exploit these specific interactions.\textsuperscript{20-32}

As mentioned before, the peptide bond has polar character. As a result, proteins, especially $\alpha$-helices, have a macromolecular dipole aligned along the helix axis that leads to a partial positive charge at the N-terminus and a partial negative charge at the C-terminus. Reduction of the helix dipole moment via reduction of the charges at the ends of the helix with side chains of opposite charge from the helix dipoles can stabilize helices considerably,\textsuperscript{33} as the helix dipole may actually be destabilizing in the formation of parallel helix bundles.\textsuperscript{34} As a result of this macrodipole, the nature of the helix-capping residues (i.e., helix boundary residues, or the first and last helical residues) is important in stabilizing a helical protein.\textsuperscript{30,35} Other studies\textsuperscript{36} have determined that helix stability may also increase with chain length because of this helix dipole. This indicates
that context effects may be as important as intrinsic propensities; local interactions may not be enough to specify fully the conformation of the protein.\textsuperscript{37}

Acids and bases tend to destabilize proteins because of the increased charge repulsion at pH values far from the isoionic point. Heat denatures proteins as well, because of the large and uncompensated increase in entropy change of folding at higher temperatures; the gain in conformational entropy of the chain overcomes the non-polar interactions in the folded protein. Chaotropic agents, such as urea and guanidinium salts, unfold proteins as well; possibly by solvating the non-polar groups (which are typically buried in a folded protein) better than water.\textsuperscript{38}

In short, most observations thus far are reconciled if the hydrophobic force is taken to drive the protein to adopt an ensemble of highly similar folded, well-packed conformations; other interactions, such as electrostatics and hydrogen-bonding, may serve to select for the native structure from this ensemble of compact conformations. The main force opposing protein folding at room temperature is the loss of conformational entropy, a quantity that is difficult to evaluate in part because of the solvent contribution. \textit{A priori}, an unfolded protein can exist in a large number of structures. The denatured protein is more flexible and has increased conformational entropy with respect to the native state. At higher temperatures, this entropic contribution can dominate the free energy of the system (given by $\Delta G = \Delta H - T\Delta S$) and lead to increased population of the unfolded state. It must be noted, however, that this statement assumes no change in enthalpy between states; this is largely an incorrect assumption. As the protein folds, the entropy becomes significantly decreased; there are few energetically favored structures
available to a highly specific folded protein chain. The side chains in the folded state lose entropy of conformation as well, if the core is considered a solid-like state.\textsuperscript{39,40} Residues with higher conformational entropies tend to remain on the protein surface, instead of being buried in the core.\textsuperscript{41} As more conformational possibilities become available to the protein (either due to the protein chain or buried amino acids becoming longer, or the $\phi, \psi$ space becoming larger), the loss in entropy upon folding will increase by $-R \times \sum p_i \ln p_i$, where $p_i$ is each available conformation. Cross-links, such as disulfide bonds, serve to stabilize the native structure by decreasing the amount of structures allowed to an unfolded chain.\textsuperscript{42} However, at least some of the favorable interactions that exist within the molecule in its folded conformation must be broken in the process of denaturation. Therefore, $\Delta H$ can be positive for the unfolding process at certain temperatures.\textsuperscript{43} As temperature increases, $\Delta H$ due to unfolding of the protein becomes increasingly positive while the entropic contribution, described as $-T\Delta S$, decreases.\textsuperscript{8} The balance between the differences in entropy and enthalpy between the unfolded and native states of the protein is what determines the stability of the protein fold; the slight gain in enthalpy of the folded state over the entropy loss at physiological temperatures is the main driving force behind protein folding. The temperature at which the thermodynamic values are being calculated is of the essence in determining whether entropy or enthalpy predominates.\textsuperscript{8}

In spite of decades of studies and the documented relationship between primary, secondary and tertiary structures, the prediction of protein structure remains an
intractable task. Merely knowing the intrinsic secondary structure propensity of a protein sequence is not enough to predict the actual fold.\textsuperscript{44} Statistical analyses show that secondary structure propensity has a high degree of positional dependence,\textsuperscript{45} making it difficult to apply this weight as a general criterion for selecting a sequence to fold to a specific structure. Heteropolymer theory, although providing a useful framework for the description of chain behavior, is not thorough enough to describe why proteins adopt a fold with high specificity and stability.\textsuperscript{46} It has also proven remarkably difficult to choose a sequence of amino acids that will adopt a tertiary structure with high specificity and stability.\textsuperscript{47} The forces driving a protein to adopt a fold, such as steric constraints and enthalpic contributions, are not always obvious in their application; a firm understanding of the importance of each parameter has not yet been achieved. In order to design proteins with specific characteristics, it is imperative to first determine the appropriate weights of each of these driving forces. It is here that computational protein design comes into play.

1.2 Definitions of Protein Design Problems

Protein design refers to the development or application of computer algorithms to the construction, either in full or in part, of proteins or protein libraries. The two main branches of protein design are termed the protein folding approach and the inverse folding approach. The protein folding approach is the determination by computational means of the secondary and tertiary structure of a protein using its sequence as the main
input. The inverse folding approach, as the name implies, involves using the fixed, native protein backbone, as described experimentally by high-resolution NMR spectroscopy or X-ray diffraction data, as input into the algorithm. The reverse-folding algorithm then decorates the backbone with side chains compatible with the fold and produces a sequence or family of sequences that should theoretically adopt the target structure. This method relies on parameters designed to mimic the folding parameters of wild-type proteins.

The usefulness of an algorithm that can predict a fold from a sequence is apparent; any high-resolution determination of protein structure could be facilitated through the use of a computationally predicted general fold as a template or guide. Most of the work involving the protein folding approach has met with limited achievement; the yearly Critical Assessment of Structure Prediction meeting (CASP) has revealed progress in the protein folding approach, but also indicates that this field is still in its infancy. The inverse folding approach, on the other hand, has been more successful. Applications of the inverse folding approach have historically focused on folds that are entirely or primary $\alpha$-helical, as these folds exhibit more short-range interactions. What is close in the primary structure of an $\alpha$-helical protein is close in the secondary structure of the fold; ($i$, $i+3$) and ($i$, $i+4$) interactions predominate. It is therefore easier to determine which amino acid pairs in the primary structure should be selected to have the most favorable electrostatic, hydrogen-bonding or hydrophobic interactions. Additionally, the Ramachandran plots of most of the amino acids indicate that helical structures are favored (to varying extents) by the steric constraints placed on
the allowed (\(\phi, \psi\)) backbone angles. \(\beta\)-sheet proteins, on the other hand, exhibit longer-range interactions that bring parts of the protein significantly separated in the sequence within close range of each other in the tertiary structure. Loss of even one interaction or shift in register could lead to a different native structure from that of the target fold. This is more difficult to design,\(^{22}\) which may explain why most protein design successes have focused on entirely or mostly \(\alpha\)-helical structures.

A point of importance is the balance between computational design and human verification. Most protein design is not fully automated; there is still a great deal of visual examination of the designed sequence or structure to ensure that the resultant protein is native-like: it is well packed and has well-dispersed electrostatic interactions and hydrogen bonds. When visual examination is given as a method for choosing one particular designed sequence or structure over another, it indicates this type of scrutiny.

1.3 Physical Parameters in Protein Design

The physical parameters in protein design include, but may not be limited to, the explicit application of statistically derived rotamers, steric filtering, the free energy of partitioning the amino acid or parts of the amino acid in water as opposed to octanol (related to the solvation parameter), the sequence entropy of the primary structure, and the enthalpy of the fold (including the hydrophobic effect, the electrostatics and the van der Waals interactions). These terms are all derived from the physical parameters
dictating protein folding as described in Section 1. The equations to calculate the energy for each term as used in this work are given in Chapter 4.

All of these explicit parameters involve the implicit ones, such as steric constraints on the protein backbone and the rigid peptide bond, as described in Section 1. As mentioned before, the wild-type protein backbone adopts a specific conformation based upon its primary structure. The amino acids, which take up a certain amount of space, are prohibitive in that they drive the backbone to adopt a structure to relieve steric hindrance; hydrogen bonds form between the backbone moieties because of the geometry of the structure. Implicit in all protein design attempts is the need to develop a sequence that has the same prohibitions on the backbone structure. If a sequence is developed for a sheet structure that can readily adopt a helical conformation, for example, the specificity of the fold can be low or, at worst, an alternative fold can be populated. These steric constraints yield a secondary structure propensity that must be mimicked in the designed protein’s sequence, and make up the repulsive part of what is known as the Lennard-Jones potential (described below).

The sequence entropy is related to the conformational entropy in mathematical terms, yet is a qualitatively different parameter. The sequence entropy is a measure of the diversity of the composition of amino acids in the primary structure. It was postulated that as sequence entropy increases, the structural specificity would also increase. As the sequence becomes more and more diverse, the low-energy structure ensemble would condense into one lowest-energy structure with high specificity. For example, a protein with multiple electrostatic interactions stemming from one or two
types of residues (such as Glu or Lys) can adopt a multitude of folds that explore these electrostatic interactions, resulting in a molten fold and low specificity. Moreover, a protein core with a high amount of identical residues can pack in a variety of different ways, again leading to low specificity. Although this is a logical conclusion, preliminary results discussed in Chapter 7 indicate that sequence entropy, at least of the core residues, may not be one of the main selective factors in determining the specificity of a protein fold.

As briefly mentioned in Section 1, the hydrophobic effect is sometimes described as the true force driving protein folding. Water molecules, when forced to hydrogen bond to each other around a hydrophobic entity, typically form a clathrate or cage-like structure\(^{52}\) that has low entropy. Although there are favorable van der Waals interactions between the water molecules and the non-polar moieties, these are greatly overwhelmed by the preference of water molecules to hydrogen bond with each other. At room temperature, the main driving force behind the hydrophobic effect is the unfavorable change in entropy when a non-polar group is exposed to water. As the enthalpic change is approximately zero at room temperature, this loss of entropy is sufficient to drive the burial of the hydrophobic groups and fold the protein. At higher temperatures, the water molecules in the clathrate structure become less ordered and exhibit weaker intermolecular attractions; at a certain temperature (extrapolated to be between 130 and 160 °C), the change in entropy between a buried and a solvent-exposed hydrophobic group is zero. However, at this same temperature, the enthalpy of transfer is more
unfavorable than at lower temperatures, and serves to define the hydrophobic force at this temperature.

The Lennard-Jones (12-6) equation is a description of interatomic pair potentials. It has a repulsive term at short interatomic distances, accounting for the steric impossibility of overlapping the atoms’ electron clouds. For this term, the atoms are approximated by hard spheres with a definable radius (called the van der Waals radius). This steric constraint, as described in Section 1, eliminates a large number of conformations. The Lennard-Jones potential also has an attractive term (the van der Waals interaction) that exists between two polar atoms (the dipole-dipole force), two non-polar atoms (the instantaneous dipole-induced dipole interaction or London dispersion interaction) or a combination of polar and non-polar atoms. This portion of the term stems from the fact that all molecules attract each other as a result of the induced or permanent polarization of the electron cloud. The optimal distance between two atoms is usually 0.3-0.5 Å greater than the sum of their van der Waals radii. Orientation is usually not considered, which is a simplification, as the polarizability of a bond is larger along the bond than perpendicular to it.¹

Salt bridges are strong electrostatic interactions that exist between a basic, positively charged residue and an acidic, negatively charged residue as per Coulomb’s law:

$$\Delta E = \frac{Z_a Z_b e^2}{Dr_{AB}}$$
where $D$ is the dielectric constant of the medium, $Z$ is the charge on each atom, $\varepsilon$ is the charge of an electron and $r_{AB}$ is the distance between $Z_A$ and $Z_B$. These interactions are a direct result of the attraction between opposite charges and the repulsion between two like charges. Electrostatic interactions are weakened by the high dielectric constant in aqueous solution (~80 $\kappa\varepsilon_0$ C$^2$ J$^{-1}$ m$^{-1}$), although they are very strong in vacuo, where $D$ is 1 $\kappa\varepsilon_0$ C$^2$ J$^{-1}$ m$^{-1}$; frequently they are utilized as a means to increase the specificity of a fold instead of one of the main driving forces behind folding. As will be shown in Chapter 4, this calculation is modified slightly for the design of proteins.

Hydrogen bonds are directional interactions; although neither donor nor acceptor need be fully charged, these interactions appear to be mostly electrostatic in nature. Hydrogen bonds between backbone atoms provide a framework for structure, and hydrogen bonds can be formed between side chain atoms because of the fold; these inter-amino acid interactions also act to specify a certain fold. The energetic contribution of hydrogen bonds to the overall enthalpy of the native state are difficult to quantify, owing to the variety of donor-acceptor pair possibilities and the uncertainty in the measurement of their strength.

Binding of cofactors or other ligands, and formation of disulfide linkages, are important determinants in maintaining a specific fold. The loss of entropy upon binding any cofactors (both entropy of hydration and structural entropy) and loss of structural entropy upon formation of a disulfide linkage must be overcome by the resultant enthalpy
of all of these processes. However, use of these factors, especially disulfide linkages, has a strong benefit in the selection of a specific fold.

It is obvious that a wide variety of physical parameters are necessary to describe the native state of a protein. The importance of each of these parameters, however, is difficult to quantify. It is through the design and characterization cycle that the weights of each parameter can be fine-tuned.

### 1.4 History and Importance of Protein Design

The history of protein design began in 1951 with Linus Pauling and Robert Corey’s proposed structures of $\alpha$-helical and $\beta$-sheet proteins. Francis Crick postulated the “knobs into holes” packing model of a two-stranded coiled-coil with an interhelical angle of 20° in 1953; this was supplemented by Cyrus Chothia’s description of “ridges into grooves” packing in 1977. Both of these observations describe interhelical packing as large residues on the exterior of one helix packing tightly into holes on the exterior of the next helix. In 1998, Lynne Regan and William deGrado reported a *de novo* design of a four-helix bundle that was shown to be stable and highly helical. The focus of protein design has since progressed from hydrophobic cores to loops, helices and simple protein motifs and the ability to design functional metalloenzymes and protein libraries, as well as *de novo*-designed structures and full sequence design. Several important milestones exist; these include the atomic solvation parameters developed by David Eisenberg and Andrew McLachlan, and Jean-
Luc Fauchère and Vladimir Pliska\textsuperscript{70}, the rotamer description of side chains based upon statistical analysis of native proteins\textsuperscript{71,72}, the importance of hydrophobic and polar pattern descriptions (HP or binary patterning) in the folding of native and designed proteins\textsuperscript{73-78} and the identification of fundamental issues of protein design, as described in later sections. In 1989, Eugene Shaknovich and Alexey Finkelstein postulated that the rate-dependent step in the folding of a protein to its native state is finding the correct side chain rotamers for tight packing of the hydrophobic core.\textsuperscript{39} Other groups identified the importance of proper packing of the hydrophobic core for a protein to adopt a specific fold with high stability.\textsuperscript{21,65,71,79} Bassil Dahiyat and Steven Mayo were the first to utilize successfully the importance of packing of the core in protein folding to design FSD-1, a designed protein that adopts the zinc finger fold.\textsuperscript{80} Additionally, a $\beta$-sheet protein was changed to an $\alpha$-helical one by mutating only 50\% of the amino acids,\textsuperscript{81} and a peptide comprised of $\beta$-amino acids was designed.\textsuperscript{82} Most interesting in light of the results presented here was the successful de novo design of a small, soluble, monomeric model $\beta$-sheet; this design utilized only natural amino acids, but involved a significant amount of human input for the design process.\textsuperscript{83} Recent reviews of protein design provide a general idea as to how far this field has come.\textsuperscript{84,85}

The applications of successful protein design algorithms are diverse. Protein design can be used to create or modify proteins that exhibit a certain multimeric state, select for sequences that adopt a certain conformation, or have a specialized function. Another important application involves the design of enzymes with a specific catalytic or binding activity. These proteins could be used in medicine, in bioremediation, as tools to
enhance nanotechnology and so forth. The more subtle application would be design of a protein that has increased stability at high temperatures or pH, or under other adverse conditions. Other driving forces behind protein design involve the quest to describe the physicochemical properties of a native, folded protein, the ability to model the denatured or unfolded state, and the description of the folding landscape. This work will focus on the application of computational algorithms to the theoretical design of thermostable proteins that obtain and maintain a single, unique fold, and the experimental characterization of these proteins.

1.5 Problems of Protein Design

The strength of protein design lies in the fact that a computer, if successfully programmed, can calculate energies of protein models and any mutations to those models much faster than a person can. However, any computational approach to an experimental situation will have inherent problems. The most obvious, and the most difficult to circumvent, is the adequate description in silico of the many different in vivo conditions. Fully realizing every physical parameter is even more complex, and determining the importance of each parameter is usually accomplished by trial and error.

Another significant problem in any computational application is that of time. If one were to consider an octapeptide, and wanted to determine the energy of every possible sequence utilizing all twenty naturally occurring amino acids on this octapeptide, the resultant number of possibilities would be $20^8$. With the added consideration of the
infinite possibilities of orientations obtainable by most of the amino acids, this situation becomes computationally intractable. The majority of protein design utilizes backbones comprising greater than twenty amino acids.\textsuperscript{86} This problem becomes so large that full sequence design, utilizing amino acids in every possible orientation, is virtually impossible. This situation can be addressed in a number of ways. The most popular methods of reducing the number of computational models involve HP patterning (either weighted by the wild-type sequence or determined with no assumptions), rotamer description of side chain orientation and steric filtering. These are all described in Section 6.

Any algorithm that identifies a “minimum energy” model is making the assumption that the “lowest energy” state is truly the global minimum. However, the algorithm frequently becomes trapped in a local minimum, resulting in a solution that does not correspond to the global energy minimum. The main method utilized to circumvent this local minimum problem is the "jumping" in energy of the current solution from the local minimum trap to a solution that is higher up in the energy well or into the isoenergetic continuum of states. It is hoped that the algorithm can then re-anneal onto the lowest energy model from this higher-energy solution. This jumping of energy is most relevant in the applications of the Monte Carlo and Genetic Algorithms, described in detail in the next section; in the former, selecting a model of higher energy than the current model is done with low probability, and in the latter, mating of models and a small probability of mutation is allowed. Both of these situations have the detriment of allowing for potentially harmful jumps away from the true minimum, but
this is compensated for by the fact that local minima might no longer serve as permanent traps.

On the computer, any sequence that exhibits the most native-like characteristics (electrostatics, HP patterning, packing, and so forth) appears to be a solution to the inverse folding problem. It is only after experimental verification, however, that the true usefulness of the algorithm can be determined. The cycle of theoretical design and experimental verification can yield appropriate weights of various physical parameters and indications of incomplete descriptions of the physical situation, as well as positive feedback. The main problem with theoretical design that can be elucidated through experimental characterization is structural specificity, on both the macroscopic and the microscopic levels. The macroscopic structural specificity issue relates to whether or not the protein adopts the intended overall fold. For example, if a coiled-coil motif is expected after sequence design and a three-helix bundle is characterized, it is clear that the theoretical design has failed in some respect. This is also evident when designing a protein that does not fold at all – it populates only the unfolded state, and the algorithm has attempted to design one that populates the folded state. This can sometimes be rectified through the correct use of various parameters, including proper weighting of secondary structure propensity of the amino acids chosen as well as solvation (burial versus exposure of hydrophobic and hydrophilic residues). Negative design is also a very useful tool in dealing with macroscopic selectivity. The use of certain residues or other methods (binding of cofactors, disulfide bridges, and so forth) can destabilize another,
competing fold to such an extent that the protein will populate the target fold with high specificity.$^{29,87}$

The microscopic specificity usually concentrates on the cooperativity or selectivity of the fold. When a designed protein has poor cooperativity of folding to the target structure and a “molten” native state, the algorithm has failed. (The term molten refers to a protein that is adopting slightly different folds or structures from that of the target fold, resulting in small fluctuations in local structure.) This can be mathematically described through the use of an energy gap, or difference in energy between the target or goal state (assuming, of course, that this target state is the state of lowest energy) and the next lowest energy state, and z-score of the target fold, as follows:

$$z - score = \frac{goal - mean}{stddev}$$

where the $mean$ and $stddev$ refer to the mean and standard deviation, respectively, of energy of the continuum of states. This microscopic selectivity issue is addressed through consideration of the alternative structures (sometimes termed decoys). If the alternative structures have a mean energy considerably higher than that of the target structure, the chance of poor microscopic selectivity is reduced. Problems arise, however, when the target and decoy structures are isoenergetic (have no energy barriers between the conformations) or, in the worst-case scenario, an alternative structure is of lowest energy, with a significant gap between itself and the continuum of energy states (which in this case includes the target fold) (Figure 1.3). Again, negative design on the
Figure 1.3. Possible decoy and target energies. The first case is the desired outcome, where all alternative folds are significantly higher in energy than the target fold. The second case results when all folds (target and alternatives) are isoenergetic. Case three involves the target structure being of higher energy than an alternative fold.
decoys is a useful method to both destabilize alternative folds and allow the protein to adopt the target fold.\textsuperscript{28}

1.6 Methods of Protein Design

The methods of protein design are diverse. They include, but are not limited to, the Monte Carlo method (with or without the Metropolis acceptance criterion, and referred to as MC), the Genetic Algorithm (or GA) and the Dead-End Elimination method (DEE).\textsuperscript{84,88} Another type of protein design involves the lattice model; however, this is usually treated as less of a design approach than a step along the way to full sequence design. The first two are algorithms that sample solutions in a near-random fashion. DEE, on the other hand, is a pruning method and is considered an exhaustive search of all possibilities.

The Monte Carlo method is a procedure that involves the use of adjustments of certain physical parameters to yield a new structure. These physical parameters can be as diverse as the $\phi$, $\psi$ backbone angles, the angles between helices in a parallel or anti-parallel helix bundle, the placement of certain residues on a structure, and so forth. The main acceptance criterion, called the Metropolis criterion,\textsuperscript{89} results in the acceptance of any move that results in a lower energy structure, and the acceptance of any move that results in a higher energy structure with the following acceptance rate:
where \( \text{ran}(seed) \) is some seeded random number. The temperature would be either set at a predetermined value or decreased with a certain frequency. The latter method would allow for high-energy structures to be selected at high temperature, with slow annealing to low-energy structures at low temperature. This method is useful mainly for small alterations in the physical parameters, but is not widely applied for full sequence design.

The GA approach\(^{90,91}\) utilizes the idea that implementation of an algorithm that can both recombine and mutate models ("evolution on the computer") would overcome the local minimum problem. The GA written for this study uses a rotamer wheel with a pre-determined number of “slots” for selection of the amino acid/rotamer combinations during the design process. These slots are filled with allowed amino acids in experimentally derived, statistically important orientations, known as rotamers or \( \chi \) angles.\(^{71}\) The filling of this rotamer wheel is done based upon a Boltzmann probability calculation:

\[
P \propto e^{-\frac{E}{RT}}
\]

where \( E \) is the energy of the amino acid/rotamer combination in Joules or kilocalories (calculated as described in Chapter 4), \( R \) is the universal gas constant in either Joules \( \text{mol}^{-1} \text{K}^{-1} \) or kilocalories \( \text{mol}^{-1} \text{K}^{-1} \), and \( T \) is the temperature in Kelvin. Typically, the
temperature is set at a very high value (around 2000 K) to allow for rotamers of high energy to be present in the roulette wheel. The models (typically 200 in this work) are built by randomly selecting an amino acid/rotamer combination from the roulette wheel for each position in the protein. The total energy of each model is determined using a set of energy calculations, and the parent roulette wheel (consisting of 200 rows) is filled using the same Boltzmann probability listed above at very high temperature, but with full sequences in each row instead of rotamer/roulette combinations. Two parents from this wheel are selected, mixed (by single-point crossover or uniform crossover), and mutated at a very low frequency with other amino acid/rotamer combinations from the first roulette wheel. Single-point crossover results in a child that is comprised of large portions of each parent, whereas uniform crossover yields a child that is a random mixture of both parents (Figure 1.4B). This cycle of selection, energy calculation, crossover and mutation is repeated a large number of times (usually 300 in this study) with a temperature factor decreasing steadily with each repeat. The decreasing temperature allows for decreased probability of selecting a model that is of high energy. A flowchart of the GA used in this study is shown in Figure 1.4.

The Dead-End Elimination method is described as an exhaustive search method utilized to find the lowest-energy sequence for a particular structure. It has drawbacks in that protein libraries cannot be designed, nor can point mutations of key residues be accurately determined. DEE designates side chains and rotamers as being either compatible or incompatible with the global energy minimum. In effect, if the lowest energy structure that contains a specific side chain rotamer is always of higher energy
Allowed amino acids in specific orientations or rotamers are individually placed on the protein backbone; Cβ centroids replace the other amino acids. Any steric clashes with another atom result in the elimination of that rotamer.

The Lennard-Jones energy of this rotamer and the Boltzmann weighted probability at high temperature \( p = e^{-\frac{\Delta E}{RT}} \) is calculated.

Rotamer wheel (portion shown above) is filled with amino acid/rotamer combinations based on Boltzmann weighted probabilities of amino acid/rotamer combinations.

Then, cycle through GA described below, decreasing temperature at each round:

- Build parent models by drawing from roulette wheel randomly at each position on the protein chain.
- Calculate energy of each parent as described in Chapter 4.
- Build parent roulette wheel, using Boltzmann weighted probabilities of each.
- Select two parents (blue and green, below) randomly from parent roulette wheel.
- Perform random crossover and mutation (in green):

Figure 1.4. Flowchart of a typical GA as used in this work.
than the highest energy structure that contains a different side chain rotamer, that first side chain rotamer can be eliminated from the design procedure.\textsuperscript{92-94} An addendum to the DEE is Goldstein’s variation, which states that if the energy of a possible structure containing a specific rotamer is always lowered by changing that rotamer to another rotamer, the first rotamer can be eliminated.\textsuperscript{95}

The lattice model utilizes a simplistic view of the protein backbone, and is usually applied to obtaining crude solutions to the protein design problem. Typically, lattice models are comprised of beads or, most recently, beads with an attached “functional group” on a chain; the beads can be either hydrophobic or polar, but have no further identity.\textsuperscript{96} The chains can adopt either 180\textdegree or 90\textdegree angles between beads, yielding three-dimensional structures that are block-like (Figure 1.5). The important forces dictating the folding of the lattice model are hydrophobic interactions, conformational freedom of the chain, and steric constraints. The complexity of the lattice increases as the length of the polymer chain utilized increases. Although this view is quite simplistic, it can yield results that assist in the solution of more complex, realistic problems. Two of the most useful derivations of the lattice model approach are the reinforcement of the importance of sequestering hydrophobic groups from solvent and exposing polar residues, which has become known as HP patterning,\textsuperscript{74,97} and what is termed the Micheletti contact score.\textsuperscript{98} This score is a determination of how exposed a certain residue is to the solvent based upon its nearest neighbors, and can be applied to the determination of the HP patterning of a backbone or the solvent exposure of a particular residue. The Micheletti contact score will be further discussed in Chapter 4.
Figure 1.5. HP lattice model conformations resembling a β-helix (top) and a four-helix bundle (bottom). In both models, the darker spheres represent hydrophobic monomers, whereas the gray spheres represent polar monomers. (Reproduced with permission.)
Most protein design efforts today concentrate on the use of the GA or DEE to arrive at a structure that is of theoretically low energy, has a high degree of fold specificity, and exhibits thermal and chemical stability as well as cooperativity of folding. Not all – in fact, few – designed structures actually fulfill more than one of these criteria. It is because of this drawback that a designed protein that adopts and maintains the proper fold is considered a success.
2.1 Protein Purification Tools

The defining characteristics of the protein of interest, such as its net charge, affinity for certain chemical groups or stability at high temperature, can sometimes be exploited to purify it away from the cellular matrix. For example, if a protein is from a hyperthermophilic organism and is stable at high temperatures, the cellular matrix can be heated to a high temperature and centrifuged. As most of the background cellular proteins in the mesophilic host utilized in this study are not stable at high temperatures, they will denature and become insoluble, thus forming a precipitate. The protein of interest, however, will remain folded and soluble (providing, of course, that it is soluble to begin with). When purifying a novel protein, whose characteristics are not yet fully known, it is wise to gradually increase the temperature at which the heat purification is performed until the purity and quantity of the protein of interest is maximized in relation to the cellular matrix.

Another method of purification exploits the affinity of the protein for a specific chemical group; a wide variety of resins with different functional groups are commercially available for this purpose. Under certain conditions, a hydrophobic
interaction column can be utilized. For example, if addition of a chemical causes the protein to alter its conformation and expose hydrophobic patches, yet does not alter the structure of much or all of the background cellular proteins, a phenyl sepharose column can be utilized for purification purposes. Phenyl sepharose consists of highly cross-linked 34 µm agarose beads with functional phenyl groups attached via uncharged ether linkages. Phenyl sepharose (as described by Pharmacia Biotech) has no ionic properties, is stable in a wide variety of pH values, and tolerates detergents and chaotropic (or denaturing) agents. It can therefore be utilized to bind noncovalently to the hydrophobic pockets in the protein described above. Ideally, the background proteins do not bind to the resin, and wash through. Elution from the column of the target protein is accomplished by running a gradient (if a significant portion of the background has adhered to the column) or a high concentration of some chemical that would allow for a conformational change in the protein, resulting in burial of the hydrophobic pockets and subsequent loss in binding affinity to the resin.

If the protein is acidic, an anionic affinity column using a strong anion exchanging resin (which typically contains positively charged functional groups, e.g. –NH₄⁺) such as QAE Sephadex A-25 can be run to purify the protein of interest. Here, the pH of the sample is adjusted such that the protein is charged negatively. The negatively charged groups on the protein interact with the positively charged functional groups on the resin, which results in binding to the column material. These interactions are overcome by introducing a salt solution of high concentration (either with or without a gradient).
If the protein is basic, a cationic column such as SP Sephadex C-25 can be used for purification purposes. The functional groups in this case are negatively charged (e.g. \( \text{SO}_3^2^- \)), and interact with the positively charged groups on the protein. Again, a high-salt solution serves to elute the protein from the resin. The cationic exchange column, as well as the anionic exchange column, must be optimized for pH, salt concentration, and, to a lesser extent, temperature – too high or low a value of any of these parameters can result in poor specificity of binding, low concentration of protein bound, high amount of impurity of the sample, and so forth.

In a different approach, various affinity tags can be added to the N- or C-terminus of the target protein to purify it from the background cellular protein matrix. One of the most popular affinity tags is the (His)_6 tag. The method involves the addition of a six-histidine stretch at the beginning or end of the protein. (The location of the (His)_6 tag is important for expression purposes. Frequently the C-terminal tags lead to the highest expression of protein.) This tag will bind with high affinity in both native and denaturing conditions to a variety of immobilized metal affinity columns (IMACs). These IMACs involve metal ions – either cobalt or nickel – coordinated to the resin matrix. The two most common IMACs are the Co TALON or Ni nitrilotriacetic acid (NTA). The Co TALON column is often preferred over the Ni NTA column for two reasons: the Ni NTA resin is slightly more toxic than the Co TALON resin, and has a higher affinity for histidines, which results in a diminished selectivity for the protein of interest. Co TALON resin, with its slightly lower affinity for His residues, has a higher selectivity for His-tagged proteins than Ni NTA. The Co TALON column has a cobalt ion (Co^{2+})
chelated to the matrix in a tetradeinate geometry, and binds with high affinity only to those proteins containing more than four His residues; any proteins bound with weak affinity will be eluted when the column matrix is washed with low concentrations (5 mM) of imidazole. After flushing the system of background proteins, the protein of interest can be eluted by subjecting the column to higher concentrations (50 mM) of imidazole. The main drawback with metal affinity columns is the fact that no chelators (EDTA, etc.) can be present in any buffer solution, as they will chelate the metal ions and render the resin useless. However, the Co TALON column is compatible with chaotrope concentrations up to 6 M, so that denaturing IMAC columns can be utilized.\textsuperscript{99} Usually, the His tag will not alter the structure of the protein; if desired, a cysteine can be added between the tag and the protein of interest for 2-nitro-5-thio-cyanobenzoic acid (NTCB) cleavage. Alternatively, other protease or cleavage methods available commercially can remove the His tag. If the positive charge of the His residues disrupts the structure of the protein, a novel affinity tag that does not utilize six His residues is offered by Clontech for Co TALON column purification.

Size-exclusion columns utilize small beads that have pores of varying diameters on their surfaces. Size-exclusion resins have specific upper and lower molecular weight cutoffs. Proteins that are too large to fit within the pores will travel around the beads and elute prior to the proteins that can gain access to the interior of the beads. These latter proteins will traverse a longer path than the larger proteins because they will travel through the beads of the resin. Size-exclusion resins are also used to remove salt (by using a desalting resin, which has very small pores) and other small molecules. In these
cases, a dye that has a molecular weight similar to the small molecule to be removed can be added if the small molecule has no other detectable characteristics. This will provide a visual indication of the location of the small molecule to ensure complete separation.

2.2 Protein Fusion Systems

There is a wide array of commercially available fusion systems for protein expression and purification. Additionally, fusion systems utilizing the expression and purification characteristics of other proteins can be invented that are useful when dealing with a system that has poor expression levels, is difficult to purify or both. This work uses two fusion systems. The first, a commercially available method called the IMPACT-CN System, involves the fusion of the target protein, either N- or C-terminal, to a self-cleavable chitin affinity intein tag. To use this method, the gene coding for the protein is cloned into an *E. coli* fusion system vector (pTYB2 or 12, depending on whether the protein is C- or N-terminally fused to the target protein, respectively) derived from pBR322. These vectors contain a multiple cloning site, are under control of the T7 promoter, are IPTG-inducible and ampicillin resistant, and self-cleave in the presence of dithiothreitol (DTT). The fusion protein can be purified using chitin resin as described in Appendix A. Briefly, the fusion protein binds chitin with high affinity; once the system is bound, cleavage buffer is introduced and the column is allowed to stand overnight. The protein of interest is washed off, leaving the fusion tag bound to the column.
The other fusion system, developed in-house, involves the C-terminal fusion of the target protein to the N-terminus of calmodulin (N-Cam). N-Cam is a calcium-binding protein that exposes hydrophobic pockets upon the binding of Ca\(^{2+}\) ions.\(^{100-102}\) This characteristic can be utilized for purification purposes, as these hydrophobic pockets bind to phenyl sepharose resin with high affinity when Ca\(^{2+}\) is present.\(^{103}\)

Washing the protein-bound column with EDTA chelates the Ca\(^{2+}\) ions from the protein, which undergoes a conformational change back to its free state. Additionally, N-Cam has a high expression level, enabling large amounts of the target protein to be obtained. This fusion system has proven to be extremely useful for the expression and purification of novel proteins, and will be described in more detail in Chapter 6.

2.3 Sodium Dodecyl-Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE)

Polyacrylamide gel electrophoresis (PAGE) is a useful method to separate proteins from each other based indirectly on their size and, to a lesser extent, charge. PAGE utilizes acrylamide and bis-acrylamide to form a sieve of varying pore size, based upon the concentration of the acrylamide in solution (the higher this concentration, the smaller the pore size will be in the gel matrix). Typically the PAGE method utilizes two types of gel matrices: a stacking matrix of low acrylamide concentration, to ensure that the samples in each well are compacted into a thin band so that no resolution based upon location within the sample is seen, and a resolving matrix of higher acrylamide concentration which resolves the proteins based on their sizes. A sample buffer
containing sodium dodecyl sulfate (SDS) plus dye (such as bromophenol blue) is added to each sample to denature the proteins. SDS is a negatively charged detergent; its hydrophobic regions bind to the hydrophobic portions of proteins and cause them to unfold. Additionally, the samples are usually heated to 95 °C for 5 minutes to ensure full denaturation. Each molecule binds large numbers of SDS molecules, which usually overwhelm their intrinsic charges and cause them to migrate to the positively charged electrode when voltage is applied. Proteins of the same size, which bind the same approximate number of SDS molecules, will tend to run within a small range of each other because of their same shape and charge. However, if a protein is highly charged, its actual location in the gel matrix can be perturbed from its theoretical one. Larger proteins, although having an increased charge, will also have an increased size, which will overpower their charge increase and cause them to run slower than smaller proteins. The proteins are loaded into individual lanes of this gel matrix along with a protein ladder of standard molecular weight proteins, which provides a rough estimate as to the molecular weight of the resultant bands. The samples are run through the SDS-PAGE gel using two different buffers – an anionic and a cationic buffer – and a gel running apparatus. This apparatus uses current (25-50 mA) to drive protein elution through the gel.\textsupercite{104,105}

Once the dye front has reached the end of the gel, the apparatus is stopped and the gel removed. It can be stained using a variety of methods – Coomassie Blue is the most popular when large amounts of protein are present; silver staining is typically used when
small amounts (greater than 10 ng) of protein are present. The gel can be photographed and stored or dried.

2.4 Mass Spectrometry

A quantitative tool in the determination of protein size is mass spectrometry, which can accurately determine the molecular weights of proteins up to 500 kDa to better than 0.01%. Mass spectroscopy involves the generation and acceleration (via an electrical field) of gas-phase ions of the sample in an ion source by one of several ionization methods as described below. These ions are ejected into a drift region that has no applied electric field, called a mass analyzer. Each ion of a given charge will have the same kinetic energy independent of size; an ion with large charge will have a higher kinetic energy than an ion with lower charge. The length of time for traversal of this drift region can be correlated to the mass-to-charge ratio \((m/z)\) of the ion (referred to as time-of-flight mass spectroscopy). The ions are detected, which results in a spectrum of abundance versus \(m/z\).

There are different types of ionization methods for analysis of protein molecular weight by mass spectrometry; the two methods utilized for analysis of the proteins in this study are matrix-assisted laser desorption/ionization (MALDI) and electrospray ionization. MALDI involves mixing the protein sample with a large excess of matrix (such as a UV-active organic acid). Typically, lasers supply the necessary photons of UV light, which is absorbed by the matrix; the matrix transfers photons to the protein
molecules, which are vaporized. The gas phase ions enter a strong electric field, which accelerates them and gives them all the same final kinetic energy. They enter a flight tube, where ions are separated in vacuo based on their mass. A detector is located after the flight tube; typically MALDI employs a time-of-flight (TOF) analyzer, which correlates the time of travel of each ion through the flight tube to the inverse of the square root of its molecular mass. MALDI is most useful for protein samples that contain a large amount of buffer molecules such as phosphate and urea. The accuracy of assignment decreases as the molecular weight of the protein increases.

Electrospray ionization (ESI) is another technique to ionize the protein molecules. In ESI, protein molecules are suspended in droplets of solution (such as CH$_3$CN in water and a volatile acid) in the presence of a high voltage, which enables the protein molecules to become multiply protonated as the solvent evaporates. The molecules are ejected by charge repulsion as a fine charged mist into the drift region. In contrast to MALDI mass spectroscopy, ESI typically uses a quadrupole mass filter for detection of the molecular mass of each ion as abundance versus $m/z$, where the charge on each ion can differ by $+n$. This results in more stable protein systems than with MALDI; however, interpretation of the spectra is more difficult as the range of charge on the protein molecules can be quite large. Additionally, ESI has low tolerance to protein samples that contain salts and detergents, and is typically used when the protein is of low molecular weight (less than 150 kDa). Typically, the proteins are HPLC-purified prior to ESI mass spectroscopy to remove the salts and detergent molecules present in the sample.
2.5 Ultraviolet/Visible (UV/Vis) Spectroscopy

Proteins and DNA molecules can interact with electromagnetic energy in the ultraviolet and visible wavelength ranges. The wavelength of maximum absorption \((\lambda_{\text{max}})\) depends on the composition of the macromolecule and, to a lesser extent, its three-dimensional structure. UV/vis spectroscopy is frequently used as a tool for determining the concentration of a sample (using Beer’s law, \(A = \varepsilon bc\), where \(b\) is the path length in centimeters, \(c\) is the concentration in mol/L and \(\varepsilon\) is the molar absorptivity, which is dependent on the composition of the macromolecule) or for following processes such as binding or conformational transitions.

The molar absorptivity \((\varepsilon)\) of a protein in the unfolded state can be calculated based upon the number of aromatic residues present in the primary structure. At 280 nm, each tryptophan has a molar absorptivity of 5600 M\(^{-1}\) cm\(^{-1}\), each tyrosine has a molar absorptivity of 1420 M\(^{-1}\) cm\(^{-1}\), and each phenylalanine has a molar absorptivity of 197 M\(^{-1}\) cm\(^{-1}\).\(^1\) The formula to calculate the molar absorptivity is therefore:

\[
\varepsilon = (\#W \times 5600) + (\#Y \times 1420) + (\#F \times 197)
\]

Once the molar absorptivity for a protein is calculated, the concentration of the protein can be determined by obtaining the absorbance of protein at 280 nm. Differences in molar absorptivity are often observed in the absorption spectra of a protein in its unfolded and native state. However, these differences are generally small and the
absorption coefficient of the unfolded state serves as an approximation for that in the native state. Experimental methods have been devised to obtain an accurate value for \( \varepsilon_N \) as well.\(^{106}\)

When UV/vis spectroscopy is utilized to follow the transition of a folded protein to an unfolded one (via heat, acid or base titration, for example), the wavelength at which the absorbance of the native and the unfolded protein most greatly differ is first determined by obtaining spectra of both states. The transition is initiated and the signal is monitored at this wavelength throughout the transition. Additionally, UV/vis has been implemented in the study of binding of ligands or small molecules.\(^{43,104}\)

### 2.6 Circular Dichroism Spectroscopy

Circular dichroism (CD) spectroscopy is a versatile tool for the biophysical characterization of proteins. CD typically yields a determination of the secondary and, to some extent, the tertiary structure of a protein. The theory behind CD spectropolarimetry involves the interactions of asymmetrical protein chromophores with right- and left-circularly polarized light. Optically active materials interact with right-circularly polarized light differently than left-circularly polarized light. This difference is sensitive to structural environment and is usually quite small, yielding an ellipticity \( \theta \) of millidegrees. Characteristically, CD spectra are reported in mean residue ellipticity.

