AN ABSTRACT OF THE DISSERTATION OF

Joshua M. Hicks for the degree of Doctor of Philosophy in Biochemistry and Biophysics presented on February 14, 2005.

Title: Analysis of Secondary Structures in Nucleic Acid Binding Proteins and Nuclear Magnetic Resonance Investigation of Helix Propagation and Residual Motions in Proteins.

Abstract approved:

Redacted for Privacy

Victor L. Hsu

This thesis involves analyses of 3_2-helices, α-helices and of the change in residual entropy of a protein upon chemical modification. Part of the long-term goal of understanding the formation, function and stability of proteins requires that all conformations be accurately assigned and structurally understood. By a statistical analysis of 258 nucleic acid binding proteins from the protein data bank, 5% of their amino acids were found to exist as 3_2-helices. The third K homology domain, Epstein-Barr nuclear antigen-1 and the homeodomain from the Drosophila paired protein contained 3_2-helices involved in nucleic acid recognition. Contact maps of these three proteins show that the 3_2-helical motif is capable of both specific and non-specific recognition and binding of nucleotides.

Once all conformational structures are accurately assigned, their natural diversity of sequence and environment in proteins can be explored to understand how they are formed. α-Helix formation in various mixtures of 2,2,2-trifluoroethanol (TFE), a known helix-inducing solvent, can be used to predict whether a sequence will form a helix at the protein core vs. the solvent-exposed protein surface. Nuclear magnetic resonance spectroscopy of a TFE solvated peptide composed of two helix-favoring pentapeptides flanking a helix-indifferent pentapeptide shows that the structure adopts N-terminal and C-terminal helices separated by two non-helical residues in the indifferent sequence. The initation of two nucleation events in this short peptide and the inability of the helix to propagate through the indifferent sequence (neither event is predicted by current models) indicates that, in addition to environment, there is a strong sequence context dependence for helix formation.

Enzymatic function can be tied directly to the internal motions identified as the residual entropy of protein structure. The use of chemical modification as a method of irreversibly trapping a mimic of the catalytic intermediate in thioredoxin function shows that these motions are substantial and are not necessarily localized to the active site. The observed increase in fast motions identifies obvious structures known to be important in the redox mechanism and several unobvious structures that could indicate regions possibly coupled to protein function. Together these results identify important properties for understanding protein structure and function.

Analysis of Secondary Structures in Nucleic Acid Binding Proteins and Nuclear Magnetic Resonance Investigation of Helix Propagation and Residual Motions in Proteins.

by

Joshua M. Hicks

A DISSERTATION

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented on February 14, 2005

Commencement June 2005

Doctor of Philosophy dissertation of Joshua M. Hicks presented on February 14, 2005.

APPROVED:

Redacted for Privacy

Major Professor, representing Biochemistry and Biophysics

Redacted for Privacy

Chair of the Department of Biochemistry and Biophysics

Redacted for Privacy

Dean of Graduate School

I understand that my dissertation will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my dissertation to any reader upon request.

Redacted for Privacy

Joshua M. Hicks, Author

ACKNOWLEDGEMENTS

I would first like to acknowledge my advisor, Victor Hsu, who provided the materials to make this research possible and gave me the latitude to explore and solve my own problems. Victor's knowledge and retention has always been a helpfully diverse source of information. Many thanks go to W. Curtis Johnson and Christopher Mathews for additional guidance in all things not NMR. Their frankness and insight were greatly appreciated. I would also like to thank P. Andrew Karplus for critical discussions and J. Randy MacDonald for computational advice.

Thanks to all my friends at Oregon State University for making my time here most enjoyable and endlessly interesting. This includes the staff of the department of Biochemistry and Biophysics and the Environmental Health and Science Center.

My biggest acknowledgements go to my family, in particular my parents, Bill and Carolyn Hicks, who have always been supportive and to my wife, Cookkai, who has always been encouraging.

CONTRIBUTION OF AUTHORS

Victor Hsu contributed to the design and analyses of each manuscript. W. Curtis Johnson and Jeanine Lawrence contributed to the design and analysis of the helix propagation studies via circular dichroism. W. Curtis Johnson provided the source code for the XTLSSTR secondary structure assignment program that was used for the study on 3_2-helices. P. Andrew Karplus provided the Globalcore/Globalmerge programs for structural alignment of the YHGH peptide.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

Page

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

Figures Page

LIST OF TABLES

DEDICATION

I dedicate this work to my parents and my wife. Their patience and understanding continue to amaze me and help me appreciate these qualities more and more each day.

Chapter 1

Introduction

1.1 Protein Folding

The ability of many proteins to spontaneously refold *in vitro* indicates that three-dimensional structural information is contained within the primary sequence. In the early 1960's, Anfinsen proved that the one-dimensional sequence of amino acid residues uniquely determined the three-dimensional conformation of Ribonuclease A (Fig. 1.1) (Anfinsen et al. 1961). However, despite the numerous efforts to identify the intrinsic properties of protein structure and function that are encoded in the primary sequence, the ability to define the final tertiary structure of a protein from its amino acid sequence has remained elusive. Termed the protein-folding problem, many models have been developed to describe the steps of protein folding. It is clear that proteins fold by a process that is both rapid and specific. Levinthal's paradox states that time estimates based on the systematic, random sampling of all conformational space cannot account for the fact that protein folding occurs on the millisecond to second timescale (Levinthal 1968). Subsequent analysis indicates that this paradox does not exist if it is assumed that protein folding occurs in much simpler conformational space (Levitt and Warshel 1975).

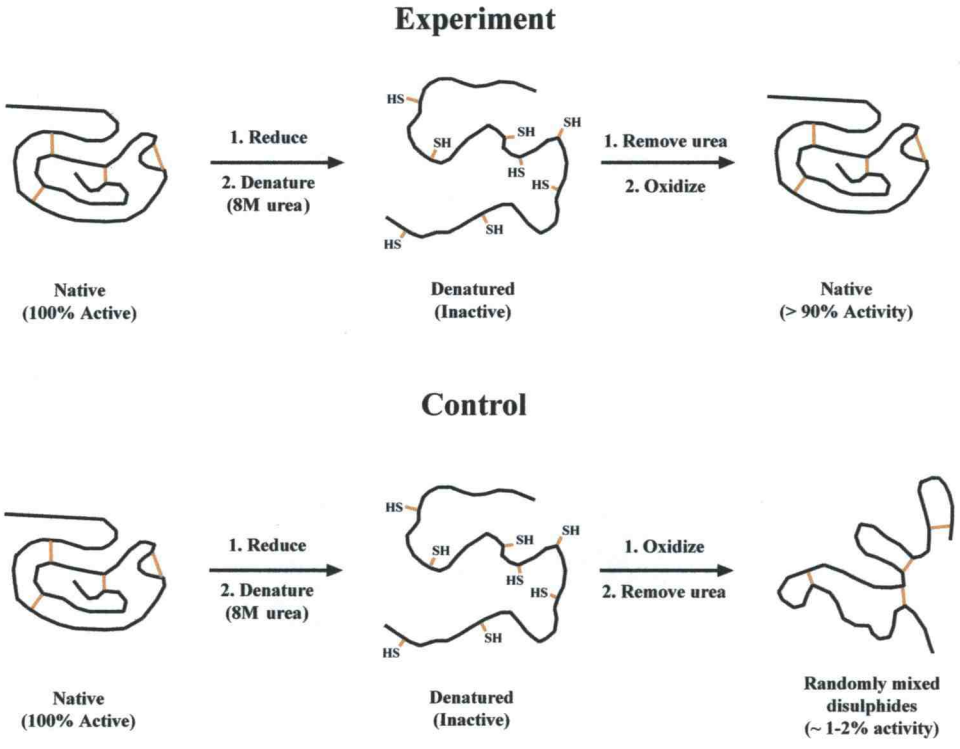


Figure 1.1 Schematic representation of Anfinsen's experiment (Anfinsen et al. 1961) describing the refolding and activity of Ribonuclease A (RNase A). Reduced and denatured RNase A was refolded under two different conditions. In the experiment, refolding then oxidizing of RNase A returned it to almost complete activity. In the control, oxidizing then refolding formed random disulphides leading to an inactive RNase A. (Adapted from the on-line biochemistry series from the Mellon College of Science).

1.1.1 The Kinetic Model

The kinetic model of protein folding bypasses Levinthal's paradox by assuming that a specific sequence of cooperative folding steps in the free energy landscape leads to the native state (Wang and Shortle 1995; Baldwin and Rose 1999; Bakk et al. 2001). This model is derived from the observation that there is an absence of a lag period in the appearance of the native structure from denatured proteins. It has been hypothesized that the initially formed, partially folded intermediates represent kinetically unstable individual folding units. These individual cooperative folding units, or foldons, are an important feature of the protein folding pathway. In a few cases, foldons have been experimentally identified as distinctive intermediate forms during protein folding (Chamberlain et al. 1996; Fuentes and Wand 1998; Maity et al. 2004). For example, of the eight helical segments (A to H) found in the native structure of apomyoglobin (apoMb), the foldon composed of helices A, G and H forms first (Jennings and Wright 1993). Stabilization of these three helices induces the formation of helix B, followed by complete folding of the entire structure (Garcia et al. 2000).

1.1.2 The Hydrophobic Collapse Model

The folding of amino acids into helices and β-hairpins takes place fast enough that secondary structure formation is capable of being a key event in the protein folding process (Williams et al. 1996; Munoz et al. 1997). However, the driving force in helix formation during protein folding is often weak, as indicated by the marginal stability of helices isolated from natural proteins (Munoz and

Serrano 1994). The hydrophobic collapse model provides a possible explanation to how secondary structures might be formed and subsequently stabilized during protein folding. Evidence that hydrophobic collapse is as fast or faster than secondary structure formation (Sadqi et al. 2003) suggests that the folding environment is much different for residues found in the hydrophobic core of the protein. In the hydrophobic collapse model, cooperative long-range interactions between hydrophobic clusters guide folding processes, and the formation of these clusters determines the folding rate of proteins (Zdanowski and Dadlez 1999; Hodsdon and Frieden 2001; Baldwin 2002; Klein-Seetharaman et al. 2002).

1.1.3 The Nucleation and Nucleation-Condensation Models

The nucleation model proposes that the tertiary structure forms as an immediate consequence of the initiation of secondary structures (Wetlaufer 1973; Abkevich et al. 1994). Nucleation of the tertiary structure occurs through the formation of native secondary structure by only a few residues. A variation of the nucleation model is the nucleation-condensation model that describes a diffuse folding nucleus that is formed and consolidated through the transition state (Jackson and Fersht 1991; Fersht 1995). The nucleation-condensation model was proposed by Jackson and Fersht (1991) to explain the experimental evidence for the folding kinetics observed in several small proteins (Jackson and Fersht 1991; Jackson and Fersht 1991).

1.1.4 The Folding Funnel Model

Figure 1.2 illustrates a more recent, general description of protein folding known as the folding funnel model, which represents the folding pathway as an energy surface (Leopold et al. 1992; Dobson et al. 1998; Dinner et al. 2000). In the folding funnel model, protein folding is characterized as a formation of the native structure by collapsing and reconfiguring of a diverse multitude of unfolded conformations at the rim. The reconfiguration cannot occur between energetically prohibitive global conformations, but geometrically similar reconfiguration of local conformations are considered, one of which is thermodynamically more stable than the rest. This leads to several pathways whereby the protein can follow the fastest pathway or pass through several local minima to the lowest energy form (Leopold et al. 1992; Dobson et al. 1998; Dinner et al. 2000). Since the native state is considered the state with the lowest free energy (Anfinsen 1973), by studying the structure and dynamics of enzymes, it is perhaps possible to understand the thermodynamic forces that guide protein folding.