Symmetric molecules have no circular dichroism; however, the transition dipole interactions between a symmetric chromophore and other parts of the molecule yield
asymmetry. The amide group or peptide bond is the major CD chromophore in proteins in the far-UV region, whereas the aromatic residues are the main chromophores in proteins in the near-UV region. CD spectropolarimetry is sensitive to conformation and environment, because the interaction of two dipoles is dependent on both distance between and orientation of these two dipoles. These properties lead to the interaction of the molecule with circularly polarized light at different wavelengths. CD bands can either be positive or negative, depending on whether the material interacts more readily with right- as compared to left-polarized light. Thus a CD spectrum is a “footprint” of the secondary and, in favorable cases, tertiary structure of protein molecules. The spectrum can be mathematically decomposed to obtain an estimate of the secondary structure content.43,104

Conformational changes involving the disruption or formation of secondary structure (and tertiary aromatic interactions) are generally detectable by CD spectropolarimetry. Thus, this method is well suited to monitor denaturation transitions. Here as well, the wavelength of maximum change is identified, and signal strength is monitored as a function of denaturation concentration or temperature. The curve can be fitted with various equations depending on the presence or absence of various intermediates along the folding pathway.43
2.7 Fluorescence Spectroscopy

Fluorescence spectroscopy involves the emission of light energy from an excited singlet state as the chromophore returns to the ground state. (If the ground state is a triplet, fluorescence will be obtained from a triplet state – there will be no change of spin state.) When an electron is excited to a singlet state, it typically undergoes rapid internal conversion to the lowest vibrational level of this singlet state. It can emit energy by fluorescence, return to the ground state via nonradiative processes, or convert to a triplet state of lower energy (which is known as intersystem crossing). All of these alternative relaxation mechanisms compete with each other and lower the quantum efficiency of the fluorescent process. The mechanism that has the shortest lifetime will dominate.

Because of the internal conversion process and subsequent release of some of the excitation energy, the wavelength of emission is longer than the wavelength of initial excitation. Fluorescent molecules, or fluorophores, have efficient absorption leading to the first singlet and, as a result, a higher probability for fluorescence. If the first singlet is populated as a result of a strong $\pi\pi^*$ transition, as is the case for the aromatic residues (especially Trp and Tyr), the lifetime for emission is shorter than any other relaxation process, and fluorescence will be evident.

Fluorescence emission is dependent upon solvent effects as well as composition of the amino acids in the protein. The solvent has two effects on the fluorescence of the protein, termed the general and specific effects. The general effect is related to the polarizability, or ability to adjust to a dipole, of the solvent, and yields a longer
wavelength of fluorescence when the dielectric constant is increased. When the fluorophore is excited, its permanent dipole typically increases in strength. This stronger dipole interacts differently with the surrounding solvent than when it is in its ground state, and the solvent molecules arrange for preferential interaction to the new dipole. This serves to reduce slightly the total energy of the system. Once the excited fluorophore fluoresces, the solvent molecules must rearrange once again, which releases energy and returns the system to the ground state. The specific effect is the result of the excited state of the fluorophore, which leads to alteration of hydrogen bonding or electrostatic interactions of this fluorophore with the solvent. The non-covalent interactions result in a new characteristic fluorescent spectrum.

Fluorescence can be used to probe the structure of the protein, quantitate the presence of protein in a solution, and, in a process called fluorescence resonance energy transfer, detail the distance between two fluorophores. Additionally, dyes such as 1-anilinonaphthalene-8-sulfonate (ANS) interact noncovalently with hydrophobic portions of proteins, resulting in an increase in the fluorescence of the sample. This dye can be used to determine the amount of hydrophobic residues in a protein that are solvent-exposed.43

2.8 Nuclear Magnetic Resonance (NMR) Spectroscopy

NMR spectroscopy is a powerful tool for structural determination of organic compounds and biomolecules in solution. Application of NMR spectroscopy to proteins
allows for characterization of structure and dynamic properties. Additionally, processes such as binding and denaturation can be studied with this technique. Two types of internuclear interactions were utilized for the purposes of this work: scalar and dipolar. An NMR active nucleus (proton, in this case) is sensitive to the spin state of other nuclei within a few covalent bonds. This scalar or J-coupling is relayed by bonding electrons and depends on the covalent structure of the molecule. The identification of J-coupled spins therefore reports on the identity of amino acids. The second interaction is termed the nuclear Overhauser effect (NOE) and provides one mechanism for relaxation back to the ground state after perturbation of spin state populations. The NOE is a dipole-dipole interaction showing a $\frac{1}{r^6}$ dependence on the internuclear distance $r$. Most multidimensional experiments involve these two types of interactions.

One more descriptor of environment of the protein nucleus is the chemical shift $\delta$. This value corresponds to the frequency of light absorbed by each spin, and is usually referenced to an internal standard. The units of $\delta$ are either Hertz (Hz) or parts per million (ppm). The value of the chemical shift is based upon both the identity of the proton (i.e., what type of atom it is bonded to) and its surrounding magnetic environment. Significant shifts can be obtained for protons that are experiencing certain environments such as ring currents from nearby aromatic residues, or bound metal ions. Additionally, backbone torsional angles appear to determine the gross pattern of the $C\alpha$ shifts (upfield for $C\alpha$ protons in a helix and downfield for $C\alpha$ protons in a sheet), although other context-dependent components of this secondary structural shift such as orientation of the
$C\alpha$ protons or location in the overall tertiary structure sometimes mask this pattern.\textsuperscript{107}

These chemical shifts can be used to identify the proton as well as its environment – a useful tool for the structural analysis of proteins.\textsuperscript{43}
CHAPTER 3

INTRODUCTION TO MOLECULAR BIOLOGY

3.1 Overview of Molecular Biology Concepts

One of the more common inducible systems employed in protein expression is the \( E. coli \)-based \( lac \) operon, which utilizes the T7 RNA polymerase, a single polypeptide of 96,000 kDa, for high-level overexpression of the protein of interest. T7 RNA polymerase was originally selected because it is inducible by IPTG. It has high selectivity for transcription at its own promoter sequence as well as efficient transcription of the gene of interest. In regards to the last property, T7 RNA polymerase is five times faster than cellular polymerases and uses up most of the NTPs in the cell, which drastically inhibits transcription of all other genes by the host polymerase. The gene coding for the T7 RNA polymerase is under the control of the \( lac \) promoter and is located within the chromosome of the bacteria used for protein expression in this study.\(^{108}\) When in its “off” state, a repressor protein is bound to the operon, prohibiting most transcription and translation of the T7 RNA polymerase. The \( lac \) repressor has a high affinity for lactose and related compounds, such as isopropyl-1-thio-\( \beta \)-D-galactopyranoside (IPTG). When IPTG is introduced into the cell culture, the repressor binds the IPTG and no longer binds the
operator, which allows for transcription and consequential translation of the T7 polymerase gene.\textsuperscript{108}

A separate plasmid-based gene used in this study has the protein of interest downstream from a T7 RNA polymerase promoter region located on either the top or bottom DNA strand, along with resistance to specific antibiotics. The promoter region for T7 polymerase is 23 nucleotides in length and indicates the direction of transcription.\textsuperscript{108} When the T7 polymerase is induced by IPTG, it, along with other necessary cellular enzymes, binds to the promoter region on the plasmid containing the gene of interest and separates the DNA double strand. The T7 polymerase binds strongly only at the promoter region. Once the T7 polymerase locates the promoter region, it binds and transcribes the gene of interest, beginning at the initiating codon (typically ATG in bacterial constructs), using one of the DNA strands as a template and ribonucleoside triphosphate monomers as building blocks of the nascent mRNA molecule. Once two RNA building blocks are found that have complementarity to adjacent sites on the single-stranded DNA molecule, a phosphodiester bond is formed between the two by the polymerase. Transcription proceeds until the polymerase reaches one or more stop codons and detaches from the DNA template. As the mRNA molecule is being synthesized, it is released from the DNA template, which reforms its double helix.\textsuperscript{105}

Translation occurs as follows. Each tRNA molecule, which uses three bases to bind to the mRNA codon (described in section 2 of this chapter), is attached via its 3’ end to the carboxy end of its associated amino acid. This attachment is of high energy and
enables the carboxy group of this incoming amino acid to easily form a peptide bond with the NH₂ group of the last amino acid incorporated by the ribosome. The ribosome complex is primed for translation and, once ready, binds an mRNA molecule, which moves through the ribosome in a stepwise fashion. The sequence of the mRNA is translated into protein through the complementarity of the tRNA molecule with the mRNA codon. The ribosome forms a peptide bond between the carboxyl group of the C-terminal amino acid on the growing peptide chain and the amine group of the N-terminal amino acid that is being integrated into the chain. The nascent protein is built from the N to the C-terminus; the carboxy end remains activated by its covalent attachment to its associated tRNA adapter. This attachment provides the activation energy for addition of the next amino acid in the chain. Once the stop codon or codons are reached, the ribosomal structure releases the newly formed peptide and the mRNA template.¹⁰⁵

The expression of lysozyme is also frequently important for protein harvest. Lysozyme is an enzyme that disrupts the glycosidic linkages within the bacterial cell walls and leads to lysis of the cell. This increases the amount of protein that can be purified from the cell, as the cell does not need to be lysed with alternative and potentially inferior techniques. There are two types of lysozyme expression utilized in this paper. One of them involves the use of the pLysS cell line, which contains a plasmid that codes for the expression of lysozyme in parallel with the expression of the protein of interest. However, this expression system leads to a pre-IPTG induction basal level of protein expression. (This basal protein concentration level is, however, lower than BL21 cells not containing the pLysS gene.) This is not harmful unless the target protein is toxic.
to the cell, in which case the basal expression of the protein could delay or disrupt log phase cell growth. This leads to a decrease in living cells for protein expression and a corresponding decrease in the amount of protein expressed by the cells. The pLysS plasmid also confers chloramphenicol resistance to the cell.108

Another type of lysozyme expression system involves the use of pLysE. This is a more stringent system, with tighter regulation of protein expression; the gene and protein expressed are the same as when using pLysS, but the level of repression of the T7 polymerase is much higher. This allows for a much lower basal level of target protein expression – a useful feature when the protein is toxic to the cell.108

Alternative methods of cell lysis, such as sonication, freeze-thaw cycles or cell presses can be utilized in place of lysozyme during overexpression of the target protein. The use of the pLysS or pLysE plasmid, however, is much easier and, when used in conjunction with one of these alternative methods, allows for a large lysis of cells and release of protein.

3.2 Codons

Every three DNA bases in the template strand make up what is referred to as a codon, which is part of a larger “code” specifying for a particular protein sequence. The number of possibilities of amino acids that are specified by a set of three DNA base pairs of four possible identities each is $4^3$, or 64. As there are only 20 naturally occurring amino acids, many of the amino acids have more than one codon associated with them.
For instance, methionine and tryptophan residues have only one codon, whereas arginine, leucine and serine have six codons each. There are also three termination, or “stop”, codons that have no cognate tRNA.\textsuperscript{105}

The third DNA base in a codon is generally termed a “wobble” base. When an amino acid has more than one codon associated with it, it is usually the last base that is changed. For example, Leu has CUU and CUC as two of its codons. The last position is the only position that is different between the two, hence the “wobble”. The biochemical reason for the wobble position is to decrease specificity of binding of the tRNA molecule to the mRNA template; this is useful if little of the tRNA with that codon is at hand, yet a large amount of another tRNA is present whose codon differs only by the last base and is attached to the same amino acid.\textsuperscript{105}

There are tables of codon frequency values for a specific expression system such as \textit{E. coli}. These describe the regularity of accurate translation from the DNA codon to the amino acid intended. The codon frequency usage for \textit{E. coli} utilized in this study was derived from Class II usage tables.\textsuperscript{109}

3.3 Plasmids

A plasmid is a circular molecule of DNA that codes for one or more different properties. Frequently, a cell can have more than one plasmid – indeed, when a protein is to be overexpressed, the gene coding for the protein is present on a plasmid that is in
addition to any other cellular plasmids, such as pLysS or pLysE (as previously discussed). Plasmid DNA is not chromosomal DNA, and should not be confused as such. Typically, plasmids for recombinant protein expression contain a promoter region, a polylinker region (where many restriction enzyme sites, discussed below, are present), a termination region (which frequently contains regulator regions such as antibiotic resistance) and the gene coding for the protein of interest. A plasmid map for pAED4 is shown in Figure 3.1. This plasmid has a T7 RNA polymerase promoter region, a polylinker region and ampicillin resistance, and is about 4.5 kilobases (kb) in length. It is an ultra-high copy plasmid, which refers to the large concentrations of plasmid that are replicated in bacterial cells. However, because of its high copy number, the expression of protein using this vector can be compromised, as a large portion of cellular machinery becomes committed in plasmid amplification.

The polylinker region provides an excellent area for cloning the insert coding for the protein of interest. This region contains a large number of restriction enzyme sites. Restriction enzymes can cleave the phosphodiester linkages, which hold together the DNA mononucleotides, at specified recognition sites along the strand. A good percentage of these enzymes are palindromic cutters, meaning that the sites at which they cleave have upper and lower strands that are palindromes of each other. For example, *Ndel* is CA/TATG in the upper strand, 5’ to 3’; the lower strand is GTAT/AC, 3’ to 5’ and CA/TATG, 5’ to 3’, where / indicates the site of the cut between the bases. The strands after digestion with *Ndel* are therefore:
Figure 3.1. Plasmid map of pAED4. T7.Pr denotes the T7 polymerase promoter sequence; T7.TT indicates the T7 transcription termination region. The polylinker is located between XhoI and ClaI, inclusive. Amp resistance is conferred as noted. Most cloning into this plasmid involved digesting pAED4 with NdeI and EcoRI or ClaI; a mutant of pAED4 was also developed that had the BamHI site located after the ClaI site for cloning purposes.

(Dr. Paul Matsudaira, Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, http://web.wi.mit.edu/matsudaira)
where overhangs left after digestion are called the sticky or cohesive ends. Some
restriction enzymes are blunt cutters, leaving no overhangs on either the 5’ or 3’ ends of
the DNA molecule. Some have a decreased specificity of digestion, and can potentially
cut at a wide array of DNA sequences. Some also have what is known as star activity,
where their specificity of digestion is altered under different conditions, such as low salt
concentration, high amount of glycerol present, or absence of other factors such as bovine
serum albumin (BSA).108

If an insert is to be cloned into the polylinker region, the plasmid is cut with the
restriction enzyme, yielding a linear vector. If sticky ends are generated by the enzyme
of choice, the insert is designed to have 5’ and 3’ sticky ends complementary to the
overhangs in the cut vector. After covalently binding the insert to the vector by a process
known as ligation (described below), the result is a circular, double-stranded plasmid.
Because of the palindromic nature of common restriction enzyme sites, it is frequently
most useful to perform a double digestion, or cutting the plasmid with two different
enzymes. Incorporation of an insert into a singly digested plasmid can result in two
products. One of these is the target product, where the insert is oriented correctly within
the vector. However, the other will have the insert in the reverse orientation. This
situation is detailed in Figure 3.2.

As mentioned previously, each restriction enzyme requires specific conditions for
Vector singly digested with \textit{BamHI}

Insert with initial \textit{BamHI} and terminating false \textit{BamHI} sites

Insert (in red) is in the correct orientation; initial \textit{BamHI} site is conserved

Insert (in red) is in the incorrect orientation; initial \textit{BamHI} site is not conserved

Figure 3.2. Pictorial representation of two possible orientations of an insert cloned into a singly-digested vector.
optimal performance. When carrying out a double digestion, conditions must be suitable for each restriction enzyme. Frequently, one set of parameters will apply to both enzymes, and the digestion reactions can be run at once on the same reaction mixture. Ready-made buffers are commercially available for specific pairs of restriction enzymes, such as One-Phor-All (OPA, Pharmacia). If the optimal digestion conditions differ greatly between each enzyme, however, this concurrent digestion cannot be performed. Instead, the enzyme that will digest in the least complex buffer is used first. Reaction components are added to mirror the preferred conditions of the other enzyme, or a precipitation of the DNA followed by resuspension in a buffer providing the new conditions is performed.

Restriction enzyme digestion in this work typically takes about three hours to complete. Most enzymes utilized in this study have optimal temperatures in the range of 25-35 °C, and the temperature is controlled accordingly. Once cutting is complete, the enzymes are inactivated by heat inactivation (heating the reaction mixture to 65 °C or higher) or purification of the DNA sample from the enzymes. This step is essential, as once the insert of interest is cloned into the cut vector, the restriction sites will be restored and the enzymes can cut the sites again, resulting in low concentration of circular product.

When the restriction enzymes digest the DNA strands, they leave behind 5’ phosphate groups on each strand that can, in the presence of an enzyme called DNA ligase, reform the phosphodiester bonds in a digested plasmid. If the insert that will be cloned into the cut vector has 5’ phosphate groups already present, or if a single digestion
has been performed, dephosphorylation of the digested vector is usually required to
prevent recircularization of the vector without insert. Dephosphorylation involves the
hydrolysis of the 5’ phosphate groups from a DNA strand, leaving a 5’-hydroxyl
group.105,108 Calf intestinal alkaline phosphatase (CIAP) is the most frequently used
phosphatase. It must be thoroughly deactivated prior to ligation, however, for obvious
reasons. Heat inactivation, followed by gel purification, typically fulfills this necessity.

3.4 Developing and Cloning the Insert of Interest into a Cut Vector

Once the vector has been digested and a linear product is present, the insert must
be constructed. This is accomplished by several methods. There are several programs
written by various investigators that utilize the E. coli codon usage table to produce a
gene that has a high probability of coding for the protein of interest. It is also sometimes
useful to incorporate as many unique restriction sites as possible within the insert of
interest through the use of several commercially available programs, as long as this does
not compromise the probability of correct translation and the GC content. If further
cloning within the gene of interest is to be performed, unique restriction sites provide an
easy method for inserting a new strand of DNA into the plasmid. (These programs only
analyze the restriction sites within the insert, which may not be unique within the entire
vector.) Alternatively, there is a program on the World Wide Web available from
Entelechon GmbH that utilizes the target protein sequence to develop the template DNA
strand. The last method involves designing the insert manually using the codon usage table. It is also the most difficult technique.

Once the insert of interest has been designed, it is verified that no repeats of either single amino acids or patterns of bases are present, as this can confound the transcription and translation cellular machinery. Additionally, the GC content should be high enough to hold the double strand structure in place. As G and C bases pair with three hydrogen bonds (as opposed to A and T, which contain only two), an increase in GC base pair content increases the stability of the double-stranded DNA duplex. The temperature at which the double strand is half-melted into the single strand ($T_m$) can be calculated using the following formula:

$$ T_m = 81.5 + 0.41 \times \left( \% GC \right) - \frac{675}{N} \times %mismatch $$

where $N$ is the length of the sequence and $\% GC$ is the percent of G and C occurrence in the DNA sequence.$^{108}$

Once the insert has been designed, its complement is determined and the sticky ends (necessary for covalently binding the insert into the cut vector) are added. The cloning sequence and its complement are divided up into small sections (60 bases or less, mainly for ease and efficiency of synthesis) and ordered from a DNA oligomer supplier (as done in this study) or synthesized.

When the oligomers of the insert are obtained, they can be purified using denaturing PAGE. This uses urea as a denaturant to remove any non-covalent structure,
along with acrylamide and bis-acrylamide to form a matrix within which the DNA molecules can travel. Because of the presence of the denaturant, the distance that the DNA molecules pass through is based on their length. The DNA samples are usually provided by the supplier as lyophilized pellets; these can be resuspended in a low salt solution and dye mix. The dye (such as bromophenol blue or xylene cyanol) serves to determine the rate of travel of the samples through the gel matrix; choice of dye is generally dependent on the length of the molecules themselves, as different dyes run at different apparent lengths of single- or double-stranded DNA molecules. These are electrophoresed through the gel matrix by PAGE and visualized under UV light against a fluorescently-active TLC plate; the top \( \frac{2}{3} \) of the fragment that absorbs the UV light and does not fluoresce is excised and placed in a tube for elution purposes. This method removes most of the shorter fragments made during synthesis. The concentration of DNA obtained after PAGE purification is slightly lower than with other purification methods such as HPLC, but the purity is generally higher. Purification is not required for most qualitative studies, but is standard in cloning procedures.\(^{108}\)

Once the DNA oligomers have been purified, the DNA insert is assembled by phosphorylating (or kinasing) the oligomers, annealing the DNA fragments together and ligating the strands to form a covalently bonded double-stranded DNA molecule which is purified (with possible repeat phosphorylation prior to purification). These steps are detailed below and illustrated in Figure 3.3.

Phosphorylation involves the addition of a phosphate group onto the 5’ end of a DNA strand. This is important for later steps involving the formation of a phosphodiester
Digestion of plasmid with appropriate restriction enzyme(s)

If external 5’ ends of oligomer will be phosphorylated, dephosphorylation of 3’ ends of plasmid

Phosphorylation of 5’ ends of internal oligomers

Annealing of all oligomers making up insert

Ligation of oligomers to form covalent attachments

Agarose gel purification

If desired, phosphorylation of external 5’ ends of insert

Ligation of insert into vector

Figure 3.3. Flowchart of mutant plasmid construction.
bond between DNA molecules, and is accomplished in this study with T4 polynucleotide kinase (PNK) and ATP. PNK removes the terminal phosphate group from ATP, producing ADP, $P_i$ and energy. It adds the $P_i$ group onto the 5’-hydroxyl end of the DNA strand. Usually only the interior pieces of DNA are phosphorylated; the 5’ pieces on either end are removed from the phosphorylation step so that concatamers of DNA are not made.\textsuperscript{108}

Following phosphorylation, the 5’ end fragments of both strands are added to the reaction mixture and the DNA single strands are annealed together. This allows for the molecules to locate their complementary “partners” by searching out the optimal number of base pairs between the other strands. It is expected that by the time annealing is complete, the majority of DNA molecules are correctly paired together.\textsuperscript{108}

Once the DNA molecules have formed a double-stranded structure maintained by noncovalent hydrogen bonds, the entire assembly is ligated. This involves the use of T4 DNA ligase, which forms the phosphodiester linkages between the 3’-hydroxyl end of the last base on one oligomer and the 5’ phosphate end of the first base on the next oligomer. This results in two DNA strands that are made of covalently linked oligomers; these strands are hydrogen-bonded together to form a double-stranded DNA segment, or insert.\textsuperscript{108}

If the vector to be used is dephosphorylated, the insert is phosphorylated yet again to attach phosphate groups at the 5’ ends of each DNA strand. This is required for formation of the phosphodiester bond between the vector and the insert. Again, CIAP is
most commonly used; the enzyme must be thoroughly deactivated following
phosphorylation, as described earlier.

Once the insert has been constructed, it is purified as described below and cloned
via ligation into the cut vector. Typically, the concentrations of insert and vector are
varied to determine the ideal amounts for complete ligation; frequently, the reaction
mixture must be saturated with one or the other to obtain the highest amount of circular
target. Ligase is responsible for ligating the 5’ phosphorylated ends to the 3’ ends of the
previous base. Cellular enzymes will close the nick between the 3’ end of the top strand
of the vector and the 5’ end of the top strand of the insert. The same is also true for the
bottom strand.\(^{108}\)

3.5 Purification of DNA

DNA purification methods that have not been mentioned yet are ethanol
precipitation and agarose gel electrophoresis. These can be used to purify both single-
and double-stranded DNA, although (in this work) gel electrophoresis is only used to
purify double-stranded DNA.

Ethanol precipitation involves the addition of ethanol (usually ice-cold 200 proof,
2.5 times the volume of the solution, \(v_s\) along with 0.1 \(v_s\) of sodium acetate and, in
some cases, 1-3 \(\mu\)L of tRNA as a template for pellet formation. In the presence of
monovalent cations such as \(\text{Na}^+\), ethanol induces a structural transition in the DNA
molecule, which leads to aggregation. The solution is cooled on ice and centrifuged at
high speed; the result is a pellet of DNA. This pellet is washed with 70% ethanol in distilled water to remove the salt ions; it is dried and resuspended in the buffer of choice. This method is most useful in the concentration of DNA samples, as well as the removal of salt and other small molecules, which remain soluble in 70% ethanol.  

Agarose gel electrophoresis is most useful in the purification of DNA from enzymes, as well as the identification and separation of DNA molecules based upon length. It is similar to the SDS-PAGE purification of proteins as described earlier. Varying types of agarose can be used; each type has different physical properties, such as the pore size of the solid gel matrix or low melting point. In most cases, the agarose is used in certain concentrations to control the degree of separation of the DNA molecules based on length. The DNA molecules are applied to the agarose gel, or “loaded”, and an electric field is generated by applying voltage at the ends of the gel. The DNA molecules travel through the agarose matrix based mostly upon their size, which is directly proportional to the amount of negative charge from their backbone. Typically, a ladder of base pair markers is utilized to identify accurately the presence and location of the DNA band of interest. 

In this work, a solution of agarose in buffer is prepared and a small amount of ethidium bromide (EtBr) is added. EtBr is a planar molecule that intercalates between the stacked DNA bases and fluoresces reddish-orange under UV light. EtBr reduces the mobility of the DNA molecule through the gel matrix; this is especially evident when electrophoresing unnicked plasmid DNA, which has a negative supercoil. (DNA twist, Tw, is defined as number of base pairs divided by the average repeat of base pairs per
turn. If the average repeat of base pairs per turn is increased, $T_w$ will decrease, and the DNA is considered unwound. The writhe $W_r$ of the DNA is correlated to $T_w$; $W_r$ is the response of the DNA molecule to the strain provided by $T_w$, and leads to supercoiling of the DNA structure. Supercoils can be either positive (positive $W_r$) or negative (negative $W_r$). The linking number $L_k$ is $T_w + W_r$, and defines the topology of the DNA. $L_k$ is fixed for a plasmid DNA molecule and can be altered only through strand cutting and resealing.) The amount and sign of this supercoil is influential when determining the mobility of electrophoresis of the plasmid through the gel; the larger the supercoil, the smaller the plasmid molecule and the larger its mobility in an agarose matrix.

Intercalated EtBr unwinds the helix and decreases $T_w$, resulting in a change in mobility from the negative supercoil value. EtBr is added after the gel is heated and cooled; immediately following addition of EtBr, the gel is poured into a gel-casting platform and a comb (to provide sample wells) is placed at the top of the apparatus. It is essential to properly dispose of all materials containing EtBr such as buffer and the gel itself, as EtBr is a suspected carcinogen.

The agarose gel is placed in a gel running apparatus, usually an electrophoresis tank, and the DNA samples are loaded into the wells. Voltage (50-200 V) is applied until the samples electrophoreses to the desired extent. The gel is visualized under UV light, and the bands can be excised or, if comprised of DNA strands less than 2 kb in length, run onto a membrane that has a high affinity for DNA. Usually membranes of this type are nitrocellulose-based; a small square of membrane is placed directly above and below the DNA band of interest, and the band is electrophoresed onto the lower membrane.
The DNA is eluted from the membrane with the use of a high-salt buffer and moderate heat.\textsuperscript{108}

A practical method of larger-scale plasmid purification is the use of commercially available plasmid purification kits. These are typically based on alkaline lysis procedures and involve the resuspension of the cell pellet in several buffers, which lysed the cell wall and digests the cellular proteins and RNA, followed by centrifugation and binding to a column. The DNA can be eluted off the column using distilled water or buffer.

The concentration of DNA is determined by diluting the DNA sample (in this work, typically 1 $\mu$L of 50-5000 ng double-stranded DNA is added to 999 $\mu$L of water or TE) and obtaining the wavelength spectrum between 200 and 300 nanometers. The spectrum typically has a $\lambda_{\text{max}}$ around 260 nm and a large peak (usually of absorbance units greater than 1) below 220-230 nm. A ratio of $\frac{A_{260}}{A_{280}}$ will give the purity of the sample; the concentration of the DNA is calculated as follows:

$$\text{Concentration} \left( \frac{\text{ng}}{\mu\text{L}} \right) = A_{260} \times \text{dilution factor} \times 50 \frac{\text{ng}}{\mu\text{L}}$$

where the dilution factor in this example would be 1000.\textsuperscript{108}
3.6 The Polymerase Chain Reaction (PCR)

PCR has become an indispensable tool for molecular biologists. There is a variety of applications that utilize PCR, as well as a wide range of methods that involve PCR and PCR-based procedures. PCR is a method used to produce large amounts of target DNA through the use of two primers (short oligomers that bind with high affinity to the top and bottom strands of the target sequence of the double-stranded DNA), dNTPs, a thermostable polymerase and the DNA template, will yield a large amount of the DNA target sequence. The template is present in low concentration, whereas the primers are supplied in vast excess of the template. The function of the polymerase somewhat mimics the process of cellular replication – it extends the primers in a 5’ to 3’ direction, yielding new strands of varying length that can be used as template for additional primer binding. The PCR amplification reaction as utilized in this study first separates the DNA double strand by incubation for less than a minute at high temperature. The temperature is lowered to promote annealing of the primers to the DNA template. After a few seconds, the temperature is raised to favor amplification of the gene of interest through binding of the PCR amplification polymerase (typically Pfu, Vent or Taq) to melt out mismatched primers and local secondary structure, as well as improve the specificity of the product. The time allowed for this amplification depends on the length of template to be amplified and is highly variable. After a high number of cycles (e.g. 15-25), the product is obtained in much greater concentration than either the
primers or the template, which is purified or further digested, depending on the final product that is desired.108

The two main applications involving PCR that are discussed in this work are amplification and mutagenesis. Both of these utilize the same set of cycles as described above, but differ in the times and temperatures as well as the enzymes used. Amplification of a target of interest utilizes primers that are completely complementary to the template and bracket the area to be amplified. Continuous cycling of dissociation, annealing and extension temperatures lead to a high concentration of the area of interest between the primers. The reaction mixture is purified on an agarose gel and the DNA fragment of interest is excised from the gel or eluted onto a DEAE nitrocellulose membrane. Frequently, PCR is used as a qualitative analysis to determine whether a specific insert has been properly cloned into the cut vector; in this case, no further purification from the agarose gel is necessary.

Mutagenesis involves amplification of almost all of the entire plasmid, plus alteration of several key base pair identities. The primers bind to the same exact region of the template (one on the top and one on the bottom strand), but they are not entirely complementary to this template. The primers differ in the area of the mutation of interest, but have long “tails” around the mutations that are completely complementary to the plasmid. The lengths of these tails are dependent on the extent of the mutation. A full codon (three base) mutation necessitates longer tails for higher binding efficiency of the primers to the template than a single base mutation. A considerably longer extension time is allowed, but it is of the utmost importance that the temperature of annealing
disfavors the enzyme’s amplifying the region where hydrogen bonds exist between the template and the primer. Failure to do so will result in DNA that is longer than the plasmid template and contains two copies of the insert. The second copy of the insert will not include the mutation.

Following mutagenic PCR, the enzyme DpnI is added to the mix to cleave the bacterially amplified template DNA. DpnI is specific for GATC sequences that contain methylated adenine residues. The digestion will only affect the template DNA amplified by E. coli, which does not contain the mutation of interest but is methylated. The PCR-amplified DNA, which contains the mutation but is not methylated, will not be cleaved. The resulting reaction mixture is transformed into DH5α. As this plasmid confers antibiotic resistance to the cell, only those cells that contain the plasmid will grow. (This procedure, supplied by Dr. Song Tan, Biochemistry and Molecular Biology Department, was based on the QuikChange method available from Stratagene.)

3.7 DNA Sequencing

All DNA sequencing performed for this work was carried out at nucleic acid facility in the Penn State University Biotechnology Institute on an ABI Hitachi 3100 Genetic Analyzer. The method used in this study involved automatic electrophoresis of the samples through the gel matrix, followed by automatic detection of the fluorescently labeled DNA. A different fluorescent tag was used for each ddNTP, which made the analysis of the results much simpler.108
3.8 Transformations of Bacterial Cells with Plasmid DNA

Once the plasmid of interest has been constructed, it is transformed into competent bacterial cells for amplification or translation into protein. Transformation involves preparing the cells to accept extracellular plasmid DNA. There are several different methods for transformation; all of them use simple salt solutions or voltage to make the cell membranes more permeable. The method used for the transformation of cells in this work involves a procedure based on the Hanahan method\textsuperscript{108}; the cells are made competent with CaCl\textsubscript{2}, MnCl\textsubscript{2} and RbCl\textsubscript{2} and are subsequently heat-shocked to allow for plasmid entry. During the competent cell procedure, the cells are grown to an optical density (OD) of 0.5-1.0 as detected by UV at 600 nm. They are harvested and resuspended in buffer containing heavy metals (here, Ca\textsuperscript{2+}, Mn\textsuperscript{2+} and Rb\textsuperscript{2+}) and allowed to equilibrate on ice. After equilibration, the cells are harvested and resuspended in another buffer containing Ca\textsuperscript{2+} and Rb\textsuperscript{2+}. Again, they are allowed to equilibrate on ice. They are aliquoted, rapidly frozen in liquid nitrogen or a dry ice and ethanol bath, and stored at –70 °C.

When the cells are required for protein expression or plasmid amplification, they are thawed on ice in the presence of the plasmid DNA. The DNA/cell mix is allowed to sit on ice for a specific period of time and the mix is heat-shocked for a few minutes, which allows the pores to open, the heavy metal ions to be displaced and the DNA molecules to enter the cells. They are placed on ice again for a few minutes,
closes the pores, and inoculated into or plated on a growth medium with a selective antibiotic and allowed to grow to the desired density.\textsuperscript{108}

There are several different types of cells that are used in various stages of the development and expression of a protein. The three most commonly used cells are DH5\textalpha, BL21 and HMS174. DH5\textalpha is specifically used for plasmid preparations, and is an \textit{E. coli} K-12 derivative with an F $\phi$80dlacZ\DeltaM15 $\Delta(lacZYA-argF)U169$ \textit{deoR recA1 endA1 hsdR17}(r $\kappa^-$, $m$ $\kappa^+$) $\text{supE44}$ \textit{thi-1 gyrA96 relA1} genotype. The presence of \textit{recA1} and \textit{endA1} increase the stability of the insert and the quality of DNA produced from miniprep by preventing recombination events. DH5\textalpha is usually highly competent; it contains no plasmid and therefore has no inherent antibiotic resistance. It is mainly used to generate high copies of the plasmid with the insert of interest and is almost never used to express protein.

BL21, on the other hand, is a B834 derivation that contains the pLysS plasmid (discussed previously). It is a general \textit{E. coli}-based protein expression system that lacks the \textit{lon} and \textit{ompT} outer membrane proteases, both of which can degrade proteins during purification. Its genotype is \textit{E. coli} B F$^-$ \textit{dcm ompT hsdSB} (r $\beta^-$ $m$ $\beta^+$) \textit{gal} $\lambda$(DE3) [pLysS Cm$^R$]. DE3 refers to the fact that the cells are infected with a copy of the T7 RNA polymerase gene under control of the \textit{lacUV5} promoter. BL21 cells are used for protein overexpression.

HMS174 is somewhat related to BL21 in functionality and stringency. It is an \textit{E. coli} K-12 derivative that is an F$^-\lambda$(DE3) [pLysS Cm$^R$ Rif$^R$] \textit{recA} $\lambda^{-}$ \textit{hsdR19} ($r^{-}$ $\kappa_{12} m^+$ $k_{12}$) genotype. The lack of the \textit{recA} gene is purported to stabilize tandem repeat
sequences. HMS174 is therefore used when the bases in the plasmid of interest have repetitious sequence identity.
CHAPTER 4

COMPUTATIONAL ALGORITHMS FOR PROTEIN DESIGN

Sequence Prediction Algorithm (SPA) and Sequence Prediction Algorithm for Numerous States (SPANS)

4.1 Introduction

Computational protein design (for the inverse folding problem) typically involves the development of an algorithm that selects for a sequence or family of sequences that theoretically adopts a target fold. An effective protein design algorithm must comprise both an atomic level scoring function for comparing different protein sequences, and efficient combinatorial reduction and sampling methods to seek low energy sequence/structure solutions to the problem defined by the template backbone structure.

Development of a protein design scoring function typically begins with the use of a previously developed molecular mechanics potential function, such as AMBER or OPLS. Such force fields comprise empirically parameterized terms representing nonbonded atomic interaction energies, including van der Waals and electrostatic interactions. Additionally, these force fields include standard geometries for interatomic distances, orientations and torsional angles, deviations from which are captured with additional scoring functions. The parameterization of these force fields was developed for consistency with high-resolution
structural data, spectroscopic measurements, and physical properties of liquids. For the most part, molecular mechanics energy functions are used for exploring the relative energetics of and dynamics between various conformational states of a given molecular entity, such as a small organic molecule, a protein or a nucleic acid. In addition, these force fields also include molecular representations of water, such that molecular simulations can explicitly incorporate interactions with solvent. However, numerous approximations also exist for implicit solvent representations.

Protein design algorithms, while drawing heavily from molecular mechanics functions, force the development of additional and/or alternative representations of atomic energetics. These changes arise for several reasons. First, most sequence design calculations result in a combinatorial explosion of possibilities. Because of this, computational efficiency is paramount, prohibiting the use of inherently complex scoring terms. Second, the sequence design problem is by definition a non-continuous problem because of frequent changes in covalent geometry and the number of atoms in the system. The problem is additionally separated at the level of the side chain orientations or rotameric states, again in the interest of computational expediency. The need for efficiency and discreteness prohibits the use of an explicit solvent representation. The separation also has unpredictable effects on the relative importance of different terms in the scoring function.

Perhaps the most dramatic difference between molecular mechanics and protein design scoring functions is that the latter must be capable of estimating the relative energetics of two protein sequences – relative to a chosen reference state – at an extremely rapid rate. This problem leads to additional terms that correct for compositional differences between
each sequence and implicitly incorporate information from the reference state (typically the unfolded ensemble). Finally, protein design algorithms should result in the creation of sequences that satisfy additional constraints such as high solubility under typical solution conditions. The novel portions of the potential function described herein, developed by the Desjarlais group, involve the application of a baseline compositional correction term, parameterization of implicit solvent representations, implicit references to the unfolded state, and ultimately, the implementation of backbone flexibility, as described in Section 10 of this chapter.

4.2 Amino Acid Baseline Corrections

Native sequences are comprised of an assortment of amino acid types. Not all amino acids need be present in a sequence, and indeed some native proteins have a frequency of occurrence of one particular amino acid that is higher than statistically prevalent throughout all known naturally occurring proteins. However, wild-type proteins do not typically have a large amount of only a few amino acids present in their primary structure. Native proteins are therefore said to generally have high sequence entropy or diversity and low sequence degeneracy. Here, sequence entropy is measured by:

$$S = -R \sum_i \frac{P_i}{\text{length}} \ln \frac{P_i}{\text{length}}$$
where $R$ is the ideal gas constant, $p$ is the normalized probability of each amino acid $i$ and $\text{length}$ is the length of the polypeptide chain. Most native proteins have a high $S$ value. Designed proteins, however, frequently have low sequence entropy – the design will be rich in a specific amino acid because of several different factors. For example, if the hydrogen-bonding term is too strong, there will be a high occurrence of amino acids that can hydrogen bond with each other, such as serine or tyrosine. If the penalty for the exposure of hydrophobes is too low, there will be a large number of hydrophobic amino acids that are solvent-exposed. Neither of these cases have a high probability of leading to a properly folded protein that is monomeric; in the first case, there are so many choices for a hydrogen-bonding partner that the native state can in fact be a continuum of low-energy states separated by low activation energy ($E_{\text{act}}$) barriers, resulting in a dynamic fold. In the second case, the protein may have large patches of hydrophobic regions on its exterior, resulting in strong intermolecular interactions. Neither of these, unless specifically desired, leads to a positive result.

Both of these situations can be corrected by arriving at the ideal parameter weight for the physical parameters discussed or, indeed, a subset of parameters. However, there are certain circumstances where the ideal weight cannot be approximated or where other ill-defined parameters make it impossible to arrive at a reasonable amino acid composition. It is here that amino acid baseline corrections can be valuable.

Amino acid baseline corrections, first addressed by Raha et al in 1999\textsuperscript{111}, take into account the sequence diversity of native proteins. Once the ideal occurrence of a certain amino acid is determined by statistical analysis of all known native proteins in all
families and from all organisms, the deviation to any great extent from this percentage, or fraction, of occurrence in a designed sequence can be penalized. In effect, the amino acid baseline correction is merely an additional energy term, utilized to provide for sequence diversity related to that of native proteins. This has proven useful to design sequences that appear native in sequence entropy. (For a more in-depth theoretical analysis of whether sequence diversity results in structural specificity, see Chapter 7.) The amino acid baseline correction term is as follows:

$$F = \sum_{x=1}^{20} n_x B_x$$

where $n_x$ is the number of times amino acid $x$ occurs in the designed sequence and $B_x$ is the baseline correction factor for this amino acid $x$. Values of $B_x$ in this study were as published. This term can be applied much as any other energy term is applied. Obviously, a designed sequence whose composition differs greatly from that of a typical native protein will have a large energy penalty, whereas a designed sequence whose composition approaches that of native proteins will have a minimal energy penalty applied towards the full energy calculation.

Another option involves the use of a sequence entropy weight. The entropy of the sequence can be calculated as discussed, weighted appropriately and applied to the energy calculation like any other parameter. This is more straightforward than the baseline corrections, but frequently does not give a native-like amino acid composition.
Other approaches utilized by different groups usually do not explicitly utilize the diversity of the sequence as a parameter for protein design. However, determination of an appropriate sequence entropy weight is difficult, and there is no verification that wild-type proteins develop so as to specifically increase their sequence diversity. This use of a sequence entropy weight will be discussed more fully in the next chapter.

4.3 Secondary Structure Propensity

The application of secondary structure propensity (ssp) is appropriate when a designed sequence appears rich in amino acids known to have high occurrence in the non-target folds. This can be determined by visual examination or calculation of the secondary structure propensity using tables of ssp values calculated for amino acids. For example, in wild-type proteins, prolines are found rarely in helices and often in turns. If a designed sequence intended to be helical is rich in Pro, it can be expected that the result will not be of the desired structure. Thus, biasing the composition against Pro in α-helices but favoring Pro in most turns will assist in designing a protein that adopts the target fold. There are certain instances where ssp values are not necessary, such as when the design has adequate constraints via other parameters to maintain the desired structure. However, if these parameters are not weighted properly, application of ssp will sometimes assist in designing a protein that folds into the intended structure.

The application of ssp as performed in this work is straightforward. The input file contains a secondary structure designator for each residue (either helix, sheet or turn,
although coil can also be specified) that is utilized, along with tables of secondary structure propensities, to adjust the energy of the models appropriately. If a region of the protein is helical, for example, the selection of amino acids of high sheet propensity will be energetically unfavorable. When an amino acid with high probability to occur in a certain type of turn is desired, the selection of an amino acid that fulfills this characteristic will be energetically favored.

The ssp should not replace other properly weighted parameters such as hydrogen bonding or electrostatic interactions. It is also unwise to utilize the ssp as the sole parameter to design a protein sequence that maintains the target fold. Protein structures are sufficiently plastic to tolerate an amino acid with low propensity to any given secondary structural element. Thus, although ssp is a useful tool to prevent the accumulation of unfavorable residues, it is not given a weight higher than 5; above this value, selection of amino acids that only have high secondary structure propensity for the target fold, yet are not ideal in terms of electrostatics, solvent exposure, or Lennard-Jones interactions becomes prevalent.

4.4 Hydrophobic/Polar (HP) Patterning

As mentioned in Chapter 1, the beginning attempts at protein design relied heavily on the HP patterning of the native protein. The computational time is greatly decreased, and the sampling efficiency increased, when the subset of available amino acids at each position is reduced. HP patterning, described in Chapter 1, serves to reduce
the permissible amino acid positions. In the core of a protein, hydrophobic amino acids are allowed, whereas on the exterior, polar amino acids are preferred. At interfacial positions, where the position is not obviously either buried or exposed, both polar and nonpolar residues are allowed. Whether the position is buried or not is frequently determined by a contact score. As discussed in Chapter 1, the Micheletti contact score is the most widely used in this work; the weight of the contact score can be adjusted as desired, much like any other parameter. The contact score is calculated as follows:

\[
C_a = \frac{1}{\sum_{i=1}^{\text{chainlength}}} \frac{1}{1 + e^{d_{a,i} - 6.5}}
\]

Here, \(d_{a,i}\) is the distance between a generic side chain \(\text{C}_\beta\) centroid \(a\) and another \(\text{C}_\beta\) centroid at position \(i\). If the generic centroid has a contact score higher than 5.5, it is considered buried and will only be allowed the H subset of amino acids [\(H = \{\text{A, F, G, I, L, P, V, W, Y}\}\)]. Anything having a contact score less than 5.5 is allowed the P subset of amino acids [\(P = \{\text{A, D, E, K, N, Q, R, S, T}\}\)]. As has been mentioned in previous chapters, burial of hydrophilic residues and exposure of large amounts of hydrophobic residues will lead to the design of a protein that does not adopt the target conformation.
4.5 Description of Rotamers

Almost all amino acids have several statistically prevalent rotameric states as observed in high-resolution protein structures,\textsuperscript{71,72} which are described as conformations of the side chain angles as dictated by the steric constraints on the residue. For example, there are three specific conformations around valine’s Cβ, labeled as \textit{gauche}+, \textit{gauche}− and \textit{trans} (Figure 4.1). A typical rotamer library is derived from statistical analyses of high-resolution wild-type protein structures. These rotamer libraries are typically backbone-independent, at least initially; after placing each amino acid in each rotamer on the backbone and filtering for steric hindrances, the resulting library is considered backbone-dependent. The \textit{trans} conformation for most residues besides Pro is 90\% favored in helices, whereas it occurs only in 72\% of the strands. Frequently, adding a certain amount of variation to the statistically derived rotamer increases the rotamer possibilities. This serves to mirror the observed properties of native proteins.

The calculation for the energy of the rotamers is as follows:

\[ E_{rot} = \sum_{\text{torsions}} \frac{V_n}{2} [1 + \cos(n\chi)] \]

Here, the sum is taken over all possible torsion angles. This is a cosine-shaped potential, with energy minima at 60°, 180° and 300° and energy maxima at 0°, 120° and 240°, to mimic the energy dependence of an sp\textsuperscript{3} hybridized bond. As mentioned above, a small
Figure 4.1. Side chain dihedral or torsion angles of valine.
amount of flexibility can be allowed around each statistically derived rotamer\textsuperscript{112}; this value is typically less than 15°.

4.6 Steric Filtering

The most important part of the pre-design setup is the filtering of rotamers of amino acids (as described in Section 4) that exhibit steric clashes with any of the backbone atoms. The Lennard-Jones (6-12) potential attributes a high energy to any atoms that are occupying the same space at the same time. Rotamers of amino acids that have a steric clash with the backbone atoms of the target fold (as provided by the high-resolution experimentally determined backbone structure) will experience a repulsive force and therefore have such a high Lennard-Jones energy. Those rotamers that experience a steric clash can result in a relaxation of the protein backbone or the population of an alternative rotameric state \textit{in vitro}. The goal of protein design in this work is to have the experimentally determined protein backbone structure maintain an rmsd of less than \(~1\ \text{Å}\) to the target fold while preserving as many of the designed rotameric states of the amino acids as possible. The relaxation of the protein backbone and the population of alternative rotameric states will thwart this goal. Steric filtering, therefore, helps in eliminating those candidates for which the Lennard-Jones energy is so high that backbone relaxation becomes unavoidable. Rotamer elimination at this stage also efficiently improves the sampling during the combinatorial stage of the design simulation.
The general method of application involves the use of steric and solvent effects. There are two stages of steric filtering. The first one involves the application of all of the rotamers of all the allowed amino acids (the set is filtered for HP preferences, if used) placed, one at a time, in a chosen position of the protein. If they exhibit a steric clash with any of the backbone atoms (or a Lennard-Jones energy greater than 20 kcal/mol), they are automatically removed from the rotamer/amino acid possibilities. This is done for all positions on the amino acid chain. The second filtration involves the application of all rotamers of all amino acids on a chosen position on the amino acid chain, as before. The extent of burial is calculated by the use of the Micheletti contact score, as described in section 4.3. Here, Cβ centroids are assigned to all other positions not under consideration; this involves defining a point at 2.9 Å from the Cα atom coordinates along the standard geometry Cα-Cβ bond vector using the optimal geometry as defined in the AMBER force field, considering the different atom types and hybridization states involved in the bond.

The contact score of the amino acid under consideration is determined. If it clashes with any centroid, it is eliminated. Additionally, in this work, rotamers of polar residues (D, E, K, N, Q, R, S, T, Y and W [hydrophilic portion]) are eliminated when the polar atoms cannot form hydrogen bonds and have a contact score greater than 5.5. If a non-hydrogen bonded polar atom is buried within the protein core in a specific rotamer, that rotamer is disallowed. Rotamers of nonpolar side chains (F, I, L, V, P and W [hydrophobic portion]) are eliminated when any of their atoms have a contact score less than 2.0. The contact score filtration overlaps with the HP patterning, as described in
section 4.3. (Ala and Gly residues are not subject to filtering. Ala has equal probabilities of being found in a buried or exposed position, and Gly has no side chain.)

The use of steric filtering does greatly increase the chance for survival of a set of amino acids that would adopt the correct protein conformation, as well as the speed of the design procedure itself. It therefore is useful in that it alleviates the computational problem by reducing the total number of possible models. However, it is sometimes overly restrictive in that it can eliminate a low-energy structure that would relieve its very slight steric hindrance (or a very small overlap of the van der Waals radii of the atoms in question) by slight adjustment of its conformation. When point mutations are made to native proteins, the resultant steric hindrance (if any) will be relieved through rearrangement of the backbone or the other side chains experiencing the hindrance. Frequently, these relaxations do not involve large motions of the structure. This is not true in most protein design attempts; all rotamers experiencing steric constraints are filtered. It is possible that, by filtering all steric clashes out of the design procedure, a sequence that would have a very slight steric clash and therefore an extremely similar structure to that of the target structure is disallowed. If there are other solutions to the problem, this is not a terrible loss. However, if all of the other solutions are of considerably higher energy or will not adopt the target structure, it is a significant problem.

There are several ways to overcome the structural relaxation problem. The use of the flexible backbone or the backbone ensemble, which will be discussed later, may serve to remove the rotamers filtered because of a near-negligible steric clash. There are also
small amounts of flexibility one can allow around each statistically derived rotameric state to mimic minor side chain adjustments in experimental proteins. Additionally, a small amount of steric clash below a set threshold can be allowed. This threshold could be increased or decreased gradually as needed. Either way may alleviate the structural relaxation problem.

4.7 Energy Calculations and the Parameters Involved

There are many different parameters that can be included in an energy calculation to mimic \textit{in vivo} and \textit{in vitro} conditions. Historically, the most frequently-utilized ones are: the Lennard-Jones energy of the side chains with each other, the Lennard-Jones energy of the side chains with the backbone, the reward for the burial of hydrophobic residues relative to their exposure in the unfolded state and the analogous penalty for the burial or desolvation of hydrophilic residues, the rewards for the exposure of hydrophilic residues and the analogous penalty for the exposure of hydrophobic residues (unless functionality and/or multimeric behavior is included in the design goal), hydrogen bonding and salt bridges. Other energy terms discussed earlier are the secondary structure propensity, the HP patterning (which is rarely a true energy term, but a guide for residue selection) and the amino acid composition correctional term.

The description of the side chain-side chain and side chain-backbone interactions involve the use of the Lennard-Jones potential energy calculation or nonbonding interaction. The Lennard-Jones potential has the following description:
\[ E = \sum_i \sum_{j>i} 4\varepsilon \left[ \left( \frac{\sigma}{R_{i,j}} \right)^{12} - \left( \frac{\sigma}{R_{i,j}} \right)^6 \right] \]

where \( \sigma \) and \( \varepsilon \) are the Lennard-Jones parameters related to the radii and well depth of the atoms under consideration, and \( R_{i,j} \) is the distance between these two atoms. The first term describes the repulsive part, and the second term refers to the attractive portion. These parameters are derived from the OPLS nonbonded parameter set.\(^{113}\) This potential is referred to as the 6-12 potential as per its dependence on \( R \). In this study, a united atom potential was used, which involved ignoring explicit hydrogen atoms but altering the size of the carbon atoms based upon the number of hydrogen atoms attached.

The burial of hydrophobic residues and the exposure of hydrophilic residues are important because water-soluble native proteins commonly fulfill these two criteria to at least some extent. The energy of the burial or exposure can be calculated using the Micheletti contact score, as discussed earlier, to determine the extent of burial or exposure of the position. The larger the contact score, the more atoms (including intraresidue atoms) with which the atom under question is in contact, and the higher the chance that the position is buried. The opposite is true for an exposed position, which does not “see” many other atoms. The contact score is used in conjunction with a partition function to quantify fully the energy of burial or exposure. The partition function is the free energy of transfer of a certain amino acid from an organic solvent to water. Usually octanol is chosen as the organic solvent, as it theoretically has a dielectric constant and a hydrogen bond-to-molecular weight ratio similar to the core of proteins.
The partition function (Q) is a mathematical description of the energy of sequestering an amino acid in octanol compared to the energy of sequestering the same amino acid in any solvent, and can be utilized to provide a means to determine the amount of hydrophobicity or hydrophilicity of the amino acid under consideration. A reference state (such as an amino acid in water) is used, and ∆E between this state and a new state (such as the same amino acid in octanol) is determined. Q can be calculated as $Q = \frac{1}{1 + e^{\frac{-\Delta E}{RT}}}$.

The partition function is also related to the solvation energetics of the system, which is evaluated as follows:

$$E = \sum_i S_i \Delta A_i$$

where $S$ is the atomic solvation parameter (or asp), which is an approximation of the true solvation free energy of the structure when summed over all atoms in the structure. These asp values are derived from partition functions of real atoms. $\Delta A$ is the change in solvent accessible surface area (sasa) for each atom between the exposure of the atom in the target structure and the average exposure of that type of atom in the unfolded protein structure (assuming an extended conformation). It is typically calculated by rolling a sphere with radius 1.4 Å (which is theoretically representative of a water molecule) over the entire surface of the atom in question. Raising or lowering the value of this radius leads to decreasing or increasing (respectively) the amount of surface area determined to be solvent-exposed. These are tabulated by residue as accessible surface areas in a Gly-
X-Gly tripeptide in an extended conformation, where X is the residue of interest.\textsuperscript{69,114}

The most buried positions in the target structure will have the smallest sasa values, whereas the most exposed positions will have the largest sasa values. Although simplistic, the solvation energetics calculation is as useful as more complex calculations for modeling the energetics in real systems. Only three solvation parameters are used – the burial of polar nitrogen and oxygen atoms, the burial of nonpolar carbon atoms and the exposure of the same nonpolar atoms. The first term relates to the fact that exposure of polar atoms allows them to be solvated and to interact in hydrogen bonding with solvent molecules, whereas burying them is energetically unfavorable, as they must become desolvated. However, this desolvation penalty can be offset by other favorable interactions such as hydrogen bonding. The second term refers to the strength of the hydrophobic effect, as mentioned previously. The third is a penalty for the forced solvation of hydrophobic atoms, and helps maintain the monomeric form of the target fold as well as the solubility of the protein.

Hydrogen bonding and salt bridges are important in specifying a target fold on the microscopic level and when dealing with buried hydrophilic atoms. Hydrogen bonding is considered present if the following has a value less than –0.3 and the distance between the hydrogen atom and the hydrogen bond acceptor atom is less than 2.5 Å:

\[ f(\theta, \phi) = \cos^2(\theta_{D,H,A}) \cdot \cos(\phi_{H,A,AA}) \]
where $D$, $H$, $A$ and $AA$ refer to the donor, hydrogen, acceptor and acceptor antecedent atoms, respectively, and $\theta$ and $\phi$ are the angles between the relevant atoms as below:

![Diagram of hydrogen bond with angles $\theta$ and $\phi$]

This equation takes into account the angular dependence of the hydrogen bond arising from the repulsive forces of the other atoms.

The weight of the hydrogen bond parameter is important; too much weight will result in the selection of too many hydrogen-bonding residues, and too little will result in a flexible protein with a poor $z$-score and energy gap (as defined in Chapter 1) between the N and U states. The hydrogen-bonding term utilized in this study was derived from the CHARMM force field method.\textsuperscript{115}

Electrostatic interactions, or “salt bridges”, were briefly discussed in Chapter 1 and can be described as two opposing charges being in close contact with each other. The interaction is calculated with Coulomb’s law, as follows:

$$E = \frac{q_i q_j}{D R_{i,j}}$$

where the $q$ values are the charges on each of the atoms $i$ and $j$, $R$ is the distance between the two atoms and $D$ is a scaling factor derived from the OPLS parameter describing the dielectric constant of the medium. The limitation of the electrostatic interaction calculation given above lies in the accurate determination of the value of the dielectric
constant of the medium. This value changes greatly depending upon the solvation of the amino acid in question, as well as the weight used to mirror the dielectric constant of the protein interior. The actual value of D for the protein core remains questionable. Additionally, more complex electrostatic representations would be necessary to increase the accuracy of capturing the multiple dielectric environment of a protein in solution.

4.8 Initialization of the Rotamer Roulette Wheel

The following sections describe the genetic algorithm utilized by the author, as developed by the principal investigator. Other groups may use a genetic algorithm that is different from the one described below.

Once steric filtering has taken place, a rotamer roulette wheel is configured. This involves developing a matrix of the same number of columns as the length of the protein. Each allowed amino acid (all but Cys, His and Met) in each available rotamer is placed on the first position, with Cβ centroids in every other position. The energy of the amino acid/rotamer identity in question with the rest of the ensemble is calculated through the use of the Lennard-Jones equation; here only the side chain-centroid and side chain-backbone nonbonded energy is determined. (Steric filtering has already taken place, so the subset of allowed rotamers of amino acids has been pruned.) This energy is used to obtain an estimate for the probability of that amino acid occurring in the lowest-energy model through the implementation of a Boltzmann-weighted probability distribution at a
temperature of 2000 K, as discussed in Chapter 1. This continues for all allowed amino acids in all rotamers, until the last possibility in this position is exhausted.

The rotamers are distributed into the roulette wheel according to their normalized Boltzmann probability as calculated above. The roulette wheel typically has 1000 slots; a rotamer with a normalized probability of 0.10 attains 100 slots, whereas another rotamer with a normalized probability of 0.25 receives 250 slots. The amino acid/rotamer combination of lowest (best) energy will therefore retain the highest number of slots. Each position is scanned, until the entire length of the protein has been considered.

The algorithm selects one amino acid/rotamer combination at random from the corresponding row in the rotamer roulette wheel for each position on the protein chain. It is obvious that the lowest-energy combination will have the highest probability of selection, as it occupies the largest number of available slots for each position on the chain. Additionally, if a variety of amino acids are determined to be of roughly equal probability to each other, the algorithm has the same rate of selection of these amino acids.

4.9 The Sequence Prediction Algorithm (SPA)

Once the rotamer roulette wheel has been initialized, the genetic algorithm can proceed. Sampling randomly from the roulette wheel at each position on the amino acid protein chain for the length of the protein creates one member; the initial population
consists of 300 of these members. The total energy of each of these members is calculated, which involves each part described in earlier sections of this chapter:

$$E = \sum_{\text{torsions}} \frac{V_e}{2} [1 + \cos(n \chi)] + \sum_i \sum_{j>i} 4\varepsilon \left[ \left( \frac{\sigma}{R_{i,j}} \right)^{12} - \left( \frac{\sigma}{R_{i,j}} \right)^6 \right] + \frac{q_i q_j}{D R_{i,j}} + \sum_i S_i \Delta A_i + \sum_{x=1}^{20} n_x B_x$$

This is also referred to as the scoring function. Once the scoring function of all 300 members is determined, the energy of each is utilized to determine their Boltzmann weighted probabilities. The temperature in this determination is set at every round according to a pre-defined diversity value. (This is set to decay linearly from 5.5 to 3.0 throughout the entire simulation, and results in a gradual decrease in temperature.) The Boltzmann weighted probabilities of the models are used to initialize another roulette wheel, this one comprised of the models instead of the amino acid/rotamer combinations. The algorithm selects for two parents, chains A and B, from the roulette wheel and performs a uniform crossover scheme. This involves flipping an unweighted coin at each position on the protein chain and selecting for the amino acid and rotamer combination from chain A or from chain B, depending on the coin flip results. As should be obvious, the initial stages of design are done at high temperature, which allows for the survival of models of high energy. As the simulation continues, however, the decreasing temperature tends to favor models of lower energy, thus annealing on a low-energy sequence for the target structure.
In addition to the uniform crossover described above, random mutation at a frequency of 0.04 is allowed in the final stages of each round. Mutation is accomplished by selecting for another amino acid or rotamer from the roulette wheel at a random position in the sequence. The energy of this new model is determined and used as one of the models in the set of 300.

This cycle of energy determination, recombination and random mutation is repeated 200 times (although this number can change, if desired). Additionally, SPA can be run several times over using a tabula rasa for each separate run. The resulting sequences from these runs can be used as input to seed a final simulation. This is done so as to avoid weighting one specific solution more than is appropriate.\textsuperscript{111}

4.10 Application of the Monte Carlo Algorithm in SPANS

The main difference between SPA and its descendent, Sequence Prediction Algorithm on Numerous States (SPANS), is the use of a backbone ensemble, or a set of backbone structures that have an rmsd of less than 0.3 Å to the target structure. The backbone ensemble is useful for two reasons. As mentioned before, the use of steric filtering is often too rigid and can allow for the dismissal of a useful sequence owing to a real but not significant steric constraint. Also, most native proteins present some extent of backbone thermal fluctuation – the deviations from average are usually not large enough to cause poor cooperativity or loss of microscopic selectivity (except for the “natively disordered” proteins described in Chapter 1), but may result in fraying of the
ends and loops of structures. It is thus useful to utilize a backbone ensemble, which both eliminates most minimal steric constraints and mimics a realistic flexible backbone. Other groups have considered the usefulness of treating the backbone as a parametric family of structures rather than a static entity.\textsuperscript{116,117}

The backbone coordinates are derived directly from a high-resolution structure obtained from either NMR or x-ray crystallography characterization. An ensemble of 30 backbone structures is generated using what is termed a Monte Carlo expansion and contraction algorithm. This involves initial random perturbation of $\phi$ and $\psi$ backbone angles at all positions in the structure ($\pm 5^\circ$) to alter the backbone structure, followed by refinement of the angles until the backbone reaches a preset maximal rmsd value relative to the reference backbone. The refinement procedure involves a perturbation of a single, randomly chosen backbone angle by $\pm 1^\circ$. If the rmsd decreases, the move is accepted; if the rmsd increases, the move is rejected. The maximal rmsd value in these designs is set to 0.30 Å; as soon as the total rmsd reaches this cutoff, the refinement procedure is stopped for that backbone model and the procedure is begun again with the wild-type backbone to obtain a different perturbed backbone model. This generates 30 backbones, all related to the initial experimentally determined backbone structure by less than 0.30 Å. Since all of the backbone structures in the ensemble are based on random perturbations of the backbone angles, however, they are nondegenerate. Figure 4.2 depicts a typical backbone ensemble.
Figure 4.2. Backbone ensemble utilized in SPANS. All backbones are less than 0.3 Å to the wild-type backbone and are developed through Monte Carlo moves of the \((\phi, \psi)\) wild-type backbone angles.
4.11 Free Energy Matrix for Amino Acid Probability Generations (SPANS)

Once the backbone ensemble has been developed, SPA is initialized on each of the 30 backbones, using the energy between every allowed amino acid in every rotamer and each backbone in the ensemble for an ESB-based rotamer roulette wheel as previously described. This ultimately results in a defined amino acid/rotamer combination for each position on each structure. An initial GA gives a sequence of reasonable energy for further calculation purposes. This sequence is used to estimate the context-dependent fitness of every allowed amino acid and its possible rotamers at every position on the protein backbone under consideration. To arrive at this estimate, the side chain identities and rotamers at all other positions not immediately under inspection are frozen to the sequence selected in the first GA. The rotamers of all allowed amino acid types are sampled exhaustively for the position of interest, and the full energy of each corresponding model is evaluated as described previously. The Boltzmann weight of each sampled side chain and rotamer combination is added to an ongoing partition function as follows. In the earlier design procedure (used for SPANS-WW1), multiple “sub-rotamer” states are stochastically sampled about each canonical rotamer (typically, 15 sub-rotamer states within 20 degrees of the central rotamer state are sampled randomly), and the Boltzmann probabilities of all of these states are combined into one super-rotamer and included in the partition function. In an alternative method (used for SPANS-WW2), each canonical rotamer is optimized by a torsion-space steepest descent minimization; this involves trying to move each of the side chain dihedral angles
separately in each direction by 5° and evaluating the effect on energy. This is done for all dihedral angles, and all the ones that lead to a decrease in energy are combined at the end.

A set of partition functions \( \{Q_{x,r,i}\} \) for all amino acid rotamers at all positions in the protein defines what is termed a probabilistic fitness matrix:

\[
Q_{x,r,i} = \sum_{m=1}^{N} \sum_s e^{\frac{-\Delta E_{x,r,s,i,m}}{RT}}
\]

where \( x \) is the amino acid type, \( s \) is a sub-rotamer state of rotamer \( r \) of amino acid \( x \), \( i \) is the position in the structure, \( m \) is the model under evaluation, and \( N \) is the total number of models used. \( E_{x,r,s,i,m} \) is the total calculated energy, according to the SPA scoring function as previously described, of the specific model under evaluation, given the current sub-rotamer of amino acid type \( x \) at position \( i \).

The partition function for each amino acid and rotamer combination is continually updated as each of the 30 backbone structures and models under evaluation are added to the simulation. Because each model contains a unique configuration of backbone structure, side chain identities, and rotamers, each rotamer state is exposed to a range of environments.

To ensure self-consistency in the final probability matrix and free energy values, three cycles of design and sampling are performed over the ensemble of 30 backbones. The \( Q_{x,r,i} \) values from each design and sampling cycle, representing the cumulative probability of each rotamer state, are used to guide the next cycle of design simulations.
by serving as a probabilistic selection matrix for amino acids and rotamers. Although the $Q_{x,r,i}$ values represent probabilities of all amino acid rotamer states (Figure 4.3), potentially useful for design of protein libraries, the single sequences designed herein are based directly on the highest probability amino acid at each position.

The design of a third WW protein (SPANS-WW3) involved the use of molecular dynamics (using the AMBER package) with an rmsd restraint on the backbone atoms and a force constant of 5 kcal/mol. This gave a backbone ensemble parallel to that resulting from the numerous states MC described for SPANS. The backbone ensemble was utilized as input into the SPA algorithm, as previously described.

As a final note, utilizing a backbone library such as one used in solving a crystal structure would lead to the design of a sequence that would have a low calculated energy on a variety of backbone conformations. As it is difficult enough to select a sequence that can adopt a single backbone conformation, this method would usually lead to the design of a protein that has a molten structure due to the low energy barriers between somewhat related conformations. This backbone library, while being powerful for structure determination, would probably not be ideal for protein design.
Figure 4.3. The sequence of lowest calculated free energy on each of the 30 backbones in the ensemble result in a probability matrix. The thickness of the amino acids directly correlates to their free energies and probabilities of occurrence in the final sequence. This figure is the result of SPANS on the wild-type WW, yielding SPANS-WW1.
CHAPTER 5

COMPUTATIONAL ALGORITHMS FOR PROTEIN DESIGN

De Novo Three-Helix Bundle Design

5.1 De Novo Design of Three-Helix Bundles – Introduction

Previous attempts at helix-bundle design have involved both the utilization of a pre-existing protein backbone\textsuperscript{118} and the \textit{de novo} design of a novel backbone.\textsuperscript{65} Other work, involving point mutation of key residues, has manipulated a coiled-coil structure to that of both parallel and anti-parallel three-stranded coiled-coils.\textsuperscript{119} Most helix-bundle designs have started out as coiled-coil-based design, with no loops or tethering units to connect the coils. Either a trimeric state is specifically engineered through the use of packing, salt-bridging and hydrogen-bonding interactions, or an alternative multimeric state is targeted but the trimeric state is the most highly populated. Other design attempts of interest have involved the manipulation of loop lengths and amino acid compositions to direct the topology of the bundles as well as their stability.\textsuperscript{57} There have been quite a few attempts at redesigning three-, four- and five-helix bundle proteins, but when this work was initiated, none had been targeted specifically at the \textit{de novo} design of a three-helix bundle protein.\textsuperscript{6,25-28,119-121}
There were two main branches of the three-helix bundle design project. One of these involved the use of SPA to redesign the Z-domain of protein A from *Staphylococcus aureus* (SpZ). This protein is a double mutant (A1V, G29A) of the IgG-binding domain B of protein A, one of the fastest-folding native proteins\textsuperscript{122}, and is responsible for binding to immunoglobulin (IgG) class I and its Fc fragment. SpZ is a three-helix bundle with up-down-up topology (PDB: 2SPZ, Figure 5.1). It is comprised of 58 amino acids; its sequence is given below, with the helical portions underlined:

VDNKFNKEQQNAFYIELHLPNLNEEQRNAFIQSLKDDPSQSANLLAEAKKLNDQAQPK

It has two connecting loops of 6 and 4 residues long, and its helices are 9, 11 and 14 residues in length; the other 14 residues make up the N and C termini. The method of this design and the results will be discussed in Chapter 7.

The other branch of three-helix bundle design was the *de novo* design of a bundle having the up-down-up topology as previously described. Native three-helix bundle proteins have a typical backbone structure that is easy to duplicate on the computer. α-Helices with 3.6 residues per turn typically have a φ angle of about $-60.0^\circ$, a ψ angle of about $-40.0^\circ$ and an Ω angle of $-180.0^\circ$. Wild-type helices range in length from five to thirty-one amino acids, with most helices having fewer than fifteen residues.\textsuperscript{123} These backbone angles and helix lengths can be used to develop a native-like three-helix bundle from first principles. In order to mimic the backbone structure of native three-helix bundles, one helix of twelve residues (initially consisting only of backbone heavy atoms)
Figure 5.1. High-resolution NMR structure of the Z-domain of protein A from *Staphylococcus aureus* (SpZ, PDB ID 2SPZ).\textsuperscript{124} Note the up-down-up topology of the helices.
was built using standard AMBER geometries. This helix was translated so that its helix axis overlaid the z-axis, and its center of mass was located on the origin. The helix was triplicated, which resulted in three identical helices that completely overlaid each other. To allow for more natural helix geometries, each helix was independently rotated around the z-axis (to twist each helix around its own axis) as well as around the x- and y-axes; they were also translated along the x- and y-axes and then along the z-axis, followed by a final rotation of each helix around the z-axis to form an ideal three-helix bundle. The second helix was rotated $180^\circ$ for an up-down-up topology. Loops were built to tether the helices together and increase the probability of maintaining the up-down-up target topology. Small moves in the backbone angles were randomly executed to decrease the rigidity of the helices, and full sequence design was initiated.

An algorithm for the \textit{de novo} design of a rigid (no bends or twists in the helices) three-helix bundle was written for this study by the author to implement the ideal backbone properties described above. A perfect $\alpha$-helix of twelve residues in length was built in a Cartesian coordinate-free space with this algorithm using standard atom geometries; at this point, only Ala was allowed. The helix axis was determined$^{125}$ by averaging the coordinates of the backbone heavy atoms of the first four residues to define point 1, and doing the same for the backbone heavy atoms of the last four residues to define point 2. The vector connecting the two points was labeled the helix axis. To move the helix axis vector so that it overlaid the z-axis, the vector’s projection on the xy-plane was determined (as $\Delta x$ and $\Delta y$), along with the angles this vector makes with the y- and z-axes. This vector was rotated about the y- and z-axes using these calculated angles.
until it overlaid the z-axis. The center of mass of the helix (which was still comprised only of backbone heavy atoms) was translated so that it was located on the origin (0,0,0). This helix was duplicated to result in three helices that completely overlay each other; the helices were labeled 1-3. Each helix was rotated by ±50° around the z-axis using a standard rotation matrix as given below. This operation was termed the first rotation around the z-axis. The helices were rotated by less than ±1° about the x-axis and less than ±7° about the y-axis to yield helices that are not completely parallel to each other; helix 2 was rotated by 180° around its center of mass to invert its axis for an up-down-up topology. Each helix was translated by small amounts in the x and z directions (less than ±1 Å) and moved away from the origin by adding less than 7 Å to the y-axis coordinates. Figure 5.2 is a pictorial representation of all of these helix manipulations.

The matrices used for rotations by θ around the x, y and z-axes when the helices are located on the z-axis with their center of mass on the origin were as follows:

For rotations around the z-axis:

\[
\begin{bmatrix}
\cos \theta & \sin \theta & 0 \\
-\sin \theta & \cos \theta & 0 \\
0 & 0 & 1
\end{bmatrix}
\]

For rotations around the x-axis:

\[
\begin{bmatrix}
1 & 0 & 0 \\
0 & \cos \theta & \sin \theta \\
0 & -\sin \theta & \cos \theta
\end{bmatrix}
\]
Start with 12-residue long helix, all A
Ψ = -60.0
Φ = -40.0
Ω = 180.0

Define helix axis and center of mass
Use first and last 4 residues of helix
Move this axis to the z-axis
Move COM of whole helix to origin
Make three helices

Rotate helices around their own axes (z-axis)
-50.0° to 50.0°

Rotate helices around x, y-axis
x = -0.5° to 0.5°
y = -10.5° to 10.5°

Figure 5.2. Helix manipulations. Continued on next page.
Move each helix a random amount in x, y and z directions
\[ x = 0.0 \]
\[ y = 5.75 \text{ to } 6.85 \]
\[ z = -0.85 \text{ to } 0.85 \]
Flip 2\textsuperscript{nd} helix 180°

Rotate helices around z-axis:
\[ \text{helix} \#1 = -2.5° \text{ to } 2.5° \]
\[ \text{helix} \#2 = 117.5° \text{ to } 122.5° \]
\[ \text{helix} \#3 = 237.5° \text{ to } 242.5° \]

Figure 5.2 continued. Helix manipulations for \textit{de novo} helix bundle design. The backbone is used as input into a fixed backbone GA as described in the text.
For rotations around the y-axis:

\[
\begin{bmatrix}
\cos \theta & 0 & -\sin \theta \\
0 & 1 & 0 \\
\sin \theta & 0 & \cos \theta \\
\end{bmatrix}
\]

Each helix was rotated around the z-axis as follows: helix 1 was not rotated at all (360°); helix 2 was rotated 120°; helix 3 was rotated 240° using the following matrix manipulations:

\[
X = \begin{bmatrix}
\Delta x & (\Delta y^2 + \Delta z^2)\beta & 0 \\
\Delta y & -\Delta y\Delta x\beta & \Delta z\gamma \\
\Delta z & -\Delta z\Delta x\beta & -\Delta y\gamma \\
\end{bmatrix}
\]

\[
R = X \begin{bmatrix}
1 & 0 & 0 \\
0 & \cos \theta & \sin \theta \\
0 & -\sin \theta & \cos \theta \\
\end{bmatrix} X' \quad \text{(the middle matrix is } Y)\]

\[
XY = Z \\
ZX' = R
\]

where

\[
\beta = \frac{1}{\sqrt{(\Delta y^2 + \Delta z^2)^2 + (\Delta y\Delta x)^2 + (\Delta z\Delta x)^2}}
\]

and
\[ \gamma = \frac{1}{\sqrt{\Delta y^2 + \Delta z^2}} \]

Matrix multiplication of the old coordinates (as given by x, y and z) resulted in the new rotated coordinates.

The final z-axis rotations were considered the second rotation around the z-axis and were slightly altered to bracket the previously described rotations by \(-2.5\) to \(+2.5^\circ\), resulting in the final rotations around the z-axis as given in Table 5.1. The product is a wild-type right-handed three-helix bundle backbone. Matrices of values sampled between the extremes given above (the two z-axis rotations, the x- and y-axis rotations, the x, y and z-axis translations) were implemented in a GA, allowing for discrete movements within the rotation and translation spaces to explore interhelical topologies followed by full sequence design. However, only Ala, Phe, Ile, Leu and Val were allowed at this stage for computational tractability and packing purposes. The interhelical angle and translation parameters were permitted only within certain values, as listed in Table 5.1. The main thrusts of this part of the design procedure were the developments of a well-packed core and the proper interhelical topology; the algorithm and modules are given in Appendix B.

An alternative program utilized a Monte Carlo algorithm with Metropolis acceptance criteria tethered to this GA to generate small, random changes in the helix geometry parameters. Here, all of the distances were altered by \( \pm 0.5 \), and all of the
<table>
<thead>
<tr>
<th></th>
<th>Helix 1</th>
<th>Helix 2</th>
<th>Helix 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Z-angle</td>
<td>-50° to +50°</td>
<td>-50° to +50°</td>
<td>-50° to +50°</td>
</tr>
<tr>
<td>X-angle</td>
<td>-0.5° to +0.5°</td>
<td>-0.5° to +0.5°</td>
<td>-0.5° to +0.5°</td>
</tr>
<tr>
<td>Y-angle</td>
<td>-5.5° to +5.5°</td>
<td>-5.5° to +5.5°</td>
<td>-5.5° to +5.5°</td>
</tr>
<tr>
<td></td>
<td></td>
<td>+180.0° for up-down-up</td>
<td></td>
</tr>
<tr>
<td>X-translation</td>
<td>0.0 Å</td>
<td>0.0 Å</td>
<td>0.0 Å</td>
</tr>
<tr>
<td>Y-translation</td>
<td>5.25 to 6.55 Å</td>
<td>5.25 to 6.55 Å</td>
<td>5.25 to 6.55 Å</td>
</tr>
<tr>
<td>Z-translation</td>
<td>-0.85 to +0.85 Å</td>
<td>-0.85 to +0.85 Å</td>
<td>-0.85 to +0.85 Å</td>
</tr>
<tr>
<td>Second Z-angle</td>
<td>-2.5 to +2.5 Å</td>
<td>117.5 to 122.5 Å</td>
<td>237.5 to 242.5 Å</td>
</tr>
</tbody>
</table>

Table 5.1. Helical angle and translation parameters.
angles were altered by ±2.5°. Whether or not the moves were accepted was based upon the Metropolis criterion. The lowest-energy structure was used as input for a set of loop design algorithms.

5.2 Design of Loops to Connect Three Helices

Loop design involved the addition of short strings of residues connecting the three helices together in an up-down-up topology. The tethers were three to five residues long to eliminate competing topology. A (ϕ,ψ) angle grid derived from a typical Ramachandran plot (-180° to +180° for both angles) was used as input for angle generation in the loop regions. Once the angle grid had been derived, energy and fit (or maximum rmsd between the C-terminus of the prior helix and the N-terminus of the loop, along with the C-terminus of the loop and the N-terminus of the following helix) values were supplied to eliminate any loops that were of high energy or poor fit. (The highest energy cutoff was set at 35.0 arbitrary units; the fit was 0.3.) The loops, which consisted of Ala residues at this point, were built onto the three helices. The energy (which only consisted of the backbone-backbone van der Waals and backbone torsional energies) was determined; no criterion for acceptance was maintained. The lowest-energy and most native-like loop models were selected and included as part of the three-helix bundle. This was done for both loops. Once a rough idea of ideal backbone angles was determined, a Ramachandran range program was applied, which gave a range around these ideal angles of ±10° in 5° increments. These ranges were used as input into a new
loop generation program, which focused on the ideal angles for loop building. In this context, the residue in each helix bracketing the loop in question was included in the loop for fit and energy determination purposes. Once low-energy loop structures had been determined, the entire file was written out as one backbone.

5.3 Backbone Flexibility

Up to this stage, the helices remained rigid. Most wild-type helix bundles, however, have a slight amount of twisting of helices evident when viewed from the top. This assists in burial of hydrophobic residues as well as optimizing the Lennard-Jones contact energy. To assist in the mirroring in silico of wild-type situations, the backbone angles of the lowest energy model after loop design were allowed to deviate slightly from the set values given above. A standard GA was run to obtain a lowest-energy sequence for the target structure. The backbone and side chain angles were moved by a small, random amount – the side chains had a $-10^\circ$ to $+10^\circ$ fluctuation allowed around the statistically derived rotamer states, and the backbone angles were adjusted by $\{\phi = -2^\circ$ to $+2^\circ; \psi = -2^\circ$ to $+2^\circ; \Omega = -1^\circ$ to $+1^\circ\}$ to the previously mentioned “perfect” backbone angles. The new model was accepted if its energy (calculated by summing the Lennard-Jones energies between the side chain atoms and other side chain atoms, backbone atoms and other backbone atoms, side chain atoms and backbone atoms, and torsion angles of the side chains and backbone) was lower than the previously determined energy or if the Metropolis criterion is fulfilled. This cycle of GA to develop a sequence for the current
structure and MC to generate a flexible backbone was repeated a set number (typically 200) of times.

After random fluctuations in the backbone angles were finished, the helices had deviated slightly from their previous, well-packed positions. To circumvent this problem, the helices were repacked somewhat by finding an rmsd between the original, rigid helix design \( C_{\alpha} \) and the new, flexible helix design \( C_{\alpha} \). This involved using a Metropolis Monte Carlo algorithm to adjust backbone angles very slightly until the rmsd between the two models approached zero. This maintained the flexible backbone of the three-helix bundle, but with better-packed helices.

Once the backbone structure had been determined with hydrophobes only, the backbone coordinates were used as input into a GA much like the one described in Chapter 4 for full sequence design. The result was a sequence that should adopt a right-handed three-helix bundle with up-down-up topology.

5.4 Negative Design

A significant problem in protein design involves macroscopic and microscopic selectivity. In this portion of the work, the goal was to design a protein that adopted a three-helix bundle topology. The macroscopic selectivity issue involves either using HP patterning modeled on native three-helix bundle proteins or, in the case of true de novo design, the use of minimized helix geometries that should lead to the \textit{in vitro} expression of a three-helix bundle topology. Microscopic selectivity, or small-scale perturbations in
the designed structure when experimentally characterized, is more difficult to avoid. It was hoped that the consideration of helix packing by using hydrophobic residues only in the initial parameter determination stage would assist in selecting and cementing the target structure. This, along with hydrogen-bonding and salt-bridging interactions, should have selected for a target structure that exhibited a significant gap and z-score to the next highest energy structure.