1.1.5 Summary

The experiments in this dissertation involve a series of investigations on three distinct aspects of protein folding, form and function. Though the three studies presented here are diverse, each one demonstrates an important aspect of understanding the physical properties of proteins. The following three sections provide a more in-depth analysis on the relationship of secondary structure to the

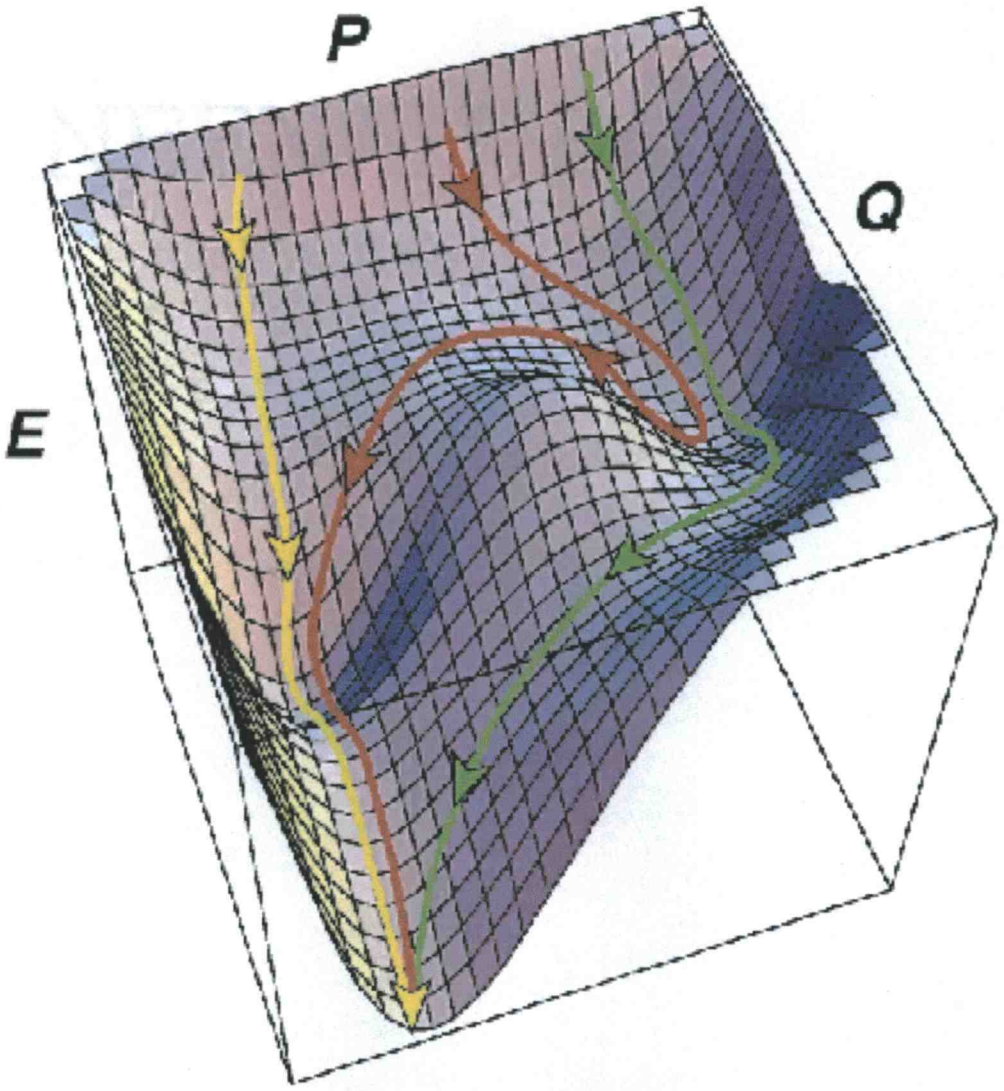


Figure 1.2 Energy surface of the protein folding funnel. The E-axis represents the energy of the system, the P-axis is a measure of the available conformational space and the Q-axis is defined as the proportion of native contacts formed. The possible folding pathways are represented by yellow (fast folding), red (slow folding, which returns to a less folded state before following the fast folding), and green (slow folding that crosses a high energy barrier) (Reproduced from Dobson et al. (1998))

function, the formation of secondary structures and the internal dynamics of a folded protein.

1.2 The 3_2-helix is an Important Secondary Structure

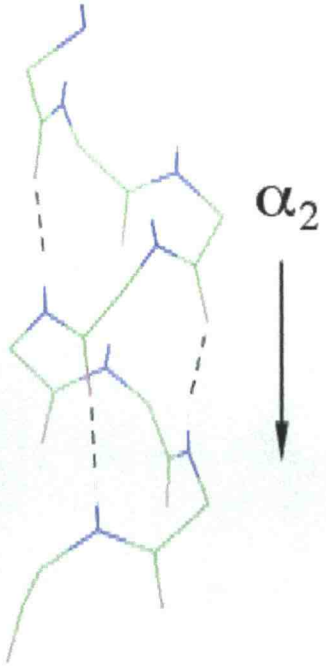
Methods for characterizing the structure of proteins have improved dramatically since the first X-ray crystal structures of proteins were reported (Kendrew et al. 1958). Physical techniques, including X-ray crystallography, nuclear magnetic resonance (NMR) and circular dichroism (CD), have all benefited from the improvement in experimental hardware and, as important, the exponential increase in computational power over the past few decades. The resulting flood of structural information arising from modern techniques of protein analysis has led to the construction of database repositories, such as the Protein Data Bank (PDB), which contain massive amounts of structural information. The PDB contains over 28,000 structures of proteins, nucleic acids and complexed macromolecules solved by a variety of methods including theoretical modeling, neutron scattering, cryogenic electron microscopy, NMR, and X-ray crystallography (Berman et al. 2000). Open access to the PDB has made it possible to investigate overlooked structural information and to further explore hypotheses resulting from improved methods for defining the physical properties of proteins.

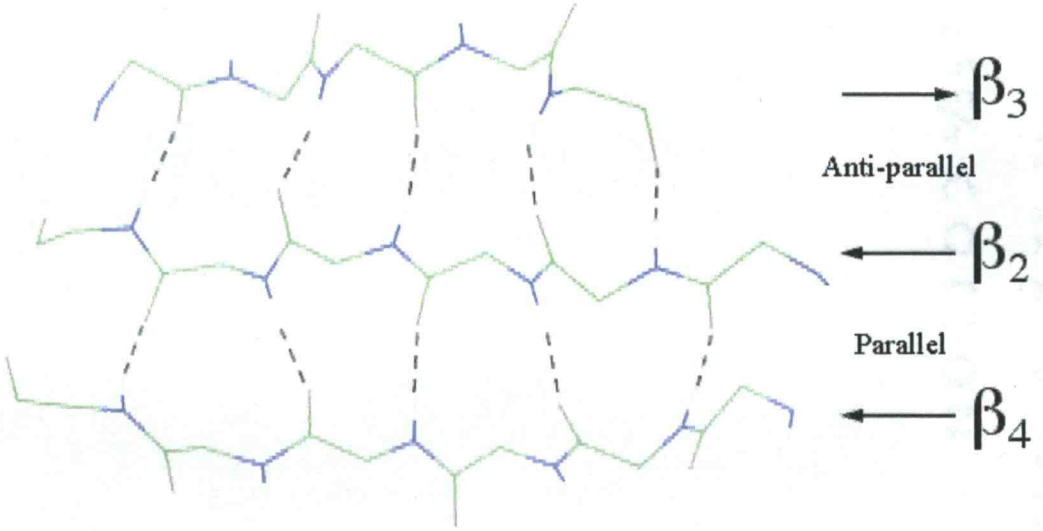
1.2.1 Secondary Structures in Proteins

Since Pauling and Corey's series of papers on hydrogen bonding patterns in secondary structures was published in 1951, the distribution of these elements in protein structure and function has been extensively documented. The identification of the formation of hydrogen bonds between backbone carbonyl oxygen groups and backbone amide protons led to the realization that the hydrogen bonding contribution to protein stability defines how the protein structure is organized. Regular, repeating hydrogen-bonding patterns between backbone atoms are characteristic of the common secondary structures annotated in proteins. For example, established hydrogen bonding between the backbone amide proton of amino acid (i+4) and the carbonyl oxygen of amino acid (i) defines a hydrogen bonded turn. Subsequent formation of identical hydrogen bonds from (i+5, i+1), (i+6, i+2) and (i+7, i+3) describes the α-helix. In fact, such patterns in absolute hydrogen bonding form the basis of the algorithm by which the secondary structure assignment algorithm DSSP (Define Secondary Structure of Protein) identifies secondary structures (Kabsch and Sander 1983).

As discussed above, the α-helix is right-handed coil composed of regular, repeating intraelement hydrogen bonds between backbone amide protons and carbonyl oxygens (Fig. 1.3). The 3_{10}-helix and π-helix are similar to the α-helix except the interval between pairs of hydrogen-bonded residues is (i+3, i) and

Figure 1.3 Models of the α-helix and β-sheet taken from the structure of *E coli* thioredoxin (1XOB). Backbone nitrogen (blue), oxygen (red), carbon (green) and proton (grey) atoms are shown with interelement hydrogen bonds (dashed lines). Helix α2 consists of residues I41-E48. Strands β2 (A22-F27), β3 (L53-L58) and β4 (T77-F81) show examples of parallel and anti-parallel orientations in β-sheets. Arrows indicate the primary sequence orientation from the N-terminal residue to the C-terminal residue.





α-Helix

β-Sheet

(i+5, i), respectively. Other, less common helices include the left-handed α_L-helix (which results in unfavorable orientation of residue side-chains at the helical axis), the Type I helix (identified as a conformation of polyglycine peptides) and the 3_2-helix (the major conformation in collagen). The second commonly identified secondary structure is the β-strand, which is the basic unit of the β-sheet. The β-strand is unstable by itself, as there are no interactions between atoms of the same strand. Organization of β-strands adjacent to other β-strands in the formation of β-sheets allows close to optimal geometry of hydrogen bonding between peptide groups that stabilize the structure. The alignment between adjacent β-strands can occur in two orientations, parallel β-sheets and anti-parallel β-sheets (Fig. 1.3). A third class of secondary structures is found in the reverse loops and turns that make up nearly one third of the residues of globular proteins. β-Hairpins link adjacent strands in anti-parallel β-sheets. If only one residue is not involved in a hydrogen-bonding pattern of the β-sheet it is termed a γ-turn, and if two consecutive residues are not involved in a hydrogen-bonding pattern it is termed a β-turn. All secondary structures can also be categorized by their characteristic backbone dihedral, or ϕ-, ψ-angles. By plotting the observed residue dihedrals in protein structure (Fig. 1.4) it is possible to identify regions associated with secondary structure (Ramachandran and Sasisekharan 1968).

1.2.2 Assigning Secondary Structures

Secondary structure assignment from three-dimensional coordinates of proteins is a major step in the description of protein structure because it forms the

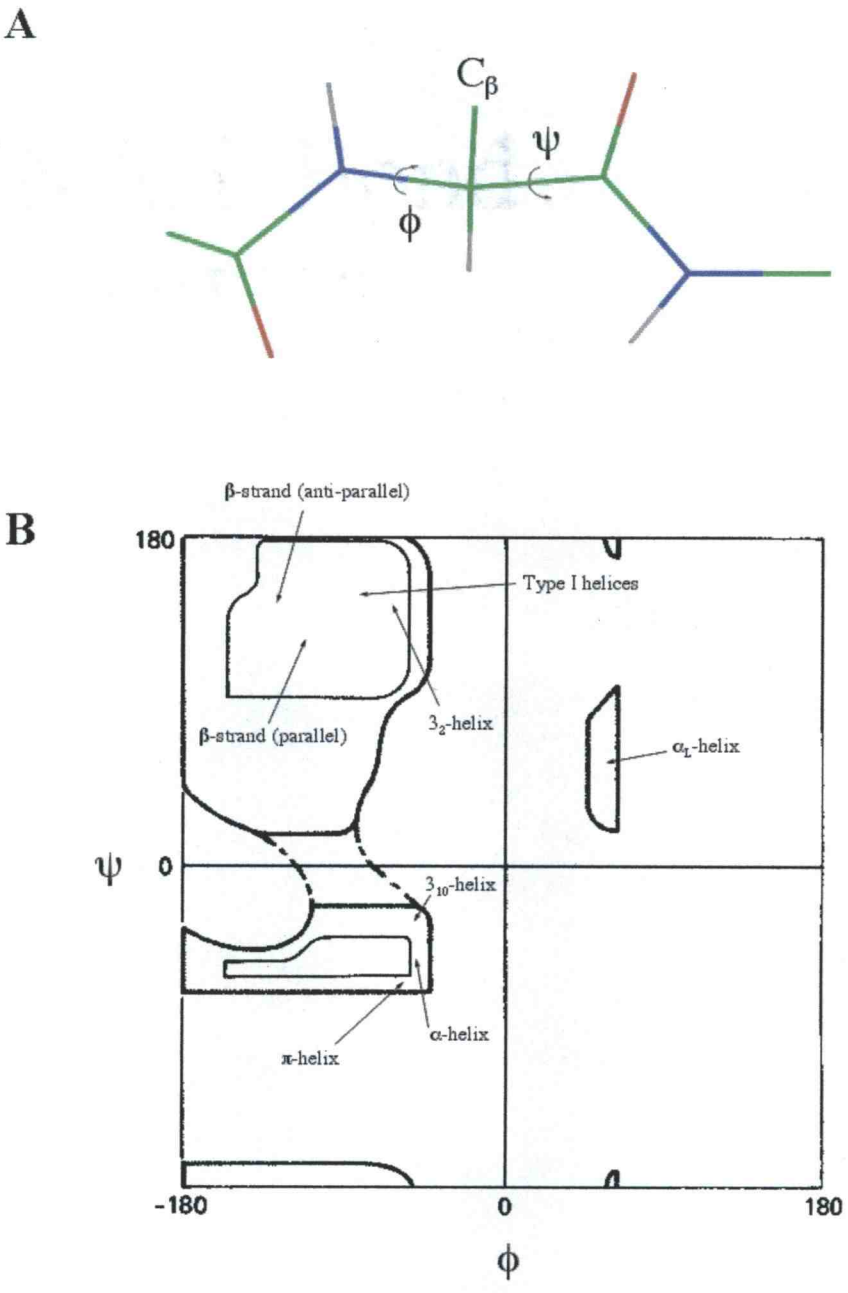


Figure 1.4 Distribution of regular secondary structures in ϕ, ψ space on a Ramachandran plot. (A) The angle ϕ defines the rotation about the N (blue) to C_α bond and the angle ψ defines rotation about the C_α to C(=O) bond. (B) Common secondary structures and their association to allowed conformations (adapted from Ramachandran and Sasiekharan (1968)).

basis for many subsequent analyses. Conceptually, our own visual notion of protein structure begins with cartoon representation of helices and strands by cylinders and ribbons, respectively. Secondary structure assignments are used for structural comparison, classification and, because secondary structure is more conserved than primary sequence, sequence alignment. Incorporation of secondary structure profiles into homology searches can increase the accuracy of sequence alignment over standard search tools using only sequence-based information (Fischer and Eisenberg 1996; Rost et al. 1997; Jennings et al. 2001).