To address the problem of microscopic selectivity, an initial MC/GA combination (to develop three rigid untethered helices) was run to select the lowest-energy helix parameters out of a set of 1000 total helix parameters (as shown in Figure 5.3). This utilized only Ala, Val, Ile, Leu and Phe for packing purposes. The parameters were determined through the MC/GA as discussed previously and the sequence that was designed for this helix bundle was stored. A set of 199 alternative structures, or decoys, consisting of small random fluctuations in the interhelical parameters and having rmsd less than 2.5 Å with the low-energy helix bundle, was developed and the energy of the stored sequence on each decoy was determined, as shown in Figure 5.4. The energy gap and z-score between the low-energy helix bundle from the first MC/GA run and the next-lowest energy decoy structure, both with this stored sequence, were determined. (The low-energy helix bundle from the first MC/GA run was confirmed to have remained the ground-state structure after the decoy runs.) If the gap and z-score were significantly high, and if all of the parameters had been accurately described and weighted, it followed that the ground-state structure should indeed have theoretically been the native structure when experimentally characterized. If this as not the case, additional runs of the MC/GA
Figure 5.3. The results from an initial MC/GA combination (to develop three rigid untethered helices); a set of 1000 total helix parameters yield 1000 structures. Only one is the lowest-energy structure.

Figure 5.4. The energy of the stored sequence on each of the 199 decoys was determined as described.
and better descriptions and weights on the parameters for energy calculations were executed.

5.5 Structural Specificity and Sequence Diversity

It has long been assumed that increasing the sequence diversity, especially in the core, should lead to structural specificity. The sequence, if highly diverse, should only have one ground-state structure that adequately fulfills the packing stringency (as discussed in Chapter 1) more completely than any competing structure. If the sequence is highly degenerate, it should have a variety of low-energy structures, resulting in a “molten” native state that is poorly cooperative. This is especially true of the core residues, which are typically very highly packed in a complimentary manner. Thus, the goal of protein design has usually been one of developing a highly specific sequence/structure relationship.

The three-helix bundle de novo design described above was used to probe this idea. If it was determined that core sequence diversity did indeed increase structural specificity, then the use of a parameter involving sequence diversity and the appropriate weight would have been necessary before further design efforts were realized.

The approach was as follows: the target structure from the first MC/GA run was stripped of its side chains, and a “full” sequence design (again utilizing A, V, I, L and F) on this fixed target backbone was initiated. The only change involved the application of a sequence diversity weight of the core residues to the energy determination. This weight
(termed the \textit{ew}) was increased from a value of 2.0 to 5.0, 10.0, 25.0, 50.0 and 100.0 and affected the sequence diversity as follows:

\[
S = ew \times \left( -\sum_i \left( \frac{n_i}{\text{length}} \right) \ln \left( \frac{n_i}{\text{length}} \right) \right)
\]

where \(i\) is each allowed residue type (A, V, I, L or F) and \(n_i\) is the number of times this residue occurs in the sequence. Increasing the \textit{ew} served to force a higher sequence diversity; full sequence diversity yielded a 20% occurrence of each allowed amino acid.

The core residues were determined by making C\(\beta\) centroids for each residue and determining the Micheletti contact score as described in Chapter 4. If the contact score is greater than a supplied contact cutoff, the residue was considered buried. This contact cutoff was slowly increased from 1 to 5 until the buried positions as determined visually were selected by the algorithm.

All of these \textit{ew} design runs were done on the target backbone from the initial MC/GA cycle, shown in red in Figures 5.3 and 5.4. Once the lowest-energy sequence obtained for each \textit{ew} run was extracted (for a total of six sequences), a decoy-target run was performed. This involved taking the target (red) state as “native” and utilizing the original set of 199 decoy structures (in blue, in Figure 5.4) that had small random changes in interhelical geometries from the native structure. The sequence determined from the \textit{ew} runs was the only sequence allowed. The lowest-energy structure was extracted. If this structure was still the target (red) structure, the gap and the z-score of
this target structure in relation to the rest of the decoys was determined. This served as a means for studying whether or not sequence diversity does indeed lead to structural specificity. A flowchart for these design approaches is shown in Figure 5.5.
Figure 5.5. Flowchart for the sequence diversity/structural specificity calculations.
CHAPTER 6

APPLICATIONS OF SPA/SPANS

WW Domain Design

Materials and Methods

6.1 Introduction to the WW Domain

WW domains are protein-protein interaction modules that are named for the two highly conserved Trp residues, one of which is functional and the other of which constitutes the majority of the hydrophobic core, separated in sequence by 20-22 amino acids. WW domains are typically between 34-40 amino acids long, which makes them one of the smallest naturally occurring monomeric protein domains, and fold without the addition of ligands, cofactors or the formation of disulfide bonds. Their secondary structure is mostly $\beta$-sheet, and they have a characteristic block of two or three sequential aromatic amino acids located between the two Trp residues. A conserved Pro is located three amino acids after the second conserved Trp. Several signaling complexes that the WW domain mediates have been discovered in muscular dystrophy, Alzheimer’s disease, Huntington’s disease and cancer. This particular system has been found in plants, yeast, worms, flies, and vertebrates, and occurs in cytoplasmic and nuclear proteins.\textsuperscript{127,128}
In recent studies, the WW domain has been divided into four categories. The first three groups recognize short proline-rich motifs, and the fourth recognizes phosphoserine (pSer) or phosphothreonine (pThr)-proline motifs. As such, the WW domain is typically considered an adaptor protein for its associated protein. The WW domains that bind poly-Pro regions typically utilize a flat binding surface and have a $K_D$ between 1-50 μM. These modules have been identified in a number of diverse proteins, including dystrophin and the human Yes-associated protein (hYap). Multiple copies of WW domains have been found in some large protein complexes; this is thought to increase the affinity of the protein-protein interaction and allow for more specific targeting. Studies on the WW domain from hYap have been published, including steady-state and folding kinetics and stability of the domain and some key point mutations; less is known about the biophysical properties of other WW domains. Additionally, a WW prototype has been experimentally designed by focusing on the network of highly conserved long-range interactions in the structure, as well as conservation of residues in the WW family.

The WW domain from human peptidyl-prolyl cis-trans isomerase (hPin1) is a characteristic WW domain comprised of 34 amino acids forming three antiparallel $\beta$ sheets. It has one solvent-exposed Trp (W29) that plays a functional role and one buried Trp (W6) participating in a hydrophobic core located within the “head” of the protein. It also has a block of aromatic amino acids situated along the central part of the protein, as well as a Pro three residues after W29 (P32). The exact physiological role of hPin1 remains unclear, although it is present in proteins responsible for signaling, regulatory
and cytoskeletal functions. The WW domain from hPin1 is responsible for targeting pSer- and pThr-proline motifs, probably in a PPII helical conformation. Additionally, the Oschkinat group has found that the WW domain from human Yes-associated protein 65 (hYap65) was only expressible as a fusion system. We have chosen the WW domain from hPin1 to test the ability of our SPA and SPANS design algorithms to redesign a β-sheet protein. As mentioned previously, most design attempts have focused on proteins that are entirely or largely α-helical. The WW domain is a β-sheet protein, and is computationally more difficult to redesign because of the preponderance of tertiary interactions as discussed in Chapter 1. This fact, along with its small size, makes it a good test case for computational redesign using the algorithms presented in Chapter 4. If the protein does indeed adopt the intended conformation, it will be one of the first successful β-sheet redesign efforts.

The backbone coordinates of the WW domain in complex with human Pin1 were extracted from the high-resolution x-ray crystallography coordinate file (1PIN, Figure 6.1). All amino acids (besides Met, His and Cys) were allowed as input into either SPA or SPANS; secondary structure propensity was not implemented at this time. Atomic solvation parameters were randomly altered until the structures had a native-like balance of hydrophobic and hydrophilic residues on the exterior of the protein when examined visually. Five separate runs of the SPA algorithms were performed with no carryover of results from one to the other, and the sequence outputs of each of those five trials were
Figure 6.1. High-resolution x-ray crystallography structure (to 1.35 Å) of the WW domain found in complex with hPin1 (PDB ID code 1PIN). The colors indicate the degree of homology with the WW family; the redder the amino acid, the more highly it is conserved. Note the two Trp residues; Trp 6 is buried in the hydrophobic core, and Trp 29 is solvent-exposed.
used as input into a final run. If the final structure was well-packed and had a good hydrogen-bonding network with no charge-charge repulsion, the protein was expressed and characterized.

6.2 The Vector – N-Cam.Y and NIa Fusion System

Unless otherwise mentioned, all DNA oligomers were purchased from Invitrogen Life Technologies (Carlsbad, CA) at a 50 nmol scale with standard desalting purification, all enzymes and enzymatic buffers were purchased from New England BioLabs (Beverly, MA), all competent cells were purchased as a glycerol stock from Novagen, except for DH5α competent cells, which were originally purchased from Invitrogen Life Technologies as a glycerol stock, and all chemicals were purchased from J.T. Baker (Philipsburg, NJ). All pH determinations in this and other chapters were performed on a Corning pH meter (model 430, Corning, NY) with a standard glass-tipped electrode (Corning 3” in 1” with replaceable junction), and adjusted with HCl (EM Science, Cincinnati, OH) and NaOH unless otherwise mentioned. All molecular weights were calculated with PAWS (Proteometrics, Inc., Rockefeller University), and pI calculations were performed with ProtParam (http://us.expasy.org/tools/protparam.html). All protocols, unless otherwise mentioned, were derived from standard protocols.104,108

Prior to development of the gene coding for the protein of interest, a fusion system involving the N-terminus of a D78Y mutant of calmodulin from Homo sapiens (N-Cam.Y) plus a C-terminal linker for proteolytic purposes (sequence ENLYFQ/GS,
where / indicates the site of protease cleavage by the mutant (His)$_6$-tagged Nla-Pro tobacco etch virus protease$^{136}$ was developed. This fusion system was chosen because of the high expression of N-Cam.Y in *E. coli*, as well as the ease of purification of N-Cam systems *via* Ca$^{2+}$-dependent binding to phenyl sepharose (as described in Chapter 2). The Nla cleavage system cleaves effectively to ~95% completion in most non-SDS buffers overnight. Further purification involved either HPLC or additional phenyl sepharose column chromatography to sequester the fragments containing the N-Cam.Y fusion system from the cleaved WW proteins. The result is a large amount of pure, soluble WW protein with two additional residues (Gly-Ser) at the N-terminus (Figure 6.2).

The gene coding for N-Cam.Y had been constructed previously, and another project (discussed in Appendix A) had developed a gene that coded for N-Cam.Y with no stop codons after the gene but with a *HindIII* site cloned between *ClaI* (within the N-Cam.Y gene) and *BamHI* (which occurred at the end of the N-Cam.Y gene). Two stop codons were located after the *BamHI* site (Figure 6.3). This construct is called the old N-Cam.Y linker.

The Nla linker, which was cloned into the *HindIII/BamHI* sites of the old N-Cam.Y linker, had *HindIII*-Nla gene-*BamHI* identity. The DNA oligomer coding for the Nla linker was designed with the protein-to-DNA tool from the Entelechon GmbH website. The typical codon preference in *E. coli* BL21 was exploited to select for the frequency of codon usage. The protein sequence was used as input into the Entelechon algorithm and the output was examined for a high number of small repeats (less than 5 bases more than twice in a row) or long stretches (greater than 6 bases) of one type of
Figure 6.2. Representative SDS-PAGE protein gels of the scheme utilized to purify WW designs. On the top and bottom figure, the red ellipses indicate presence of a WW protein; the red ellipse on the middle figure indicates the presence of the N-Cam Y/Nla fusion protein (top band) and a WW protein (bottom band).
Figure 6.3. The gene developed for expression of WW proteins as a fusion system with the N-terminus of a calmodulin mutant. The italicized restriction enzymes are located directly beneath vertical lines indicating their relative location of digestion activity. The pale yellow box represents the gene coding for N-Cam.Y; note the internal *HindIII* site. The gene coding for NIa is located in the region labeled NIa. The *BamHI* site located prior to the WW gene but after the NIa linker was maintained after cloning as an endonuclease site. The “*BamHI*” line designates the altered *BamHI* site, which was no longer cleavable by *BamHI*. This fusion system was located within pAED4.
base. If either of these two characteristics were discovered, the area of the sequence with the characteristic was altered to select for another preferential codon. Once the DNA sequence was designed, an online complementary sequence program available through Harvard Medical School was used to determine the complementary sequence. The final result was a double-stranded piece of DNA that coded for the protein of interest. The *HindIII* and *BamHI* restriction enzyme sites (cohesive ends only) were added to the N- and C-termini (respectively) of the linker.

The two oligos (top and bottom strands) were purified by a PAGE method as follows. A total of 24 mL of a 30% acrylamide/0.8% N,N’-methylenebisacrylamide solution was added to 29 g of urea (VWR, Bridgeport, NJ) and 6 mL of 10x TBE (Tris base, boric acid and EDTA, pH 8) in a 125-mL Erlenmeyer flask. This solution was heated gently under running hot water or with short bursts of heat in a standard microwave oven until the urea became completely soluble. The solution was cooled to ~30 °C under running water and filled to a volume of 60 mL with distilled water. 300 µL of freshly prepared 10% ammonium persulfate (APS) in TE and 30 µL of N,N,N’,N’-tetramethylethylenediamine (TEMED) were added and the mixture was gently swirled. This solution was poured between two prepared, NaOH-cleaned, clamped glass plates. A comb was added and the gel was allowed to polymerize at a ~15° angle until set. During this time, the samples were centrifuged to ensure that the freeze-dried DNA pellet was at the bottom of the tube and resuspended in enough 1x TE (10 mM Tris-HCl, 1 mM EDTA, pH 8) to result in ~1nmol/µL of DNA. 5 µL of a solution of dye (25% glycerol, 25% distilled water, 50% TE, pH 8.0, 0.5% w/v bromophenol blue [International
Biotechnologies, Inc., New Haven, CT]) was added to 25 µL of the resuspended oligo solution in a clean Eppendorf tube. The plate apparatus was washed thoroughly to remove any acrylamide overflow, the comb was gently removed and the gel running apparatus was set up with 1x TBE in the top and bottom reservoirs. The gel wells were washed thoroughly with buffer to remove any urea that had eluted from the gel. The samples were loaded onto the gel and run at a voltage of 250 V until the blue dye was ~¾ through the gel. The gel running apparatus was stopped, the plates removed and pried apart, and the gel itself was placed on saran wrap on top of a plate coated with fluorescently active material. The gel was viewed under long-wave UV light and the top ⅔ of each DNA oligo (visible under UV as dark smudges or triangles, and slightly yellow to the naked eye) were excised with a razor blade. Only the top ⅔ was excised, as the portion of the DNA sample that had the highest probability of not having any single- or multiple-base deletions due to incorrect synthesis tended to run faster through the gel matrix. The gel slices were placed in individual labeled Eppendorf tubes and thoroughly crushed with a pipet tip. 500 µL of TE was added and the tubes were placed on a 37 °C heat block overnight.

After ~15 hours at 37 °C, the samples were centrifuged at 14,000 RPM, the supernatant was removed and the acrylamide fragments were discarded. This was repeated twice to eliminate the presence of acrylamide from the supernatant. An ethanol precipitation was performed by adding 2.5 times the volume of liquid of ice-cold ethanol (EtOH, 200 proof, Pharmco, obtained from the Pennsylvania State University Chemical Stockroom, Whitmore Laboratory) and 0.1 times the volume of liquid of 3 M sodium
acetate (NaOAc, pH 5.2), along with 1 µL of tRNA (Invitrogen Life Technologies) as a template for seeding the precipitation reaction, and inverting the tube several times. This reaction mixture was placed on ice for ten minutes and centrifuged at high speed for another ten minutes, resulting in a pellet of DNA. The liquid was removed and the pellets were air-dried to allow any remaining liquid to evaporate, in particular EtOH, which could have inhibited future enzymatic reactions. Once the pellets were dry, they were completely resuspended in 25 µL TE. 1 µL was removed for UV concentration determination at 260 nm (in 999 µL distilled water) by using Beer’s law,

\[
\text{concentration} = A_{260} \times \text{dilution factor} \times 30 \mu g / mL, \quad \text{where} \ A \ \text{was absorbance at 260 nm.}
\]

The samples were diluted with TE to a final concentration of 1 mg/mL and stored at −20 °C. To anneal the oligomers to the correct double-stranded orientation, 1 µL of each oligomer was removed and placed in an Eppendorf tube with 23 µL of distilled water, the tube was placed on a 90 °C heat block and the heat block was turned off. The temperature returned to room temperature after ~3 hours. No further purification was necessary.

Digestion of the vector involved placing 10 µL of old N-Cam.Y linker (0.2 µg/mL) in an Eppendorf tube with 2 µL 10x NEBuffer 2, 7 µL of distilled water and 1 µL of HindIII (20 u/µL). This reaction mixture was placed on a 37 °C heat block for 2-3 hours. Double digestion can sometimes be performed concurrently without alteration of conditions; however, BamHI has star activity at low salt concentration, which can result in poor specificity of digestion. To remove star activity, the salt concentration was increased and bovine serum albumin (BSA, New England BioLabs) was added as
follows. A 3.3 M NaCl solution with 2-3 µL of 100x BSA was prepared, and 1 µL of this solution plus 1 µL of *BamHI* (20 u/µL) was added to the digestion reaction and placed on the 37 °C heat block again for 2-3 hours.

To purify the digested vector, a 1% Sea Plaque gel was prepared as follows: 0.3 g of SeaPlaque (FMC BioProducts, Rockland, ME) in 30 mL of TE was microwaved until boiling. This solution was swirled until the SeaPlaque completely dissolved, and cooled slightly under running tap water. 5 µL of EtBr (EM Science) was added and the solution was mixed thoroughly. It was poured into a horizontal gel tray and one eight-well comb was placed near the top of the tray. This was allowed to cool until completely polymerized. The comb was removed and 6 µL of 5X loading buffer (1.5 g Ficoll [Sigma, St. Louis, MO], 0.02 g bromophenol blue, 0.02 g xylene cyanol [EM Science] in 20 mL distilled water) was added to the restriction digestion mix. This was run at 80 V in one lane with a control comprised of 2 µL uncut linker vector, 6 µL 5X loading buffer and 2 µL distilled water. As the uncut vector was supercoiled and therefore more compact than the cut linker, it ran the fastest. When the leading dye (bromophenol blue) approached the end of the gel, the voltage was removed and the gel visualized under long-wave UV light. The band of interest was excised using a sharp scalpel and placed in an Eppendorf tube.

Ligation of the unphosphorylated insert and phosphorylated vector was performed with increasing volumetric ratios of insert:vector of 0:2, 1:2 and 10:2 using the following protocol:
<table>
<thead>
<tr>
<th></th>
<th>0:2</th>
<th>1:2</th>
<th>10:2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insert (Annealed Nla Linker)</td>
<td>0 µL</td>
<td>1 µL</td>
<td>10 µL</td>
</tr>
<tr>
<td>10x Ligation Buffer</td>
<td>2 µL</td>
<td>2 µL</td>
<td>2 µL</td>
</tr>
<tr>
<td>Vector (Digested old N-Cam.Y linker)</td>
<td>2 µL</td>
<td>2 µL</td>
<td>2 µL</td>
</tr>
<tr>
<td>Distilled water</td>
<td>16 µL</td>
<td>15 µL</td>
<td>6 µL</td>
</tr>
<tr>
<td>T4 DNA Ligase (400 u/µL)</td>
<td>1 µL</td>
<td>1 µL</td>
<td>1 µL</td>
</tr>
</tbody>
</table>

These reactions were left at room temperature for 2-3 hours. When the reactions were complete, 15 µL of each reaction was transformed into competent dH5α cells as follows. The cell aliquots (~100 µL total) were removed from the –70 °C freezer and thawed on ice. During this time, the ligation reactions were placed on a 65 °C heat block for five minutes to melt any gel material present, and incubated on a 37 °C heat block. Once thawed, 15 µL of each ligation reaction was gently mixed into an aliquot of competent cells. The cells were immediately placed on ice for twenty minutes and LB plates (10 g tryptone, 5 g yeast extract, 15 g Bacto agar (all from Difco, Franklin Lakes, NJ) and 10 g NaCl per liter of plate solution) with 100 mg/L ampicillin (USB, Cleveland, OH) were warmed to room temperature and labeled. The cells were heat-shocked at 37 °C for five minutes and placed on ice again for five minutes. Each cell mix was aliquoted onto the corresponding plate and the mixture was spread over the plate surface with glass beads. The plates were placed in a 42 °C incubator overnight.
After incubation for ~16 hours, up to twenty-three colonies were chosen for PCR analysis to test for the presence of the target plasmid. A PCR reaction mixture was prepared with 100 µL of 10x *Taq* PCR buffer, 30 µL of 10 mM MgCl₂, 25 µL of 10 mM dNTPs (all three chemicals obtained from Invitrogen Life Technologies), 820 µL of distilled water and 10 µL of each primer. The primers utilized in this PCR amplification reaction corresponded to regions of the vector (pAED4) located outside the insert region and have the sequences ATAGGGAGACCACAACGG (for primer #90, which is completely complementary to this segment of DNA) and GAGTTGGCTGCTGCCACC (for primer #91, again with complete complementarity to this sequence). (All primers were resuspended in TE, EtOH-precipitated, washed with 70% EtOH at room temperature, centrifuged for 5 minutes at 14,000 RPM, dried and resuspended in TE again to remove salt. No further purification was performed.) A total of 30 µL per each PCR reaction plus an additional 30 µL to compensate for pipet error was removed from this stock, and 3 µL *Taq* (5 u/µL, Invitrogen Life Technologies) per 280 µL of stock solution was added. This mix was kept on ice until needed. One 200-µL PCR tube for every reaction was labeled and 30 µL PCR stock + *Taq* was aliquoted into each one. A small amount of each colony under inspection was removed from the plates with a pipet tip and completely crushed on the bottom of the PCR tube. One control colony, if any were present (from the 0:2 plate) and 1 µL of check (undigested pAED4 plasmid) were also included for background and positive controls, respectively. PCR amplification was run as follows: the sample was heated to 95 °C for 1 minute; a cycle of 95 °C for 5 seconds, 59 °C for 5 seconds and 72 °C for 45 seconds was repeated 25 times, and the
amplification was finished with a three minute period at 72 °C. After amplification, one or two 3% NuSieve/1% SeaKem agarose gels (both from FMC Bioproducts) were prepared in TE with 5 µL EtBr added after cooling slightly, as described previously. Two eight-well gel combs were added to each gel tray – one near the top, one in the center – to increase the loading capacity of each gel from eight samples to sixteen. 5x DNA loading buffer was added to each tube to a final concentration of 1x. The samples were loaded, along with 7 µL of a prestained DNA ladder (pBR322 MspI digest, 1 µg/µL, New England BioLabs) in one lane on each comb level, and run at 110 V until the lowest dye (bromophenol blue) reached the bottom of the sample row. The gels were visualized by long-wave UV light to determine the length of the PCR amplified fragments.

All colonies that appeared positive were prepared for sequencing by inoculating 5 mL of LB and 20 µL of 25 mg/mL Amp with a small portion of the colony. This was shaken overnight at 37 °C and 200 rpm in an orbital shaker. The plasmids were purified with a Qiagen plasmid prep kit (Qiagen, Valencia, CA) as per the supplied instructions. The concentration of each sample was determined (using the Beer’s law equation given above, but with 50 µg/mL as the conversion factor for double-stranded DNA) by UV at 280 nm. The samples thought to be positive for presence of insert were selected for further PCR analysis using primer #90 and the N1a bottom insert itself as primers. Presence of a band indicated the likelihood that the colony contained the gene coding for the N1a protein. The samples of highest concentration that showed presence of this band were sent to the sequencing facility for forward sequencing using a primer with complementarity to the T7 promoter region. The positive samples were digested with
BamHI (using BamHI buffer and BSA) and, if so desired, dephosphorylated by adding 1 µL of CIAP (10 u/µL), placing the sample on a 37 °C heat block for 30 minutes, heat inactivating the sample at 65 °C for 15-20 minutes, and running a 1% SeaPlaque agarose gel for purification.

6.3 Obtaining the Gene Coding for the Designed Proteins

While the preparation of the vector was taking place, genes coding for the WW domain designs of interest were developed as per the protocol given above, with a few minor changes. Using the Entelechon GmbH program, a sequence was developed that was used as input into an online restriction enzyme analysis program and the unique restriction sites were noted. No attempt to maximize the unique sites was performed. The complement was obtained and restriction enzyme sites were added as follows: a typical BamHI site was placed at the beginning of the gene, and an altered BamHI-like sequence was placed at the end. (BamHI cleaves at G/GATCC; an altered BamHI-like sequence would not be digested by BamHI and has the sequence C/GATCC.) This was necessary because of the two orientations of insertion of this gene into the plasmid digested with only BamHI; restriction enzyme analysis gave an approximate idea as to whether or not the insert was oriented in the correct orientation. If it was in the correct orientation, the digestion fragment was of the N-Cam.Y/Nla length; if it was in the incorrect orientation, the digestion fragment was of the N-Cam.Y/Nla/WW length, a full 100 base pairs longer than it should have been (see Figure 6.3 for clarification). This
digestion, plus PCR amplification analysis, provided a positive or negative answer for
gene presence and direction. The insert was divided into oligomers of less than 60 bases
in length (two on top, labeled 1A and 2A, and two on bottom labeled 1B and 2B). The
oligomers were purified by PAGE as described previously, and diluted until their
concentration was 1 mg/mL.

The construction of the insert was as follows. 1 µL each of 2A and 2B was placed
in an Eppendorf tube with 4 µL of 5x forward reaction buffer (Invitrogen Life
Technologies, 350 mM Tris-HCl (pH 7.6), 500 mM KCl, 50 mM MgCl2, 5 mM 2-
mercaptoethanol), 2 µL of 10 mM ATP (Invitrogen Life Technologies), 11 µL of distilled
water and 1 µL of T4 phosphonucleotide kinase (10 u/µL). This reaction mixture was
placed on the 37 °C heat block for 30 minutes and at 65 °C for 20 minutes for enzyme
inactivation. 1 µL each of 1A and 1B was added and annealing took place as described
earlier. 1 µL of ligase and 2.5 µL of distilled water were added and the reaction was left
at room temperature for two to three hours. If the vector was to be dephosphorylated
after digestion, the insert was phosphorylated by adding 1 µL of kinase and 3 µL of
10 mM ATP after heat inactivation of the ligase for 10 minutes at 65 °C. The
phosphorylation reaction proceeded as described. This step was unnecessary if the vector
was not dephosphorylated after digestion. Purification of the inserts using a 3%
NuSieve/1% SeaKem agarose gel was performed as previously described; again, pBR322
MspI digest was used as a marker.

The bands of interest were removed from the gel using a DEAE membrane
procedure. This involved visualizing the gel under long-wave UV light, photographing
the gel for documentation purposes, excising any bands above the band of interest, and cutting below the band of interest to the side of the gel. A small piece of NA 45 DEAE cellulose membrane, pore size 0.45 µm (Schleicher and Schuell, Keene, NH) was trimmed to the appropriate size. It was soaked with distilled water and inserted into the cut beneath the band of interest. Its position was confirmed and the gel was run at 180 V for five minutes. Complete capture of the band of interest onto the membrane was verified; the membrane was trimmed and placed into a clean Eppendorf tube. A sufficient volume of DEAE NET elution buffer (8 µL of 0.5 M EDTA; 800 µL of 1.0 M Tris base, pH 8; 2.34 g NaCl; ~40 mL of distilled water) was placed into the tube to cover the membrane, and the DNA was eluted at 37 °C for 16-18 hours. (An alternative protocol involved elution at 65 °C for 2-3 hours.) The membrane was inspected under UV light; if the DNA had eluted entirely off the membrane, an ethanol precipitation was performed. If some DNA remained, the membrane was replaced into the same tube and put on the 65 °C heat block for 20-30 minutes, followed by ethanol precipitation. The pellet was resuspended in 25 µL of TE and stored at −20 °C.

A ligation of insert into the digested vector was executed as discussed. The ligation mixtures were transformed and plated, and PCR amplification was carried out on 21 or fewer of the resultant colonies to determine the presence or absence of the insert. The PCR positives were grown to an OD_{600} greater than 1.4, and the plasmids were extracted. The concentrations of the resultant DNA solutions were determined by UV and confirmed using 1% SeaPlaque gel analysis (with a known concentration plasmid as control). This confirmation became necessary because the RNase I of the Qiagen kit
degraded over time, and a significant fraction of the UV signal at 260 nm was eventually 
due to undigested RNA. Adding a small amount of RNase I (Sigma) to the Qiagen buffer 
P1 eliminated this problem.

Owing to the possible bi-directional orientation for this insert, a restriction 
enzyme digestion was performed on the positive samples. This involved digesting 2 µL 
of each sample with NdeI and BamHI, as previously discussed and running the digests on 
a 3% NuSieve/1% SeaKem gel. As the selected samples were all PCR positives, the 
insert was expected to be present. Only those with the correct orientation were expected 
to yield the shorter length fragment (265 bases long); anticipated fragment size for those 
with the incorrect orientation was 370 bases. The samples testing positive for the correct 
orientation were sequenced for final confirmation.

6.4 Point Mutations using PCR Mutagenesis

In one of the constructs selected through implementation of the above procedure, 
insertion of an additional base was detected in the coding region. This defect, which 
would have led to expression of non-target protein, was removed by PCR mutagenesis. 
This method was also applied to generate mutant WW designs differing by one amino 
acid to the template proteins. These variants were used in studies of the physical 
properties of the designed proteins.

To remove the base insertion, the entire WW gene was amplified using primers 
coding for the elimination of the base insertion. Because there were no restriction
enzyme sites present near the 3’ end of the insertion, the second BamHI restriction site was restored by mutagenesis as well. These primers were ordered and prepared as discussed and PCR mutagenesis was carried out as follows. 80.5 µL of distilled water was placed in a PCR Eppendorf tube, along with 1 µL of each primer, 2.5 µL of 10 mM dNTPs, 1 µL of Pfu Turbo (Stratagene, Cedar Creek, TX; chosen for its 12-fold increased fidelity compared to Taq, as well as its 3’-5’ exonuclease activity), 5 µL of template and 10 µL of 10x Pfu Turbo buffer. The PCR mutagenesis program involved the following steps: the sample was heated to 95 °C for 1 minute, followed by 25 cycles of heating the sample to 95 °C for 5 seconds, lowering the temperature to 59 °C and maintaining it for 5 seconds, and increasing the temperature to 72 °C for 45 seconds, and ending with 3 minutes at 72 °C. After the program terminated, a 3% NuSieve/1% SeaKem gel served to separate the PCR product from the background DNA, and the band of interest was run onto DEAE membrane and eluted overnight. This was followed by ethanol precipitation and resuspension. The resulting insert was digested with BamHI followed by electrophoresis via a 3.5% NuSieve agarose gel for increased separation of cut and uncut bands. These were also extracted using the DEAE membrane procedure and diluted to 1/10, 1/100 and 1/1000 their original concentration with distilled water. The undiluted insert and dilutions were ligated into cut vector, transformed into DH5α and plated. PCR amplification was performed using three primers – #90, #91 and the bottom primer of the mutagenesis reaction. All of them should have one band present (due to #90 and #91 amplification), but only the ones with the mutation should have a strong second band due to #90 and bottom primer amplification. The positive colonies were inoculated into 5 mL
of LB with 100 mg/L Amp, harvested, subjected to plasmid preparation and sequenced. This procedure was sufficient to eliminate the base insertion.

The method of PCR mutagenesis involved the use of primers with high complementarity to the surrounding regions of the gene but alterations in the region of the codon to be mutated. The primers were designed to maximize the T_m (via increasing the GC content) and minimize the amount of noncomplementarity of the primer with the template. The length of the primer segments flanking the site of the mutation was increased based upon the size of the mutation site. Reaction mixtures were prepared of 35.7 µL of distilled water, 5 µL of 10x ThermoPol buffer, 5 µL of 10 ng/µL plasmid template, 1.4 µL of each primer at 10 µM, and 0.5 µL of Vent DNA polymerase, 2 u/µL. The PCR program consisted of 2 minutes at 95 °C, and 16 cycles of 30 seconds at 95 °C followed by 55 °C for 8.5 minutes. The reaction mixtures were cooled to 4 °C. The number of cycles and time at 55 °C was dependent on the number of bases that constituted the mutation and the size of the vector, as this was a full plasmid amplification (see Chapter 3). This mixture, minus 2 µL for transformation purposes, was digested with 1 µL of 20 u/µL DpnI and incubated at 37 °C for 1-2 hours. 2 µL of this reaction mixture and 2 µL of the undigested reaction mixture were transformed into DH5α and plated. The undigested plate contained more colonies than the digested plate because of cleavage of the template DNA and subsequent reduction in number of circular plasmids present for transformation. Several of the colonies from the digested plate were selected and inoculated into 5 mL LB with 100 mg/L ampicillin as described. After
plasmid preparation, the samples were sequenced for the presence or absence of the mutation in question.\textsuperscript{137}

6.5 Expression in BL21 or HMS174 Bacterial Cells

If the sequencing of any of the PCR-mutated or constructed colonies proved positive, 1 µL of various dilutions of plasmid in distilled water (1, 1/10, 1/100, 1/1000) was transformed into 50 µL of BL21 or HMS174 competent cells (using the same protocol for transformation in DH5\textalpha) and plated. After ~22 hours, one colony was selected for either small-scale (10 mL) or large-scale (2-4 L) expression. The main goal of small-scale expression was to determine the expression level of the protein and the best time for harvest of the cells. This information guided a large-scale growth for expression of large amounts of protein for biophysical characterization.

In either case, the colony was inoculated into 10 mL LB containing 10 µL of 25 mg/mL chloramphenicol (USB) and 40 µL of 25 mg/mL ampicillin. This was placed in an orbital shaker at 37 °C and allowed to shake overnight at 150 rpm. The procedures differ slightly from this point depending on the scale of the expression. For the small-scale expression, 100 µL of this culture was inoculated into 5 mL of LB with the same concentrations of antibiotics after ~16 hours. It was allowed to shake at 250 rpm until the OD\textsubscript{600} was 0.5-0.8. (Generally, each 0.1 OD unit is roughly equivalent to 10\textsuperscript{8} cells/mL.\textsuperscript{108}) It was induced with 50 µL of 100 mM IPTG (Promega, Madison, WI) and allowed to shake for 3-4 hours following induction. At every time point (including the
zero hour time point immediately following IPTG induction), 100-500 µL of sample was collected, placed in an Eppendorf tube and centrifuged at high speed. The supernatant was removed and the pellet frozen overnight. After ~16 hours, a protein gel was run using the samples harvested from the day before. The method for preparing denaturing protein gels was derived from standard protein gel protocols. In brief, the plates for each protein gel were cleaned with 70% EtOH and placed, along with appropriate spacers, in a vertical gel apparatus. The resolving gel solution for four gels (5.0 mL of 30% acrylamide/0.8% bis-acrylamide, 3.4 mL gel buffer [3 M Tris base, pH 8.45 and 0.3% SDS] and 2.0 mL of 50% glycerol) and stacking gel solution (0.8 mL of 30% acrylamide/0.8% bis-acrylamide, 1.6 mL gel buffer and 3.9 mL distilled water) were made up. The polymerization of the resolving gel was initiated by the addition of 100 µL 10% APS in TE and 10 µL TEMED; this solution was pipetted into the plate apparatus to a level of ¾ from the bottom. The remainder of the volume was filled with distilled water, and the gel solution was allowed to set. Once polymerization was complete, the water was decanted and the polymerization of the stacking gel was initiated in the same manner. This solution was pipetted on top of the resolving gel to the level of the top of the plate apparatus, and a comb was introduced. This was again allowed to polymerize. The comb was removed, the exterior of the plates was washed and the protein gel was set up in a vertical gel electrophoresis unit. Cathode buffer (100 mM Tris base, 100 mM Tricine [USB], 0.1% SDS, pH 8.25) was poured behind the gel’s top, and anode buffer (100 mM Tris base, pH 8.9) was poured into the bottom of the tank. The cell pellet time points were resuspended in 50 µL TE, and 50 µL 2x protein loading buffer (8% SDS, 100
mM Tris base, 24% glycerol, bromophenol blue, pH 6.8) was added. The pellets were placed at 65 °C for five minutes and sheared on a vortexer with a small glass pipet. The samples were loaded onto the gel, along with a prestained low-molecular weight protein ladder (BRL Life Technologies), and run under constant voltage at 25 mA until the dye had run through the stacking gel. The current was increased to 50 mA and applied until the dye reached the bottom of the gel plates. The plates were removed and pried open, and the gel was placed into a container with staining agent (45% methanol [Burdick and Jackson, Muskegon, MI], 45% distilled water, 10% acetic acid [EM Science], 2.5 mg/mL Coomassie Brilliant Blue G250). It was allowed to stain on a rocker for 20 minutes. After this time, the staining agent was removed and a destaining agent (45% methanol, 45% distilled water, 10% acetic acid) was applied. Once the bands were evident, a picture was taken of the gel for future reference.

For a large-scale expression, 5 mL of the overnight growth was inoculated into 1 L of LB containing 0.5 mL of 25 mg/mL chloramphenicol and 4 mL of 25 mg/mL ampicillin. This was allowed to shake at 250 rpm until the OD$_{600}$ was 0.5-0.8. IPTG was added to the same final concentration given above (5 mL of 100 mM IPTG per liter of cells), and 500 µL of 0-hour and immediately pre-harvest protein samples were taken. The cells were harvested by centrifugation (5,000 g for 10 minutes) after at least 3-4 hours post-induction. The cell pellets (separated into thirds) were resuspended in 20 mL TE per liter of LB. A crystal of DNase I (USB) was added and the cell suspensions were frozen overnight for increased cell lysis. After ~16 hours, the cell paste was thawed in warm water and stirred for 30 minutes. The cell paste was homogenized using a standard cell
homogenizer, and 0.117 g NaCl and 0.041 g MgCl₂ (USB) was added to each sample third and the cell paste was again stirred for 20 minutes. The cell suspension was centrifuged at 12,000g for 20 minutes, and the supernatant saved for column purification. Until it was clear that the protein was present in the soluble fraction and not in inclusion bodies, the pellet was also retained and processed. A protein gel of the 0- and 3- or 4-hour time points, along with the soluble fraction of the large-scale expression, was usually run at this time to determine the solubility of the protein as well as the large-scale expression level.

6.6 Purification and Cleavage of WW Fusion Systems

The main method of purification of the WW fusion systems involved column chromatography using phenyl sepharose resin. Two other methods, utilized only when this primary method did not work, will also be described. As mentioned previously, N-Cam.Y will bind to phenyl sepharose in the presence of Ca²⁺, and will elute when that Ca²⁺ is chelated out of the protein matrix. The column was run at high salt to reduce the charge of the protein and maximize the exposure of the hydrophobic pockets for phenyl sepharose binding. The ability to purify N-Cam systems with this method proved very useful for the purification of the WW fusion systems.

After removing the supernatant from the cell solution as described above, the appropriate volume (~20 mL) was made 5 mM Ca²⁺ and 500 mM Na⁺, doubled in volume with high-salt Ca²⁺ wash buffer (50 mM Tris base, 500 mM NaCl, 1 mM CaCl₂,
pH 7.5), placed on ice and applied to a newly-poured and washed phenyl sepharose column (~20 mL pre-swelled resin volume, Pharmecia, Peapack, NJ, CL-4B) at low pressure using a BioRad LP system. The system was washed with the high-salt Ca\(^{2+}\) wash buffer until the UV absorbance returned to the pre-loading value. (Until it was verified that the protein had affinity to the column resin, the flow-through was retained throughout the procedure.) The protein was eluted with 5 mM EDTA buffer (50 mM Tris base, 500 mM NaCl, 5 mM EDTA, pH 7.5). The column was cleaned with one bed volume of 6 M guanidine (ICN Biomedicals) and five bed volumes of distilled water, or until the UV absorbance returned to the pre-load value. The samples of interest were collected and a small amount was reserved for protein gel analysis. This amount was diluted two-fold with TE and run on an SDS-PAGE protein gel. The samples containing the desired protein were pooled and reserved for either HPLC or direct proteolytic cleavage.

Another purification method involved running an anionic exchange column using QAE Sephadex A-25 (Pharmecia) as the resin. This again concerned diluting the sample two-fold with the low-salt buffer (50 mM Tris base, pH 7.5; 50 mM NaCl), loading the sample and washing the column with the same low-salt buffer until the absorbance returned to the pre-load value. A 50% gradient of increasing high salt concentration was run using the low-salt buffer and a high-salt buffer (50 mM Tris base, pH 7.5; 1 M NaCl) to a maximum salt concentration of 25 mM Tris and 500 mM NaCl. Additionally, a lower-salt concentration anionic exchange column was run to attempt to maximize the affinity of protein for the column resin; this would involve using the previously described
low-salt buffer in place of the high-salt buffer in the last procedure, and using a lower-salt buffer (10 mM Tris base, pH 7.5; 10 mM NaCl) as the loading and washing buffer. A denaturing anionic exchange column was also attempted, wherein 6 M urea was added to both the low- and high-salt buffers and the column was run as described.

Another method did not utilize column chromatography at all. Because of the high $T_m$ and thermal stability of N-Cam.Y, a thermal stability experiment was performed to purify the fusion protein partially. This involved centrifuging the samples to remove any precipitates, heating the samples at the desired temperature (which ranged from 40 °C to 90 °C) for 10 minutes, centrifuging the samples for three minutes, removing the supernatant, discarding the pellets and running an SDS-PAGE protein gel. The goal of this step was to eliminate, by unfolding and precipitation, the majority of the proteins less thermally stable than N-Cam.Y. Upon centrifugation, these proteins formed a pellet at the bottom of the tube, whereas N-Cam.Y and its fusion systems remained soluble. After heat purification, HPLC or other methods could have been performed to further purify the fusion system away from the background proteins had the thermal stability purification worked.

Once the fusion system was purified from the background proteins, it was cleaved by the NIa protease method. A small-scale cleavage was performed to determine the minimal amount of time necessary for complete cleavage of the fusion system. Protein (about 50 µM) was incubated at room temperature with protease in a 200:1 protein:protease ratio (initially obtained from Dr. Song Tan, Biochemistry and Molecular Biology Department, and later expressed and purified in-house with a construct obtained
from Dr. Jennifer Doudna, Yale University) at the supplied concentration. A small amount of this sample was allowed to react for 0-3 hours; a larger sample was allowed to cleave for ~16 hours. A protein gel was run to determine the extent of cleavage at 1, 3 and ~16 hours as compared to the uncleaved sample. After 3 hours, cleavage was ~75% complete; ~100% cleavage was reached overnight. ~5 mL of purified fusion protein (~100 µM) was cleaved with 5 µl protease after two to three days at room temperature.

Following cleavage, it was necessary to remove the uncleaved fusion protein as well as the N-Cam.Y/Nla fusion system from the WW protein solution. There were two methods utilized to accomplish this. The first involved another phenyl sepharose column, which bound any of the proteins containing the N-Cam.Y portion, but did not bind any cleaved WW proteins or background proteins not able to bind phenyl sepharose. The protein sample was placed in a Falcon tube with approximately 9.5 mL of cleaned phenyl sepharose. A total of 1.25 mL of 100 mM CaCl₂ was added to induce conformational change in N-Cam.Y and concomitant binding to phenyl Sepharose. This was rocked on a platform for one hour at room temperature to allow for equilibration, and was then poured back into the column. Approximately 5 mL of the sample was collected under atmospheric pressure; about 10 mL of high-salt buffer was run through under low pressure at 2 mL/min to remove the N-Cam.Y bound to the resin. The first collection was dialyzed into 100 mM NaCl, 1 mM CaCl₂ and 10 mM MOPS (Sigma), pH 7.2 with a Spectra/Por membrane (Spectrum, Rancho Dominguez, CA; MW cutoff of 3,500 Da) for characterization purposes.
Alternatively, the cleaved sample was HPLC-purified. This involved filtering the sample with a 0.45 µm Acrodisc filter (Pall Corporation, East Hills, NY) and loading 5 mL aliquots onto a reverse-phase HPLC system washed with buffer A (100% HPLC-grade water with 0.1% TFA). A gradient of 90% acetonitrile (Burdick and Jackson)/10% HPLC-grade water with 0.1% TFA (buffer B) was initialized; as the run progressed, increasing concentration of buffer B was mixed with buffer A and washed through the column at high pressure. The peaks were collected and lyophilized. The resulting protein powder was resuspended in distilled water and analyzed by SDS-PAGE to determine which peaks contained the protein of interest. The HPLC conditions were adjusted to optimize the purification of each protein; in some cases, the protein of interest eluted at a low acetonitrile concentration. Because the salts present in the sample eluted in the same solvent conditions, the introduction of acetonitrile was delayed in order to separate the salt peak from that of the protein of interest. The percentage of acetonitrile in the solvent was also decreased to improve the separation between the protein and salt peaks, as well as the target protein and the background protein peaks. This was mainly to ensure that the fusion system background eluted significantly after the WW protein of interest.

6.7 Experimental Characterization of the WW Proteins

All DNA plasmids in this section were sequenced by the DNA Sequencing Facility on an ABI Hitachi 3100 Genetic Analyzer (Applied Biosystems, Foster City,
PCR was performed using an MJ Research minicycler. Gel pictures were taken at every step involving agarose or PAGE analysis using a Kodak digital camera with a UV filter, and were saved on a Power Macintosh computer. All ultraviolet spectroscopy was carried out on a Shimadzu BioSpec-1601 spectrometer in a 1-cm path length quartz SUPRASIL UV cell (Hellma, Plainview, NY). Column purification was performed on a BioRad LP Econo gradient pump system (Hercules, CA). HPLC utilized a Shimadzu HPLC (Columbia, MD) equipped with an LC-10AT liquid chromatograph, an SCL-10A controller, an SPD-10AV UV detector, and a c18 silica 10x250 mm 5 µm 300 Å VYDAC HPLC column (The Nest Group, Southborough, MA). All lyophilization was performed using a Labconco Freeze Dryer 4.5 (Kansas City, MO). All CD spectra were collected on an Aviv CD model 62 DS (Lakewood, NJ). Additionally, all fluorescence data were obtained using a fluorometer from Fluorolog (Jobin Yvon-Spex, Horiba group, Edison, NJ) with rough temperature control provided by an external water chiller (Neslab, Portsmouth, NH). After SDS-PAGE confirmed the weight of the designed protein and extent of cleavage from the fusion system, a small amount (about 50 µL) was sent to the Intercollegiate Mass Spectroscopy Facility for molecular weight determination on a Perseptive Biosystems Mariner mass spectrometer (Framingham, MA) via either MALDI or ESI in positive ion mode. This was typically done immediately following a second round of HPLC, wherein the location of the peak containing the protein of interest was already known. If the mass spectroscopy results were as expected for the WW protein of interest, further biophysical characterization was performed.
The concentration of WW protein solutions was determined by UV spectroscopy. A spectrum was recorded between 200 and 300 nm in a 1 cm quartz cuvette (Hellma), at a concentration expected to give rise to absorbance values between 0 and 1. The solvent was TE buffer. The absorbance of this sample was determined at 280 nm ($A_{280}$) and converted to concentration as $\frac{A_{280} \times \text{dil. factor}}{\varepsilon}$ using the expected $\varepsilon$ value calculated from the primary structure.

Circular dichroism spectropolarimetry was utilized to examine the 2° (and, to some extent, 3°) structure and thermal stability of each of the proteins. All CD spectra for WW proteins were collected in a 2 mm cuvette (Hellma) with 7-10 seconds averaging time, 0.1 minute equilibration time, 1 nm step size and 2.5 nm slit width, unless otherwise noted and converted to mean residue ellipticity (MRE) using the following equation:

$$MRE = \frac{\text{signal}}{10 \times \text{conc} \times \text{pathlength} \times N_{AA}}$$

where the signal was in millidegrees, the concentration was in moles per L, the path length was in cm, and $N_{AA}$ indicated the number of amino acids in the protein. The resulting units were mdeg cm$^2$ mol$^{-1}$. A buffer (background) spectrum was obtained from 200-270 nm or 200-300 nm for near-UV CD characterization. Buffer consisted of either 100 mM NaCl and 10 mM Na MOPS, pH 7.2 or 10 mM KH$_2$PO$_4$ (EM Science), pH 7.0 and 100 mM NaCl. Protein concentration was 50 µM unless otherwise noted to yield sufficient signal intensity. Conditions for one of the designed proteins were altered to include concentrations of trimethylamine N-oxide (TMAO, Sigma) ranging from 500
mM to 2 M (calculated from the mass dissolved and the formula weight of TMAO) in distilled water and 20% glycerol (v/v). Wavelength spectra at 2 °C of buffer and of sample were obtained, followed by wavelength spectra at 98 °C of same. A third set of wavelength spectra was obtained at 2 °C to determine roughly the reversibility of the thermal denaturation of the protein. The first 98 °C condition prior to thermal denaturation was necessary in the case of one of the designed proteins for annealing to a correctly folded structure.

Thermal denaturation experiments involved a 0.1-1.5-minute equilibration time plus a 7-90 second averaging time at each new temperature, depending on the protein under study. The thermal denaturation signal at 230 nm was obtained from 2 °C to 98 °C with 1.5-2 °C step size. As before, the bandwidth was 2.5 nm, and a 2 mm cuvette was used. The spectra were saved to a 3.25” disk, uploaded onto a PC and plotted with Microsoft Excel. The spectra were curve-fitted with a non-linear least-squares fitting program (Nfit, University of Texas, Galveston), utilizing the standard Gibbs-Helmholtz free energy equation:

\[ Y = \left( Y_N + (Y_U - Y_N) \right) \times \frac{1}{\Delta H_m^0 + \Delta C_p (T-T_m) + \frac{\Delta H_m^0}{T_m} + \Delta C_p \ln \left( \frac{T}{T_m} \right)} \times \frac{1}{1 + e^{\frac{\Delta C_p (T-T_m) + \frac{\Delta H_m^0}{T_m} + \Delta C_p \ln \left( \frac{T}{T_m} \right)}{RT}}} \]

where \( Y_N \) is the folded baseline, \( Y_U \) is the unfolded baseline, \( T \) is the temperature in Kelvin, \( T_m \) is the temperature at which the population of unfolded molecules equals the population of folded molecules, \( \Delta C_p \) is the change in specific heat, and \( \Delta H_m^0 \) is the
standard molar enthalpy. Typical equations for straight lines were utilized \( Y = a_1 x + b_1 \), where \( a_1 \) is the slope and \( b_1 \) is the y-intercept of the line, or the y-value at \( x = 0 \) K) in place of the folded and unfolded baseline parameters in the above equation.

Fluorescence spectroscopy was performed with 1 µM protein samples in a buffer containing 20 mM \( \text{PO}_4^{3-} \), pH 7.0, 100 mM NaCl and either distilled water or 6 M urea (ICN Biomedicals, Aurora, OH) to fill the solution volume to 1 mL. The cell was 1 mm in path length (10x4 quartz SUPRASIL, Hellma). A blank was also prepared of buffer without protein. The integration time was 1 second with a 3-minute delay time between points and a 2-minute equilibration time. Two spectra at 0 M (10 °C) and 5.26 M urea (20 °C) were obtained. A urea denaturation titration utilizing 6 M urea (as calculated by determining the necessary weight of solid urea to be resuspended in distilled water) was also performed with the same parameters at 10 °C. The urea was manually added incrementally to a maximum final concentration of 3.5 M; the sample was stirred for about 1.5 minutes after each addition for equilibration purposes. The stirrer was turned off and the spectrum obtained with an emission wavelength of 346 nm and an excitation wavelength of 295 nm. Raw data were exported as ASCII X/Y and viewed on a PC running Microsoft Excel.

\(^1\text{H}\) NMR spectroscopy was utilized to characterize further one of the designed proteins. Neat protein solution post-lyophilization and resuspension was raised in pH from the post-lyophilization value of approximately 2.8 to 6.8-7.2 in an Eppendorf tube using NaOH. The pH was measured with a Beckman \( \phi \)71 pH meter equipped with a small-bore probe (Mettler Toledo, Columbus, OH). Several annealing steps, involving
placing the sample on a 90 °C heat block for five minutes, followed by removing the heat
block, allowing the system to cool to room temperature, and adjusting the pH again, were
necessary once the pH was raised above the pI. Both 1D and 2D ^1H spectra were
collected at 600 MHz on a Bruker DRX spectrometer (Billerica, MA, 14.1 T, operating at
a ^1H frequency of 600.13 MHz). The samples were prepared in 90% H2O/10% D2O, as
well as 100% D2O.

For the 1D spectra in 90% H2O/10% D2O, the protein concentration was about
700 µM, pH 6.8. Sodium azide (approximately 3 mM, Sigma) was added for inhibition
of bacterial growth. The temperature of the probe was varied between 5 and 25 °C. For
the 1D ^1H spectrum, data were collected over a spectral width of 10 kHz (4096 complex
points, 256 transients). Suppression of the water signal was achieved with a 1.2-second
low-power saturation pulse. Data were processed with a squared cosine bell window. A
variable temperature study was performed in 95% D2O with 20 mM PO_4^{3-}. The
temperature of the probe was decreased from 35 °C to 5 °C in 5 °C increments. Between
256 and 4000 complex points were collected. All other parameters were identical to the
90% H2O data set described above.

The concentration of the protein sample in D2O was about 900 µM, with the post-
lyophilization pellet resuspended in D2O to a final volume of approximately 600 µL and
adjusted to pH 7.4* with DCI and NaOD (where the * indicates that isotope effects were
neglected). 20 mM PO_4^{3-} was added for solubility purposes. All homonuclear 2D data
sets were obtained at 5 °C for maximum population of the folded state. NOESY^{138},
DQF-COSY^{139} and TOCSY^{140} data were collected with TPPI quadrature detection in the
indirect dimension. For each of these data sets, data size was 2048 x 512. The mixing time for the NOESY was 120 ms and 45 ms for the TOCSY. Either a WATERGATE with 3-9-19 binomial suppression sequence\textsuperscript{141} or a low-power presaturation pulse was applied to reduce the water signal. All experiments involved the collection of 64 transients per time increment. Data sets were processed using xwinmr (Bruker) to produce a 2048 x 2048 real matrix. All data were subjected to a square sine-bell window shifted by 45° in both dimensions. Chemical shifts were referenced indirectly through water after correction for temperature.\textsuperscript{142}

6.8 Ni\textsubscript{a} Protease Purification Procedure

A significant amount of the mutant Ni\textsubscript{a}-Pro protease was obtained from Dr. Song Tan, Biochemistry and Molecular Biology Department. However, some of the latter cleavage reactions were performed using Ni\textsubscript{a} protease prepared in-house.\textsuperscript{136} Dr. Jennifer Doudna (Yale University) generously supplied the gene coding for the mutant Ni\textsubscript{a}-Pro protease mutant from the tobacco etch virus. This was transformed into BL21 cells as usual, but the cells were induced with IPTG at an OD\textsubscript{600} = \sim 1. The cells were harvested at 4-5 hours post-IPTG expression and resuspended in 50 mM Tris base, pH 8.0, and 300 mM NaCl. A crystal of DNase was added and the cell paste was frozen overnight. After \sim 16 hours, the cell suspensions were thawed in cool water and adjusted to 20 mM MgSO\textsubscript{4} and 2 mM CaCl\textsubscript{2}. These solutions were stirred on ice for 30 minutes and homogenized as usual. 0.26 mL of Triton X-100 (USB) was added to each liter of cells, and the mixture
was stirred on ice for an additional 20 minutes. The cells were centrifuged as usual, and the pellets were stored at −20 °C. When needed, the pellets were resuspended in 25 mL of resuspension buffer (6 M guanidine HCl, 100 mM Na₂HPO₄, 10 mM Tris-HCl, pH 8.0) per liter’s worth of cells and incubated at 65 °C for 10 minutes. They were then centrifuged for 20 minutes at 15,000g.

Co TALON resin (Clontech, Palo Alto, CA) was washed with 10 column volumes of equilibration buffer (6 M urea, 100 mM Na₂HPO₄, 10 mM Tris-HCl, 10% glycerol, 300 mM NaCl, 5 mM 2-mercaptoethanol, pH 8.0). The supernatant was loaded onto the column and the column was washed with four column volumes of equilibration buffer and six column volumes of wash buffer (equilibration buffer, pH 6.3, plus 5 mM imidazole). One column volume of elution buffer (equilibration buffer, pH 4.5, plus 150 mM imidazole) was flowed through the column; the stopcock was closed, the pump turned off, and the column was incubated for five minutes at room temperature. The stopcock was opened and the pump turned on, and four column volumes of elution buffer were used to wash the column. All solutions were kept on ice during the loading, and all fractions were collected after initializing the column with elution buffer.

A high-molecular weight SDS-PAGE protein gel, utilizing a prestained high-molecular weight protein ladder as a standard (BRL Life Technologies), was utilized to determine the presence of protein in the fractions; those fractions containing protein were pooled, adjusted to pH 8.5 with 10 N NaOH, and dialyzed for 4-8 hours into storage buffer (100 mM Tris base, pH 8.5, 500 mM NaCl, 50% glycerol, 5 mM DL-dithiothreitol [DTT, Sigma], 0.5 mM EDTA) using the Spectra/Por dialysis membrane as previously
described. The sample was centrifuged at 15,000g for 20 minutes, and the pellet and supernatant were stored at –20 °C until needed. The pellet was resolubilized in storage buffer when required for use.
CHAPTER 7

APPLICATIONS OF THREE-HELIX BUNDLE DESIGN

Backbone-Dependent and De Novo Design

Materials and Methods

Results, Discussion and Conclusion

7.1 Redesigning SpZ

Unless otherwise noted, all chemicals, enzymes, buffers, equipment and methods were the same as given in Chapter 6.

The three-helix bundle design project took two paths. The first was the complete redesign of SpZ using the experimentally determined backbone coordinates. The second, a de novo attempt, involved the design of both the backbone itself and the sequence.

The redesign of SpZ employed SPA and varying weights of the secondary structure propensity (ssp). These ranged from 0 to 2.5 and, with the increasing strength, guided the algorithm to select for amino acids that are typically found in $\alpha$-helices of wild-type proteins. The final weight selected for the designed protein was ssp = 1.0, which was a moderate constraint on the selection of amino acids that are statistically found in helices of native proteins.

The gene coding for wild-type SpZ (a necessary control) was first developed as follows. Oligos were designed through the use of an in-house algorithm for maximization of restriction sites and codon usage in BL21. A (His)$_6$ tag [(CATCAC)$_3$]
was added at the 3’ end of the insert for protein purification purposes, and two stop
codons were placed after the tag. Finally, *Ndel* and *EcoRI* restriction sites were added at
the 5’ and 3’ ends of the insert, respectively. Bases coding for a (His)$_6$ tag were also
added prior to the two stop codons. The insert top and bottom were both divided into
four oligomers. The oligos were resuspended in 50 µL TE for a final concentration of ~1
nmol/µL; 25 µL was removed and PAGE-purified as described earlier. The insert was
cloned into a pAED4 vector digested with *Ndel* and *EcoRI* as described in Chapter 6.

Two colonies were identified by PCR amplification as possibly containing the
gene of interest. These were sequenced by the nucleic acid sequencing facility. The
results confirmed that both were positive for the presence of the SpZ gene. One of these
plasmids was transformed into BL21 for a small-scale expression of protein. The
resultant samples indicated good expression of target protein at the three-hour post-IPTG
time point when qualitatively analyzed by SDS-PAGE. A 4-L growth was performed as
described in Chapter 6 and harvested after three hours post-IPTG induction; the other
liter was not harvested until 17 hours after the addition of IPTG, to determine if a longer
time of expression produced a greater amount of target protein. Both samples were
resuspended in Co TALON extraction buffer (50 mM Na$_2$HPO$_4$, 300 mM NaCl, pH 7.0)
instead of TE as described in Chapter 6. This was due to the chelation activity of EDTA,
which would remove the functional column-bound Co atoms necessary for binding the
His tag. The 17-hour post-IPTG sample had to be discarded because of its high viscosity.
The 3-hour post-IPTG sample was not prohibitively viscous, and was prepared for
affinity purification as described in Chapter 6. After the final centrifugation, the His-
tagged protein was purified by Co TALON low-pressure column chromatography as follows. The column was washed with 5-10 bed volumes (20-50 mL) of extraction buffer at 4 mL/min flow rate. The sample was loaded onto the column at 2 mL/min flow rate. The column was washed with 10-20 bed volumes of wash buffer (50 mM Na phosphate, 300 mM NaCl, 5 mM imidazole, pH 7.0), or until the UV absorbance returned to the pre-load value. (The load and wash were collected throughout, to ensure the protein was not lost if it had low affinity to the Co TALON column.) The target protein was eluted with 5-10 bed volumes of elution buffer (50 mM Na phosphate, 300 mM NaCl, 150 mM imidazole, pH 7.0), with all fractions being collected and the UV absorbance recorded. The column was washed with four bed volumes of 6 M guanidine (pH 5.0) and rinsed with five bed volumes of distilled water, or until the UV absorbance returned to the pre-load value. The column was stored under 30% EtOH at 4 °C to inhibit bacterial growth.

An SDS-PAGE gel was run to determine the presence and approximate molecular weight and concentration of SpZ in each column fraction. If within normal range, the appropriate fractions were pooled and a small sample (~20 µL post-column) was analyzed by mass spectroscopy. Further purification was done by HPLC in the first expression, but was deemed unnecessary for later expressions. Desalting was performed with a G-10 Sephadex column as follows. A molecular weight indicator (about 10 mg of xylene cyanol) was added to a small volume (less than 7 mL, for maximum protein-salt separation) of the protein sample and mixed thoroughly. Approximately 50 mL of G-10 Sephadex resin was prepared as described in the product literature and poured into a clean column. The SpZ sample was loaded on the column at a rate of 2 mL/min.
Distilled water was used to wash the column at the same flow rate. The fractions of interest were collected and the remainder of the column was washed with distilled water to remove the dye and salts and allow the UV absorbance to return to the pre-load value. SDS-PAGE was utilized to determine the presence or absence of SpZ in each fraction, and the appropriate fractions were pooled. MALDI-LR mass spectroscopy confirmed the presence of a protein whose molecular weight was 7463.1 Da, indicating that the SpZ protein was present but that the initiating Met had been cleaved from the protein by a cellular protease.

The concentration of SpZ was determined by UV spectroscopy in 6 M guanidine (ICN Biomedicals) as described in Chapter 6. A 50 µM SpZ sample was prepared for CD analysis with 10 mM phosphate, pH 7.0 and 100 mM KCl as the buffer. The CD spectrum in a 1 mm cuvette was obtained from 200-270 nm in 1 nm increments with 7 second averaging time, 2.5 nm slit width and no equilibration time. Thermal denaturation by CD was performed in the same buffer with the detection wavelength set to 222 nm, at which the maximum ellipticity difference between folded and unfolded protein was detected. The averaging time was 30 seconds, the temperature step size was 2 °C, the slit width was 2.5 nm and the equilibration time was 2.5 minutes. The thermal denaturation was performed in a 2 mm cuvette. The signal at 2 °C prior to the denaturation was compared with the signal at 2 °C following the denaturation for qualitative determination of reversibility. No further biophysical characterization was completed on this protein.

SpZ overexpressed to a high level and was readily purified. An approximate volume of 8 mL of 0.40 mM protein was obtained from 0.5 L of LB cell growth. The CD
spectrum showed minima at 208 and 221 nm, as expected for a highly $\alpha$-helical protein. The fitted thermal denaturation curve (using Nfit and the Gibbs-Helmholtz equation described in Chapter 6) gave a $T_m$ of 82.1 °C. The CD spectrum and thermal denaturation curve are shown in Figure 7.1.

7.2 Computationally Redesigning SpZ

The high-resolution SpZ backbone was used as input for SPA with an ssp weight of 1.0. One of the designed sequences on this backbone was the most native-like when visually inspected for packing, electrostatic interactions and hydrogen bonding networks, and lowest in energy as calculated by the energy function described in Chapter 4. The sequence of this protein, termed HisSSP (Figure 7.2), was:

MDEGFITNNIKDALKDLYALDSLGGDVFKKEALDKFKAIWDDIKDFKQVIIEIKRSYAN(H)$_6$

The gene coding for this protein was developed using the nested ligation of overlapping oligonucleotides as described in Chapter 6, and the presence of the insert cloned within the vector verified, but no expression of the protein was performed.
Figure 7.1. (Top) CD spectrum of (His)6-tagged SpZ. Buffer conditions were as described in the text. (Bottom) Thermal denaturation profile of (His)6-tagged SpZ as followed by ellipticity at 222 nm. The thermal denaturation profile was fitted to the Gibbs-Helmholtz equation as described in the text with a $\chi^2$ value of 0.011 and an $R^2$ value of 1.00.
Figure 7.2. SpZ designed sequence and rotamers on the wild-type SpZ backbone.
7.3 De Novo Design

All of the algorithms previously discussed in Chapter 5 were utilized to design a three-helix bundle protein from first principles. Loops of four or five residues in length were necessary to connect the helices in an up-down-up topology. Typical, well-packed helices after initial helix design had the parameters given in Table 7.1.

These parameters were used as input into the loop-building algorithm to obtain two loop structures, one with a length of five residues and the second with a length of four residues. Both loops involved an energy cutoff of 50.0 arbitrary units and a fit (rmsd between the ends of the helices and the beginnings of the loops) cutoff of 0.3. This structure was used as input into the backbone flexibility algorithm to generate random fluctuations of the backbone dihedral angles. Because of the resultant poor packing after the application of backbone flexibility, the protein was repacked by minimizing the rmsd between the flexible-backbone and the rigid-backbone designs. This backbone was used as input into an old version of SPA, which allowed for all amino acids (besides His, Met and Cys) after filtering for HP patterning and steric constraints. This algorithm utilized all of the previously described parameters, including a secondary structure propensity weight. The set of sequences so obtained exhibited low composition entropy, as the weight for this property was set to zero. The sequence results from the ten lowest energy structures, after inspection of each structure for tight packing, hydrogen-bond geometries and acceptable electrostatic interactions.
Table 7.1. *De novo*-designed three helix bundle parameters. Model 634, $E_{\text{total}} = 0.73$, $E_{\text{SB}} = -63.18$, $E_{\text{SS}} = -55.58$.

<table>
<thead>
<tr>
<th>Helix number</th>
<th>Distance or Angle #1 – z-axis</th>
<th>Distance or Angle – x-axis</th>
<th>Distance or Angle – y-axis</th>
<th>Distance or Angle #2 – z-axis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.000</td>
<td>6.397</td>
<td>-0.678</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>39.405</td>
<td>0.429</td>
<td>0.582</td>
<td>-0.858</td>
</tr>
<tr>
<td>2</td>
<td>0.000</td>
<td>6.628</td>
<td>-0.328</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>14.912</td>
<td>-0.300</td>
<td>180.046</td>
<td>118.577</td>
</tr>
<tr>
<td>3</td>
<td>0.000</td>
<td>6.604</td>
<td>0.659</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>21.761</td>
<td>-0.007</td>
<td>-0.263</td>
<td>241.780</td>
</tr>
</tbody>
</table>
were used as seed for another, final design attempt. All results from the individual steps are shown in Figure 7.3, and the protein that resulted from the first de novo attempt is given in Figure 7.4. As can be seen, the helices twist slightly around each other, much like some wild-type three-helix bundles and coiled-coils. The sequence of this protein was:

```
KQQEKKIKRELKRLQEERGEEKARRILQKIEKIGAGKRYLYKYRRKIREIK
```

The gene coding for this protein was designed with the use of the previously mentioned in-house program to maximize the unique restriction sites present. Two stop codons were added to the 3’ end of the insert; Ndel and Clai restriction sites were added to the 5’ and 3’ ends, respectively. The top and bottom strands of the insert were divided into three oligos each and ordered. PAGE purification was performed as described in Chapter 6, and the insert was constructed and ligated into the cut pAED4 vector. However, no plasmid that coded for the protein of interest was obtained.

7.4 Further Goals – SpZ Redesign and De Novo Design of Three-Helix Bundles

The difficulties met in the construction of the gene could not be explained but they were sufficiently severe to require a redirection of the project. It is clear that if the SpZ design goal were to be pursued, this first step should be approached with improved molecular biology tools. As is obvious from inspection of the sequence, the de novo-
Figure 7.3. Results of individual steps towards the *de novo* design of a three-helix bundle.
Figure 7.4. *De novo*-designed three-helix bundle protein. Note high amount of charged residues with long side chains.
designed protein was quite highly charged. This would be corrected in future design
tries through adjustments of the parameter weights, especially the electrostatic and
hydrogen-bonding terms. It would be worthwhile to use the de novo-designed backbone
as input into the latest version of SPA, instead of SPANS, to determine if the alternative
method of de novo backbone flexibility described in Chapter 5 would have been useful.
Owing to the novel de novo design approach, high-resolution structure determination
would be an absolute necessity, provided that the qualitative probes of structure (such as
CD) indicate that a helical structure is populated. Additionally, some method of
determining the oligomeric state of the protein, such as equilibrium ultracentrifugation,
would be of utmost importance, as previous two-, three- or four-helix bundle design
attempts by other researchers have resulted in multimeric behavior that was not
intentionally designed.

Other future goals of this project include the overexpression and characterization
of the HisSSP designed protein. Two properties of immediate interest are its 2° structure
and thermal stability. If either or both of these criteria were acceptable, further high-
resolution spectroscopy, such as NMR or x-ray crystallography, would be performed to
determine a high-resolution structure for this design. If the design did not adopt a three-
helix bundle conformation or if the thermal stability was poor, the weights of the
parameters would have been adjusted until these issues were resolved. A redesign of this
protein using SPA would also have elucidated the ability of this design attempt to design
a fully folded helical protein using a backbone ensemble, as its usefulness in designing a
β-sheet protein has been proven to some extent (as will be demonstrated in Chapter 8).
Currently, the algorithms have been partially rewritten to utilize the new SPA instead of an older genetic algorithm as described in Chapter 4. Further development of the updated algorithms is necessary to adequately exploit the usefulness of the current SPA algorithm and to streamline the design process. All algorithms (in their old formats), complete with brief introductory descriptions, are included in Appendix B.

7.5 Sequence Diversity and Structural Specificity Results

This body of theoretical experiments involved modeling sequences of increasing diversity upon a de novo-designed three-helix bundle protein backbone. The ideal contact cutoff (shown in Figure 7.5 and described in Chapter 5) for the buried position calculation in the initial MC/GA cycle was determined to be 4.0. An optimized structure/sequence pair (referred to as structure A/sequence A, Figure 7.6) was designed in this cycle that had an \( ew \) of 25.0 applied to the buried positions and a calculated core entropy of 1.6 as described in Chapter 5. The gap between sequence A on structure A and sequence A on any other structure in the 199-member decoy ensemble was 18.5 and the z-score was 0.89. Increasing \( ew \) weights (from 0 to 25) were applied to the core residues for design of 200 sequences using the backbone of structure A; the resulting core sequence diversity ranged from 0.9 to 1.6, as described in Chapter 5. Each ground state sequence was used as input into an algorithm designed to calculate the energy of this
Figure 7.5. The ideal contact cutoff for the buried position calculation in the initial MC/GA cycle was determined to be 4.0. The contact cutoff was defined in Chapter 5.

Figure 7.6. Sequence A and structure A, the optimized sequence/structure pair as described in the text. Core residues are in bold.
sequence on the 199 decoys and structure A. The ground-state structure from each one of these calculations was always structure A, as shown in Table 7.2. The energy gap and z-score of each sequence of different core diversity on structure A did not show any discernable trend as the diversity of that sequence was increased (Table 7.3, Figure 7.7 top and bottom). Additionally, the energy of any of these increasing-core entropy sequences on any of the decoy structures did not correlate with the rmsd of the decoy structures to structure A, as shown in Figure 7.8. These initial results indicate that the initial MC/GA cycle was sufficient to identify the structure and sequence combination of lowest energy. Additionally, increasing sequence diversity weight on the core positions does not correlate with an increase in structural specificity, as measured by the gap or the z-score of the native structure A. However, this result might in part have been due to the inability of the design procedure (which utilized a GA) to maintain both a sequence of high diversity and low energy for the structures involved. Therefore, a good deal of further design, including random MC switches of amino acid identities in the core, is necessary before it can be said that sequence diversity does not truly increase structural specificity. In addition, full sequence design using these studies is necessary, as is experimental characterization of the resultant proteins, to ascertain the dependence of specificity on diversity of the core and entire amino acid sequence.
Table 7.2. The results of placing sequences of increasing core diversity on the target (1st model) and the 199 alternative folds. The target fold was always selected as the fold of lowest energy, independently of the core sequence diversity weight.

<table>
<thead>
<tr>
<th>ew</th>
<th>Lowest E</th>
<th>1st model?</th>
<th>Gap</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>z-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0, 2</td>
<td>1.082</td>
<td>yes</td>
<td>19.44</td>
<td>211.85</td>
<td>202.95</td>
<td>1.038</td>
</tr>
<tr>
<td>5</td>
<td>2.386</td>
<td>yes</td>
<td>19.43</td>
<td>199.38</td>
<td>212.93</td>
<td>0.925</td>
</tr>
<tr>
<td>8-20</td>
<td>2.671</td>
<td>yes</td>
<td>16.96</td>
<td>196.51</td>
<td>193.32</td>
<td>1.003</td>
</tr>
<tr>
<td>25</td>
<td>4.668</td>
<td>yes</td>
<td>18.65</td>
<td>143.47</td>
<td>154.93</td>
<td>0.896</td>
</tr>
</tbody>
</table>

Table 7.3. The energy gap and z-score of each sequence of different core diversity on structure A. Note that no discernable trend in gap or z-score is evident as the core diversity was increased.
Figure 7.7. Increasing sequence diversity of the core residues did not correlate with an increase in structural specificity, as measured by the z-score (top) and energy gap (bottom) of the native structure.
Figure 7.8. The energy of any of the increasing-core entropy sequences on any of the decoy structures did not correlate with the rmsd of the decoy structures to structure A (the target or ground state structure).
CHAPTER 8

WW DOMAIN DESIGN

Results and Discussion

8.1 Wild-Type WW Domain from Human Pin1

The WW domain from human Pin1 (hPin1WW) has the following sequence:

KLPPGWEKRMSRSSGVRYYFNHITNASQWERPSG (Structure in Chapter 7)

where the underlined regions indicate location in a sheet. Two additional residues (GS) were located on the N-terminus of the protein after proteolysis of the fusion system for a molecular weight of 4168 Da and a calculated pI of 10.4. Wild-type WW overexpressed to a sufficient level at 3 hours post-IPTG induction. SDS-PAGE analysis of the supernatant after centrifugation of the cell pellet indicated that the protein was found in the soluble fraction. SDS-PAGE analysis of the Phenyl Sepharose fractions showed a large amount of wild-type WW protein. The last few fractions collected from each column purification were ~95% pure. The fusion system eluted off the HPLC column at 54% and 66% buffer B, although the 54% fraction contained more background proteins. The fusion system resuspended post-lyophilization to a final concentration of 50 µM.
Cleavage of this protein with Nla protease for 2 days at room temperature was sufficient to cleave the fusion system to ~90% completion. Cleaved wild-type WW protein eluted from the HPLC column at 40% buffer B. This sample was clean (better than 90% by SDS-PAGE analysis) and was resuspended post-lyophilization to a final concentration of 410 µM. The MALDI-LR mass spectroscopy results showed two peaks at 4168.18 Da and 4185.02 Da.

Data were collected on wild-type WW in the native state (100 mM NaCl, 10 mM MOPS, pH 7.2, 50 µM protein) to obtain a CD spectrum. The equilibration time was 0.5 minutes in a 2 mm cuvette at 2 °C. The spectrum showed a positive ellipticity with a maximum centered around 227 nm (Figure 8.1, top). The mean residue ellipticity was ~7300 at this point. The pre- and post-thermal denaturation signals at 227 nm had only changed by ~1% in ellipticity; this was a qualitative indication of a reversible thermal denaturation. The ellipticity at 230 nm was followed as a function of temperature (from 2 °C to 98 °C) to obtain a thermal denaturation profile (Figure 8.1, bottom). When this thermal denaturation profile was fitted with Nfit to the Gibbs-Helmholtz equation given earlier, the T_m was determined to be ~55 °C. No further thermodynamic parameters were determined, as only one thermal denaturation profile in this buffer was obtained. An additional CD spectrum was obtained in 10 mM PO_4^{3-}, 100 mM NaCl, pH 7.0 with 50 µM protein. These conditions led to a CD spectrum with a positive ellipticity centered around 226 nm and slightly decreased intensity compared to the spectrum of native WW in MOPS buffer. The shape of the peak was also slightly different, with a greater loss of ellipticity at higher wavelengths of the peak compared to that at lower wavelengths. It is
Figure 8.1. (Top) CD spectrum of wild-type WW; conditions were as described in the text. (Bottom) Thermal denaturation profile of wild-type WW as followed by ellipticity at 230 nm. Data were fitted to a Gibbs-Helmholtz equation with a $\chi^2$ value of 0.005.
unclear why this was so; perhaps the protein structure underwent a rearrangement in the phosphate buffer.

The wild-type WW appeared thermally stable compared to other small proteins. The ~230 nm peak observed here is characteristic of the WW fold\textsuperscript{129,132} and was utilized as a probe for WW-like structure in further designs.

\section*{8.2 SPA-Designed WW Protein}

The sequence prediction algorithm (SPA) was run as mentioned using hPin1 backbone coordinates. The resultant sequence, referred to as SPA-WW1, was:

\begin{verbatim}
DLPDGWKVLVVAGSAVIYFFNEASGIKQKTRPTS
K P EKRMSRS GRV Y HITNAS WE SG
\end{verbatim}

(Figure 8.2) (WW hPin1, 32.4\% identity)

According to the purification requirements, an additional GS was included at the N-terminus of this protein, which would be retained after proteolysis of the fusion system. The intended product had a molecular weight of 3839 Da and a calculated pI of 8.4. The gene coding for the protein of interest developed by nested ligation of overlapping oligonucleotides had a base insertion in the coding region due to improper oligomer purification by PAGE. This extra base was removed by PCR mutagenesis. The resulting protein was overexpressed to a sufficient level at 3 hours post-IPTG induction. SDS-PAGE analysis on the supernatant after centrifugation indicated that the protein was found in the soluble fraction. Several protocols for purification of SPA-WW1 were
Figure 8.2. SPA-WW1 designed sequence and rotamers on the wild-type WW backbone.

The residues colored orange retained the identity of the wild-type WW.
attempted without success. Initially, a native Phenyl Sepharose column was run; a large number of samples were collected, but none of them contained the designed protein. All of the protein that was loaded on to the column was present in the load and wash, indicating that SPA-WW1 had no discernable affinity to Phenyl Sepharose under the conditions chosen. An anionic exchange column was run, as N-Cam.Y has a high affinity for anionic resin. However, neither a “typical” anionic exchange column, using 50 mM Tris and 50 mM NaCl, pH 7.0 as the low-salt wash, nor a low-salt anionic exchange column (10 mM Tris, 10 mM NaCl, pH 7.0 as the wash) led to adequate purification of the designed protein (Figure 8.3). An additional anionic exchange column, this time including 6 M urea for denaturation of the proteins, was run. The protein failed to bind under these conditions as well.

As N-Cam.Y is stable at high temperatures (and therefore a fusion system containing N-Cam.Y may also be thermostable), a heat gradient experiment, which involved heating the samples from 40 °C to 90 °C as mentioned previously, was performed. However, there was a significant amount of background proteins that decreased the usefulness of this method. After multiple purification efforts, the protein was considered to be binding to the N-Cam.Y portion of the fusion system, leading to the loss of affinity for both Phenyl Sepharose and anionic exchange resin. Owing to the probability that this protein, if it was indeed interacting with N-Cam.Y, was not properly folded (or was completely or mostly unfolded), this design was abandoned.
Figure 8.3  (Top) SDS-PAGE protein gel of the attempted purification of SPA-WW1 using anionic exchange.  Note thick band in load + wash (l+w) was protein of interest, exhibiting low binding affinity for the resin.  (Bottom) SDS-PAGE protein gel of the attempted purification of SPA-WW1 using anionic exchange with a reduced ionic strength.  Again, thick band in load + wash was protein of interest; circled bands indicate slightly increased binding affinity of protein to resin in lowered salt concentration.
8.3 Glycine-Baseline WW Design

The preponderance of glycines, an amino acid that confers flexibility in the backbone, in previous WW designs had led to the idea that the Gly baseline was not adequate (greater than 4 Gly in a 34-amino acid protein). A stronger penalty against Gly overselection was implemented by slowly increasing the energetic penalty for selecting Gly via adjustment of the Gly baseline value. The resultant protein, referred to as SPA-G71, had the following sequence:

\[
\text{SLPSGWFLRKSAGAENYYYNKESNEKQATRPTD (Figure 8.4)}
\]
\[
\text{K P EKRMSRSSGFR HIT AS WE SG (WW hPin1, 32.4% identity)}
\]

As for all of the WW proteins, this protein had a GS tag on the N-terminus, leading to a molecular weight of 4054 Da and a calculated pI of 8.3. The only gene obtained for this protein through the methods described in Chapter 6 had a base deletion, due most probably to incomplete oligomer purification prior to gene construction. No further attempts to correct this problem were completed.
Figure 8.4. The glycine-baseline designed sequence and rotamers on the wild-type WW backbone. The residues colored gray retained the identity of the wild-type WW.
8.4 SPANS-WW1

The implementation of the numerous states or backbone ensemble in the algorithm (SPANS) required for experimental verification of its usefulness. The first design, termed SPANS-WW1, had the following sequence:

\[
\text{NLPSGWTPRTKSGASENYYYNKETNEVTNTRPTD} \quad (\text{Figure 8.5})
\]
\[
\text{K P EK MSRSSGRV F HI ASQWE SG} \quad (\text{WW hPin1, 35.3\% identity})
\]

As usual, a GS tag was present on the N-terminus of this protein after cleavage. The resultant protein had a molecular weight of 4036 Da and a calculated pI of 6.3. Overexpression for four hours post-IPTG yielded sufficient amount of soluble protein as evidenced by SDS-PAGE analysis. The protein eluted from the Phenyl Sepharose column at a lower volume of elution buffer than the wild-type WW, which could have indicated a decreased affinity for Phenyl Sepharose. This may have pointed to a molten or unfolded protein that interacted with the N-Cam.Y fusion system. Presence of purified protein was confirmed by SDS-PAGE analysis. About 3 mL of this protein was cleaved with N1α protease for two days at room temperature. The concentration of this sample was 100 µM; another purification with Phenyl Sepharose led to pure SPANS-WW1 at a concentration of 70 µM. This was dialyzed into 100 mM NaCl, 1 mM CaCl₂, 10 mM MOPS, pH 7.2, with a final protein concentration of 60 µM. The remainder was HPLC-purified and eluted from the HPLC column at less than 30-32% buffer B. This resuspended to a final protein concentration of 180 µM. MALDI mass spectroscopy
Figure 8.5. (Top) SPANS-WW1 designed amino acids and rotamers on the wild-type WW backbone. The residues colored gray retained the identity of the wild-type WW. (Bottom) The homologies of the selected amino acids with the WW family; the redder the residue, the more highly it is conserved. The right view shows the same structure as on the left after rotation by 180° about a vertical axis. Note Pro 8, which has never been found in any known WW domains. Also note the low degree of homology of this designed protein in the loop region to the WW family.
showed presence of SPANS-WW1 at 4040.62 Da (the difference from the calculated molecular weight was thought to be caused by inadequate calibration of the instrument), along with protein 275.8 Da larger than the calculated molecular weight for SPANS-WW1. This was probably due to proteolysis by a cellular protease at a site located two residues N-terminal to the Nla protease cleavage site; instead of the GS tag, a tag of FQGS N-terminal to the SPANS-WW1 protein was most likely present. The addition of the FQ residues would have led to a ~270 Da molecular weight increase. This might have also explained why the protein eluted off of the HPLC column in a broad peak: this peak may have actually been two peaks overlapping almost completely. Mass spectroscopy results indicated that almost half of the protein molecules had this additional tag present.