A variety of strategies have been incorporated into algorithms for assigning secondary structures in proteins. Hydrogen bonding patterns identified by the DSSP algorithm are used to assign α-, 3_{10}- and π-helices, β-bridges, β-strands and β-turns. More recently, an improved version of DSSP (DSSPcont) includes continuous secondary structure assignment to account for structural variations and thermal fluctuations (Carter et al. 2003). Frishman and Argos (1995) developed another algorithm called "Stride" for neural network-based assignment of α-helices and β-strands that can reproduce secondary structure assignments created by crystallographers. Stride uses a product-weighted hydrogen bonding energy in its force field and a sum of terms involving statistically-derived backbone torsion angles (Frishman and Argos 1995). A different approach is used by the "Define" algorithm that was written to assign α-helices, 3_{10}-helices, β-strands, β-turns and ω-loops using distance matrices of idealized inter-Cα distances (Richards and Kundrot 1988). Another method uses the so-called "P-curve" algorithm to assign various types of helices and β-strands

from backbone curvature (Sklenar et al. 1989). More recently, the XTLSSTR (X-ray crysTaL Secondary STRucture) algorithm was developed to identify secondary structure content in proteins in a way that more closely correlates to actual CD spectroscopic measurements of secondary structure (King and Johnson 1999). Unlike algorithms using absolute hydrogen bonding, XTLSSTR assigns secondary structures via intercarbonyl vectors first described by Woody (1992), resulting in a better description of overall chain conformation. The advantages of the XTLSSTR, Define and P-curve algorithms are that they are not limited by the definition of hydrogen bond patterns, allowing for the identification of nontraditional secondary structures left unassigned by other algorithms.

1.2.3 The 3_2-helix as an Overlooked Secondary Structure

A common secondary structure that lacks main chain hydrogen bonds such as those found in α-helices, γ-turns and β-pleated sheets is the 3_2-helix. First described in collagen (Pauling and Corey 1951c), the 3_2-helix is a well-defined, extended left-handed helix that is essential in several biological interactions important for signal transduction, transcription and the immune response (Kelly et al. 2001). Raman spectroscopy has also identified the formation of 3_2-helices in amyloidogenic prefibril intermediates (Blanch et al. 2000). Estimates of the statistical presence of 3_2-helices in proteins has varied from 2% (Stapley and Creamer 1999), to 5% (Woody 1992; Adzhubei and Sternberg 1993), to as much as 20% (Adzhubei et al. 1987b) of the residues in proteins. Extensive analysis of 3_2-helices in proteins has been limited by the lack of definitive annotation in

reported structures, despite recurring evidence of their existence. This can be attributed to the fact that structures submitted to repositories like the PDB only include notation for α-helices, 3_{10}-helices, π-helices, β-turns, β-bridges and extended β-strands. As a result, approximately 25% of the protein residues within the PDB are unassigned by secondary structure assignment algorithms (Sreerama and Woody 1994). These unassigned regions are typically designated "random coil", though they are often neither random nor coiled (Richardson 1981). Extensive analysis detailing the conformation of these unassigned regions in all protein structures has, in the past, been limited by the lack of a rapid, efficient method for correctly assigning all secondary structures and an underdeveloped understanding of the composition, function and stability of alternate secondary structures like the 3_2-helix.

1.2.4 The 3_2-helix Study Reported in Chapter 2

By themselves, individual protein structures may not yield overwhelming evidence about 3_2-helix structure/function, but when massive data sets are analyzed this secondary structure can be put into a relevant context. Computational bioscience focuses on evaluating large amounts of data collected from solved structures without rigorous individual investigation of each protein. To determine the level of occurrence of 3_2-helical structures and their role in macromolecular interactions, XTLSSTR secondary structure assignments of nucleic acid (NA) binding proteins in the PDB were analyzed. NA binding

proteins were chosen because of initial reports of 3_2-helices in the TAR-binding region from the HIV-1 Tat protein fragment (Kanyalkar et al. 2001).

The statistical analysis in Chapter 2 indicates that 5% of all residues found within a dataset of 258 NA binding proteins are in the 3_2 conformation. Most of the 3_2-helices are located within the so-called unassigned regions of the proteins, with a few examples of 3_2-helices actually assigned as other secondary structures by DSSP due to their hydrogen bonding patterns. Three proteins (the third K homology domain, Epstein-Barr nuclear antigen-1 and the homeodomain from the Drosophila paired protein) that contain 3_2-helices serving as the NA recognition and/or binding motif were described and illustrated. Mapping the contacts between 3_2-helical residues and the respective NA substrate of these three proteins showed the 3_2-motif can form specific interactions with the bases and/or non-specific interactions with the sugar-phosphate backbone of nucleotides. In the corresponding literature (Wilson et al. 1995; Bochkarev et al. 1998; Lewis et al. 2000), none of these residues were identified as being in a regular repeating structure but, rather, were referred to as unordered or random. The previously unassigned 3_2-helices found in this function-directed sampling of the PDB proves that a significant amount of this secondary structure exists and can contribute to the intrinsic function of the protein.

1.3 α-Helix Initiation and Propagation

1.3.1 Helix-coil Transition Theory

The early assembly of secondary structures in protein folding and the formation of secondary structure elements by peptide fragments indicate that the initiation of secondary structures is one of the earliest steps in protein folding. Development of the Zimm and Bragg (1959) and Lifson and Roig (1961) statistical mechanics theories of helix-coil initiation-propagation has made α-helix initiation and propagation the most extensively studied event in the initiation of protein structure formation. The initiation-propagation model proposes that the formation of a nucleation center occurs by random conformational fluctuation of the unfolded protein. Structures such as helices form an initial structure, or "seed", that would be small enough so that it can be formed randomly, and its formation would be limited by the rate of random fluctuations along the protein backbone. The location of these initiating sequences would be identified by the presence of residues that exhibit a strong probability (propensity) to adopt a particular secondary structure and initiate its formation. Folding, then, would be expected to cooperatively extend, or propagate, between neighbors if adjacent residues are in the helical conformation (Zimm and Bragg 1959; Lifson and Roig 1961). Propagation would continue in either or both directions through sequences that are neither helix initiators nor helix breakers (indifferent to helix formation) until helix breakers are encountered (Zimm and Bragg 1959; Chou and Fasman 1974; Jaravine et al. 2001). Chou and Fasman (1974) attempted to calculate the

propensity of each amino acid to form an α-helix, β-strand or turn. Their method involved conducting statistical analysis of solved structures and then applying this information to amino acids in proteins whose secondary structure was unknown. The Chou and Fasman prediction method was designed to eliminate complex computer calculations by assigning a secondary structure to amino acids based on the amino acid's propensity.

Recently, the use of chemically synthesized peptides has led to numerous strategies for determining the intrinsic helical propensity of individual amino acids (Marqusee et al. 1989; Storrs et al. 1992; Chakrabartty et al. 1994; Doig and Baldwin 1995; Myers et al. 1998; Krittanai and Johnson 2000; Goch et al. 2003). These host-guest studies involve substituting amino acids into known helix-forming sequences and measuring the change in helical content. Helical peptides are designed to include large amounts of alanine because of its properties as a strong α-helix former in solution (Marqusee et al. 1989). Individual amino acid propensities determined by host-guest studies are calculated from the relative increase or decrease in secondary structure content as measured by physical techniques like CD spectroscopy. The resulting propensity describes the ability of an amino acid to propagate an α-helix in the absence of any possible stabilizing, or destabilizing, side-chain to side-chain interactions and without consideration for a residue's neighboring sequence. The use of such a simplified model cannot inclusively describe the cooperative features of folding/unfolding transitions. However, revisions of these helix-coil transition theories are being developed to identify the intrinsic tendencies of individual amino acid residues to adopt helical

vs. nonhelical conformations. The ability of these statistical mechanics models (Zimm and Bragg 1959; Lifson and Roig 1961) to predict the secondary structures from the primary sequence alone has been limited to approximately 70% accuracy among non-homologous sequences (Holley and Karplus 1989; Stolorz et al. 1992; Rost and Sander 1993; Avbelj and Fele 1998).

Various organic solvents are commonly used to mimic the hydrophobic environment of the protein core in an effort to understand secondary structure formation of residues buried in the protein structure. These solvents represent a general class of compounds, which can accelerate protein folding (Lu et al. 1997) and stabilize tertiary structures (Nelson and Kallenbach 1986; Bienkiewicz et al. 2000). For example, the levels of the protective osmolyte trimethylamino-N-oxide (TMAO) builds up in many organisms and cells as a result of external stress such as high salinity, freezing and dehydration (Yancey et al. 1982). TMAO is known to counteract the denaturing properties of urea that may accumulate and cause deleterious intracellular effects (Yancey and Somero 1979; Celinski and Scholtz 2002). The alcohols methanol and 2,2,2-trifluoroethanol (TFE) are often used to study how amino acid propensities change with solvent environment (Rohl et al. 1996; Myers et al. 1998; Krittanai and Johnson 2000; Lawrence and Johnson 2002).

1.3.2 Solution State Studies via NMR

Spectroscopic techniques like CD and NMR are precise physical methods for providing conformational information on the solution state of

macromolecules. Of these, NMR is the only method for resolving structures at the amino acid-specific level. The development of Fourier transform (FT) pulsed NMR spectroscopy (Ernst and Anderson 1966) and its expansion into multiple dimensions (Aue et al. 1976; Jeener et al. 1979; Goldman and Porneuf 1994) signaled the conversion of NMR spectroscopy from a tool of the organic chemist into a method for elucidating solution structures previously inaccessible due to molecular weight limits and spectral overlap. Techniques for selective ^{2}D, ^{13}C and ^{15}N labeling of samples and the availability of higher field instruments have decreased the experimental time and dramatically increased the sensitivity for solving macromolecular structures and complexes above 50,000 Daltons (Pervushin et al. 1997). Pulse program sequence development and the strategies implemented to solve macromolecular structures are commonly used to measure conformation (Vuister and Bax 1994; Wang and Bax 1996), function (Akke 2002; Metcalfe et al. 2004; Revington et al. 2004) and dynamics (Abragam 1961; Lipari and Szabo 1982). Because of its ability to detail residue-specific changes in macromolecular structure and environment, NMR is commonly used to identify multimers in protein complexes (Altieri et al. 1995; Johnson 1999; Danielsson et al. 2002), map protein-protein or protein-NA interactions (Clore and Gronenborn 1998; Gross et al. 2003) and identify intermediates in the refolding pathways of proteins in solution (Roder and Wuthrich 1986; Udgaonkar and Baldwin 1988; Englander and Mayne 1992).

1.3.3 The Helix Formation Study Presented in Chapter 3

In Chapter 3, the ability of helices to propagate through a polypeptide sequence that is neither helix forming nor helix breaking (an indifferent sequence) is investigated in TFE, a known helix-inducing solvent that favors helix elongation (Storrs et al. 1992). Circular dichroism studies of the sequence YVAEAKTSGSRVAEAK in TFE indicate that the percent helicity is lower than when all other amino acid residues but aspartic acid are substituted for the central glycine residue (Waterhous and Johnson 1994; Lawrence and Johnson 2002). NMR spectroscopy was used to distinguish between the three possibilities: one central helix with nonhelical ends, a terminal helix, or an N-terminal and C-terminal helix with a central break. Data show that the sequence forms an N-terminal and a C-terminal helix with a break occurring in the region of the central glycine residue. The NMR solution structure of the YVAEAKTSGSRVAEAK peptide was solved in 80%, 90% and 100% TFE, and residue-specific secondary structure determinations were made by NMR. The elucidated NMR structures consist of an N-terminal and a C-terminal helix separated by the non-helical central residues G9 and S10. The inability to propagate through the TSGSR sequence is particularly interesting because, even though glycine is considered a "helix breaking" residue, glycine has been shown by CD spectroscopy to allow propagation within the structurally conserved VAGAK sequence (Waterhous and Johnson 1994; Lawrence and Johnson 2002). Theoretical calculations for predicting two distinct nucleation events under the most favorable conditions require a peptide greater than 20 residues long (Zimm and Bragg 1959; Werner et

al. 2002). The ability to induce a second nucleation event in a short peptide and the inability to propagate through an indifferent sequence indicates that consideration of sequence context and environment are integral to determining the helical propensity of an amino acid.