The CD characterization of SPANS-WW1 (50 µM) was performed in a 2 mm cuvette as previously described. Another spectrum, with 0.1-minute equilibration time in 2 nm steps, was obtained for quantitative determination of thermodynamic parameters. The spectra obtained before and after thermal denaturation were almost identical, although slightly decreased in ellipticity (possibly due to degradation of the MOPS buffer). The CD wavelength spectrum looked quite unlike the CD wavelength spectrum of the wild-type WW. It appeared much more random coil-like, with no discernable peak at ~230 nm and with increased negative ellipticity below 210 nm (Figure 8.6, top). The thermal melt was not cooperative and had no true upper and lower baselines. Because of this fact, the thermal denaturation data were not fitted; however, the Tm was most likely below 25 °C, as shown in Figure 8.6 (bottom). Addition of osmolytes (1 M TMAO, 20%
Figure 8.6. (Top) CD spectra of SPANS-WW1 and wild-type WW; buffer conditions were as described in the text. (Bottom) Thermal denaturation profile of SPANS-WW1 (following ellipticity at 230 nm) in the absence of osmolyte.
glycerol, or 40% glycerol), which are thought to destabilize the unfolded state over the
native state via preferential burial of the polypeptide backbone, led to a slight increase
in the ~230 nm peak. However, the ellipticity at this wavelength was still not comparable
to the wild-type WW ellipticity (data not shown). Thermal denaturation with 40%
glycerol yielded no discernable folded baseline. The CD wavelength spectrum in 10 mM
\( \text{PO}_4^{3-} \), 100 mM NaCl, pH 7.0 with no osmolyte showed a slight positive ellipticity
centered around 230 nm, but it was still not analogous to the intensity of the wild-type
WW.

No firm evidence of WW-like structure was obtained. The thermal melt indicated
a low or non-existent population of the native state. Owing to the low solubility of the
protein, no further characterization was performed.

The presence of Pro in the 8th position of SPANS-WW1, along the center of the
first \( \beta \)-strand, was initially a cause for concern. Pro restricts the number of degrees of
freedom of the backbone (\( \phi, \psi \)) angles and is never found in wild-type WW proteins in
the strand regions. However, as will be discussed in section 7, this was probably not the
confounding factor in allowing SPANS-WW1 to adopt a WW-like structure.

8.5 SPANS-WW2

A different treatment of the rotamers using the SPANS algorithm (see Chapter 4)
led to the design of another protein called SPANS-WW2. The sequence of this protein
was:
SLPSGW\_TQLTKASDDTTYYYNKTTDVVT\_NTRPTD (Figure 8.7)

K P EKRMSR SGRV F HITNASQ\_WE S G (WW hPin1, 32.4% identity)

The GS fusion tag was present at the N-terminus of the protein, giving a molecular weight of 3985 Da and a calculated pI of 4.8.

Overexpression of SPANS-WW2 for 3 hours post-IPTG induction led to a sufficient amount of soluble protein as evidenced by SDS-PAGE analysis. The presence of fusion protein in the fractions obtained from Phenyl Sepharose purification was confirmed by SDS-PAGE analysis, which also showed that a band with a molecular weight within error of SPANS-WW2 was present in the load and wash solutions. This pointed toward a small amount of basal proteolytic cleavage of the fusion system prior to addition of protease. 10 mL of the post-Phenyl Sepharose sample was cleaved with NIa protease for two days at room temperature. HPLC purification of this cleavage mixture led to pure SPANS-WW2 (better than 90% by SDS-PAGE analysis) that eluted from the HPLC column at ~37% buffer B. The post-lyophilization pellet resuspended completely in 500 µL distilled water, yielding a concentration of 260 µM. Another preparation of protein yielded 770 µM protein; heating this sample to 90 °C for five minutes and cooling to room temperature resulted in a protein concentration of 1.06 mM. Resuspension of another lyophilized sample of SPANS-WW2 in a limiting amount of distilled water led to a maximum concentration of 1.85 mM. MALDI-LR mass spectroscopy of this protein showed two peaks at 3986.32 Da and 4261.67 Da. The latter peak was most likely due to the additional N-terminal FQ as seen for SPANS-WW1. This may have explained the
Figure 8.7. (Top) SPANS-WW2 designed amino acids and rotamers on the wild-type WW backbone. The residues colored gray retained the wild-type WW identity. (Bottom) The homologies of the selected residues to the WW family as in Figure 8.4; the redder the amino acid, the more highly it is conserved.
basal proteolysis of the fusion system prior to Phenyl Sepharose and Nla protease cleavage. The Nla linker was designed to be flexible for protease accessibility; perhaps there was a background cellular protease that was capable of cleaving this linkage over time. Again, almost half of the protein molecules had these two additional residues present N-terminal to the proteolytic tag. The product in the load and wash after column purification was not analyzed; whether or not the number of protein molecules containing the extra two residues was dependent on how promptly the cells were processed was not determined.

CD spectra of SPANS-WW2 (50 µM) in 10 mM PO_4^{3-} and 100 mM NaCl, pH 7.0 were obtained in a 2 mm cuvette. The averaging time was 7 seconds; the equilibration time was 0.1 minute; the wavelength step was 1 nm. The spectrum at 2 °C prior to thermal denaturation showed a peak with positive ellipticity centered around 231 nm. Following thermal denaturation, the spectrum at 2 °C showed a peak with positive ellipticity centered around 230 nm. Interestingly, the intensity of the peak had actually increased by more than 1.8 mdeg cm^2 dmol^{-1} after thermal denaturation (Figure 8.8). This indicated that the protein was probably misfolded prior to thermal denaturation, and that the fold was corrected with one cycle of annealing. The thermal denaturation prior to annealing had an increased slope in the transition region compared to the thermal denaturation after annealing and a longer baseline at low temperatures. Two thermal denaturation profiles of SPANS-WW2 were obtained for quantitative determination of thermodynamic parameters using an annealed sample of protein; the first monitored the ellipticity as the temperature was raised, whereas the second monitored it as the
Figure 8.8. (Top) CD spectra of SPANS-WW2 before and after annealing assuming protein concentration of 50 µM in both. Note the marked change in mean residue ellipticity at ~230 nm. Buffer conditions were as described in the text. (Bottom) Thermal denaturation profile (unfolding N to U and folding U to N) of annealed SPANS-WW2, following the ellipticity at 230 nm. Data were fitted to the Gibbs-Helmholtz equation given in Chapter 6 with a $\chi^2$ value of 0.003.
(Figure 8.8, cont.) CD spectra of wild-type WW and SPANS-WW2 (after annealing).

Note change in mean residue ellipticity between the two proteins. Actual concentration of SPANS-WW2 was most likely not 50 µM due to the use of unannealed sample; therefore, its calculated MRE is suspect.
temperature was returned to the initial value. These two curves were superimposable, indicating that the process was reversible (Figure 8.8). The thermal stability of this protein was still not as high as that of the wild-type WW; fitting of the thermal denaturation data using Nfit led to a $T_m$ of $\sim 28 \, ^\circ\text{C}$. The absence of a reliable folded baseline prevented the determination of thermodynamic parameters.

Although there was a positive ellipticity at 230 nm for SPANS-WW2 after annealing, it still was not comparable to the wild-type WW. Other studies have shown that the 230 nm peak in the WW domain from human Yes-associated protein is due, at least partially, to the aromatic interactions of the solvent-exposed Trp with the aromatic residues located near it. SPANS-WW2 did not contain the solvent-exposed Trp; this may explain the discrepancy between its spectrum and that of wild-type WW domains. The issue of aromatic interactions leading to an increased 230 nm CD ellipticity will be addressed in a later portion of this chapter. It was also interesting to note that addition of 2 M TMAO to an annealed sample of SPANS-WW2 actually led to loss in ellipticity. Increasing the pH to 8.0 also led to a loss in ellipticity (prior to annealing).

A limited fluorescence spectroscopy study was performed on this protein as well. A spectrum of (unannealed) protein at 1 $\mu$M concentration was collected at 10 °C with no denaturant. Another spectrum was collected at 20 °C in the presence of 5.3 M urea (concentration estimated from the dilution of a fresh stock solution). The temperature was increased for this latter spectrum to avoid urea crystallization. This precaution was later found out to be unnecessary. A $\sim 2.2$-fold decrease in fluorescence intensity coupled with a $\lambda_{\text{max}}$ red-shift by 9 nm was observed upon addition of urea (Figure 8.9, top). A
manual stepwise urea titration of SPANS-WW2 was also performed at 10 °C to a final concentration of 5.7 M urea. As urea was added, a red-shifted peak appeared that decreased and then increased in intensity (Figure 8.9, bottom). The origin of this peak was unclear; most chemical denaturation fluorescence profiles show a red-shifted peak that decreases in intensity with the addition of chaotropic agent. Two hypotheses can be considered: 1) the aggregation state of SPANS-WW2 changed as a function of urea concentration and temperature, or 2) the initial fold was incorrect and rearranged throughout the titration. In view of later experiments, this second hypothesis was favored.

NMR spectroscopy methods were applied to investigate the conformational properties of SPANS-WW2 in solution. The 1D 1H spectrum collected at 25 °C exhibited little chemical shift dispersion and supported an ensemble of rapidly exchanging conformations. Lowering the temperature to 5 °C led to the appearance of resolved signals in the upfield and downfield regions. These signals confirmed the formation of stable secondary and tertiary structure at low temperatures. Based on these results, 2D data were collected at 5 °C to characterize the three-dimensional features of the fold. Homonuclear 2D NMR spectroscopy (NOESY, DQF-COSY, and TOCSY data) provided high-quality information and allowed for the assignment of several residues. For example, Trp 6, the only Trp in this structure, presented a typical pattern of scalar and dipolar connectivities (shown in Figure 8.10) and served as the basis for further analysis. In the aromatic region, other signals were readily attributed to the rings of the
Figure 8.9. (Top) Fluorescence spectra of SPANS-WW2 in 0 M urea (10 °C) and 5.5 M urea (20 °C). SPANS-WW2 had not been annealed prior to the study. Conditions were as described in the text. (Bottom) Fluorescence spectra of SPANS-WW2 at 10 °C with manual titration of urea. As urea was added, $\lambda_{\text{max}}$ became red-shifted and the intensity at $\lambda_{\text{max}}$ first decreased and then increased.
Figure 8.10. The pattern of scalar and dipolar connectivities of the Trp 6 in SPANS-WW2. Conditions were as described in the text.
three tyrosines found at positions 18, 19, and 20. The ring of one of these tyrosines had NOEs to the β protons of one of the others, while the β protons of the third were found in contact with Trp 6. Contacts between the ring protons of two of these Tyrs and the α and indole protons of Trp 6 were present as well. Tentative assignments were derived from these interactions and backbone connectivities, and on the assumption that the structural model held for this conformation of the protein. Also in contact with Trp 6 and, to a lesser extent, Tyr 19 were protons arising from a Pro residue. These protons were shifted upfield, which confirmed their location in the shielding region of the indole ring of Trp 6. Additional upfield shifted resonances (due again to proximity with the Trp ring) were consistent with an Asn side chain; one of the β protons of this Asn was shifted ~2.8 ppm farther upfield than the other β proton, indicating high population of a specific rotamer of this Asn. The α proton of this Asn was in contact with the α proton of Trp 6 and the ring of Tyr 19. By reference to the structural model, these two residues were assigned to Pro 32 and Asn 21. The network of interactions matched the expectations illustrated in Figure 8.11A and 8.11B.

Pro 32 and the three tyrosines could be used for further assignments. Protons arising from a Val residue were in contact with Tyr 18 and Tyr 20. Based upon the structural model, this residue was assigned to Val 27 (Figure 8.11C). Other protons arising from Val 26 were assigned by reference to the model and based upon their NOEs to the indole ring of the Trp. Only one of the γ methyls from Val 26 had an additional NOE to the upfield-shifted β of Asn 21, again indicating high population of specific rotamers of these residues as depicted in Figure 8.11D. Protons with scalar connectivities
consistent with a Leu residue were found in contact with protons from Pro 32 and the Tyr 19 ring as well as one of the β protons from Trp 6; these protons were assigned to Leu 2. One of the δ methyl groups of Leu 2 was in contact with the δ3 proton of Pro 32 and not the δ2 proton of the same Pro, indicating high population of a specific rotamer of this Leu (Figure 8.11E). Additional contacts involving protons from a Pro residue (identified by virtue of its scalar connectivities) were observed for one of the δ methyl groups of Leu 2, the indole protons of the Trp, and protons from Pro 32; based upon the structural model, this new proline was assigned as Pro 3. The α proton of Pro 3 was in contact with the β protons of Pro 32, which was not predicted in the structural model (Figure 8.11F).

Protons with scalar and dipolar connectivities consistent with a Gln residue had NOEs to the protons from Tyr 18 and 19, including a strong NOE between the α protons of Tyr 19 and this Gln. An additional strong NOE was present between the β proton of this Gln residue and Leu 2. Based upon the structural model, these dipolar connectivities suggested assignment to Gln 8 (Figure 8.11G). Additional scalar connectivities consistent with the specific residue types, along with reference to the structural model and dipolar connectivities with the α protons of Tyr 18, assisted in the assignment of Asn 29 (Figure 8.11H). Protons from a Thr were in contact with Trp 6 and the ring protons of Tyr 20. Referring to the model, these protons were assigned to Thr 7. Thr 7 exhibited contacts between its β and backbone amide protons, inconsistent with the rotamer present in the structural model (Figure 8.11I). The protons of Gly 5 were assigned based on its scalar connectivities; neither α proton of Gly 5 was in contact with any other residue. Protons with scalar connectivities consistent with a Lys residue were assigned to Lys 22.
Figure 8.11. Portions of the SPANS-WW2 structural model utilized to assign the resonances discussed in the text.
Figure 8.11 continued. Portions of the structural model utilized to assign the resonances discussed in the text.
based upon their contacts to the ring protons of Tyr 20. Protons from a Thr residue were in contact with the Trp indole ring and the ring of Tyr 19, and were assigned to Thr 28. Protons from another Thr were assigned to Thr 23 based upon the contacts present between its $\alpha$ proton and the amide proton of Lys 22, the weak contacts between this Thr and Asn 21, and the lack of contacts between the Thr side chain and any other proton in the protein (Figure 8.11J).

A total of 13 of the assigned C$\alpha$ proton shifts in the region surrounding Trp 6 were in the 4.5-6.0 ppm range. There were a number of dipolar connectivities between the $\alpha$ proton of one residue and the amide proton of the next residue. Additionally, several of the $\alpha$ protons from individual residues were in contact with $\alpha$ protons on other residues as predicted in the structural model. 1D NMR variable temperature studies (Figure 8.12) clearly showed evidence of the population of the target fold at temperatures below 15 °C, and that ~90% of the molecules populated the folded structure at 5 °C.

These results, as shown in Figures 8.13-16, indicated the presence of a protein containing $\beta$-sheet topology. If the assumptions made in assignment of the protons were correct, the majority of the protein molecules adopted a structure that appeared remarkably similar to the target structure, including (in some cases, where determinable) the predicted rotamers of the amino acids. The region surrounding Pro 3 exhibited the most marked difference between the actual structure and the model; dipolar connectivities existed between protons of this Pro and the protons of Pro 32 that were not bolstered by the proposed structure. However, the dipolar connectivities in the region surrounding Pro 3 indicated that this deviation was not extreme.
Figure 8.12. 1D variable temperature spectra of SPANS-WW2. Note that a significant amount of the protein is unfolded above 20 °C, as evidenced by the absence of peaks in the regions shown. Conditions were as described in the text. The intensities are of arbitrary scale.
Figure 8.13. Backbone NH-CαH section of the NOESY spectrum of SPANS-WW2 in H₂O (the fingerprint region). Peaks of interest are labeled. V27 αH to T28 NH (A); Q8 αH to Y20 NH (B); Y18 αH to Y19 NH (C); W6 αH to T7 NH (D); Y19 αH to Y20 NH (E); G5 α3H to T23 NH (F); G5 α2H to T23 NH (G); N21 αH to K22 NH (H); V26 αH to V27 NH (I) and to N21 NH (J); T20 αH to N21 NH (K). Contacts B, F, G and J indicate that the protein adopts the target conformation. The other contacts are sequential and due to the β-strand secondary structure. All contacts indicate population of a WW-like fold. Conditions were as described in the text.
Figure 8.14. Section of the NOESY spectrum of SPANS-WW2 in D₂O. Peaks of interest are labeled. W6 ζ2H to V26 γ2H₃ (A), V26 γ1H₃ (B) and T28 γ2H₃ (C); W6 ε3H to T28 γ2H₃ (D); W6 η2H to V26 γ2H₃ (E) and γ1H₃ (F), T28 γ2H₃ (G) and V26 βH (H); Y20 δ1H and δ2H to V27 γ2H₃ (I) and γ1H₃ (J); W6 ζ3H to T28 γ2H₃ (K), V26 βH (L) and N21 βH₂ (M); Y20 δ1H and δ2H to V27 βH (N); Y18 δ1H and δ2H to V27 γ2H₃ (O) and γ1H₃ (P), T28 γ2H₃ (Q) and V27 βH (R); Y19 δ1H and δ2H to L2 δ2H₃ (S) and δ1H₃ (T), T28 γ2H₃ (U), Q8 βH₂ (V), Q8 γH₂ (W) and P32 δ2H (X); Y19 ε1H and ε2H to L2 δ2H₃ (Y) and δ1H₃ (Z) and T28 γ2H₃ (AA); Y18 ε1H and ε2H to V27 βH (AB); W6 δ1H to V26 βH (AC); W6 ζ3H to V26 γ1H₃ (AD) and γ2H₃ (AE). All of these contacts indicate the presence of a WW-like fold. Conditions were as described in the text.
Figure 8.15. Section of the NOESY spectrum of SPANS-WW2 in D$_2$O corresponding to contacts between the aromatic amino acids (Trp 6, Tyr 18-20) and some of the C$\alpha$ protons. Four peaks of interest are labeled. Y20 $\delta$1H and $\delta$2H to V27 $\alpha$H (A); Y19 $\delta$1H and $\delta$2H to Y18 $\alpha$H (B); Y19 $\delta$1H and $\delta$2H to Q8 $\alpha$H (C); Y29 $\varepsilon$1H and $\varepsilon$2H to K22 $\alpha$H (D). All four contacts indicate presence of the target structure. Conditions were as described in the text.
Figure 8.16. Section of the NOESY spectrum of SPANS-WW2 in D₂O corresponding to contacts between some of the Cα protons. Peaks of interest are labeled. V27 αH to Y20 αH (A); Q8 αH to Y19 αH (B); Y18 αH to N19 αH (C); N21 αH to W6 αH (D). All four contacts indicate the presence of a WW-like fold. Conditions were as described in the text.
8.6 SPANS-WW3

The wild-type WW backbone was used to seed a molecular dynamics study using parameters as described in Chapter 4. The ensuing backbone ensemble, which had less than 0.3 Å rmsd to the wild-type backbone, was used as input into SPA as described above. The resultant protein, termed SPANS-WW3 had the following sequence:

ELPGWTPRKKDGASEFYYNENTSEVQTPRTD (Figure 8.17)
K P EK MSRSSGRVY F HI NAS WE SG (WW hPin1, 32.4% identity)

As was typical, a GS tag N-terminal to the above sequence was present after proteolytic cleavage with the NIa protease, yielding a protein with a molecular weight of 4052 Da and a calculated pI of 4.8. The protein was soluble and present in sufficient amount when overexpressed for 3 hours post-IPTG induction as indicated by SDS-PAGE analysis. Phenyl Sepharose purification of this protein yielded pure protein (better than 90% by SDS-PAGE analysis). All post-Phenyl Sepharose fractions containing SPANS-WW3 were pooled and cleaved with NIa protease for 3 days at room temperature. HPLC purification yielded protein that eluted from the HPLC column at 26-35% buffer B. Post-lyophilization samples resuspended to a final protein concentration ranging from 1.45-1.61 mM.

CD spectroscopy was performed on SPANS-WW3 (50 µM) in a 2 mm cuvette with 10 µM PO$_4$$^{3-}$, pH 7.0 and 100 mM NaCl. The wavelength spectrum of this protein had a slightly positive peak in ellipticity centered at 230 nm (Figure 8.18, top). The
Figure 8.17. SPANS-WW3 designed sequence and rotamers on the wild-type WW backbone. The residues colored gray are those that retained the wild-type WW identity.
amount of ellipticity at 230 nm was larger than SPANS-WW1 (Figure 8.18) and was comparable to a SPANS-WW1 N17F mutant (as discussed in Section 7), but was not analogous to SPANS-WW2.

Thermal denaturation of this protein showed a noncooperative transition with no upper or lower baselines, as shown in Figure 8.18 (bottom). Lowering the ionic strength of the sample to 10 mM PO$_4^{3-}$, pH 7.0 resulted in no change in the CD spectra (data not shown). These factors indicated that, although some structure may be present as evidenced by the positive ellipticity at 230 nm, the protein molecules did not populate the folded state and the unfolding transition had low cooperativity.

As the main difference between SPANS-WW2 and SPANS-WW3 was the method of designing a flexible backbone, it is tempting to point towards this as the defining characteristic that made SPANS-WW3 a poorly folded protein. However, SPA was designed to locate a sequence that would theoretically adopt a specific fold, whether or not that fold was obtained from a native backbone, a Monte Carlo backbone ensemble or a backbone ensemble obtained through molecular dynamics simulations. Because all of the SPA parameters were identical between the SPANS-WW2 and WW3 designs, it remains difficult to pinpoint any obvious factors that could have played a role in diminishing the stability and structural cooperativity of SPANS-WW3.
Figure 8.18. (Top) CD spectra of wild-type WW, SPANS-WW3 and (for comparison) SPANS-WW1. Note that although the MRE for SPANS-WW3 is more positive at 230 nm than the MRE for SPANS-WW1 at the same wavelength, its value is still much lower than the MRE of the wild-type WW. Conditions were as described in the text. (Bottom) Thermal denaturation profile of SPANS-WW3, following ellipticity at 230 nm.
8.7 Point Mutations to the WW Proteins

As mentioned before, the presence of Pro in the 8th position of SPANS-WW1, which is not located in this region in any protein in the WW family, was cause for concern. Pro 8 was in the middle of the first strand, and was originally hypothesized to restrict the ability of SPANS-WW1 to adopt the WW-like fold. The wild-type WW from human Pin1 has a Lys in position 8. This Lys was mutated to Pro (K8P) via the QuikChange method; the primers were PAGE purified for increased purity. The success rate of mutagenesis was ~50% as evidenced by DNA sequencing of four of the colonies. The resultant protein had a molecular weight of 4137 Da and a calculated pI of 10.3. Harvesting the cells 3 hours post-IPTG yielded a sufficient quantity of soluble protein as evidenced by SDS-PAGE analysis. Phenyl Sepharose purification produced pure protein (better than 90% by SDS-PAGE analysis). Ten mL of the pooled fractions containing K8P was cleaved with N1a protease for 2 days at room temperature to ~90% completion. HPLC purification generated pure protein that eluted from the HPLC column at 32% buffer B. After lyophilization and resuspension, the concentration by UV spectroscopy was 0.46 mM. MALDI-LR mass spectroscopy showed one peak at 4139.35 Da, which corresponded to the calculated molecular weight of K8P; the other peak was present at 4414.99 Da, a 275.64 Da increase from the calculated molecular weight. This was again indicative of possible background cellular proteolytic activity prior to N1a protease cleavage, which would leave an additional pair of residues (FQ) at the N-terminus. Approximately equal proportions of proteins starting with FQGS and GS were obtained.
One CD spectrum was obtained for K8P (100 µM) in 10 mM PO$_4^{3-}$, 100 mM NaCl, pH 7.0 in a 1 mm cuvette. This typical WW-type positive peak was present and centered around 270 nm (Figure 8.19, top). The pre-thermal denaturation ellipticity of K8P was increased from that of the wild-type WW. This was probably due to the increased aromatic interactions in the protein. As Pro 8 was located near the solvent-exposed Trp, it may have partaken in these interactions, leading to an increased ellipticity. This was indicative that the wild-type WW secondary and tertiary structures of this protein were maintained with this mutation. The thermal denaturation of this protein was performed for an approximation of the $T_m$ and fitted with Nfit for a value of 53 °C (Figure 8.19, bottom).

These results indicate that Pro in the 8th position of SPANS-WW1 was not the sole contributor to the loss of WW-like structure and stability of this designed protein. In the SPANS-WW1 protein matrix, P8 could be detrimental; however, in the matrix of the WW native, P8 was not harmful (at least as qualitatively indicated by CD spectroscopy).

Another mutation was made to the SPANS-WW1 protein to remove a potentially detrimental electrostatic interaction. Asn 17, in the rotamer predicted by SPANS, had its partially negative oxygen directly interacting with the partially negative oxygen of Thr 10. Phe was predicted by SPANS, however, to be almost equally energetically favored in this position as compared to Asn. This Asn was mutated to Phe (N17F) via the QuikChange method; the primers were PAGE purified for increased purity. The success rate of mutagenesis was ~100% as determined by DNA sequencing of four of the colonies. N17F had a molecular weight of 4069 Da and a calculated pI of 6.3.
Figure 8.19. (Top) CD spectra of wild-type WW, K8P (mutant of wild-type WW) and (for comparison) SPANS-WW1. All conditions were as described in the text. (Bottom) Thermal denaturation profile of K8P, following ellipticity at 230 nm. The curve was fitted to the Gibbs-Helmholtz equation with a $\chi^2$ value of 0.002. The inset shows the thermal denaturation profile of wild-type WW; note that the overall shape of the curves is conserved.
Harvesting the cells 3 hours post-IPTG induction yielded a sufficient amount of soluble protein for characterization as evidenced by SDS-PAGE analysis. Phenyl Sepharose purification led to the presence of pure N17F (better than 90% by SDS-PAGE analysis). A small amount of this purified protein was cleaved with NLa protease for 2 days at room temperature. This cleaved sample was HPLC-purified and lyophilized; SDS-PAGE analysis indicated the presence of pure, cleaved N17F. The protein pellet resuspended in distilled water post-lyophilization to a final concentration of 1.75 mM.

A CD spectrum of N17F (50 µM) at 2 °C was obtained in 10 mM PO₄³⁻ (pH 7.0) and 100 mM NaCl in a 2 mm cuvette with 7 seconds averaging time, 0.1-minute equilibration time and 2.5 nm bandwidth. This showed a positive peak in ellipticity centered around 230 nm (Figure 8.20, top). This ellipticity had increased noticeably from that of SPANS-WW1, but was still not comparable to the ellipticities of either wild-type WW or SPANS-WW2. The thermal denaturation of N17F showed no discernable baselines and an uncooperative unfolding transition (Figure 8.20, bottom). This data was not fitted and no T_m was estimated.

It is possible that removing the potentially harmful electrostatic interactions between Thr 10 and Asn 17 enhanced the WW-like fold of SPANS-WW1. However, this protein was still not cooperatively folded or thermally stable. Addition of osmolytes may have assisted in obtaining a more WW-like fold; 40% glycerol yielded an increase in positive ellipticity at 230 nm of SPANS-WW1, and may have done so as well to N17F, which was a mutant of SPANS-WW1.

As mentioned before, a portion of the 230 nm positive ellipticity in the CD
Figure 8.20. (Top) CD spectra of wild-type WW, N17F (mutant of SPANS-WW1) and (for comparison) SPANS-WW1. All conditions were as described in the text. (Bottom) Thermal denaturation profile of N17F, following ellipticity at 230 nm.
spectrum of the wild-type WW protein was due to the aromatic interactions of the solvent-exposed Trp 29 and the other aromatics in the region resulting from the tertiary structure of this protein. As this residue was solvent-exposed and not contacting much of the rest of the protein matrix, however, no constraints were placed on the algorithm that would have led to the selection of a Trp in this position. Therefore, the Trp would not have been selected by SPA in position 29, and the wild-type WW aromatic interactions would not have been conserved. It was necessary, therefore, to mutate this solvent-exposed Trp in the WW native to determine if the 230 nm peak was completely lost or if a small amount remained due to the secondary structure specific to the WW domain fold. Ala was selected as a conservative mutation; Ala is not charged and has equal probabilities of being located in solvent-exposed and buried positions. The resulting protein was termed W29A. The Trp was mutated to Ala via the QuikChange method; the primers were PAGE purified for increased purity. The success rate of mutagenesis was ~75%. The resultant protein had a molecular weight of 4052 Da and a calculated pI of 10.4. Harvesting the cells at 3 hours post-IPTG led to the acquisition of a sufficient amount of soluble protein for characterization as evidenced by SDS-PAGE analysis. Phenyl Sepharose purification yielded pure protein (better than 90% by SDS-PAGE analysis). The fractions containing protein were pooled and cleaved with Nla protease for 2 days at room temperature. HPLC purification yielded pure protein at 31 and 36% buffer B as confirmed by SDS-PAGE. After lyophilization, the sample was resuspended in distilled water to a final concentration of 1.97 mM, as determined by UV spectroscopy at 280 nm; mass spectroscopy confirmed the presence of W29A after HPLC purification.
CD spectroscopy of W29A (50 µM) was performed in 10 mM \( \text{PO}_4^{3-} \), pH 7.0 and 100 mM NaCl in a 2 mm cuvette. The CD spectrum showed a significant decrease (by about 7 deg cm\(^2\) dmol\(^{-1}\)) in ellipticity as compared to the native WW. The positive peak was centered on 226 nm and had an ellipticity maximum comparable to SPANS-WW2 as shown in Figure 8.21 (top). Thermal denaturation of W29A showed a transition and baselines comparable to the native WW thermal denaturation. The fit indicated a \( T_m \) of \(~50 \, ^\circ\text{C} \) (Figure 8.21, bottom). No thermodynamic parameters were determined.

The thermal denaturation of W29A indicated a well-folded protein that had a cooperative transition from native to unfolded protein. This was comparable to the wild-type protein, which indicated that the mutant protein had maintained the stability and structure of the wild-type. However, the \(~230 \, \text{nm} \) positive ellipticity had significantly decreased, indicating that the hypothesis of the 230 nm peak being mostly attributable to the aromatic interactions in the native protein was correct (Figure 8.21). The fact that the ellipticity of W29A was comparable to SPANS-WW2 (although blue-shifted) gave good reason to believe that SPANS-WW2 was most likely folded into a WW-type structure when characterized by CD spectroscopy. It also strengthened the reasons for utilizing CD and the 230 nm peak to determine, at least qualitatively, the amount of WW-type fold present in each of the designed proteins.

For cross-comparative purposes, Figure 8.22 provides an overlay of the CD spectra of all of the WW proteins at 2 \(^\circ\text{C} \).
Figure 8.21. (Top) CD spectra of wild-type WW, W29A (mutant of wild-type WW) and (for comparison) SPANS-WW2. All conditions were as described in the text. (Bottom) Thermal denaturation profile of W29A, following ellipticity at 230 nm. Data were fit to the Gibbs-Helmholtz equation with a $\chi^2$ value of 0.025. The inset shows the thermal denaturation profile of wild-type WW; note overall shapes of curves are similar.
Figure 8.22. Overlay of all CD spectra of WW proteins. Conditions were as described in the text.
CHAPTER 9

WW DOMAIN DESIGN

Conclusions

9.1 Secondary Structure Predictions of Designed WW Domains

The sequences of all WW designs and wild-type WW from hPin1 were used as input into PSIPRED to determine a qualitative prediction of the secondary structural elements. The results are shown in Table 9.1. It is interesting to note that only SPANS-WW2 had a small region with predicted helical content; the location and confidence level of amino acids with this predicted helix propensity were similar to the wild-type WW. None of the other designed proteins exhibited this helical propensity. Additionally, SPANS-WW2 had the highest amount of residues with predicted sheet propensity out of all the SPANS designs; SPANS-WW2 had an additional small region of sheet propensity (at residues 26 and 27), which was also present (at residues 27 and 28) in the wild-type WW domain (both with low confidence levels). SPANS-WW1 had the lowest region of residues with predicted sheet propensity, potentially indicating a protein with only a small amount of β-sheet structure.

Based on GenTHREADER, a sequence profile based fold recognition method,
<table>
<thead>
<tr>
<th>Protein</th>
<th>Conf:</th>
<th>Pred:</th>
<th>AA:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild-type WW from hPin1</td>
<td>9887303131068984899974888231017999</td>
<td>CCCCCCHHEEECCCEEEEEEECCCCCEEECCCCCC</td>
<td>KLPPGWEKRMSRSSGRVYYFHITNASQWERPSG</td>
</tr>
<tr>
<td>SPA-WW1</td>
<td>98887078999835547887604446423466889</td>
<td>CCCCCEEEEEEECCCEEEEEEECCCEEECCCCCC</td>
<td>DLPGWKVLVAGSAVIYFFNEASGIKQKTRPTS</td>
</tr>
<tr>
<td>SPANS-WW1</td>
<td>9888877655677655412315432225667889</td>
<td>CCCCCCCCCCCCCCCCCCEEECCCCCCCCCCC</td>
<td>NLPSGWTPTKSGASENYYNKETNEVTNTRPTD</td>
</tr>
<tr>
<td>SPANS-WW2</td>
<td>987510221026776222665031100002577889</td>
<td>CCCCCHHHHHCCCCCCCEEEEEEEECCCEEECCCCCC</td>
<td>SLPSGWTQLTKASDDTYYYYNKTTDVVTNTRPTD</td>
</tr>
<tr>
<td>SPANS-WW3</td>
<td>98888886554676320266537620014677889</td>
<td>CCCCCCCCCCCCCCCCCCEEEEEEEECCCEEECCCCC</td>
<td>ELPSGWTPKRKGASEFFYNYNENTSEVQTRPTD</td>
</tr>
</tbody>
</table>

Table 9.1. PSIPRED results of WW designs and wild-type WW. “Conf” indicates the confidence level of the secondary structure prediction (0-9, where the higher numbers indicate a greater level of confidence in the prediction); “Pred” refers to the actual predicted secondary structure element (C = coil, H = helix, E = sheet); “AA” is the amino acid composition of the protein.
the wild-type protein of highest fold homology to SPANS-WW2 was identified as the hPin1 WW domain (utilizing the high-resolution structure of the entire domain).

SPANS-WW2 had 32.4% identity with the wild-type WW; the score was 0.481. SPANS-WW1 had a 35.2% identity with the same hPin1 wild-type WW; the score was 0.478. SPANS-WW3 had a 32.4% identity with the wild-type WW; the score was 0.486. These latter two proteins were identified as having a predicted fold most like the wild-type WW domain used in this study (in complex with hPin1). In contrast, SPA-WW1 had the highest fold homology with the WW domain in the absence of hPin1. This protein (PDB ID: 1I6C) was synthesized using a peptide synthesizer; 5 additional residues (NSSSG) were placed on the C-terminus. The structure of the synthesized WW domain is not identical to the wild-type WW domain as complexed with hPin1, which was the protein backbone utilized in this study (Figure 9.1). It is interesting to note that Trp 6 in the synthesized WW domain is no longer buried in the hydrophobic core. SPA-WW1 may adopt a fold with high homology to the fold of this synthesized protein; perhaps the rearrangement of the structure of SPA-WW1 yielded a patch of solvent-exposed hydrophobic surface area capable of interacting with the hydrophobic patches exposed by N-Cam.Y in the presence of Ca$^{2+}$.

A BLAST search on the SPANS-WW2 sequence showed no matches; the Pfam alignment did identify the wild-type WW domain from hPin as a candidate for structural homology with an E-value of $1.4 \times 10^{-4}$, indicating a high amount of confidence in the prediction. None of the other designed proteins had any hits after subsequent Pfam searches. Furthermore, BLAST examinations on the other designed
High-resolution NMR structure of synthesized WW domain from hPin1 (not in complex), PDB ID 1I6C.\textsuperscript{142} Note exposure of Trp 6 (in yellow) to solvent. The last 5 residues (which are not present in wild-type hPin1) have been removed.

High-resolution x-ray structure of the WW domain of hPin1 in complex with the isomerase, PDB ID 1PIN.\textsuperscript{132} Trp 6 is in yellow. This backbone was utilized for design throughout the study.

Figure 9.1. Comparison of the high-resolution structures of the synthesized WW domain (not in complex with hPin1) and the WW domain in complex with hPin1.
proteins did not identify any wild-type WW proteins as potential structural similarity matches.

For reference, GenTHREADER predicted the wild-type WW secondary structure to be homologous to that of a dystrophin WW domain fragment from human (score = 0.513). However, it had the second greatest secondary structure homology to itself, with a score of 0.510 and a low confidence level. This indicated that, although GenTHREADER and other secondary structure prediction tools provide a useful qualitative description of the potential overall fold, the results obtained from these tools should not be viewed as a quantitative description of the actual structure of the protein.

9.2 Protein Solubility

It is of interest that the solubility of the proteins in TE, pH 8 at 25 °C covered such a wide range. The solubility limit of the native protein post-cleavage and HPLC purification was 410 µM, and the K8P point mutation to the native protein had a solubility limit of 460 µM. Mutating the exposed Trp 29 to an Ala (W29A) increased the solubility 5-fold and allowed for 1.97 mM samples to be prepared. This increase probably stemmed from the fact that Trp has a higher hydrophobicity than Ala. In comparison, SPANS-WW1 had a solubility of only 180 µM under the same conditions. The N17F point mutation to SPANS-WW1 raised the limit to 1.75 mM. SPANS-WW2 (annealed) and SPANS-WW3 both had high solubilities; the former had a solubility of 1.85 mM, whereas the latter had a solubility of 1.45-1.61 mM. SPANS-WW1 had the
lowest solubility of all of the characterized designed proteins (see Table 9.2). However, the native state of the wild-type WW domain has a low solubility as given above; this is not surprising since this domain contains exposed hydrophobic residues necessary for function. As demonstrated by the data in Table 9.2, there is no correlation between target fold and solubility.

It is worthwhile to note that the concentration limit of SPANS-WW2 increased by at least 1.4-fold after annealing. This pointed to an alternative fold or misfold relative to the intended target fold, populated at room temperature at a pH higher than the pI. It is likely that raising the temperature disrupted this alternative structure and allowed the chain to populate the WW fold upon cooling. This behavior might explain the abnormal fluorescence results. In those studies, the protein was not annealed prior to collecting the data.

The entire annealing process and its necessity is worthy of a closer look by kinetics studies, equilibrium ultracentrifugation, and possibly solid-state NMR. Since the WW domain is a $\beta$-sheet protein, the ability to resolubilize the protein via a heating and cooling cycle is intriguing. Prion proteins, which can become insoluble, are also $\beta$-sheet proteins that form fibrils or amyloids under certain conditions. The fact that the insolubility of the SPANS-WW2 protein was completely reversible indicates that perhaps an amyloid-like structure was being formed that, under extreme heat, decomposed into monomers that then refolded to their soluble, target structure. This alternative fold or misfold may hold some important keys to prion protein insolubility and possible methods for resolubilization.
Table 9.2. Solubilities of each of the WW proteins as calculated with Beer’s law using absorbance at 280 nm in TE, pH 8 at 25 °C. Last column details solubility of proteins or, if not characterized, their status.
9.3 The Usefulness of the Positive Ellipticity at 230 nm by CD

As previously mentioned in Chapter 8, WW domain proteins typically have a positive ellipticity at or around 230 nm. The degree of positive ellipticity depends on the fold; the WW domain from hYap had a very slight 230 nm peak by CD (50 µM protein in 10 mM PO₄³⁻, 100 mM NaCl, 0.1 mM DTT, 0.1 mM EDTA, 0.02% NaN₃, pH 6.0), akin to that of SPANS-WW3 or N17F. The WW domain from hPin1 used in this study had an increased positive ellipticity at 230 nm relative to any of the computationally designed proteins. It is apparent that this 230 nm peak is an important property of WW domain-like folds. As discussed in Chapter 8, the amount of ellipticity of the 230 nm peak in the wild-type protein was due not only to the WW-like fold, but also to the aromatic interactions provided via the tertiary structure of the protein. The WW domain has two Trp residues – one that is solvent-exposed for functional purposes, and one that is buried in the core – three Pro residues, two Tyr residues and a Phe. The Pro residues are clustered around the buried Trp, whereas most of the other aromatics are located on a groove of the protein. The aromatic rings in the native structure are oriented perpendicularly to each other. This geometry may favor energy transfer from one ring to another. When the solvent-exposed Trp was mutated to Ala (W29A), a large portion of this positive ellipticity at 230 nm was lost. It is possible that the remainder of the 230 nm ellipticity was due to the other aromatic rings interacting with each other; however, as this energy transfer is dependent on 2° and 3° structure, it is a useful probe for a WW-like fold.
9.4 Stability of the WW Domain Designs

Of all of the designs, SPANS-WW2 had the most cooperative unfolding transition. However, even this curve had no distinct upper baseline, indicating that the protein could have had a decreased enthalpy or stability as compared to the wild-type WW from hPin1. The wild-type WW and its point mutants (K8P, W29A), on the other hand, had cooperative unfolding transitions and discernable TmS. This protein has a larger number of exposed hydrophobes than the designed proteins – the WW module packs up against the Pin1 isomerase to a certain extent, which accounts for some of its hydrophobic character. It is not a highly charged module; it has sixteen possible hydrogen bonds when viewed by WebLab ViewerLite (Molecular Simulations Inc., San Diego, CA). SPANS-WW1, on the other hand, is slightly more charged and has over nineteen possible hydrogen bonds when viewed by WebLab ViewerLite. It remains unclear why the wild-type protein has a high thermal stability for such a small protein, and why only one of the designed proteins had a discernable population of folded protein. It is here where the next design attempt should focus.