1.4 Structural and Dynamical Studies of a Chemically Adducted Thioredoxin

1.4.1 Chemical Modification of Thioredoxin

Chemical adduction provides a strategy for observing changes in protein structure and function that result from physical changes to residues without altering the primary sequence. Irreversible chemical adduction of proteins arises as a result of reactive byproducts from cellular detoxification of exogenous compounds such as halogenated alkanes (Dekant et al. 1989; Erve et al. 1995; Erve et al. 1996), isothiocyanates (Temmink et al. 1986) and quinones (Mertens et al. 1990; Lavoie and Hastings 1999). Generally, residue modification has a direct effect on protein stability, protein folding and unfolding profiles (Radford et al. 1991; Eyles et al. 1994), and *in vivo* modification of macromolecules has been shown to be both mutagenic and carcinogenic (Humphreys et al. 1990; Richardson et al. 1994). For example, 1,2-dichloroethane (DCE) is a *bis*-halogenated ethane known to irreversibly adduct the N-terminal cysteine residue in the active site C-G-P-C motif required for thioredoxin function (Erve et al. 1995; Kim et al. 2002).

Originally identified as an electron donor in the function of ribonucleotide reductase (Laurent et al. 1964), thioredoxin has since been well-characterized biochemically and biophysically (Jeng et al. 1994; Chivers and Raines 1997; Dyson et al. 1997; Dillet et al. 1998). Thioredoxin consists of an active site dithiol that mediates several oxidation-reduction reactions in a variety of cellular and viral reactions (Holmgren 1985; Holmgren 1989). The redox function of thioredoxin involves the formation of an intermediate consisting of a mixed disulphide between the thioredoxin active site cysteine (C32) and the electrophilic cysteine sulfur involved in a disulphide bond in the target protein (Jeng et al. 1995). The acidity of C32 makes it both susceptible to oxidation and reactive to electrophiles such as DCE. The final structure resulting from chemical modification by DCE provides a means for studying the salient dynamical features in a transient state mimic of thioredoxin function. By measuring residue-specific properties of the adducted thioredoxin, inferences can be drawn about structurally/functionally important amino acids that accommodate the rearrangements that occur during the redox mechanism of thioredoxin.

1.4.2 Theoretical and Experimental Studies of Protein Dynamics

X-ray crystallography is a technique that provides information on the average position of atoms and the amplitudes (B-factors) of their displacements from these average positions. B-factors are proportional to the mean square amplitude or oscillation of an atom around the position in the model. After the molecular dynamic simulations of bovine pancreatic trypsin inhibitor

(McCammon et al. 1977), plots of B-factors vs. primary sequence became a standard part of structural analysis, and have been interpreted as being indicative of internal motions of the protein (Artymiuk et al. 1979; Frauenfelder et al. 1979). Since then, B-factors have been used, for example, to calculate spatial fluctuations between atom pairs (Kundu et al. 2002). Molecular dynamics simulations of these structures describe the physical basis of the function and structure of proteins by studying individual particle motions as a function of time (Karplus and McCammon 2002). These simulations involve the use of complicated potential functions and significant computational resources.

Another technique for studying the motions in proteins involves directly measuring residue-specific motions by NMR (Abragam 1961; Lipari and Szabo 1982). The diversity and sensitivity of NMR measurements enables one to study proteins in widely varying solvation conditions and characterize residue-specific motions and dynamics in weakly interacting systems over vast timescale ranges. NMR is a powerful tool for studying residue-specific dynamical changes (Palmer 1997; Akke 2002) and has been used to characterize the residual entropy in proteins, which are believed to be important for protein function and mechanism of action (Feher and Cavanagh 1999; Stock 1999; Ishima and Torchia 2000; Wand 2001).

1.4.3 The Structural and Dynamical Study Presented in Chapter 4

The NMR analysis in Chapter 4 of the cysteine-ethyl glutathione-C32 adduct of thioredoxin (CEG-thioredoxin) shows that the modification has little

effect on the overall fold and structure. This corroborates the previous prediction of the structure of CEG-thioredoxin and the orientation of the glutathione moiety at the active site region (Kim et al. 2001). The backbone dynamics of CEG-thioredoxin indicate an overall increase in the amide motions in the picosecond to nanosecond time scale compared to the reduced and oxidized forms. Several regions show a larger than average increase in fast motions relative to the reduced and oxidized forms of thioredoxin. The active site and the two loops that make up the hydrophobic domain important for protein-protein interactions have a larger than average increase in fast motions. Interestingly, a significant number of the residues with a larger than average increase in these fast motions are found within defined secondary structure elements spatially and sequentially distant from the active site. These secondary structures are unified by their location in the secondary structures that lead into and out of the active site region. The increase in the backbone mobility identified by these measurements indicates that the change in protein plasticity due to active site rearrangements as a result of the adduction is not derived from a change in overall structure, but rather is a result of subtle changes in the stabilizing interactions between elements.

Chapter 2

The Extended Left-Handed Helix: A Simple Nucleic Acid Binding Motif

Joshua M. Hicks and Victor L. Hsu

Published in *Proteins: Structure, Function and Bioinformatics*

Elsever Science, New York, NY, USA

2004, 55: 85-92

2.1 Summary

The poly-proline type II extended left-handed helical structure is well represented in proteins. In an effort to determine the helix's role in nucleic acid recognition and binding, a survey of 258 nucleic acid-binding protein structures from the Protein Data Bank was conducted. Results indicate that left-handed helices are commonly found at the nucleic acid interfacial regions. Three examples are used to illustrate the utility of this structural element as a recognition motif. The third K homology domain of NOVA-2, the Epstein-Barr nuclear antigen-1, and the Drosophila paired protein homeodomain all contain left-handed helices involved in nucleic acid interactions. In each structure these helices were previously unidentified as left-handed helices by secondary structure algorithms but rather were identified as either having small amounts of hydrogen bonding pattern to the rest of the protein or as being "unstructured". Proposed mechanisms for nucleic acid interactions by the extended left-handed helix include both non-specific and specific recognition. The observed interactions indicate that this secondary structure utilizes an increase in protein backbone exposure for nucleic acid recognition. Both main-chain and side-chain atoms are involved in specific and non-specific hydrogen bonding to nucleobases or sugar-phosphates, respectively. Our results emphasize the need to classify the left-handed helix as a viable nucleic acid recognition and binding motif, similar to previously identified motifs such as the helix-turn-helix, zinc fingers, leucine zippers and others.

2.2 Introduction

The physical basis of nucleic acid (NA) recognition by proteins is often hypothesized via the identification of commonly classified secondary structures. Interestingly, approximately 25% of the amino acid residues of proteins whose structures are reported in the Protein Data Bank (PDB) are unidentified by secondary structure assignment algorithms (Sreerama and Woody 1994). It would therefore be reasonable to expect that not all of the possible NA binding motifs have been elucidated. The poly-proline type II (PPII) extended left-handed helix is an unreported secondary structure in the PDB, whose documentation contains standard notations for α-helices, 3_{10}-helices, β-strands and turns. This study identifies left-handed helices that are either incorrectly assigned by secondary structure hydrogen bonding patterns or are found within the structurally unassigned regions of the protein, yet are responsible for NA recognition and binding.

Within the PDB, classification of these structurally unassigned regions is believed to be made up of unordered peptide segments (Sreerama and Woody 1994). Assignment of these regions as random coils prevails because most automated algorithms identify α-helices, β-strands, and other common secondary structures via hydrogen bonding parameters (King and Johnson 1999). Until recently, the poly-proline type II (PPII) left-handed helix (also known as: P_{II}, P_2, extended left-handed or left-handed 3_1 helices) has been an underreported secondary structure in the literature. This left-handed helix is a well-defined

extended structure, but lacks intraelement hydrogen bonds such as those found in α-helices and γ-turns, or repeating interchain hydrogen bonds between main chain hydrogen bond donors and acceptors to form higher ordered structures such as the β-pleated sheet. Contrary to what its name suggests, the presence of proline is not obligatory (Adzhubei and Sternberg 1993). In order to limit the confusing and dated nomenclature that "PPII" holds concerning proline content, this secondary structure element should more correctly be referred to by its 3_2-helix symmetry nomenclature (Van Holde et al. 1998). Although the descriptor "left-handed 3_1-helix" is correct, because the conventions of symmetry are based on right-handed Cartesian coordinates, the more accurate term is "3_2-helix".

The 3_2-helix was initially identified in proteins containing disproportionately large amounts of proline, such as collagen, where 3_2-helices comprise the major conformation (Pauling and Corey 1951c). Attempts at locating and quantifying 3_2-helices in existing solved structures are based on either conformational searches (Adzhubei et al. 1987; Adzhubei et al. 1987; Adzhubei and Sternberg 1993; Sreerama and Woody 1994; Stapley and Creamer 1999) or searching for protein sequences with proline-rich regions (PRRs) that are believed to form left-handed helices *in vivo* (Williamson 1994). Examples of PRRs role in protein-protein interactions include those essential for profilin function, which involves the binding of proline-rich regions at preorganized sites that recognize 3_2-helical structures (Mahoney et al. 1997; Ostrander et al. 1999). Deletion studies on the proline-rich carboxyl-terminal domains (CTD) of both RNA polymerase II (Allison et al. 1985) and bacteriophage pf1 gene V protein

(Fox et al. 1999) have shown significant effects on NA recognition and protein function. Interestingly, T-cell antigen recognition by major histocompatible class II (MHC II) requires no proline residues, yet the antigen still adopts the 3_2-conformation (Jardetzky et al. 1996; Reinherz et al. 1999).

Perhaps the most well studied proline-rich recognition element is the Src-Homology 3 (SH3) domain, a β-barrel composed of 60 amino acids which binds a core sequence of four amino acids (PxxP) (Ren et al. 1993; Yu et al. 1994; Mahoney et al. 1997). SH3 domains are found linked to proteins required for tyrosine kinase signaling and have been identified in several aspects of protein-protein interactions involving neutrophil cytochrome oxidase and cytoskeletal structure (Yu et al. 1994). A pseudo two-fold symmetry of the 3_2-helix allows two possible binding orientations of the core sequence by the SH3 domain (Feng et al. 1994).

Recently, there has been great interest in these structures, as 3_2-helices have been implicated in several conformation-based diseases including the identification of large amounts of 3_2-helices found in amyloidogenic prefibril intermediates (Blanch et al. 2000). 3_2-helices have also been shown to be essential in several biological processes such as signal transduction, transcription, cell motility and the immune response (Kelly et al. 2001). The possibility of a 3_2-helix involved in NA binding was hypothesized for the TAR binding region of the N-terminal HIV-1 Tat peptide fragment (Kanyalkar et al. 2001).

Circular dichroism spectroscopy measurements by Tiffany and Krimm (Tiffany and Krimm 1968) reported 3_2-helices as an important conformation in

short poly-lysine and poly-glutamic acid peptides, and left-handed helices have been identified in poly-alanine peptides (Chou and Fasman 1974; O'Neil and Degrado 1990; Chakrabartty et al. 1994). In a conformational survey of 68 globular proteins by Adzhubei and coworkers (Adzhubei et al. 1987), approximately 20% of the total residues were identified in an extended left-handed conformation (designated the M-conformation) similar to the 3_2-helix. However, this study was limited to individual residues regardless of the conformation of adjacent residues. In a study of 3_2-helices composed of four or more consecutive residues, Adzhubei and Sternberg (Adzhubei and Sternberg 1993) found 96 occurrences of 3_2-helices comprising approximately 5% of the residues. Sreerama and Woody (Sreerama and Woody 1994) reported in a conformational survey of 16 high-resolution structures that approximately 10% of these residues are in the 3_2-conformation with half of the residues being isolated. Lower estimates from a conformational survey of 274 non-homologous sequences indicates that the number of residues in 3_2-helices (of at least four residues in length) is approximately 2% of the total residues (Stapley and Creamer 1999).

The canonical 3_2-helix is located in ϕ, ψ-space at -78 , $+146$ (Adzhubei et al. 1987; Sreerama and Woody 1994; Stapley and Creamer 1999) and contains the *trans*-peptide bond ($\omega = 180$). This conformation of the left-handed helix creates a regular, repeating peptide unit from n to $n\pm3$ (Fig. 2.1). The extended structure of the 3_2-helix alters the backbone exposure over the average for residues in other secondary structures by increasing the solvent accessible area of polar surfaces by 60% and in nonpolar surfaces by 50% (Stapley and Creamer

1999). As a structural amino acid, proline has obvious entropic (conformational limitation) and enthalpic (lack of amide proton) folding benefits. The additional presence of the large C_δ moiety in the prolyl ring restricts the backbone dihedral conformation at the *n-1* residue to the regions centered at (ϕ, ψ) = (-61 , -35) and (-65 , +150) through locally driven steric limitations (Creamer 1998). The prevailing strategy for substitution of proline residues in 3_2-helices is the incorporation of residues capable of forming stabilizing interactions between polar side chain groups and main chain atoms, such as between the side-chain amide and the main-chain carbonyl in glutamine (Stapley and Creamer 1999).