Even in hindsight, the origin of the design flaws is not clear. SPANS-WW2 is rich in Thr residues compared to other water-soluble proteins. This could lead to a loss in structural specificity caused by excessive formation of branched hydrogen bonds. Yet all of the designed proteins were inspected and appeared to have an adequate availability of hydrogen bonds and electrostatic interactions, as well as a high degree of packing in the interior and excellent HP patterning. However, the question is whether the network of
hydrogen bonds is unique, not whether there is a high availability of these bonds. Removing a potentially harmful electrostatic repulsion in SPANS-WW1 did not increase the stability, although a small increase in ellipticity at 230 nm becomes evident. The designed proteins are somewhat related to each other; it is impossible to discern which residues were important for SPANS-WW2 to obtain the target structure, and which should have been mutated to other residues in the SPANS-WW1 and SPANS-WW3 designed proteins. It is here that SPANS, with its probability matrix of rotamers and residues, would come into play in future design efforts. This matrix gives the probability of occurrence of every allowed rotamer and residue combination on the target structure based on its energy and overall Boltzmann probability. If a residue and rotamer combination is selected with a very slightly higher probability than another residue and rotamer combination, the first combination will be chosen for the final design regardless of the rest of the population. The probability matrix could therefore indicate a potentially useful point mutation at such a position; perhaps the parameter weights for SPANS were slightly off, good enough to discern the sequences of low energy but not weighted properly enough to calculate precisely the energy of each sequence. In this case, the residues and rotamers that SPANS labels nearly degenerate in energy and Boltzmann probability would make interesting point mutations to the designed proteins to determine if their structure and stability could be augmented.
9.5 Nuclear Magnetic Resonance Spectroscopy of SPANS-WW2

The initial NMR correlation spectroscopy results from SPANS-WW2 have proved extremely promising. Almost all of the identified NOEs indicated that the experimental structure resembled highly the target structure. In some cases, it was obvious that the algorithm had correctly predicted the actual rotameric state of individual side chains. Although not all of the protein resonances have been assigned, those that have been indicate that the algorithm correctly designed a sequence adopting a WW-like fold. The residues that have not been assigned are primarily located in the loop region, which is typically flexible and is not as important as the core for positive feedback of the usefulness of SPA. The near absence of unexpected NOEs for the residues participating in the core supports that the loop residues do not interact strongly with the rest of the protein.

The NMR assignment process was rendered difficult by the fact that only ~90% of the protein was folded at 5 °C. A few peaks were detected that corresponded to the minor form in slow exchange with the folded form, most notably from Trp 6. The 1D variable temperature studies showed that the conformational entropy of this side chain increased (i.e., the Trp 6 peak broadened and became less dispersed) at higher temperatures. Asn 21, which was significantly shifted downfield because of ring current effects from Trp 6 at 5 °C, lost this interaction when the temperature was increased. It was possible to identify some of the portions of the protein that became unfolded first using variable temperature NMR spectroscopy. However, it was sufficient for this study
to note that a significant amount of the protein’s structure was lost at temperatures above ~25 °C.

9.6 Future Experiments

A number of additional experiments could be performed with this WW system. Several point mutations, including a Gly 13 mutation to SPANS-WW1 (as SPANS-WW2 and the wild-type WW both have Ser in position 13), are in order to clarify the extent of importance of these residues for stability and foldability of the proteins. Additionally, it would be interesting to repeat the fluorescence spectroscopy studies of SPANS-WW2, utilizing annealed protein, for a better approximation of the thermodynamic parameters as well as the stability of SPANS-WW2 in chaotropic solutions. A labeled growth of SPANS-WW2 (\(^{13}\)C, \(^{15}\)N) would assist in assigning NMR resonances to certain parts of SPANS-WW2, and may additionally facilitate a high-resolution NMR structure determination of the protein. The areas of the protein that differ significantly from the proposed structure would be indicative of potential problems; it is in those areas that point mutations could be focused.

To investigate the algorithm’s capability of design, G71 should be expressed and characterized. If this protein cannot be purified, as the SPA-WW1 could not, this may be indicative that the backbone ensemble is useful in designing proteins that at least somewhat adopt the target fold. Additionally, more designs, with different weights on the computational parameters, should be completed to adequately probe each parameter’s
usefulness in the design procedure; inclusion of a secondary structure propensity term in the WW design might also be constructive. Kinetic analysis of each protein’s folding transitions would assist in determining the population or existence of intermediate states. (Although the WW domain from hYap has been described as a 2-state folder, it is not clear if this holds for other native WW domains.)

Additional design experiments would also utilize the backbone of the high-resolution structure of the synthesized WW domain without hPin1 present. This protein has a different structure than the WW domain complexed with hPin1; the environment of Trp 6, for example, is completely different in the synthesized protein. It would also be interesting to obtain a high-resolution structure of the WW domain from hPin1 utilized in this study that was obtained via molecular biology methods; the synthesized WW domain has a 5-residue tag C-terminal to the wild-type WW domain sequence, which, while solvent-exposed and not exhibiting many interactions with the rest of the structure, may or may not have caused the overall fold of the protein to change.

It would also be worthy of note to design a functionally active WW protein. This may be as easy as simply mutating the 29th residue of SPANS-WW2, which is an Asn, to a Trp, and determining the level of activity in the resultant mutant; the three-residue stretch of exposed aromatic residues present in the WW native is maintained in the SPANS-WW2 design. However, other mutational analyses of this position indicate that the solvent-exposed Trp is vital for functional activity in the hPin1 WW matrix. Therefore, functional design would probably involve redesigning the wild-type, keeping the functionally active residues in their current identities and rotamers and designing only
the subset of residues that are not involved in targeting pSer, pThr and Pro residues. A
design module could be implemented into the computational algorithms that would allow
for construction of a subset of residues that provides for the same activity as the WW
native. This is a far-reaching goal, however, as functional design is still in its infancy.

Another future project involves the further study of the off-pathway folding of
SPANS-WW2 using various techniques such as binding of Congo Red. Most prion
proteins are known to exhibit specific binding to Congo Red; the dye binds to the
intermolecular interaction region between antiparallel β strands.\textsuperscript{150,151} When Congo Red
bonds to prion proteins, an orange color is evident by light microscopy as well as green
birefringence dispersion under plane-polarized light due to anisotropic alignment of the
Congo Red molecules upon binding and noncovalent interactions between Congo Red
molecules and the prion protein.\textsuperscript{152} Use of other biophysical methods, such as
equilibrium ultracentrifugation at different pH or salt environments, would yield some
clues as to what leads to the insolubility of SPANS-WW2 at pH conditions above its pI.

In addition to the WW domain work, the three-helix bundle project mentioned in
Chapters 5 and 6 has potential for further improvement. The SpZ project utilized SPA to
obtain a sequence that theoretically codes for the target SpZ structure. A SPANS study
utilizing the same backbone is necessary as well to further explore whether SPA and/or
SPANS is dependent at all on the secondary structure, and whether the algorithms are
applicable to a wide range of protein design problems. Continual determination of the
limitations of the algorithm would assist in perfecting it; this includes redesign of a
variety of target folds.
9.7 Summary

We reported here the first successful fully computational redesign of a β-sheet protein. The design procedure utilized a novel backbone ensemble to mimic backbone flexibility and relieve steric constraints. CD spectroscopy qualitatively indicated the presence of a WW-like fold, the target structure of the design attempt. Initial 1D and 2D NMR spectroscopy confirmed these results. This bodes well for the field of computational protein design. Further attempts at design resulted in a range of successes and indicated several interesting point mutations to increase the extent of accomplishment. In addition, several interesting properties of the designed protein were discovered, including a reversible insolubility formed when the pH is raised above the pI. This has potential consequences in the field of prion proteins, which also form insoluble clusters of protein molecules under certain conditions.

The design of a protein that adopts a specific fold with high stability is not simple. Wild-type proteins are linear heteropolymers comprised of L-amino acids that are guided by various parameters to adopt their native fold. Anfinsen’s hypothesis that sequence dictates fold still holds true with few exceptions, and identifying the various properties of the individual amino acids that cause a specific fold to be adopted and maintained is the true aspiration of computational protein design. Specificity and stability of the fold are all dependent on these parameters; identifying these parameters is a difficult undertaking, and properly determining the importance of each parameter to provide for a stable protein has proven to be complicated. Wild-type proteins adopt a specific structure with certain
stability; some wild-type proteins are disordered for functionality or until binding of a cofactor or other molecule. The fact that very few designed proteins mimic the properties of a native protein proves that protein design has still not accomplished its goals of clarifying the importance and relevance of each parameter to develop a native-like protein.

Many theoretical and experimental studies have elucidated the relative importance of several parameters. This design procedure has utilized the combined applications of the Lennard-Jones potential, the penalties against exposure of hydrophobic surface area and burial of polar atoms that are not in compensating noncovalent interactions with other polar atoms, and an implicit secondary structure term within the steric constraints imposed upon the protein backbone by the side chain identities. SPANS-WW2 has proven to be a protein that specifically adopts the target fold, but is not very stable. This indicates that, while the parameters listed above are important, they may not be properly weighted or complete enough to develop a sequence that codes for a specific fold and maintains this structure at elevated temperatures. The field of protein design is relatively young. Most successes have been in the inverse protein design field, which attempts to design a sequence that folds to a target structure. Continued successes may lead to the ability to design proteins that adopt a certain fold and have a desired functionality; the tools utilized for inverse design may also be applied to the design field, making it possible to predict computationally the structure of a specific sequence. This may benefit the field of proteomics, as well as the high-resolution determination of the structure of novel proteins. This work, along with others in the field, may eventually assist in various
aspects of drug design, bioremediation and other areas that require novel enzymes that maintain a target fold in physiological or extreme conditions.


139. Rance, M., Sorensen, O. W., Bodenhausen, G., Wagner, G., Ernst, R. R. &


142. Wishart, D. S., Bigam, C. G., Yao, J., Abildgaard, F., Dyson, H. J., Oldfield, E.,


146. Wintjens, R., Wieruszeski, J. M., Drobecq, H., Rousselot-Pailley, P., Buee, L.,

147. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. &


APPENDIX A

CspB Hyperthermophilic Variant Design

Introduction, Materials and Methods

Results, Discussion and Conclusions

A.1 Introduction

The cold-shock protein from *Bacillus subtilis* (CspB)\(^1\) is a member of the bacterial stress response protein family.\(^2\) It is a mesophilic β-sheet protein, with a published T\(_m\) value of 60 °C and a ΔG\(^0\) of –3.5 ± 0.1 kcal/mol at 25 °C.\(^3\) It is comprised of five antiparallel strands that make up two subdomains.\(^4\) It is an ideal system for protein design because it is a small protein of 67 amino acids, it has no disulfide bridges or prosthetic groups, and its structure, along with several of its homologues, has been determined through x-ray crystallography\(^4\) (1CSP) and NMR\(^5\) (Figure A.1). It is not very stable but folds in a two-state transition with no populated intermediates observed.\(^1\) Although it does dimerize, this property can be removed through the use of a phosphate buffer.\(^6\) Maintenance of tertiary structure and function can be determined through monitoring the binding activity of the protein to a specific sequence of single-stranded DNA.\(^7\) Finally, a variety of spectroscopic methods can be utilized to determine the extent
Figure A.1. High-resolution x-ray crystallography structure (to 2.50 Å resolution) of the cold-shock protein from *Bacillus subtilis* (CspB), PDB ID 1CSP.⁴
of unfolding of the protein, including UV absorbance, fluorescence and circular
dichroism.\(^8\)

A homologue of CspB from *Thermotoga maritima* (with 66% homology to
CspB) is hyperthermophilic\(^1\); this cold-shock protein (referred to in this work as CspT)
has a published \(T_m\)^9 of \(\sim 87\) °C and a \(\Delta G^\circ\) of approximately \(-6.3\) kcal/mol at 25 °C.\(^1\)
CspT has a strongly increased conformational stability but similar folding kinetics to
CspB.\(^1\) The conserved residues between CspB and CspT are distributed evenly
throughout the protein. The sequences for both are:

CspB    MLEGKVKNSEKPGFGI-EVEGQDDVFVHPSAIQQEGFKTL0EGQAVSFIEVEGRRPGQANVTKEA
CspT     -R      D K  Y   TKD  G-     W   EM      K   V E   Q  KK     H – VVE

To assess the usefulness of the algorithms in selecting for a sequence with high
stability, the sequences of CspB and CspT were used as input for the design procedure.
The algorithms selected an amino acid from either CspB or CspT at every position on the
CspB backbone; the resultant proteins theoretically had an increased stability compared
to that of the mesophilic CspB, while maintaining a measurable affinity for DNA.
Determining which particular amino acids convey thermal and chemical stability to a
related homologue, along with clarifying the least amount of mesophilic-to-
hyperthermophilic mutations that are necessary to preserve this stability, would assist in
guiding the algorithms to select for a sequence that codes for and maintains a specific fold.

A.2 Materials and Methods; Results and Discussion

Two methods existed for assessing the applicability of the algorithms. The first involved a measure of how well the algorithm chose the more appropriate residue for each position from a small set of allowed residues. A secondary structure propensity using sheet preference values\textsuperscript{10} was applied to all locations in the primary structure that had a $\phi$ angle $> 100^\circ$. In total, 18 mutations from CspB to CspT were allowed in the design process. The structure determined to have the lowest energy from the first design cycle had 11 mutations, chosen in more than 10 of the 20 output sequences, from the mesophile to the hyperthermophile out of the allowed 18 (61%). The resultant protein was termed CspB11. The sequence with 6 mutations from CspB to CspT present in all of the output sequences was also selected for synthesis; this was referred to as CspB6.

The second method implemented for assessing the usefulness of the algorithm involved the design of a somewhat \textit{de novo} protein on the CspB backbone, where the program was allowed as input all polar amino acids at locations along the CspB primary structure that were interfacial or fully exposed as determined by a solvent-exposure calculation described in Chapter 4. An increasing hydrophobic exposure penalty was applied as well until the resultant structure appeared more native-like. The structure
chosen by the algorithm to be the most stable was termed novo. The surface aromatics responsible for binding single-stranded DNA were constrained to those of the wild-type CspB to probe the structural specificity of the designed protein through single-stranded DNA binding assays.

These sequences are listed below:

<table>
<thead>
<tr>
<th></th>
<th>cspB</th>
<th>MLEGKVWFNSEKFGFPI-EVEGQDDVFVHSAlAQEGKTL6GQAVSFEIVEGNRPQANVTKEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>cspb</td>
<td>MLEGKVWFNSEKFGFPI-EVEGQDDVFVHSAlAQEGKTL6GQAVSFEIVEGNRPQANVTKEA</td>
<td></td>
</tr>
<tr>
<td>csphyp</td>
<td>-R</td>
<td>D K Y TKD G- W EM K V E G Q KK H- V VE</td>
</tr>
<tr>
<td>cspb11</td>
<td>R</td>
<td>D K Y - W K V K H V</td>
</tr>
<tr>
<td>cspb6</td>
<td>R</td>
<td>- W V H V</td>
</tr>
<tr>
<td>novo</td>
<td>Q Q Q NDQ</td>
<td>- D SQ Q Q N SQ QR K K Q DD Q SQ</td>
</tr>
</tbody>
</table>

As described in Chapter 4, the basic process involved using only the fixed backbone coordinates from the crystallographic structure determination of CspB. The side chains were removed and the amino acid identities of CspB and CspT at each position were used as input for the algorithm. (The deletions present in CspT were not maintained; the amino acid possibilities were wild-type where CspB and CspT were identical or where CspT showed a deletion relative to CspB, and both CspB and CspT where they differed.) A rotamer library of dihedral angles based upon statistical analysis was applied to determine the rotamer possibilities of each side chain. A small amount of flexibility was added by allowing the dihedral angles of the rotamers to vary slightly (± 15°). An initial filtering step was used to eliminate energetically unfavorable side chain rotamer combinations. A genetic algorithm (GA) was then utilized as described in Chapter 4.
A calculation of solvent accessible surface area was coupled with atomic solvation parameters to describe the effect of solvent upon the stability of the model in question. A harsh penalty (+5.0 kcal/mol) was applied when a polar atom was fully buried but not hydrogen bonded. Steric interactions, good packing of the core and ionic interactions also played important roles in the algorithms utilized. All chemicals, equipment and methods were as described in Chapter 6, unless otherwise mentioned.

The insert was constructed for CspB11 using nested ligation of overlapping synthetic nucleotides and cloned into pAED4 that had been digested with NdeI and Clal. This construct was sequenced as described in Chapter 6, expressed in E. coli BL21 and harvested 3 hours post-IPTG induction. CspB11 was initially purified with a strongly acidic cationic exchange column (Econo-Pack S cartridge, BioRad) and a gradient to a final salt concentration of 500 mM NaCl, 25 mM Tris base, pH 7.5. The presence of CspB11 was analyzed with SDS-PAGE; all fractions containing protein were HPLC-purified, lyophilized and resuspended in distilled water. The resultant samples were again analyzed by SDS-PAGE for presence of protein. An additional cationic exchange purification of another sample of CspB11 after cell harvest utilized SP Sephadex, C-25 (Pharmacia) as the resin, but the protein bound weakly. To determine whether heat purification could have been utilized, 100 µL of flow-through from this purification method was centrifuged at high speed, placed on a 65 °C heat block for 10-15 minutes, and centrifuged again to remove all thermally denatured proteins. SDS-PAGE analysis confirmed the presence of protein in the resultant solution. This heat purification method was scaled up and applied to a 1 L growth of CspB11. The protein eluted from the HPLC
column at 36% buffer B; the resultant solution was lyophilized and resuspended to a final concentration of 1.51 mM in TE as determined by UV spectroscopy at 280 nm. (UV spectroscopy of this same protein in 6 M guanidine yielded a final concentration of 0.86 mM at 280 nm.)

CD spectroscopy of 50 μM CspB11 in 50 mM PO₄³⁻, pH 7.5 was performed as described in Chapter 6. The resultant spectrum is shown in Figure A.2 (top, with apologies for the poor quality of the graphic). A guanidine titration of 5 μM CspB11 in 50 mM PO₄³⁻ (Mallinkrodt AR, Philipsburg, NJ), pH 7.5 to a final guanidine concentration of 6.6 M was accomplished utilizing fluorescence spectroscopy on an Aviv ATF 105 fluorometer (Lakewood, NJ) with an excitation wavelength of 280 nm and emission monitored at 352 nm. The guanidine stock was 5 μM CspB11, 50 mM PO₄³⁻, 7.2 M guanidine, pH 7.0 and was added in 20 μL increments with an automatic titrator (Hamilton Microlab 500 Series, Reno, NV) to a 1 x 1 cm cuvette (Hellma) containing a stir bar. Additional chemical denaturation was performed using fluorescence spectroscopy in a buffer containing 50 mM cacodylic acid (Sigma), pH 7.0 under the same fluorescence conditions as described above.

The thermal denaturation of CspB11 by CD at 230 nm (Figure A.2, bottom) was not fitted using the Gibbs-Helmholtz equation as described in Chapter 6 because of the lack of an unfolded baseline. The Tm was estimated to be ≥ 92 °C. Comparison of the Tm of this designed protein to the published values for the Tm's of the mesophile and the hyperthermophile indicated significant increase in thermal stability from the mesophile of ~32 °C¹ and a slight increase from the hyperthermophile of ~5 °C. The chemical
Figure A.2. (Top) CD spectrum of CspB11 at 2 °C. Conditions were as described in the text. The spectrum had to be scanned in, as the original spectrum was lost. (Bottom) Thermal denaturation profile of CspB11 as followed by ellipticity at 230 nm. Conditions were as described in the text. Signal is in millidegrees.
denaturation data obtained by fluorescence at 25 °C (signal in arbitrary units as a function of temperature in Kelvin) was fitted with Kaleidagraph (version 2.1.3, Abelbeck Software, Reading, PA) to the following equation:

$$
\Delta G^\circ (T) = -RT \ln \left( \frac{y_F - y}{y - y_U} \right)
$$

where $y$ is the signal at the emission wavelength, $y_F$ is the signal of the folded protein, and $y_U$ is the signal of the denatured protein. These data, along with the fitting parameters, are shown in Figure A.3. The $\Delta G^\circ$ of CspB11 was determined to be $-3.8 \pm 0.1$ kcal/mol; in comparison, the $\Delta G^\circ$ of wild-type CspB is $-3.5 \pm 0.1$ kcal/mol. The $\Delta G^\circ$ of CspB11 had not increased in relation to CspT, whose $\Delta G^\circ$ is approximately $-6.3$ kcal/mol.

The gene coding for CspB6 was constructed as described above and transformed into BL21. A significant amount of proteolysis was evident after 1-2 hours post-IPTG induction. Heat purification as described above yielded a low concentration of protein, along with a significant amount of background cellular protein as determined by MALDI-LR and electrospray mass spectroscopy. An anionic exchange column was utilized to purify CspB6, with improved results (> 60% purity by SDS-PAGE analysis). HPLC purification yielded protein at 66% buffer B, as determined by SDS-PAGE analysis post-lyophilization. No further biophysical characterization was performed.
Figure A.3. Chemical denaturation of CspB11, monitored using intrinsic tryptophan fluorescence. Conditions were 5.0 μM protein and 50.0 mM KH₂PO₄, pH 7.0. $\lambda_{\text{ex}} = 280$ nm; $\lambda_{\text{em}} = 352$ nm. The fitted curve does not appear smooth because of the resolution of the figure.
Expression and purification of novo resulted in insoluble protein partitioned in inclusion bodies. An inclusion body preparation was performed using urea and Triton X-100 as per standard protocol. The protein remained soluble when dialyzed into dilute (0.6 and 0.3 M) guanidine solution in TE, but the resultant protein could not be purified by HPLC. A gene with N-Cam.Y-Cys-novo identity was constructed for 2-nitro-5-thiocyanobenzoic acid (NTCB, Sigma) cleavage prior to the Cys residue in the presence of chaotropic agent as follows. The previously developed novo construct in pAED4 was PCR-amplified with specific primers to obtain \textit{HinDIII}/Cys-Met-novo/\textit{BamHI} identity. The PCR reaction was digested with \textit{HinDIII} and \textit{BamHI} and purified using a 3.5% NuSieve agarose gel. This was cloned into a complementary digested N-Cam.Y-derived vector that had a \textit{HinDIII} site cloned immediately prior to the terminal \textit{BamHI} site but after the internal \textit{ClaI} site (Figure A.4). This system yielded high amounts of soluble protein, which were applied to a phenyl sepharose column in the presence of calcium as discussed in Chapter 6. The protein showed high affinity to the phenyl sepharose resin, but the samples were not very pure; a significant amount of background protein was detected by SDS-PAGE analysis.

An alternative column purification method involving an anionic exchange column with a salt gradient as described previously in Chapter 6 was utilized as well; the protein did not bind to this resin, as evidenced by SDS-PAGE analysis. A phenyl sepharose column with a 50% EDTA gradient was performed, which yielded a high amount of protein with increased purity (> 75% as shown by SDS-PAGE analysis). All fractions containing protein from this last phenyl sepharose purification were pooled and cleaved
A. Amplification of the gene of interest with primers designed to introduce new bracketing restriction enzyme sites (XXX, YYY) and digestion with these endonucleases.

B. Cloning of a segment DNA into the new vector to introduce a new internal endonuclease site.

The mutant plasmid is then digested with restriction enzymes X and Y, and the insert, which has complementary overhanging base pairs, is ligated into this vector.

Figure A.4. Schematic of a PCR amplification method to introduce new endonuclease sites at either end of an insert (A) for cloning into a plasmid that has been altered to have the same sites (B).
with 25 mM NTCB in the presence of 6 M guanidine as follows: 20 mL of the protein sample was mixed with the appropriate volume of 8 M guanidine and 0.114 g NTCB. This sample was placed at 37 °C for 30 minutes, and the pH was raised to 9.5. It was then incubated at 37 °C for ~16 hours, filtered, and either dialyzed into 1 L of 20 mM Tris, pH 7.5 for one hour for a total of four times or run through a Sephadex G-10 size-exclusion column (Pharmacia) at a flow rate of 2 mL/min. The sample volume from the dialysis step was doubled with 20 mM Tris, pH 7.5, and the resultant reaction mixture was loaded onto an anionic exchange column. All fractions that were positive by SDS-PAGE analysis were injected into the HPLC port, which yielded no discernable protein. The protein sample obtained after size-exclusion chromatography had a peak at 7485.6 Da as evidenced by mass spectroscopy analysis (152.5 Da larger than the calculated molecular weight of novo, which accounted for the NTCB moiety) but had too many background proteins and salts present to be useful for biophysical characterization.

Development of the gene coding for CspB wild-type initially involved cloning the constructed insert into pAED4 as previously described. _KpnI_ was added to the insert:vector ligation reaction to digest any uncut plasmids. (_KpnI_ was present in the polylinker region, which should have been removed if the plasmid was properly cut.) The protein appeared toxic to the cells, as an OD₆₀₀ > 0.2 was never reached.

The previously developed CspB insert in pAED4 was amplified by PCR with specific primers developed to introduce an _NheI_ site prior to and a _BamHI_ site after the insert. N-Cam.Y had an _NheI_ site introduced via PCR mutagenesis immediately prior to the terminal _BamHI_ site but after the internal _ClaI_ site. This new N-Cam.Y was digested
with *NheI* and *BamHI*, and the insert was cloned into this vector. No colony with this gene present was ever obtained. Consequentially, the next attempt utilized the IMPACT-CN system (New England BioLabs), which involved the following: the insert coding for CspB was cloned into two separate plasmids (pTYB2 or pTYB12) digested with *NdeI* and *EcoRI*, (for expression C- or N-terminal to a chitin-binding domain, respectively). This domain, as described in Chapter 2, had self-proteolytic activity in the presence of salts and DTT. A gene coding for CspB C-terminal to the fusion system was obtained and transformed into BL21 at 37 °C as previously described. The cells were induced with IPTG at 30 °C and harvested after 6 hours. 1 L of cell paste post-harvest was resuspended in 50 mL of lysis buffer (20 mM Tris HCl, 1 mM EDTA), a crystal of DNase was added and the samples were frozen. After ~16 hours, the pellets were thawed and adjusted to 500 mM NaCl and 0.1% Triton X-100. This solution was homogenized and centrifuged for 20 minutes, and the supernatant was loaded onto a chitin column that had been equilibrated with column buffer (20 mM Tris HCl, 500 mM NaCl, 1 mM EDTA). After loading was complete, the column was washed with ~10 bed volumes of the column buffer at 2 mL/min. After the absorbance units returned to the pre-load value, the column was quickly flushed with cleavage buffer (20 mM Tris-HCl, 500 mM NaCl, 1 mM EDTA, 50 mM DTT) to promote self-cleavage of the fusion system. The stopcock was closed and the column was allowed to equilibrate for ~16 hours at room temperature. The target protein was washed from the system with column buffer, and SDS-PAGE was utilized to determine presence of protein in the individual fractions. The column was
stripped of the remainder of the bound fusion proteins with stripping buffer (20 mM Tris-HCl, 500 mM NaCl, 1% SDS). No protein was obtained using the IMPACT-CN method.

The pTYB12/CspB and pAED4/CspB vectors were transformed into BL21 concurrently with transformation into ER2566 (New England BioLabs), an E. coli derivative with an F' \( \lambda \) fhuA2 [lon] ompT lacZ::T7 genEl gal sulA11 \( \Delta(mcrC-mrr)114::IS10 \) R(mcr- 73::miniTn10)2 R(zgb-210::Tn10)1 (TetS) endA1 [dcm] genotype. The pTYB12/CspB transformants produced a good amount of protein; the pAED4/CspB transformants produced little (ER2566) to no (BL21) protein. As the ER2566 line had no pLysS or pLysE plasmid, the cells were not lysed upon IPTG induction. Addition of lysozyme after induction was not possible, as lysozyme was not compatible with chitin. pAED4/CspB was transformed into 100 µL of ER2566 and immediately inoculated into 5 mL of LB with Amp. This was shaken at 37 °C for < 12 hours and inoculated into 1 L of LB with Amp. After reaching an OD<sub>600</sub> = 0.5, the cells were induced with IPTG at 25 °C. After 17 hours post-induction, the cells were harvested and resuspended in TE. About 10 mg of lysozyme (USB) was added and the cells were frozen for one hour. A freeze-thaw cycle was utilized for lysis of the cells as follows: an EtOH and dry ice bath was made in a large beaker. The protein sample to be lysed was placed in a sterile 50 mL Falcon tube and submerged in the bath for five minutes. It was then immediately plunged into a 37 °C water bath for five minutes. This cycle of cold and hot baths was repeated three times. The sample was then homogenized, salted and centrifuged as described in Chapter 6. This produced a good amount of soluble protein that was readily purified with an anionic exchange column as shown by SDS-PAGE analysis. The protein was no
longer present after HPLC purification in any lyophilized sample as evidenced by MALDI-LR mass spectroscopy. Heat purification as described above yielded a CspB sample of very low purity (< 5% as shown by SDS-PAGE analysis).

Because of these purification problems, a Cys-(His)$_6$ tag was introduced after the CspB insert in pAED4 by PCR amplification with the proper primers. Once the gene was obtained, it was transformed into BL21 and ER2566 separately and directly inoculated into 5 mL of LB without plating. Again, the ER2566 cell line produced a higher amount of protein than did the BL21 cell line. The protein resulting from the ER2566 growth was desalted with a G10 size-exclusion column using xylene cyanol as a molecular weight standard. After desalting, a Co TALON column was utilized, which purified to a limited extent the protein of interest from the cellular background (approximately 50% as shown by SDS-PAGE analysis). Owing to the lower than expected purity, the sample was run through another size-exclusion column of higher molecular weight cutoff (Superdex 30 prep grade, Pharmacia). A small amount of DTT was added after this purification, but SDS-PAGE and positive electrospray mass spectroscopy analyses indicated a high concentration of dimer relative to monomer due to poor reduction of the intermolecular disulfide linkage. PCR mutagenesis was used to remove the Cys prior to the His tag, but the gene was not utilized for protein expression.
A.3 Conclusions

The identification of 11 residues from the hyperthermophilic homologue of CspB (CspT) out of an allowed 18 that increased the thermal and chemical stabilities relative to the mesophile was quite promising. This protein, termed CspB11, also showed improved thermal stability when compared to the hyperthermophile. It was noted that many of these mutational selections resulted in a higher sheet propensity, which may indicate the necessity for a secondary structure parameter in the design process. Ideally, characterization of CspB6, which selected 33% of the allowed mutations from CspB to CspT, would be necessary to quantify the size of mutations required to convey a larger increase in thermal stability to the mesophilic protein. A related experimental study\textsuperscript{12} utilized mutagenesis to determine the necessary subset of mutations to the CspB matrix from another, thermophilic homologue of CspB (the cold-shock protein from \textit{Bacillus caldolyticus}, whose $T_m$ is 77 °C\textsuperscript{1}). Through inspection of the sequences of these two homologues, two important mutations (G3R and G66L) were identified that increased the $T_m$ of CspB to 74.6 ± 1 °C. This parallel study indicates the residues necessary for stability comprise only a small subset of the nonconserved residues in thermophilic homologues. Additionally, functionality using a DNA binding assay would be imperative, to ascertain whether thermal and chemical stability has been gained at the cost of functional activity.

These results indicate that the algorithm is well-parameterized for the selection of residues important for thermal stability, but is not yet fully capable of identifying those
mutations that would increase chemical stability. Further comparison of the results obtained to the theoretical results will lead to a system of determining the usefulness and subsequent alteration of the algorithm. The ability to increase the transition temperature of a functional protein while still maintaining the same level of functionality has obvious ramifications in the fields of enzyme catalysis and bioremediation.
BIBLIOGRAPHY


APPENDIX B

Fortran 90 Programs Written for the Design of a
Three-Helix Bundle Protein
This is the main program for developing a \textit{de novo} three-helix bundle backbone.

Program MCGA_3helix
Use atomtypes
Use restraints
Use comparisons
Use genetic
Use hel3sub2
Use transformation
Use hel_axis
Use read
Use m_helix
Use random
Use read_dihedrals
Use Builder
Use Energies
Use moves
Use write_pdb
Use write_movie
Implicit none
Integer,parameter::setsize=200
Real,Dimension(setsize,cl)::phi,psi,ome
Real,Dimension(cl,5)::bndlen,bndang
Character,Dimension(setsize,cl)::Sequence
Real,Dimension(setsize,4,cl)::Rotamer
Real,Dimension(cl,7,3)::Backbone,HELIX0,HELIX
Real,Dimension(cl,30,3)::Sidechains
Real::x1,y1,z1,x2,y2,z2
Real,Dimension(setsize,3,3)::d
Integer::helixnum,length,chainstart,chainend
Real,Dimension(setsize,3,4)::anglei
Character*50::rotamerfile
Real,Dimension(20)::STD
Real,Dimension(setsize,2)::Ranked
Real, Dimension(cl,3)::CA
Integer::rnum, mutations, chainlength, opennumber, pos, pos1, pos2, posb, k, i, totalsteps, mcstep
Integer, Dimension(cl)::sstype
Character, Dimension(cl,20)::Use
Character, Dimension(cl)::Native
Character*80::dihedrals, library1, library2, std_state, moviefile, bndfile, helixfile, statfile, outfile
Character*80::paramfile
Character::star, sim*2, ss
Logical, Dimension(cl)::Open, mask=.false., virtual=.false.
Logical::metropolis, input=.true.
Real, Dimension(setsize)::EHB, ESB, ESS, EBB, EST, EBT, ECOM=0.0, Etotal, EI
Real::D
Real::scw, To, Tf, T, ran, mf, bmf, smf, gw, ssw, dis, Ecomstart
Integer::bl, s, lnum, choice, time, best, moviestep, spot, bigloops, start, n, set
Integer, Dimension(20)::List

seed=2*time( )+1 !initialize random number seed
Print*, seed

Call readtypes !reads in atomic parameters and names
Call Read14 !reads in list of nonbonded interaction types (1-2 bond, 3bond, 4 or more)
Call Readstab

Read(5, fmt='(i3)') opennumber
Do i=1, opennumber
Read(5, fmt='(i3, a1, a1, a1, 2x, 20(a1, 1x))') pos, star, ss, Native(i), (Use(pos, k), k=1, 20)
If(star.eq.'*') mask(pos)=.true.
Select Case(ss)
  Case('h')
    sstype(i)=1
  Case('s')
    sstype(i)=2
  Case('t')
    sstype(i)=3
Case Default
   sstype(i)=4
   End Select
Open(pos)=.true.
End Do

Read(5,fmt='(i10)')totalsteps
Read(5,fmt='(a80)')helixfile
Read(5,fmt='(i3)')start
Read(5,fmt='(i3)')n
Read(5,fmt='(f8.1)')To
Read(5,fmt='(f8.1)')Tf
Read(5,fmt='(a80)')library1
Read(5,fmt='(a80)')library2
Read(5,fmt='(a80)')std_state
Read(5,fmt='(a80)')moviefile
Read(5,fmt='(a80)')outfile
Read(5,fmt='(f4.3)')bmf
Read(5,fmt='(f4.3)')smf
Read(5,fmt='(f5.3)')ssw
Read(5,fmt='(i3)')bigloops
Read(5,fmt='(a80)')paramfile
If(paramfile(1:3).eq.'   ')input=.false.

statfile=moviefile !name statistics file based on moviefile prefix
Do k=1,80
   If(statfile(k:k+2).eq.'xyz')statfile(k:k+3)='stat'
End Do

moviestep=INT(bigloops*totalsteps/1000)

chainlength=opennumber
Open(33,file=moviefile,status='replace')
Open(44,file=outfile,status='replace')
Open(34,file=statfile,status='replace')
If(input)then
Open(42,file=paramfile,status='old')
   Do helixnum=1,3
      Read(42,fmt='(3x,3f9.3)')(d(1,helixnum,k),k=1,3)
      Read(42,fmt='(3x,4f9.3)')(anglei(1,helixnum,k),k=1,4)
   End Do
End If

Call Readrotamers(library1,Open,Use,opennumber)
Call Calculate_STDS(library1,std_state,STD)

phi=-60.0
psi=-40.0
ome=180.0
sequence='A'

Call Amber_bonds(bndlen,bndang)
length=opennumber/3.0
Call Buildbackbone(bndlen,bndang,phi(1,:),psi(1,:),ome(1,:),length,Sequence(1,:),HELIX0)

!obtain helix axis; move this onto z-axis and move c.o.m. to origin
Call helix_axis(length,start,n,x1,y1,z1,x2,y2,z2,HELIX0)
Call move_helix(x1,y1,z1,x2,y2,z2,length,HELIX0)

If(.not.input)then
   Do s=1,setsize
      !generate random differences in helix parameters (translation, angles, etc.); store the info
      Do helixnum=1,3
         anglei(s,helixnum,1)=100.0*(ran(seed)-0.5)
         anglei(s,helixnum,2)=10.0*(ran(seed)-0.5)
         anglei(s,helixnum,3)=10.0*(ran(seed)-0.5)
         d(s,helixnum,1)=0.0+2.0*(ran(seed)-0.5)
      End Do
   End Do
End If
d(s,helixnum,2)=6.0+2.0*(ran(seed)-0.5)
end do
anglei(s,2,3)=180.0+10.0*(ran(seed)-0.5)
anglei(s,1,4)=0.0+10.0*(ran(seed)-0.5)
anglei(s,2,4)=120.0+10.0*(ran(seed)-0.5)
anglei(s,3,4)=240.0+10.0*(ran(seed)-0.5)
!
use 3helix program to generate a backbone (all three helices)
call helix3sub2(HELIX0,Backbone,anglei(s,:,:),d(s,:,:),length)
!
use backbone to run a fixed backbone genetic algorithm
call FixedGA2(1.0,1,Mask,chainlength,Backbone,Sequence(s,:),Rotamer(s,:),STD,ssw,sstype,.false.)
call Buildsidechains(Backbone,chainlength,Sequence(s,:),Rotamer(s,:),Sidechains)
call Backbone_Energies(mask,chainlength,Backbone,ome(s,:),EBB(s),EBT(s))
call Sidechain_Energies(mask,chainlength,Sequence(s,:),ssstype,Rotamer(s,:),Sidechains,Backbone,&
ESB(s),ESS(s),EST(s),EI(s),EBH(s))
Etotal(s)=EBB(s)+EBT(s)+ESB(s)+EI(s)+ESS(s)+EST(s)+EBH(s)
do pos=1,chainlength
etotal(s)=etotal(s)-STD(rnum(Sequence(s,pos)))-ssw*sstab(sstype(pos),rnum(Sequence(s,pos)))
end do
print*, 'Total energy = ', etotal(s), ESS(s)
!
write movie file for each of trial structures
atoms=0
do pos=1,chainlength
  call Mainmovie(1,Sequence(s,pos),Backbone(pos,:,:),pos,33,atoms)
call Sidemovie(1,Sequence(s,pos),Sidechains(pos,:,:),pos,33,atoms)
end do
write(33,fmt=('a15,i4,3x,a4,f7.2'))'trial structure',s,'E = ',etotal(s)
write(33,fmt=('i12'))'atoms
do pos=1,chainlength
  call Mainmovie(2,Sequence(s,pos),Backbone(pos,:,:),pos,33,atoms)
call Sidemovie(2,Sequence(s,pos),Sidechains(pos,:,:),pos,33,atoms)
Call Readrotamers(library2,Open,Use,chainlength)

Do bl=1,bigloops
  scw=1.0
  !ssw=5.0-(5.0-1.0)*(bl-1)/bigloops
  T=To-(To-Tf)*mcstep/totalsteps
  Call helix3sub2(HELIX0,Backbone,anglei(1,:,:),d(1,:,:),length)
  Print*,'Genetic algorithms, round',bl
  Call FixedGA2(0.0,10,Mask,chainlength,Backbone,Sequence(1,:),Rotamer(1,:,:),STD,ssw,sstype,.false.)
  Call Buildsidechains(Backbone,chainlength,Sequence(1,:),Rotamer(1,:,:),Sidechains)
  Call Backbone_Energies(mask,chainlength,Backbone,ome(1,:),EBB(1),EBT(1))
  Call Sidechain_Energies(mask,chainlength,Sequence(1,:),sstype,Rotamer(1,:,:),Sidechains,Backbone,& ESB(1),ESS(1),EST(1),EI(1),EHB(1))
  Etotal(1)=EBB(1)+EBT(1)+ESB(1)+EI(1)+2*ESS(1)+EST(1)
Do pos=1,chainlength
E_{\text{total}}(1) = E_{\text{total}}(1) - \text{STD}(\text{rnum(Sequence}(1, \text{pos}))) - ssw \times \text{sstab}(\text{sstype}(\text{pos}), \text{rnum(Sequence}(1, \text{pos})))