On the basis of our own interest in the TAR binding region from the Tat protein fragment and the proline-rich region of the RNA polymerase II carboxyl-terminal domain, the 3_2-helix content in NA binding proteins from crystal structures in the PDB was investigated. Unfortunately, most commonly used secondary structure assignment algorithms, such as the Database of Secondary Structure in Proteins (DSSP) (Kabsch and Sander 1983), identify secondary structures by hydrogen-bonding patterns of main-chain atoms, which 3_2-helices do not contain. Thus, the computer program XTLSSTR that analyses protein structures based on interresidue carbonyl vectors for assigning secondary structure elements (King and Johnson 1999) was used to identify 3_2-helices as defined by Sreerama and Woody (1994). Of the 258 NA-binding proteins for which crystal structures have been solved at 2.70 Å resolution or better, 219 had at least one region of 3_2-helix with three or more consecutive residues (constituting at least one turn of the helix).

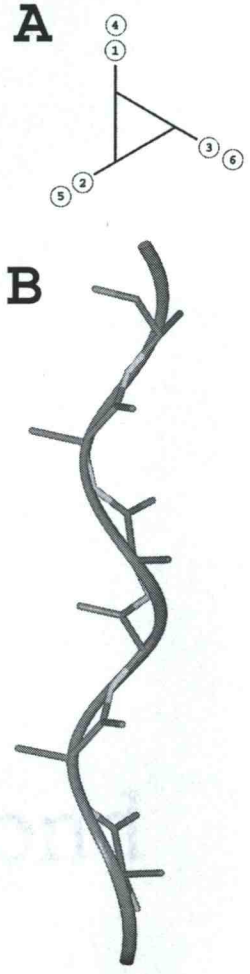


Figure 2.1 The 3_2-helix of a poly-alanine chain (A) Helical Wheel viewed down the helix axis. (B) Left handed 3_2-helix; $\phi = -78$, $\psi = +146$, and $\omega = 180$. This figure was prepared with InsightII (Biosym/Msi 1994).

The 3_2-helix as a NA binding motif is illustrated by three examples of protein-nucleotide complexes identified within the database. The third K homology domain (KH3) of NOVA-2, the Epstein-Barr nuclear antigen-1 (EBNA-1), and the homeodomain from the Drosophila paired (Pax) protein structures were the only structures in this NA database that are found in both the free and bound state. Recognition and binding are characterized by using the full capability of the motif, incorporating the exposed peptide backbone for sugar-phosphate recognition and, in some cases, side chain groups for additional nucleotide-specific or sugar-phosphate backbone interactions. The 3_2-helix possesses the ability to facilitate all necessary contacts for non-specific and/or specific NA interactions, and emphasizes its role as a new NA recognition motif whose structural simplicity allows for more diverse methods of recognition than previously identified motifs.

2.3 Results and Discussion

2.3.1 Statistics

Table 2.1 identifies 3_2-helices as reported by Stapley and Creamer (Stapley and Creamer 1999) and from our analysis of the data from 258 NA binding proteins. Additionally, the data from the KH3, the EBNA-1, and the Pax NA binding proteins was included to characterize the nature of 3_2-helices in our examples. 3_2-helix content in the bound structures uses non-mutant structures of

the third K homology (KH3) domain, the Epstein-Barr nuclear antigen-1 (EBNA-1), and the homeodomain from the Drosophila paired (Pax) protein. The structures of a fourth protein, the sex-lethal protein (PDB codes 3SXL and 1B7F) from *D. melanogaster*, were also determined in the free and bound form, respectively. However, due to the absence of residues in the model of the free conformation, some of which are subsequently identified as 3_2-helices involved in NA recognition in the bound form, the sex-lethal protein was not included.

3_2-helix content within the 258 NA binding protein data set agrees with estimates by both Adzhubei and Sternberg (1993) and Sreerama and Woody (1994). Comparison of 3_2-helices in NA binding proteins to 3_2-helical statistics from of survey of 274 non-homologous structures (Stapley and Creamer 1999) indicates a similar pattern of residue types found in the 3_2-conformation. Of the residues observed in 3_2-helices in this study, proline is the most abundantly represented individual amino acid at approximately 16%. Because searches for poly-proline rich regions (PRR) are based on regions with more than 33% proline content (Williamson 1994), it explains how these helices would have been unrecognized by PRR search strategies. The most abundant type of 3_2-helix residues in the 258 NA binding proteins is the hydrophobic amino acids alanine (A), valine (V), leucine (L) and isoleucine (I), which comprise over one quarter of residues in this conformation. This is in stark contrast to the KH3, EBNA-1, and Pax protein data in which only 11% of the 3_2-helix residues are alanine, valine, leucine, and isoleucine. Although the KH3, EBNA-1, and Pax proteins are not a representative database of all 3_2-helix recognition elements, these numbers reflect

the properties of 3_2-helices involved in NA recognition. Hydrophobic residues are under-represented in 3_2-helices involved in NA recognition compared to 3_2-elements represented in all NA binding proteins. Alternatively, the most abundant type of 3_2-helix residues in the KH3, EBNA-1, and Pax protein data are the polar serine (S) and threonine (T) residues (24%). Only 11% of the 3_2-helices in the NA binding protein database contain serine and threonine. The serine and threonine residues identified in the three proteins provide 3_2-helices with the ability to solvate the structure, form hydrogen bonds to nucleic acids, and stabilize the 3_2-helix backbone through side chain to main chain hydrogen bonds. In both the KH3, EBNA-1, and Pax protein data and the NA binding protein data, 3_2-helices contain a substantial amount of the basic lysine (K), arginine (R), and histidine (H) residues. The charged side chains of these basic residues are ideal for NA recognition and have been identified in the structures presented in this paper. Less than eleven percent of the total 3_2-helix composition is composed of the bulky and sulfur containing amino acids (cysteine, methionine, tryptophan, phenylalanine, and tyrosine). The remaining 10-15% of the residues are the polar amino acids glutamic acid (E), aspartic acid (D), glutamine (Q), and asparagine (N).

It is interesting to note that in the KH3, EBNA-1 and Pax protein data glycine residues are favored in 3_2-helices that recognize nucleic acids. Arguments concerning the conformational freedom of glycine residues in the absence of steric constraints suggest a higher entropic penalty for folding compared to all the other amino acids. However, Ramachandran and coworkers (1966) identified

Table 2.1 3_2-helix amino acid composition in NABP with ≤ 2.70 Å resolution

	Average amino acid occurrence		
	KH3, EBNA-1, and Pax proteins	NA binding proteins (258 structures)	Stapley and Creamer (1999) (274 Non-homologous structures)
Average protein length	105	182	228
Number of 3_2-helix residues *	12.3 (11.7)	9.4 (5.2)	4.5 (2.0)
3_2-Helix residues*			
A/V/L/I	1.3 (10.9)	2.6 (28.1)	1.1 (24.0)
M/C	0.5 (4.1)	0.3 (3.5)	0.1 (3.1)
F/W/Y	0.5 (4.1)	0.7 (7.5)	0.3 (5.9)
K/R/H	2.6 (20.8)	1.5 (15.6)	0.6 (12.3)
D/E	0.3 (2.7)	1.1 (10.1)	0.5 (10.0)
N/Q	0.7 (5.6)	0.5 (4.9)	0.4 (8.0)
S/T	2.9 (23.5)	1.0 (10.8)	0.5 (10.5)
Proline	2.0 (16.3)	1.5 (16.0)	1.1 (24.0)
Glycine	1.3 (10.9)	0.3 (3.5)	0.4 (3.5)

* Numbers in parentheses are the percent of occurrences

higher order structures of poly-glycine type II (PGII) helices as conformationally identical to 3_2-helices in asymmetric fiber formation. If recognition and binding are facilitated by intermolecular interactions involving main chain atoms, the presence of glycine at interfacial regions is ideal. The lack of a large group attached to the α-carbon allows complete backbone exposure. This type of interaction has been identified in the Epstein-Barr Nuclear Antigen-1 (Fig. 2.3) and is discussed below.

In order to address the possibility that the 3_2-helix numbers in the 258 NA binding protein dataset are inflated by the identification of β-strands as 3_2-helices, secondary structure assignments using the XTLSSTR program were compared to results from DSSP analysis (Kabsch and Sander 1983). Some of the 3_2-helix regions identified by XTLSSTR were classified by DSSP as being short (three residues or less) extended strands (E), beta bridges (B) or as being parts of turns (T or S). Visual inspection indicates that these cross-identified structures are either: 1) terminal residues of β-strands, 2) linkers (three residues or less) between non-parallel β-sheets, 3) β-twists (three residues or less), or 4) results of hydrogen-bonding patterns of turns. These hydrogen bond patterns are key to main chain stabilization, but clearly there is important conformational information that is lost when searching only for standard secondary structure hydrogen bonding patterns. Since 3_2-helices are often located in solvent exposed regions or on protein surfaces (Adzhubei and Sternberg 1993), these hydrogen bond interactions with the core of the protein and/or the resulting backbone solvation effects are common strategies for stabilizing functional 3_2-helices.

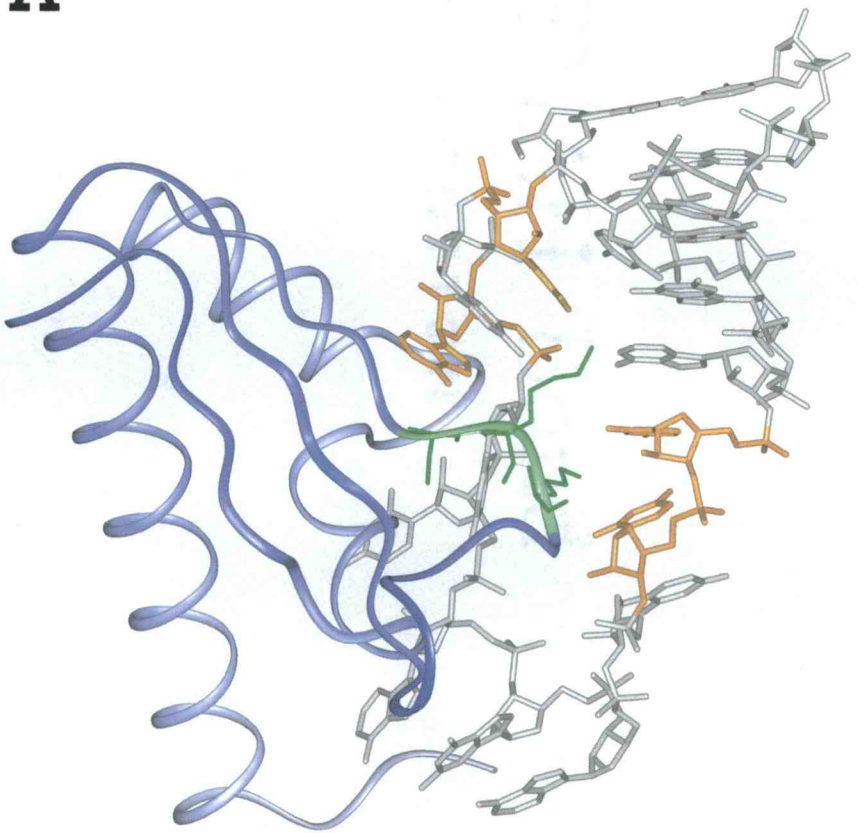
2.3.2 Third K Homology Domain of NOVA-2

The first example of NA recognition by a 3_2-helix is the third K homology (KH3) domain of NOVA-2, a sequence-specific single stranded (ss) RNA binding domain implicated in fragile X syndrome and paraneoplastic disease (Lewis et al. 1999; Lewis et al. 2000). The K homology (KH) motif is a key component of regulators of RNA metabolism in neurons. First isolated in sera from cancer patients, KH3 is a conserved region that is repeated three times in NOVA-2 isolated from human sources, but can be found with as many as 15 repeats in other KH proteins (Lewis et al. 1999; Lewis et al. 2000). The human NOVA-2 protein has been crystallized and has had its structure determined both free (PDB code 1DTJ (Lewis et al. 1999; Lewis et al. 2000)) and in a ssRNA complex (PDB code 1EC6 (Lewis et al. 2000)).