\text{End Do}

mf = 0.0
sim = 'mc'

\text{Write(34, fmt='(i3,1x,a2,1x,2f8.1,1x,60a1)') bl,'GA',E_{\text{total}}(1),E_{\text{SS}}(1),(Sequence(1,k),k=1,\text{chainlength})}
\text{Write(6, fmt='(i3,1x,a2,1x,2f8.1,1x,60a1)') bl,'GA',E_{\text{total}}(1),E_{\text{SS}}(1),(Sequence(1,k),k=1,\text{chainlength})}
\text{Print*,E_{\text{total}}(1),E_{\text{SS}}(1)}

! set up for monte carlo
\text{Sequence(2,:)=Sequence(1,:)}
\text{Rotamer(2,:,:)=Rotamer(1,:,:)}
\text{d(2,:,:)=d(1,:,:)}
\text{anglei(2,:,:)=anglei(1,:,:)}

\text{Print*,'Monte Carlo Simulation, round',bl}
\text{Do mcstep=1,totalsteps}
\text{T=To-(To-Tf)*mcstep/totalsteps}
\text{Print*,mcstep,'T=',T}

!! generate random helix parameters to make a test structure
\text{Do helixnum=1,3}
\text{If(ran(seed).lt.bmf)d(2,helixnum,1)=d(2,helixnum,1)+0.5*(ran(seed)-0.5)}
\text{If(ran(seed).lt.bmf)d(2,helixnum,2)=d(2,helixnum,2)+0.5*(ran(seed)-0.5)}
\text{If(ran(seed).lt.bmf)d(2,helixnum,3)=d(2,helixnum,3)+0.5*(ran(seed)-0.5)}
\text{If(ran(seed).lt.bmf)anglei(2,helixnum,1)=anglei(2,helixnum,1)+3.0*(ran(seed)-0.5)}
\text{If(ran(seed).lt.bmf)anglei(2,helixnum,2)=anglei(2,helixnum,2)+3.0*(ran(seed)-0.5)}
\text{If(ran(seed).lt.bmf)anglei(2,helixnum,3)=anglei(2,helixnum,3)+3.0*(ran(seed)-0.5)}
\text{If(ran(seed).lt.bmf)anglei(2,helixnum,4)=anglei(2,helixnum,4)+3.0*(ran(seed)-0.5)}
\text{End do}

\text{Call helix3sub2(HELIX0,Backbone,anglei(2,:),d(2,:),length)}
! run GA on backbone - just do 1 ga
Call Buildsidechains(Backbone, chainlength, Sequence(2,:), Rotamer(2,:,:,:), Sidechains)

!! evaluate test structure
Call Backbone_Energies(mask, chainlength, Backbone, ome(2,:), EBB(2), EBT(2))
Call Sidechain_Energies(mask, chainlength, Sequence(2,:), sstype, Rotamer(2,:,:,:), Sidechains, Backbone, &
ESB(2), ESS(2), EST(2), EI(2), EHB(2))

Etotal(2)=EBB(2)+EBT(2)+ESB(2)+EI(2)+2*ESS(2)+EST(2)

Do pos=1, chainlength
Etotal(2)=Etotal(2)−STD(rnum(Sequence(2,pos))−ssw*sstab(sstype(pos),rnum(Sequence(2,pos)))
End Do
Print*, 'test E = ', Etotal(2)

If(metropolis(Etotal(2),Etotal(1),T)) then
Etotal(1)=Etotal(2) ! update structure info if accepted
Sequence(1,:)=Sequence(2,:)
Rotamer(1,:,:)=Rotamer(2,:,:)
ESS(1)=ESS(2)
d(1,:,:)=d(2,:,:)
anglei(1,:,:)=anglei(2,:,:)
Else ! reset values if not accepted
Sequence(2,:)=Sequence(1,:)
Rotamer(2,:,:)=Rotamer(1,:,:)
d(2,:,:)=d(1,:,:)
anglei(2,:,:)=anglei(1,:,:)
End If

Print*, mcstep, 'Current E = ', Etotal(1), 'ESS = ', ESS(1)
Write(6, fmt='(i3,1x,a2,1x,2f8.1,1x,60a1)') bl, 'GA', Etotal(1), ESS(1), (Sequence(1,k), k=1, chainlength)

If(Real(mcstep)/moviestep.eq.INT(Real(mcstep)/moviestep)) then
! make three helices based on current helix parameters
Call helix3sub2(HELIX0, Backbone, anglei(1,:,:), d(1,:,:), length)
Call Buildsidechains(Backbone, chainlength, Sequence(1,:), Rotamer(1,:), Sidechains)

!!write movie file
atoms=0
Do pos=1, chainlength
Call Mainmovie(1, Sequence(1, pos), Backbone(pos,:,:), pos, 33, atoms)
Call Sidemovie(1, Sequence(1, pos), Sidechains(pos,:,:), pos, 33, atoms)
End Do
Write(33, fmt='(a2, i3, lx, a8, i5, 3x, a4, f7.2)') 'bl', bl, 'mc cycle', mcstep, 'E = ', Etotal(1)
Write(33, fmt='(i12)') atoms
Do pos=1, chainlength
Call Mainmovie(2, Sequence(1, pos), Backbone(pos,:,:), pos, 33, atoms)
Call Sidemovie(2, Sequence(1, pos), Sidechains(pos,:,:), pos, 33, atoms)
End Do

!!update pdb file
Open(45, file='temp.pdb', status='replace')
atoms=0
Do pos=1, chainlength
Call Writemain(Sequence(1, pos), Backbone(pos,:,:), pos, 45)
Call Writeside(Sequence(1, pos), Sidechains(pos,:,:), pos, 45)
End Do
Write(45, fmt='(a3)') 'TER'
Close(45)
End If
End Do  ! end mc loop

Write(34, fmt='(i3, lx, a2, 1x, 2f8.1, 1x, 60a1)') bl, 'MC', Etotal(1), ESS(1), (Sequence(1, k), k=1, chainlength)

!build final structure
!make three helices based on final helix parameters
Call helix3sub2(HELIX0, Backbone, anglei(1,:,:), d(1,:,:), length)
Call Buildsidechains(Backbone, chainlength, Sequence(1,:), Rotamer(1,:), Sidechains)

Write(44, '(a5,6x,i3)') 'MODEL', bl
atoms=0
Do pos=1, chainlength
Call Writemain(Sequence(1,pos), Backbone(pos,:), pos, 44)
Call Writeside(Sequence(1,pos), Sidechains(pos,:), pos, 44)
End Do
Write(44, fmt='(a3)') 'TER'
Write(44, fmt='(a6)') 'ENDMDL'

End Do !end big loop

End Program MCGA_3helix
This is the main module for MCGA_3helix. It calls all of the subroutines for helix manipulations.

Module hel3sub2
Implicit none
Contains
Subroutine helix3sub2(HELIX0,HELIX3,angle,d,length)
Use transformation
Use hel_axis
Use m_helix
Implicit none
Real,dimension(:,:,:):::HELIX0,HELIX3
Real,dimension(size(helix0,1),size(helix0,2),size(helix0,3))::HELIX
Real,dimension(:,:):::angle
Integer::i,helixnum,length
Real,dimension(:,:):::d
   do helixnum=1,3
     HELIX=HELIX0
     call rotate(HELIX,angle(helixnum,1),'z')
     call rotate(HELIX,angle(helixnum,2),'x')
     call rotate(HELIX,angle(helixnum,3),'y')
     call translation(HELIX,d(helixnum,1),d(helixnum,2),d(helixnum,3))
     call rotate(HELIX,angle(helixnum,4),'z')
     if (helixnum.eq.1) then
       HELIX3(1:length,:,:)=HELIX(1:length,:,:)
     end if
     if (helixnum.eq.2) then
       HELIX3(length+1:2*length,:,:)=HELIX(1:length,:,:)
     end if
     if (helixnum.eq.3) then
       HELIX3(2*length+1:3*length,:,:)=HELIX(1:length,:,:)
     end if
   end do
end subroutine helix3sub2
end module hel3sub2
This module determines the helix axis.

Module hel_axis
Implicit none
Contains
Subroutine helix_axis(length,start,n,x1,y1,z1,x2,y2,z2,HELIX)
Implicit none
Integer::k,length
Integer::start,end,start2,end2,n
Real::x,y,z,x1,y1,z1,x2,y2,z2
Real,Dimension(:,:,:)::HELIX
Real,dimension(3)::new,new2

end=start+3
start2=n-1
end2=n+2

    do k=1,3
        new(k)=SUM(HELIX(start:end,1:4,k))/16
    end do
    do k=1,3
        new2(k)=SUM(HELIX(start2:end2,1:4,k))/16
    end do

x1=new(1)
y1=new(2)
z1=new(3)
x2=new2(1)
y2=new2(2)
z2=new2(3)

end subroutine helix_axis

end module hel_axis
This module contains all of the matrices necessary for rotation around all three axes.

Module transformation
Implicit none
Contains
Subroutine rotate(coords, angle, axis)
Implicit none
Real, dimension(:, :, ::) :: coords
Real :: angle
Character :: axis
Integer :: i, j, pos
Real, dimension(3, 3) :: matrix1

! Initialize matrix1
10 select case(axis)
    case ('z', 'Z')
        matrix1(1, 1) = cosd(angle)
        matrix1(1, 2) = sind(angle)
        matrix1(1, 3) = 0
        matrix1(2, 1) = -1 * sind(angle)
        matrix1(2, 2) = cosd(angle)
        matrix1(2, 3) = 0
        matrix1(3, 1) = 0
        matrix1(3, 2) = 0
        matrix1(3, 3) = 1
    case ('x', 'X')
        matrix1(1, 1) = 1
        matrix1(1, 2) = 0
        matrix1(1, 3) = 0
        matrix1(2, 1) = 0
        matrix1(2, 2) = cosd(angle)
        matrix1(2, 3) = sind(angle)
        matrix1(3, 1) = 0
        matrix1(3, 2) = -1 * sind(angle)
        matrix1(3, 3) = cosd(angle)
case('y','Y')
    matrix1(1,1) = cosd(angle)
    matrix1(1,2) = 0
    matrix1(1,3) = -1*sind(angle)
    matrix1(2,1) = 0
    matrix1(2,2) = 1
    matrix1(2,3) = 0
    matrix1(3,1) = sind(angle)
    matrix1(3,2) = 0
    matrix1(3,3) = cosd(angle)
    case default
        print*, 'Incorrect axis designation!'
        print*, 'Try again.'
        read(5, fmt='(a1)') axis
        goto 10
    end select
    do pos=1, size(coords,1)
        do i=1, size(coords,2)
            coords(pos,i,:) = MATMUL(matrix1, coords(pos,i,:))
        end do
    end do
end subroutine rotate

Subroutine translation(coords, dx, dy, dz)
    Implicit none
    Real::dx, dy, dz
    Real, dimension(:, :, ::)coords
    ! Initialize translation
    coords(:, :, 1) = coords(:, :, 1) + dx
    coords(:, :, 2) = coords(:, :, 2) + dy
    coords(:, :, 3) = coords(:, :, 3) + dz
end subroutine translation

end module transformation
This module moves the helix axis so that it overlays the z-axis. Additionally, it moves the center of mass of the helix (using heavy atoms only) to the origin.

Module m_helix
Implicit none

Contains

Subroutine move_helix(x1,y1,z1,x2,y2,z2,length,HELIX)
Use transformation
Implicit none
Real::dx,dy,dz
Character::x,z
Real::x2,x1,y2,y1,z2,z1
Real,dimension(3)::newcoords
Real,dimension(:,:,:)::HELIX
Real::theta,phi
Integer::i,length

! Determine the projection of the helix vector on the xy-plane

dx = x2-x1
dy = y2-y1
dz = z2-z1

   HELIX(:,:,1)=HELIX(:,:,1)-x1
   HELIX(:,:,2)=HELIX(:,:,2)-y1
   HELIX(:,:,3)=HELIX(:,:,3)-z1

! Now obtain the angles between the y-axis (phi) and the z-axis (theta)

theta = atand((SQRT((dx**2)+(dy**2)))/dz)
phi = atand(dx/dy)
! Initiate rotation loops

call rotate(HELIX,phi,'z')
call rotate(HELIX,theta,'x')

! Move center of mass of helix to the origin

newcoords(1) = (SUM(HELIX(1:length,1:4,1)/(length*4)))
newcoords(2) = (SUM(HELIX(1:length,1:4,2)/(length*4)))
newcoords(3) = (SUM(HELIX(1:length,1:4,3)/(length*4)))

HELIX(1:length,:,1) = (HELIX(1:length,:,1) - newcoords(1))
HELIX(1:length,:,2) = (HELIX(1:length,:,2) - newcoords(2))
HELIX(1:length,:,3) = (HELIX(1:length,:,3) - newcoords(3))

end subroutine move_helix

derm module m_helix
This program generates random movements of the backbone angles.

Program mcbackbone
Use atomtypes
Use restraints
Use comparisons
Use genetic
Use random
Use read_dihedrals
Use Builder
Use Energies
Use moves
Use write_pdb
Use write_movie
Implicit none
Integer, parameter:: setsize=2
Real, Dimension(setsize, cl):: phi, psi, ome
Real, Dimension(cl, 5):: bndlen, bndang
Character, Dimension(setsize, cl):: Sequence
Real, Dimension(setsize, 4, cl):: Rotamer
Real, Dimension(cl, 7, 3):: Backbone
Real, Dimension(cl, 30, 3):: Sidechains
Real, Dimension(20):: STD, ASTD
Real, Dimension(setsize, 2):: Ranked
Real, Dimension(cl, 3):: CA
Integer:: rnum, mutations, chainlength, opennumber, pos, pos1, pos2, posb, k, i, totalsteps, mcstep
Integer, Dimension(cl):: sstype
Character, Dimension(cl, 20):: Use
Character, Dimension(cl):: Native
Character*80:: dihedrals, library1, library2, std_state, moviefile, bndfile, stdlib, statfile, outfile
Character:: star, sim*2, ss
Logical, Dimension(cl):: Open, mask=.false., virtual=.false.
Logical:: metropolis
Real, Dimension(setsize):: ESB, ESS, EBB, EST, EBT, ECOM=0.0, Etotal, EI, EHB
Real::To,Tf,T,ran,mf,bmf,smf,gw,D,dis,Ecomstart
Integer::s,lnum,choice,time,best,moviestep,spot,bigloops,bl
Integer,Dimension(20)::List

seed=2*time( )+1   !initialize random number seed
Print*,seed

Call readtypes !reads in atomic parameters and names
Call Read14   !reads in list of nonbonded interaction types (1-2 bond, 3bond, 4 or more)
Call Readsstab

Read(5,fmt='(i3)')opennumber
chainlength=opennumber
Do i=1,opennumber
Read(5,fmt='(i3,a1,a1,a1,2x,20(a1,1x))')pos,star,ss,Native(i),(Use(pos,k),k=1,20)
If(star.eq.'*')mask(pos)=.true.
   Select Case(ss)
   Case('h')
   sstype(i)=1
   Case('s')
   sstype(i)=2
   Case('t')
   sstype(i)=3
   Case Default
   sstype(i)=4
   End Select
   Open(pos)=.true.
End Do

Read(5,fmt='(i10)')totalsteps
Read(5,fmt='(f8.1)')To
Read(5,fmt='(f8.1)')Tf
Read(5,fmt='(a80)')dihedrals
Read(5,fmt='(a80)')library1
Read(5, fmt='(a80)') library2
Read(5, fmt='(a80)') std_state
Read(5, fmt='(a80)') moviefile
Read(5, fmt='(a20)') outfile
Read(5, fmt='(f5.3)') bmf
Read(5, fmt='(f4.3)') smf
Read(5, fmt='(f5.3)') ssw
Read(5, fmt='(f5.3)') hw
bigloops=20
moviestep=INT(bigloops*totalsteps/1000)

statfile=moviefile ! name statistics file based on moviefile prefix
Do k=1,80
If(statfile(k:k+2).eq.'xyz') statfile(k:k+3)='stat'
End Do

Open(33, file=moviefile, status='replace')
Open(44, file=outfile, status='replace')
Open(34, file=statfile, status='replace')

Call Readphipsi(dihedrals, phi(1,:), psi(1,:), ome(1,:), chainlength)
Call Amber_bonds(bndlen, bndang) ! set default bond lengths and angles

Call Readrotamers(library1, Open, Use, chainlength) ! read in initial rotamer configuration
stdlib='/wallace/people/johnd/fcode/protein_data/library.dat'
Call Calculate_STDS(stdlib, std_state, STD)
Call Jump_Sidechains(1.0, chainlength, Sequence(1,:), Rotamer(1,:,:), 0.0)
!(initialize with randomly chosen side chains)

phi(2,:)=phi(1,:); psi(2,:)=psi(1,:); ome(2,:)=ome(1,:)
Sequence(2,:)=Sequence(1,:)
Rotamer(2,:,:)=Rotamer(1,:,:)
Call Readrotamers(library2,Open,Use,chainlength) !read in initial rotamer configuration

Do bl=1,bigloops
If(bl.gt.1)then
Call FixedGA(10,Mask,chainlength,Backbone,Sequence(1,:),Rotamer(1,:),STD,ASTD,sstype,.false.)
End If
Call Buildbackbone(bndlen,bndang,phi(1,:),psi(1,:),ome(1,:),chainlength,Sequence(1,:),Backbone)
Call Buildsidechains(Backbone,chainlength,Sequence(1,:),Rotamer(1,:),Sidechains)
Call Backbone_Energies(mask,chainlength,Backbone,ome(1,:),EBB(1),EBT(1))
Call Sidechain_Energies(mask,chainlength,Sequence(1,:),sstype,Rotamer(1,:),Sidechains,Backbone,&
ESB(1),ESS(1),EST(1),EI(1),EHB(1))

Etotal(1)=EBB(1)+EBT(1)+ESB(1)+EI(1)+ESS(1)+EST(1)+hw*EHB(1)
Do pos=1,chainlength
Etotal(1)=Etotal(1)-STD(rnum(Sequence(1,pos)))-ssw*sstab(sstype(pos),rnum(Sequence(1,pos)))
End Do
Print*,'Starting energy =',Etotal(1)
Write(34,fmt='(i3,1x,a2,1x,2f8.1,1x,60a1)')bl,'GA',Etotal(1),(Sequence(1,k),k=1,chainlength)

Do mcstep=1,totalsteps
T=To-(To-Tf)*mcstep/totalsteps
!generate MC moves on backbone and sidechains (small moves in both)
Call Creep_Backbone(hmf,chainlength,phi(2,:),2.0,psi(2,:),2.0,ome(2,:),1.0)
Call Creep_Sidechains(0.1,chainlength,Rotamer(2,:),10.0)

Call Buildbackbone(bndlen,bndang,phi(2,:),psi(2,:),ome(2,:),chainlength,Sequence(2,:),Backbone)
Call Buildsidechains(Backbone,chainlength,Sequence(2,:),Rotamer(2,:),Sidechains)
Call Backbone_Energies(mask,chainlength,Backbone,ome(2,:),EBB(2),EBT(2))
Call Sidechain_Energies(mask,chainlength,Sequence(2,:),sstype,Rotamer(2,:),Sidechains,Backbone,&
ESB(2),ESS(2),EST(2),EI(2),EHB(2))
E_{total}(2) = \sum_{i=2}^{n} (E_{BB}(i) + E_{BT}(i) + E_{ES}(i) + E_{EI}(i) + E_{ISS}(i) + E_{EST}(i) + hw \cdot E_{HB}(i)) \\
\text{Do } \text{pos}=1, \text{chainlength} \\
\text{E}_{total}(2) = \text{E}_{total}(2) - \text{STD}(\text{rnum(Sequence}(2,\text{pos}))) - ssw \cdot sstab(sstype(pos), \text{rnum(Sequence}(2,\text{pos}))) \\
\text{End Do} \\

\text{If}(\text{metropolis(E}_{total}(2), E_{total}(1), T)) \text{then} \\
\text{E}_{total}(1) = \text{E}_{total}(2) \quad \text{!update structure info if accepted} \\
\phi(1,:) = \phi(2,:) \\
\psi(1,:) = \psi(2,:) \\
\omega(1,:) = \omega(2,:) \\
\text{Sequence}(1,:) = \text{Sequence}(2,:) \\
\text{Rotamer}(1,:) = \text{Rotamer}(2,:) \\
\text{Else} \\
\quad \text{!reset values if not accepted} \\
\phi(2,:) = \phi(1,:) \\
\psi(2,:) = \psi(1,:) \\
\omega(2,:) = \omega(1,:) \\
\text{Sequence}(2,:) = \text{Sequence}(1,:) \\
\text{Rotamer}(2,:) = \text{Rotamer}(1,:) \\
\text{End If} \\

\text{Print}, mcstep, 'Current E = ', E_{total}(1) \\

\text{If}(\text{Real(mcstep)/moviestep.eq.INT(Real(mcstep)/moviestep)}) \text{then} \\
\text{Call Buildbackbone(bndlen, bndang, phi(1,:), psi(1,:), omega(1,:), chainlength, Sequence(1,:), Backbone)} \\
\text{Call Buildsidechains(Backbone, chainlength, Sequence(1,:), Rotamer(1,:), Sidechains)} \\
\quad \text{!write movie file} \\
\text{atoms}=0 \\
\text{Do } \text{pos}=1, \text{chainlength} \\
\text{Call Mainmovie(1, Sequence(1,pos), Backbone(pos,:), pos, 33, atoms)} \\
\text{Call Sidemovie(1, Sequence(1,pos), Sidechains(pos,:), pos, 33, atoms)} \\
\text{End Do} \\
\text{Write}(33, \text{fmt}='(i3,1x,a5,i5,3x,a4,f7.2)') \text{bl, 'cycle', mcstep, 'E = ', E_{total}(1)}
Write(33,fmt='(i12)')atoms
Do pos=1,chainlength
   Call Mainmovie(2,Sequence(1,pos),Backbone(pos,:,:),pos,33,atoms)
   Call Sidemovie(2,Sequence(1,pos),Sidechains(pos,:,:),pos,33,atoms)
End Do
End If
End Do  !end mc loop
Write(34,fmt='(i3,1x,a2,1x,2f8.1,1x,60a1)')bl,'MC',Etotal(1),(Sequence(1,k),k=1,chainlength)

!build final structure
Call Buildbackbone(bndlen,bndang,phi(1,:),psi(1,:),ome(1,:),chainlength,Sequence(1,:),Backbone)
Call Buildsidechains(Backbone,chainlength,Sequence(1,:),Rotamer(1,:),Sidechains)
Write(44,'(a5,6x,i3)')'MODEL',bl
atoms=0
Do pos=1,chainlength
   Call Writemain(Sequence(1,pos),Backbone(pos,:,:),pos,44)
   Call Writeside(Sequence(1,pos),Sidechains(pos,:,:),pos,44)
End Do
Write(44,fmt='(a3)')'TER'
Write(44,fmt='(a6)')'ENDMDL'
End Do !end big loop
End Program mcbackbone
This program gives a range of backbone angles around pre-existing values for finer sampling during loop building.

```
Program ramarange
Implicit none

Real,Dimension(100)::phi,psi
Character*5::model
Integer::test,number,k,l,m,looplength
Character*80::dataline,datafile,outfile
Real,Dimension(5,100)::rama,rama2

Print*, 'Enter the ramachandran data file to be used:'
Read(5,fmt='(a80)')datafile

Print*, 'Enter the output file name:'
Read(5,fmt='(a80)')outfile

Print*, 'What model number do you want to use?'
Read(5,fmt='(i8)')number

Print*, 'What is the length of the loop?'
Read(5,fmt='(i4)')looplength

! This program inputs a ramachandran angle file outputted after loops1 has run.
! It then provides angles in a range of 25 degrees around a specific model's
! phi and psi angles and outputs these new angles to a file.

Open(12,file=datafile,status='old')
Open(13,file=outfile,status='replace')

Do
  Read(12,fmt='(a80)')dataline
  If(dataline(1:5).eq.'MODEL')Read(dataline,fmt='(10x,i5)')test
```
If (test.eq.number) then

! Read in old phi and psi angles for variation
Do k=1,looplength
print*,k
i=0
Read(12,fmt='(f6.1,f7.1)')phi(k),psi(k)
    Do l=-10,10,5
        Do m=-10,10,5
            i=i+1
            rama(k,i)=l+phi(k)
            rama2(k,i)=m+psi(k)
        End Do
    End Do
End Do
exit
End if
End Do

Do m=1,i
Write(13,100)(rama(k,m),rama2(k,m),k=1,looplength)
End Do

100 format(5(f6.1,f7.1,2x))

End program ramarange
This program utilizes the finer sampling grid developed in ramarange for building finer-tuned loops than was possible with loops1.

Program LOOPS3
Use Builder
Use moves
Use write_pdb
Use read_dihedrals
Use read
Use energies
Use analysis
Implicit none
Real::maxE
Integer,parameter::setsize=300
Real,Dimension(5000,5)::testphi,testpsi
Real,Dimension(5000)::testfit,testEBB
Real,Dimension(cl)::phi,psi,ome=180.0
Real,Dimension(cl)::phi1,psi1,ome1=180.0
Character,Dimension(5000,setsize,cl)::testseq
Character,Dimension(setsize,cl)::Sequence
Real,Dimension(setsize,4,cl)::Rotamer
Real,Dimension(cl,7,3)::Backbone,Loopstructure,Testloop
Real,Dimension(cl,30,3)::Sidechains
Real,Dimension(setsize,2)::Ranked
Real,Dimension(cl,3)::CA
Real,Dimension(cl,5)::bndlen,bndang
Real,Dimension(100,2)::rama1,rama2,rama3,rama4,rama0
Real::fit,EBB,EBT,EHB,EAB,userfit
Integer::loopcount,lstart,lend,looplength,chainstart,chainlength,pos,pos1,pos2,posb,k,i,totalsteps
Integer::lpos,bpos,m,p,match
Integer,Dimension(5)::n
Integer::p0,p1,p2,p3,p4,model
Character*80::outfile,ramachandran,phipsifile
Character*80::pdbfile,dihedrals,bndfile
Logical::metropolis, matched, first
Logical, Dimension(c1)::Mask=.false.
Integer, Dimension(20)::Counts

Print*, 'Enter the loop length:'
Read(5, fmt=’(i2)’), looplength

Print*, 'Enter starting residue:'
Read(5, fmt=’(i3)’), lstart
Print*, 'Enter ending residue:'
Read(5, fmt=’(i3)’), lend
Print*, 'Enter the name of the phi-psi angle file:'
Read(5, fmt=’(a80)’), ramachandran
Print*, 'Enter the highest energy cutoff value:'
Read(5, fmt=’(f5.1)’), maxE
Print*, 'Enter the highest fit cutoff value:'
Read(5, fmt=’(f5.2)’), userfit
Print*, 'Enter the name of the input pdb file:'
Read(5, fmt=’(a80)’), pdbfile
Print*, 'Enter the name of the output pdb file:'
Read(5, fmt=’(a80)’), outfile
Open(44, file=outfile, status=’replace’)          
Print*, 'Enter the name of the output phipsi angle file:'
Read(5, fmt=’(a80)’), phipsifile

Open(12, file=ramachandran, status=’old’)        
Open(13, file=phipsifile, status=’replace’)

n(1)=0
Do
   n(1)=n(1)+1
   Read(12, fmt=100, end=10) rama0(n(1),1), rama0(n(1),2), rama1(n(1),1), rama1(n(1),2), rama2(n(1),1), rama2(n(1),2), & rama3(n(1),1), rama3(n(1),2), rama4(n(1),1), rama4(n(1),2)
End Do
10     n(1)=n(1)-1        Print*,n(1)
Call readtypes
Call Amber_bonds(bndlen,bndang) !set default bond lengths and angles

Sequence='A'
!build a helix of appropriate length to get an estimate of low energy configuration
phi=-60.0;psi=-40.0;ome=180.0
Call Buildbackbone(bndlen,bndang,phi,psi,ome,looplength+8,Sequence(1,:),Testloop)
Call Backbone_energies(mask,looplength+8,Testloop,ome,EBB,EBT)
Print*, 'perfect helix energy = ',EBB
phi=-130.0;psi=130.0;ome=180.0
Call Buildbackbone(bndlen,bndang,phi,psi,ome,looplength+8,Sequence(1,:),Testloop)
Call Backbone_energies(mask,looplength+8,Testloop,ome,EBB,EBT)
Print*, 'perfect extended energy = ',EBB

Call readpdb(pdbfile,backbone,sidechains,Sequence(1,:),chainstart,chainlength)
Sequence='A'
Call phipsi(Backbone,lend+1,lend+4,phi1,psi1,ome1)
phi(lstart+looplength:lstart+looplength+3)=phi1(lend+1:lend+4)
psi(lstart+looplength:lstart+looplength+3)=psi1(lend+1:lend+4)
ome(lstart+looplength:lstart+looplength+3)=ome1(lend+1:lend+4)
Loopstructure(1:lstart,:,:)=Backbone(1:lstart,:,:)

n=n(1)
n(looplength+1:5)=1
loopcount=0
model=0
first=.true.
testphi=0.0
testpsi=0.0
testfit=100.0
p=1

Do p0=1,n(1)
  phi(lstart)=rama0(p0,1)
  psi(lstart)=rama0(p0,2)
  Do p1=1,n(2)
    phi(lstart+1)=rama1(p1,1)
    psi(lstart+1)=rama1(p1,2)
    Do p2=1,n(3)
      phi(lstart+2)=rama2(p2,1)
      psi(lstart+2)=rama2(p2,2)
      Do p3=1,n(4)
        phi(lstart+3)=rama3(p3,1)
        psi(lstart+3)=rama3(p3,2)
        Do p4=1,n(5)
          phi(lstart+4)=rama4(p4,1)
          psi(lstart+4)=rama4(p4,2)

      loopcount=loopcount+1
  Call Buildloop(bndlen,bndang,lstart,lstart+looplength+3,phi,psi,ome,Sequence(1,:),Loopstructure)

    fit=0.0
    Do pos=1,looplength
      bpos=lend+pos-1
      lpos=lstart+looplength+pos-2
      Do k=1,4
        fit=fit + Distance(Backbone(bpos,k,:),Loopstructure(lpos,k,:))**2
      End Do
End Do
fit=SQRT(fit)/20.0

If(Real(loopcount)/10000.0).eq.Real(loopcount)/10000.0) Print*,loopcount,'fit =',fit
! if good fit, write out coordinates
If(fit.eq.userfit) then
Testloop(1:looplength+7,:,:)=Loopstructure(1start-3:1start+looplength+3,:,:)
omel(1:looplength+7)=ome(lstart-3:lstart+looplength+3)
Call Backbone_energies(mask,looplength+8,Testloop,ome1,EBB,EBT)
   !Do pos1=1,looplength+8
   ! Do pos2=pos1+1,looplength+8
   ! EHB=EHB+6*HBond(0,0,'B','B',Testloop(pos1,:,:),Testloop(pos2,:,:))
   ! If(Distance(Testloop(pos1,4,:),Testloop(pos2,4,:)).lt.3.5)EAB=EAB+100.0
   ! End Do
! End Do
Print*,'fit =',fit,'Backbone energy =',EBB
If(EBB.maxE) then
   matched=.false.
   Do i=1,p
      match=0
      Do m=1,looplength
         If(phi(lstart+m-1).eq.testphi(i,m)) then
            match=match+1
         End if
         If(psi(lstart+m-1).eq.testpsi(i,m)) then
            match=match+1
         End if
      End Do
      If(match.eq.(2*looplength-1)) then
         Do m=1,looplength
            If(abs(phi(lstart+m-1)-testphi(i,m)).eq.5.0.or.abs(psi(lstart+m-1)-testpsi(i,m)).eq.5.0) matched=.true.
         End Do
      End If
   End Do
If(matched.and.fit.lt.testfit(i))then
   Do m=1,looplength
      testphi(i,m)=phi(lstart+m-1)
      testpsi(i,m)=psi(lstart+m-1)
   End Do
   testseq(i,1,:)=Sequence(1,:)
   testfit(i)=fit
   testEBB(i)=EBB
   End if
End if   !  End of match test loop
End Do   !  End of i loop                   
If(.not.matched)then                      
   If(first)then
      first=.false.
   Else
      p=p+1
   End if
   Do m=1,looplength
      testphi(p,m)=phi(lstart+m-1)
      testpsi(p,m)=psi(lstart+m-1)
   End Do
   testseq(p,1,:)=Sequence(1,:)
   testfit(p)=fit
   testEBB(p)=EBB
   End if
End if     !  End of EBB/Emax test loop
End If     !  End of big good fit loop
   End Do
   End Do
End Do         !  End of nested rama loops
Do i=1,p
Print*,i,'fit =',testfit(i),',EBB=',testEBB(i),',EHB=',EHB
Write(13,fmt='(a5,2x,i8,2x,a5,f8.3,2x,a5,f8.3)')'MODEL',i,'fit =',testfit(i),'
&
'EBB =',testEBB(i)
Do m=1,looplength
Print*,testphi(i,m),testpsi(i,m)
Write(13,fmt='(f6.1,f7.1)')testphi(i,m),testpsi(i,m)
phi(lstart+m-1)=testphi(i,m)
psi(lstart+m-1)=testpsi(i,m)
End Do
Sequence(1,:)=testseq(i,1,:)
Call Buildloop(bndlen,bndang,lstart,lstart+looplength+3,phi,psi,ome,Sequence(1,:),
&
Loopstructure)
Write(44,'(a5,6x,i5)')'MODEL',i
atoms=0
Do pos=lstart-3,lstart+looplength+3
   Call Writemain(Sequence(1,pos),Loopstructure(pos,:,:),pos,44)
End Do
Write(44,fmt='(a3)')'TER'
Write(44,fmt='(a6)')'ENDMDL'
End Do

Close(44)

100 format(f6.1,f7.1,2x,f6.1,f7.1,2x,f6.1,f7.1,2x,f6.1,f7.1,2x,f6.1,f7.1)
End Program LOOPS3
This program repacks the helices after the backbone angles were randomly altered with mcbackbone.

Program helix_compare2
Use random
Use write_pdb
Use comparisons
Use read_dihedrals
Use read
Use builder
Use write_movie
Implicit none

Character*80::filename,outfile,dihedrals,moviefile,finalangles
Real,Dimension(100,5)::bndlen,bndang
Logical::metropolis,input=.true.
Real,Dimension(2)::fit
Real::T,To,Tf,bmf,ran
Integer::k,pos,loopstart1,loopend1,loopstart2,loopend2,i,j,chainlength,chainstart,chainend
Integer::mcstep,count,totalsteps,time,moviestep
Real,Dimension(2,100)::phi,psi,ome
Character,Dimension(100)::Sequence
Real,Dimension(100,7,3)::Backbone1,Backbone,Newbackbone
Real,Dimension(100,30,3)::Sidechains
Real,Dimension(100,3)::CA1,CA2

seed=2*time( )+1 ! Initialize random number seed
Print*,seed

Read(5,fmt='(a80)')filename
Read(5,fmt='(a80)')dihedrals
Read(5,fmt='(i3)')loopstart1
Read(5,fmt='(i3)')loopend1
Read(5,fmt='(i3)')loopstart2
Read(5,fmt='(i3)')loopend2
Read(5,fmt='(i8)')totalsteps
Read(5,fmt='(f8.1)')To
Read(5,fmt='(f8.1)')Tf
Read(5,fmt='(f4.3)')bmf
Read(5,fmt='(a80)')outfile
Read(5,fmt='(a80)')moviefile
Read(5,fmt='(a80)')finalangles
moviestep=totalsteps/1000
Open(33,file=moviefile,status='replace')
Write(33,fmt='(i12)')1000
Call readtypes
Call readpdb(filename,Backbone1320,sidechains,Sequence,chainstart,chainend)
Call Readphipsi(dihedrals,phi(1,:),psi(1,:),ome(1,:),chainlength)
Call Amber_bonds(bndlen,bndang) !set default bond lengths and angles
Call Buildbackbone(bndlen,bndang,phi(1,:),psi(1,:),ome(1,:),chainlength,Sequence,Backbone)

count=0
Do i=1,chainlength
  If((i.ge.loopstart1.and.i.le.loopend1.or.i.ge.loopstart2.and.i.le.loopend2)then
    Continue
  Else
    count=count+1
  Newbackbone(count,:,:)=Backbone(i,:,:)
End if
End Do

Open(44,file=outfile,status='replace')
atoms=0
Do pos=1,count
    Call Writemain(Sequence(pos),Newbackbone(pos,:,,:),pos,44)
  End Do
Write(44,fmt='(a3)')'TER'

fit(1)=CAfit(48,Backbone1320(:,2,:),Newbackbone(:,2,:))
Print*,'first fit=',fit(1)

! Set up for monte carlo
phi(2,:)=phi(1,:)
psi(2,:)=psi(1,:)
ome(2,:)=ome(1,:)

Print*, 'Monte Carlo Simulation, round', mcstep

Do mcstep=1,totalsteps
    T=To-(To-Tf)*mcstep/totalsteps
    Print*, mcstep, 'T=', T

    ! Generate random dihedral angle parameters to make a test structure
    Do j=1,chainlength
        If(ran(seed).lt.bmf) phi(2,j)=phi(1,j)+2.0*(ran(seed)-0.5)
        If(ran(seed).lt.bmf) psi(2,j)=psi(1,j)+2.0*(ran(seed)-0.5)
        If(ran(seed).lt.bmf) ome(2,j)=ome(1,j)+1.0*(ran(seed)-0.5)
    End Do

Call Buildbackbone(bndlen,bndang,phi(2,:),psi(2,:),ome(2,:),chainlength,Sequence,Backbone)

  count=0

Do i=1,chainlength
    If(i.ge.loopstart1.and.i.le.loopend1.or.i.ge.loopstart2.and.i.le.loopend2) then
        Continue
Else
    count=count+1
    ! Print*,count
    Newbackbone(count,:,:)=Backbone(i,:,:)
    ! Write(6,fmt='(3f8.3)')(Newbackbone(count,2,k),k=1,3)
End if
End Do

! Evaluate fit of test structure
fit(2)=CAfit(48,Backbone1320(:,2,:),Newbackbone(:,2,:))
Print*,'test fit =',fit(2)

If(metropolis(fit(2),fit(1),T))then
    fit(1)=fit(2)
    phi(1,:)=phi(2,:)
    psi(1,:)=psi(2,:)
    ome(1,:)=ome(2,:)
Else
    phi(2,:)=phi(1,:)
    psi(2,:)=psi(1,:)
    ome(2,:)=ome(1,:)
End if

Print*,mcstep,'Current fit =',fit(1)

If(Real(mcstep)/moviestep.eq.INT(Real(mcstep)/moviestep))then
    Call Buildbackbone(bndlen,bndang,phi(1,:),psi(1,:),ome(1,:),chainlength,Sequence,Backbone)

    ! Write movie file
    atoms=0
    Do pos=1,chainlength
        Call Mainmovie(1,Sequence(pos),Backbone(pos,:,:),pos,33,atoms)
    End Do
    Write(33,fmt='(a8,i5,3x,a6,f10.6)')'mc cycle',mcstep,'fit = ',fit(1)
Write(33, fmt='(i12)') atoms
   Do pos=1, chainlength
      Call Mainmovie(2, Sequence(pos), Backbone(pos,:,:,:), pos, 33, atoms)
   End Do
End if

End Do  ! End mc loop

! Update pdb file

Call Buildbackbone(bndlen, bndang, phi(1,:), psi(1,:), ome(1,:), chainlength, Sequence, Backbone)

   Open(13, file=finalangles, status='replace')
   Do pos=1, chainlength
      Write(13, 100) pos, Sequence(pos), phi(1,pos), psi(1,pos), ome(1,pos)
   End Do

100 format(i4,3x,a1,2x,4f12.1)

   Open(44, file=outfile, status='replace')
   atoms=0
   Do pos=1, chainlength
      Call Writemain(Sequence(pos), Backbone(pos,:,:,:), pos, 44)
   End Do
   Write(44, fmt='(a3)') 'TER'

End program helix_compare2
CURRICULUM VITAE
Christina M. Kraemer Pecore

EDUCATION

1997-2002 Ph.D., Biophysical Chemistry  
*The Pennsylvania State University, University Park, PA*  
- Research Title: Computational Design and Experimental Characterization of Protein Domains

1992-1996 B.S., Chemistry and Environmental Chemistry (ACS-certified)  
*State University of New York, College at Plattsburgh*  
*Plattsburgh, New York 12901*

AFFILIATIONS

- The American Institute of Chemists, Inc.
- Biophysical Society
- Protein Society
- Phi Kappa Phi National Honor Society
- American Chemical Society

RECENT AWARDS

- Miller Graduate Student Research Award (2001)  
- Graduate Student Travel Award (2001)  
- Roberts Graduate Fellowship (1997-1999)  
- Nominated for the Dan Waugh Teaching Award (1999)  
- American Institute of Chemists Foundation Award (1996)  
- E. Yale Clarke Endowment Fund Award (1996)

RECENT PUBLICATIONS AND PRESENTATIONS

- C.M. Kraemer-Pecore, J.T.J. Lecomte, B. Bregar and J. Desjarlais, "Redesigning the WW Domain," manuscript in preparation.