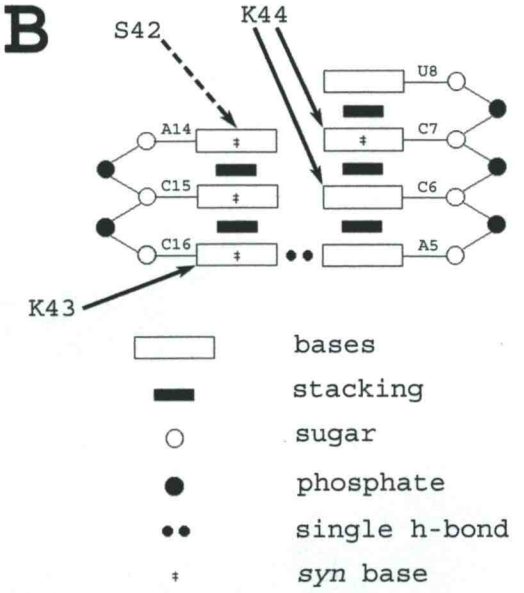
Figure 2.2 shows the contacts of a 3_2-helix in NOVA-2 complexed to the ssRNA sequence. Residues S42, K43, and K44 located at the C-terminal end of a β-strand adopt a 3_2-helical conformation, but show no main chain hydrogen bonding (DSSP analysis, Fig. 2c). The terminal lysine (K44) main chain atoms participate in hydrogen bonds formed by the folding back of subsequent residues, where close inspection shows that the identification of the turn is based on backbone hydrogen bonding at the n+2 residue, but K44 is still in the 3_2-conformation. In contrast, the corresponding residues in the structure of the free protein are in a different conformation and have B-factors >70 along the main chain. The protein maximizes its interaction with the RNA by inserting the side

Figure 2.2 Structure of the Nova KH Domain. (A) Ribbon diagram shows the color-coded 3_2-helix (green) motif; RNA is included as color-coded atomic stick (gray) with recognition nucleotides (gold). (B) Schematic drawing of RNA complex 1 (1EC6:D) showing residue-specific contacts using main-chain (dashed line) and side-chain (solid line) atoms. (C) Sequence representation of the RNA recognition sequence (1EC6:D); alignment of 3_2-helix amino acids (S42-K44) in the free (1DTJ:A) and bound (1EC6:A, B) to DSSP (T, hydrogen-bonded turn) and XTLSSTR (T, hydrogen-bonded turn; P, 3_2-helix) secondary structure identification algorithm. This figure was prepared with InsightII (Biosym/Msi 1994).

A



B



C

```
1EC6:D    GAGGACCUAG  AUCACCCCUC

1DTJ:A    SKK         1EC6:A,B    SKK
DSSP      -TT                     --T
Xtlsstr   -TT                     PPp
```

chains of the 3_2-helical residues directly into the core ssRNA base sequence while utilizing the 3_2- helical motifs' exposed polar backbone for sugar-phosphate recognition. The side chains of both K43 and K44 are involved in base-specific recognition where the N_ζ-H are capable of forming stable hydrogen bonds to the exocyclic carbonyl oxygens on cytosine bases 16 and 7, respectively. Base interactions are accentuated by the ability of the main chain carbonyl oxygen of S42 to be stabilized by the exocyclic amino ($-NH_2$) group on adenine-14. These specific interactions are indicated in Figure 2.2b.

2.3.3 Epstein-Barr Nuclear Antigen 1

A second binding example is the comparison of the free (PDB code 1VHI (Bochkarev et al. 1998)) and complexed (PDB code 1B3T (Bochkarev et al. 1998)) forms of Epstein-Barr nuclear antigen 1 (EBNA-1) dimers. The EBNA-1 dimer binds the latent origin of DNA replication (oriP) from the Epstein-Barr virus (EBV), activating DNA replication initiation once every cell cycle. The dyad symmetry element of the oriP site consists of an 18-basepair palindromic sequence recognized by EBNA-1.

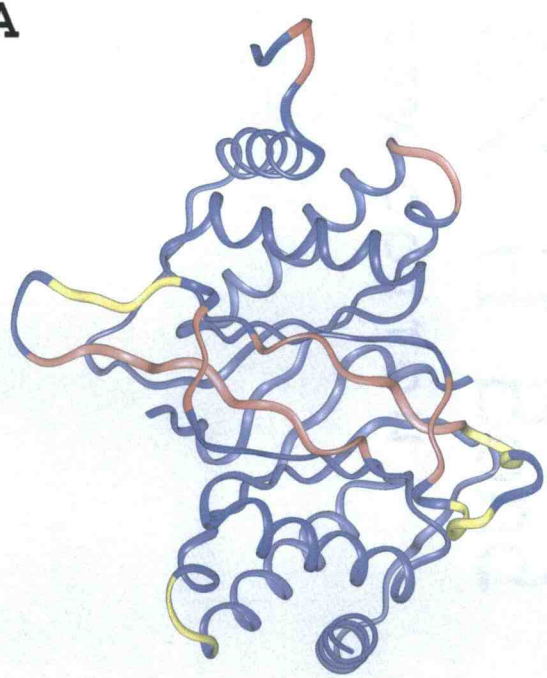
Figure 2.3 shows the alignment of secondary structure identification (the N-terminal residues of 1B3T:B have been included to illustrate the symmetry of DNA recognition by this protein). Both the bound and unbound forms contain some identical 3_2-helices (red) including one long 3_2-helix in each subunit. The bound form exhibits a four-residue lengthening (green) of this long 3_2-helix in one of the asymmetric units as well as formation of additional 3_2-helices proximal to

the recognition site. As in the example above, a comparison to the DSSP results indicates how parts of the 3_2-helices are stabilized. However, in the existing literature, none of the residues currently identified as 3_2-helices were characterized as being a regular repeating secondary structure but rather were identified as either having small amounts of hydrogen bond patterns to the rest of the protein or as being "unstructured". This lack of identification in bound form was not due to the limitations of the experimental data, as the average B-factors in these regions were all below 22 with most of the residues having main-chain atom B-factors of approximately ten.

Figure 2.3 shows a comparison of the structures and a backbone representation of the macromolecular interactions. The core binding recognition domains (residues 472-474, 534-541 and 555-557) are maintained in both the free protein and complexed protein (red), while other chains shift upon binding (green). There is no identified reason for the ordering of some parts of the protein, such as the expansion of the two "lobes" of EBNA-1 upon binding. Whether these regions are created for other interactions after DNA binding or are just a result of crystal packing is not known. Inspection of the uncomplexed dimer shows that this motif is bi-directional with symmetry related contacts in each of the monomers. Because oriP is palindromic, binding is independent of 5' to 3' upstream/downstream orientation along the DNA axis. The pseudo two-fold axis of binding at the active site is required for reversible upstream/downstream orientation along the DNA, as illustrated in repeated structural implementation in proteins that bind to palindromic sequences (Bochkarev et al. 1998).

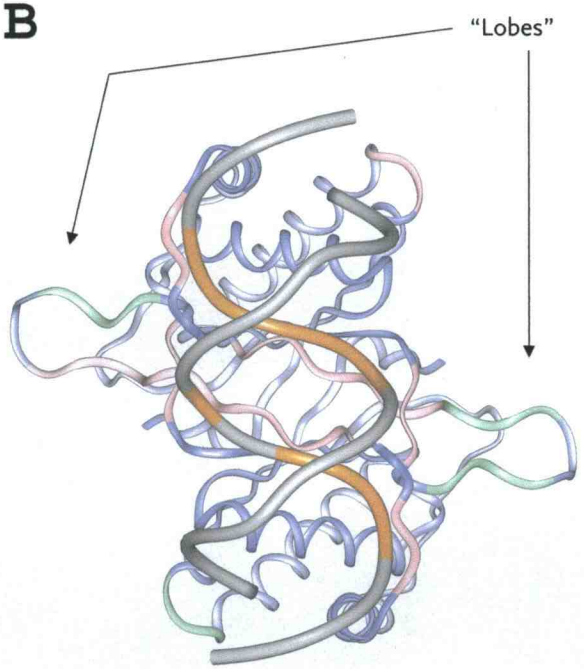
Figure 2.3 Ribbon structures of the Epstein-Barr nuclear antigen 1 (EBNA-1) dimers; residue comparison of secondary structure assignments. 3_2-helices found in both states are shown in red. Secondary structure alignment using DSSP (T, hydrogen-bonded turn; S, bend; E, extended strand; B, β-bridge) and XTLSSTR (T, hydrogen-bonded turn; P, 3_2-helix; G, 3_{10}-helix). (A) Free form of EBNA-1 (1VHI), residues not in the 3_2-helix in absence of DNA are shown in yellow. (B) EBNA-1 with DNA backbone (1B3T); residues that adopt the 3_2-helix upon recognition are shown in green, DNA backbone contacts are shown in gold. This figure was prepared with InsightII (Biosym/Msi 1994).

A



Residues	472	474	534	545	549	552	555	557	586	588
1VHI:A	GGS		TPLSRLPFGMAP		PQPG		RES		KPA	
DSSP	TS-		---EE-B--SS-		SS--		-B-		TTT	
Xtlsstr	PPp		PPPPPPPp----		-PpG		PPp		PPp	
1VHI:B			TPLSRLPFGMAP		PQPG		RES		KPA	
DSSP			---EE-B----S		----		-B-		TTT	
Xtlsstr			PPPPPPPPPPPp		---T		PPp		-Pp	

B



	472	474	534	545	549	552	555	557	586	588
1B3T:A	GGS		TPLSRLPFGMAP		PQPG		RES		KPA	
DSSP	---		---EE-B----S		----		-B-		TTT	
Xtlsstr	PPp		PPPPPPPPPPPp		PPPp		PPp		PPp	
1B3T:B	GGS		TPLSRLPFGMAP		PQPG		RES		KPA	
DSSP	---		---EE-B----S		----		-B-		S-T	
Xtlsstr	PPp		PPPPPPPPPPPp		PPPp		PPp		PPp	

Residues G472 and G473 illustrate the role of glycine in the 3_2-helix and interaction with the ligand as described in the statistical results above. Amide protons of the main chain associated with G472 and G473 are oriented to make hydrogen bonds to oxygen atoms (O1P) on phosphate groups of the sugar-phosphate backbone of nucleotides cytosine-113 (and the symmetric cytosine-213) and cytosine-114 (and the symmetric cytosine-214), respectively. Lack of a side chain on G472 prevents unfavorable steric interactions with the sugar-phosphate backbone and enables G472 and G473 to form hydrogen bonds to consecutive backbone phosphate groups. The short 3_2-helix of residues 586-588 (KPA) introduces the K586 side chain to the DNA backbone; however, the premature end to the DNA helix in the crystal structure prevents elucidating the complete role of the 3_2-helix's interactions in this region. The backbone amide of L536 in the central recognition element is aligned with O1P of guanine-112 (and the symmetric guanine-212). R538 side chain interactions include positioning along the sugar-phosphate backbone for Nε-H to O1P of thymine-111 (and the symmetric thymine-211). Additional interactions between $N\zeta^1$-H and $N\zeta^2$-H to O5' of adenine-110 (and the symmetric adenine-210) completes the characterization of the 3_2-helix. All non-specific interactions by the 3_2-helices are facilitated through various parts of the protein backbone and are augmented by side chain atoms.

The non-specific binding aspects are analogous to proposed mechanisms based on observations of the carboxy-terminal domain (CTD) of RNA polymerase II and pf1 gene V protein. The CTD of RNA polymerase II is a highly conserved

motif containing the heptapeptide YSPTSPS sequence repeated 17 to 52 times. Deletion studies of RNA polymerase II CTD show significant length dependence in mammalian cells on growth and lethality (Allison et al. 1985). The RNA polymerase II CTD possesses a large number of amino acids (S/T) with side chains capable of interacting with main chain hydrogen bond acceptors/donors. Although the repetition of proline every three to four residues suggests a required periodicity in the motif to facilitate its formation, helices containing more than 33% proline are not likely to be involved in NA recognition. The lack of backbone amide protons and polar side-chain groups would leave only the backbone carbonyl oxygen atoms as a possible participant in hydrogen-bond formation to NA (i.e. every third residue). As a result, proline residues would only be functional as acceptors in positions where the prolyl ring is not oriented towards the NA. Additionally, due to the close proximity of the exposed backbone to the NA and the fixed position of the prolyl ring, the presence of proline residues at any other position of a 3_2-helix involved in NA binding would potentially create unfavorable intermolecular steric interactions.

Gene V CTD of bacteriophage pf1 contains many prolines, but is also characterized by a large number of alanine (A) and glutamine (Q) residues. Conformational CD studies of gene V CTD show the formation of 3_2-helices in hydrophobic environments or in the presence of single-stranded (ss) DNA (Fox et al. 1999). Deletion studies also indicate that the CTD of gene V is not required for affinity, but does cause a change in kinetics (Fox et al. 1999). The absence of a

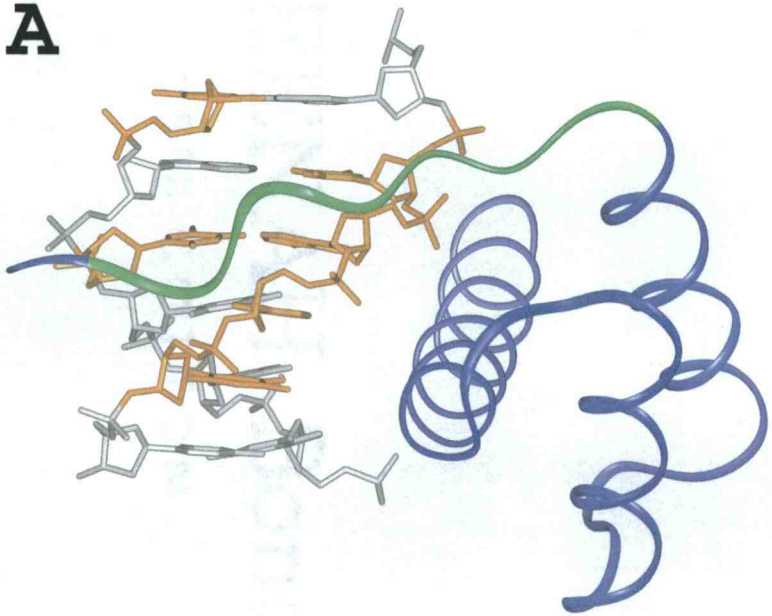
non-specific DNA recognition motif could explain the previous observations on how deletion studies of these regions affect affinity, but not specificity.

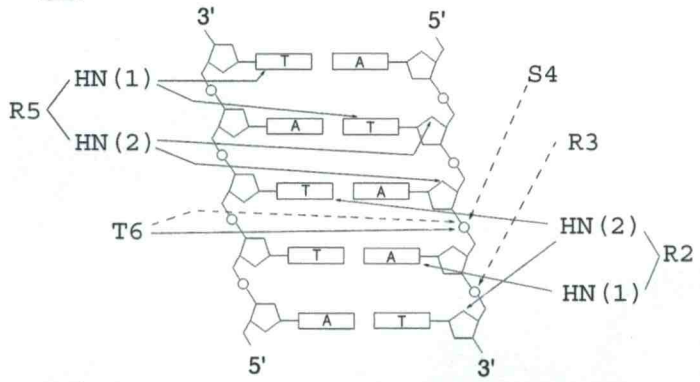
2.3.4 Homeodomain from Drosophila Paired Protein

The third example of protein binding to nucleic acids via a 3_2-helix, the homeodomain from the Drosophila paired (Pax) protein, is used to illustrate how 3_2-helices can be used in multiple recognition mechanisms using a minimal number of residues. The paired homeodomain (PDB code1FJL (Wilson et al. 1995)) is a double stranded (ds) DNA binding domain that plays a critical role in developmental differentiation and development; it forms cooperative dimers at a palindromic "P3" sequence found in the promoters/enhancers of genetic targets of the paired gene (Wilson et al. 1995). The "recognition arm" of the N-terminal residues is similar to other homeodomains and is conformationally identical to the Oct-1 (PDB code 1OCT) homeodomain arm (Klemm et al. 1994) indicating that its conformation is not altered by its interactions. However, none of the previous studies identified the recognition arm as a regular repeating structure. Instead this extended structure was reported as not conforming to any regular secondary structure. Unlike the two examples above, the recognition arm of Pax does not have any regular pattern of main chain interactions to the rest of the protein and DSSP does not identify any structure in this domain (Fig. 2.4c).

Non-specific recognition involves both backbone and main chain polar groups on residues R2, R3, S4, R5 and T6. Interaction with the sugar phosphate backbone include alignment of the $N\zeta^2$-H group of R2 to the O4' of thymine-6

Figure 2.4 Structure of the homeodomain from Drosophila paired (Pax) protein. (A) Ribbon diagram of 1FJL:A show the color-coded 3_2-helix (green) motif; DNA recognition sequence is included as color-coded atomic stick (grey) with recognition nucleotides (gold). (B) Contact map to the DNA complex show residue specific interactions using main-chain (dashed line) and side-chain (solid line) atoms. (C) Sequence representation of the DNA sequence (1FJL:D, E); secondary structure alignment using DSSP and XTLSSTR_(P, 3_2-helix). This figure was prepared with InsightII (Biosym/Msi 1994).





C

```
          1         7                1         9
1FJL:D  5' A A T A A T C 3'   1FJL:A  QRRSRTTFS
1FJL:E  3'   T A T T A G 5'   DSSP     ---------
            14       9        Xtlsttr  PPPPPPPPp
```

and backbone amide of R3 to the phosphate group of thymine-6. Backbone amide groups of both S4 and T6 and the side chain hydroxyl of T6 interact with the phosphate group of adenine-5. The side chain protons on $N\zeta^2$ of R5 hydrogen bond to consecutive sugar O4' groups on thymine-3 and adenine-4. Specific recognition is facilitated by interaction of the side chain amides of residues R2 and R5. The $N\zeta^2$-H group of R2 is oriented towards the exocyclic oxygen on thymine-12 and the $N\zeta^1$-H group hydrogen bonds to the nucleobase N3 of adenine-5. The side chain amide $N\zeta^1$-H of R5 makes two contact points with the exocyclic oxygens of thymine-3 and thymine-13.

The smaller size of the motif compared to other more complex motifs allows closer access to the DNA minor groove (Fig. 2.4). The recognition arm has an average main chain B-factor of 30.6 and interactions (Fig. 2.4b) determine how the arm is ordered along the minor groove. This analysis helps identify two properties of this motif. The periodicity of the helix allows repeated insertion of side chain groups every three residues (R2 and R5 of Pax), while intermediate residues have backbone amides oriented towards the DNA. Recognition of consecutive phosphate groups by backbone amides, similar to the interactions of G472 and G473 in the EBNA1 example above, occurs through backbone exposure and orientation.

2.4 Conclusion

This study has presented the illustrations of three structures to represent the diversity in how the 3_2-helix binds to nucleic acids. Non-specific recognition is facilitated through main chain amide and side chain interactions to the NA backbone. Sequence-specific base recognition occurs through intermolecular interaction networks involving protein side-chain atoms and where helix symmetry creates a repeat in side chain and main chain atom alignment every three residues. The lack of complexity and overlapping elements, such as in helix-turn-helix domains, allows the utilization of the exposed backbone groups of the 3_2-helix for recognition and binding. The ability to recognize areas not easily accessible, such as the minor groove, is enhanced by the large number of potential interactions with a minimal number of residues. Clearly 3_2-helices are important elements for protein-protein interactions (as seen in the SH3 domain) and it is likely that a closer inspection of the proteins within the PDB will reveal other functional utilization of this motif. We have shown that 3_2-helices are also vitally important in facilitating interactions with nucleic acids and as such should be recognized as a nucleic acid binding motif.

2.5 Materials and Methods

A search of the Brookhaven PDB for "DNA binding", "DNA-binding", "RNA binding", and "RNA-binding" yielded 258 crystal protein structures

resolved at 2.70 Å or better. The Brookhaven four letter identification codes for the NA binding protein structures are: 1914, 1a02, 1a0p, 1a1f, 1a32, 1a3c, 1a3q, 1a4x, 1a62, 1a6i, 1a6y, 1a73, 1a74, 1a7g, 1a7w, 1a8h, 1a8v, 1aa7, 1ad2, 1ad6, 1ae2, 1ae3, 1ail, 1ais, 1akh, 1ako, 1ali, 1am9, 1anv, 1apl, 1aqi, 1aqj, 1au7, 1awc, 1axc, 1azp, 1b01, 1b34, 1b3t, 1b67, 1b72, 1b7f, 1b8i, 1b8z, 1b9m, 1baz, 1bc8, 1bdh, 1bdt, 1bf4, 1bgf, 1bgw, 1bhl, 1bhm, 1bi0, 1bi4, 1bia, 1bj0, 1bjt, 1bjz, 1bl0, 1bl3, 1bm8, 1bm9, 1bnz, 1bxe, 1by4, 1by9, 1c1k, 1c9b, 1c9s, 1cdw, 1cf7, 1ci4, 1cit, 1cjs, 1cmb, 1cmc, 1cuk, 1cx0, 1cyq, 1cz0, 1d1l, 1d1m, 1d2e, 1d3b, 1d3u, 1d3y, 1dbq, 1dct, 1dd9, 1dde, 1dew, 1dg1, 1di2, 1div, 1dj0, 1dm9, 1dml, 1dp7, 1dps, 1drz, 1dsz, 1dt4, 1dtj, 1du7, 1dux, 1dvk, 1ec6, 1ecr, 1efc, 1ega, 1egw, 1enh, 1euj, 1evx, 1fip, 1fjl, 1f1z, 1f5a, 1fby, 1fia, 1fm6, 1fos, 1g5y, 1g6n, 1gcb, 1gkh, 1gpc, 1gvp, 1ha1, 1hcq, 1hcr, 1hei, 1hta, 1hus, 1huu, 1hwt, 1ign, 1ipp, 1itg, 1jhg, 1jmc, 1jnk, 1ksp, 1lat, 1lbh, 1lmb, 1mb1, 1mnm, 1noy, 1noz, 1odd, 1ois, 1opc, 1pcf, 1pcz, 1pdn, 1per, 1pfs, 1piq, 1pjr, 1plq, 1pnr, 1pue, 1qaw, 1qhh, 1qhg, 1qmh, 1qpz, 1qqb, 1quq, 1qvc, 1ram, 1rep, 1rl2, 1rl6, 1rmd, 1run, 1ruo, 1rss, 1sei, 1sfe, 1skn, 1svc, 1tc3, 1tro, 1tsr, 1up1, 1urn, 1vhi, 1vok, 1vol, 1vpw, 1vqb, 1vqc, 1wap, 1wet, 1whi, 1wrp, 1xbr, 1xxa, 1ycq, 1yhb, 1yrn, 1zaa, 1zay, 1zdh, 1zdi, 1zii, 1zqa, 2a8v, 2abk, 2adm, 2bam, 2bop, 2cgp, 2dgc, 2dtr, 2fok, 2gli, 2gn5, 2hap, 2hdd, 2hts, 2irf, 2itg, 2lbd, 2nll, 2paw, 2pud, 2pug, 2reb, 2tdx, 2trt, 2wrp, 3bam, 3fis, 3gcb, 3hdd, 3hts 3sxl, 3ull, 4fis, 4lbd, 5cro, 6pax and 9ant.

3_2-Helix identification was accomplished by analysis of the secondary structure assignment output by the XTLSSTR program (King and Johnson 1999) which uses a predefined 3_2-helix secondary structure (Sreerama and Woody

1994). XTLSSTR identifies secondary structure elements using the relative spatial orientation of sequential carbonyl vectors. The program identifies secondary structures based on the analysis of two vectors and three distances. The angle zeta is created by measuring the angle between consecutive carbonyl vectors (n and n-1) from C to O and the angle tau is the angle between inter-α carbon vectors Cα (n) to Cα (n-1) and Cα (n) to Cα (n+1). The three distances measured are O (n) to N (n+3), O (n) to N (n+4), and C (n) to N (n+3). Secondary structure output files from XTLSSTR were searched using Awk scripts. The scripts were created to locate sequences of three or more residues in the 3_2-helix conformation and were reported along with protein primary sequence.

Selection of the proteins required the satisfaction of two primary criteria. First, only crystal structures were used. Because the solution-state measurements that specify 3_2-helical characteristics required for structure elucidation have only been recently described (Lam and Hsu 2003), no NMR-derived structures currently in the PDB database were determined using constraints based on these parameters. Therefore, NMR-based structures were not included in our analysis. Second, a 2.70 Å resolution cutoff was chosen, lowering the resolution cutoff below 2.70 Å would have minimized the sampled pool of structures. Structures that fit these criteria resulted in an R-value less than 23%; this was used as an indicator of structural integrity. In order to avoid over sampling from multiple structures contained in a PDB file, structures found more than once within a single PDB four-letter code (i.e. asymmetric unit) had the secondary structure results assigned to the residue(s) with the lowest B-factors.

Criteria for identifying the 3_2-helix regions were defined as three consecutive 3_2 residues forming one complete turn. Statistically, this represents a $< 0.14\%$ chance of randomly forming such regions if it is assumed that there are three possible orientations of the ϕ and ψ-dihedral angles. Because of the method of identification, traditional limits of ϕ, ψ angles were not important in discriminating 3_2-helices from β-strands. Although β-strands are close in ϕ, ψ space to the 3_2-helix, a hierarchical order of conformation over hydrogen bond criteria was defined. This strategy was adopted for three reasons: first, to identify structures that were previously overlooked because of hydrogen bonding patterns used for identifying secondary structure. Second, the use of the XTLSSTR program more closely correlates secondary structure quantification to molecular spectroscopic methods such as CD. Third, because this method utilizes interresidue carbonyl vectors and not hydrogen bonding potentials, the interresidue parameters yield a unique data set based on residue-to-residue orientation and thus a more complete picture of the overall chain conformation.

The third K homology domain, the Epstein-Barr nuclear antigen-1 and the homeodomain from Drosophila paired protein NA binding proteins were a result of initially eliminating all proteins based on two requirements in addition to the methods outlined above. When comparing, both the bound and unbound forms must be from the same species and, therefore, have the same primary sequence. Second, only the residues found in the free and bound structures (i.e. deletions) were compared (except the N-terminal residues of one monomer of EBNA-1 which were retained in the bound state to emphasize the symmetry of the NA

binding domain); those residues with high temperature factors (> 80) or no electron density in either form were not included in the corresponding structure or analysis. The 3-structures with bound ligand are the third K homology domain (1EC6), Epstein-Barr Nuclear Antigen-1 (1B3T bound and 1VHI free), and the homeodomain from Drosophila paired protein (1FJL). Visualization was carried out in Insight II from Accelrys (Biosym/Msi 1994).

Acknowledgements

We thank all those who have deposited their structural coordinates in the PDB, W. Curtis Johnson Jr. and his laboratory for making the code for XTLSSTR available to the public, J. Randy MacDonald for programming advice, and W. Curtis Johnson and P. Andrew Karplus for critical readings of the manuscript. This publication was made possible by grant number P01ES00040 and P30ES00210 from the National Institute of Environmental Health Sciences (NIEHS), NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS, NIH.

Chapter 3

Helix propagation cannot always overcome the helix propensity of an indifferent sequence

Joshua M. Hicks, W. Curtis Johnson and Victor L. Hsu

Formatted for Submission

3.1 Summary

An understanding of the nucleation and propagation of secondary structure is of vital importance to the study of protein folding and structure prediction. The secondary structure nucleation-propagation hypothesis suggests that only a few consecutive residues with a strong propensity for a secondary structure are required to initiate its formation. Once a secondary structure is initiated, propagation from this structure is favored over the formation of a second nucleation site. Propagation occurs indifferently in both directions until amino acids with low propensities are encountered. Using a previously described sequence (YVAEAKTSGSRVAEAK) composed of two helix-favoring pentapeptides (VAEAK) flanking a helix indifferent pentapeptide (TSGSR) (Lawrence and Johnson 2002), circular dichroism spectroscopy revealed that regardless of 2,2,2-trifluoroethanol (TFE) concentration the sequence does not attain the same amount of helicity as a sequence with most other amino acids substituted for the central glycine residue. NMR spectroscopy was used to distinguish between the three possibilities: one central helix with nonhelical ends, a terminal helix, or an N-terminal and C-terminal helix with a central break. Data show that the sequence forms an N-terminal and a C-terminal helix with a break occurring in the region of the central glycine residue. These results indicate that both sequence context and environmental effects must be considered when identifying secondary structure nucleation sequences and that nucleation is not as unfavorable as previously predicted.

3.2 Introduction

Protein folding *in vivo* and the ability of many proteins to spontaneously refold *in vitro* indicates that three-dimensional structural information is contained within the primary sequence. Formation of secondary structure elements by peptide fragments and the early assembly of secondary structures in protein folding indicate that the initiation of secondary structures is one of the earliest steps in protein folding. Predicting which residues will form a helix based on their constituent amino acid propensities has been problematic and is complicated by the fact that all amino acids are found in all types of secondary structure elements. One possible explanation is the helix nucleation hypothesis, which states that initiation occurs if a few consecutive residues with high helical propensity form a helix "seed", which is followed by propagation through other helical or "indifferent" residues until strong "helix-breakers" are encountered (Zimm and Bragg 1959; Chou and Fasman 1974; Jaravine et al. 2001). The initiation event in small peptides has been hypothesized to be the formation of a helix occurring in the middle of a peptide sequence (Bierzynski and Pawlowski 1997) after which propagation may proceed in either or both directions (Chou and Fasman 1974; Jaravine et al. 2001). In this hypothesis, helix elongation is a result of the propagation of a short helix instead of the fast formation of adjacent short helices, which requires more random events (Brook 1996). Thermodynamically, propagation is predicted to be more favorable than initiation (Doig 2002), and factors such as the "end separation effect" (Bierzynski and Pawlowski 1997) and

long range dipole-dipole interactions (Aqvist et al. 1991; Bierzynski and Pawlowski 1997; Goch et al. 2003) can aid in helix propagation and stability, respectively.

The development and application of the statistical mechanics theory of helix-coil transitions for predicting helix initiation and propagation by Zimm and Bragg (1959) and Lifson and Roig (1961) has made helix formation the most extensively studied protein folding event. Host-guest studies that substitute residues into known helix-forming sequences have been used for measuring the helical propensity of a residue (Chakrabartty et al. 1994). Improvement of the two parameter treatment has been expanded to include consideration for capping effect (Doig and Baldwin 1995; Munoz and Serrano 1995), charge-dipole interactions (Aqvist et al. 1991) and side-chain to side-chain interactions (Munoz and Serrano 1995). Although this has yielded a great deal of insight into the helix-forming tendencies of the amino acids, none of these approaches have been able to completely describe the helix-coil transition.

Evidence that sequence context is a key factor in determining the secondary structure of a sequence has significant implications for understanding the helix-coil transition and protein folding. In a study of nonlocal amino acids, Maeda and coworkers (2003) demonstrated that the presence of different oligopeptide sequences can cause either helical stabilizing or destabilizing effects and that these effects vary even when the oligopeptides have the same amino acid composition but different primary sequences. Additionally, the positioning of residues has a significant effect on the propensity of a residue when found on a

hydrophobic versus a hydrophilic face of the helix (Zhou et al. 1993). These aspects of secondary structure formation have also been studied in several solvents that favor specific conformations. Solvent can affect the propensity of a residue for a particular secondary structure (Lawrence and Johnson 2002) or can even induce secondary structures that are not predicted (Waterhous and Johnson 1994). In particular, 2,2,2-trifluoroethanol (TFE) is a known helix-inducing solvent that facilitates helix elongation (Storrs et al. 1992), increases the effect of sequence context (Myers et al. 1998), and has been shown to stabilize the ribonuclease S-peptide α-helix (Nelson and Kallenbach 1986) and increase the folding rate of lysozyme (Lu et al. 1997). TFE destabilizes the unordered state and favors helix propagation as a result of less effective hydrogen bonding of the backbone amides to the solvent (Storrs et al. 1992).

To determine if environmental effects alter the intrinsic conformational properties of a residue, circular dichroism (CD) measurements have been used to measure helicity and monitor propagation in TFE titrations (Rohl et al. 1996; Myers et al. 1998; Krittanai and Johnson 2000; Lawrence and Johnson 2002). Solvent effects on helix propagation were followed in a synthetic peptide previously described (Lawrence and Johnson 2002) and its structure was determined in various TFE/H_2O solvent mixtures using nuclear magnetic resonance (NMR) spectroscopy. The peptide sequence consists of two helix-favoring segments, VAEAK (abbreviated "H" for helix), flanking the indifferent TSGSR (abbreviated "G" for the glycine residue) sequence (Waterhous and Johnson 1994; Lawrence and Johnson 2002). An N-terminal tyrosine (Y) residue

was used to monitor peptide synthesis and purification, and for determining peptide concentration, but had no effect on secondary structure. The VAEAK sequence, originally derived from a 13-residue sequence of EcoRI, satisfies the Chou and Fasman criteria for helix formation (Chou and Fasman 1974). The central TSGSR sequence was designed with amino acids that are indifferent to helix formation, and are a conservative substitution in side chain character compared to the H sequence at all positions except for the central residue (G). Insertion of this sterically similar, but functionally different G sequence between two intervening VAEAK sequences was done to determine if any helix propagation can occur in the helix-favoring solvent TFE.

Circular dichroism (CD) studies of YHGH indicate that regardless of 2,2,2-trifluoroethanol (TFE) concentration YHGH does not attain the same amount of helicity as when most other residues are substituted for the central glycine. The amount of helical content in the YHGH sequence is 80% of other substituted sequences in 100% TFE. This indicates that only 13 of the residues are in the helical conformation in the YHGH peptide. Because the only difference is the central glycine residue, this would suggest that the peptide consists of two short helices interrupted by residues at the central sequence, so the helix is not propagated through the indifferent TSGSR sequence. However, residue-specific secondary structure assignments cannot be made from CD data.

Using NMR spectroscopy, the solution structures of the YHGH sequence in 100%, 90% and 80% TFE were solved. NOE secondary structure patterns and simulated structures identify the peptide structure as being composed of an N-

terminal and a C-terminal helix separated by the nonhelical GS sequence. The percent helicity in the NMR-solved structures is in agreement with the percent helicity calculated from CD measurements. These results show that propagation is not always favored over initiation and is dependent on sequence context. Furthermore, propagation does not always occur through an indifferent sequence even when placed in a helix favoring solvent and inserted between two helices.

3.3 Results and Discussion

The aim of this work was to determine if the YHGH peptide in higher concentrations of TFE consisted of one central helix with ill-defined termini, a longer N or C-terminal helix or two short helices. CD data indicate that the structure of the YHGH in 100% TFE is 80% helical (Lawrence and Johnson 2002), decreasing the number of helical residues by three. This indicates that not only did the glycine substitution destabilize part of the helix, but also that the destabilization was synergistic with sequence effects. The effect on helix formation is unexpected since none of the amino acid substitutions (excluding proline) into the central TSXSR sequence is expected to be a helix breaking sequence in TFE. Importantly, all amino acids except proline were shown to be capable of helix initiation and propagation when substituted in the structurally conserved VAXAK sequence (Krittanai and Johnson 2000). The inability of TFE to induce propagation through the G sequence is unusual since the design of

peptides thus far has produced few occurrences of peptides that don't form helices in TFE.

Table 3.1 shows the sequential alignment of the YHGH sequence with the calculated propagation free energies at various TFE concentrations and includes the N-capping and C-capping free energy (in water) for a residue located at the N-terminus or C-terminus, respectively, of an α-helix (Doig and Baldwin 1995). At 40% TFE, propagation statistics from sequences derived from helical sequences in proteins are closer in agreement than to those measured in alanine-based peptides. Peptides for measuring the propensity of an amino acid were originally designed to include large amounts of alanine due to its α-helix forming tendencies in solution (Marqusee et al. 1989). The resulting propensity describes the ability of an amino acid to propagate in the absence of any helix-stabilizing, or destabilizing, interactions from side chains, but these results tend to overestimate amino acid propensities (Qian and Chan 1996; Myers et al. 1997). For example, glycine has been identified as a strong helix breaker (Rohl et al. 1996), but this can only be applied to alanine-based peptides in 40% TFE. Based on the propensities presented in Table 3.1, at low TFE concentrations the central SGS sequence would likely not propagate in water and at the most form terminal residues of helices. As the concentration of TFE increases, the ability to propagate through this sequence increases, regardless of direction, and would readily propagate through the serine residues in 90% TFE.

Table 3.1 Sequence alignment with the calculated helical propagation free energy for amino acids as reported from several different laboratories in various TFE concentrations and the N-capping and C-capping free energies (ΔG°)

	90% TFE[a]	40% TFE[b]	40% TFE[c]	40% TFE[d]	0% TFE[e]	N/C-capping[f]
Y	-0.55	-0.30	-0.75	-	-1.28 to –1.11	-1.5/-2.2
V	-0.76	-0.64	-0.24	-0.05	-0.83	-0.1/0.9
A	-0.90	-0.66	-0.52	-0.52	-1.88	0
E	-0.40	-0.28	-0.32	-0.27	-1.20	-0.7/-0.5
A	-0.90	-0.66	-0.52	-0.52	-1.88	0
K	-0.81	-0.17	-0.52	-0.06	-1.52	0.1/-0.1
T	-0.29	-0.14	-0.20	0.27	-0.56	-0.7/NA
S	-1.28	-0.18	-0.36	0.30	-1.10	-1.2/0.8
G	-0.11	0.32	0.01	1.30	0	-1.2/0.1
S	-1.28	-0.18	-0.36	0.30	-1.10	-1.2/0.8
R	-0.40	-0.20	-0.03	-0.08	-1.67	-0.1/-0.4
V	-0.76	-0.64	-0.24	-0.05	-0.83	-0.1/0.9
A	-0.90	-0.66	-0.52	-0.52	-1.88	0
E	-0.40	-0.28	-0.32	-0.27	-1.20	-0.7/-0.5
A	-0.90	-0.66	-0.52	-0.52	-1.88	0
K	-0.81	-0.17	-0.52	-0.06	-1.52	0.1/-0.1

a Lawrence and Johnson (2002)
b Krittanai and Johnson (2000)
c Myers, Pace and Scholtz (1998)
d Rohl, Chakrabartty and Baldwin (1996)
e Chakrabartty, Kortemme and Baldwin (1994)
f Capping energies were taken from uncharged termini, all residues are standardized to alanine in water. Doig and Baldwin (1995)

3.3.1 NMR Analysis of YHGH

Figure 3.1 shows the NMR analysis of the NOE hydrogen bonding patterns of YHGH in 100%, 90% and 80% TFE where thick lines represent strong correlations (~2.5-3.5 Å), intermediate lines represent medium correlations (up to 5 Å) and thin lines represent weak correlations (up to 6 Å). The most definitive indication of helical structure is the strong, sequential HN(i) to HN(i+1) NOE cross-peaks and weak or absent Hα(i) to HN(i+1) NOE cross-peaks that are typically strong in β-strands. The strong HN(i) to HN(i+1) NOE cross-peaks limit the number of possible backbone psi (ψ) conformations when taken in context with experimentally determined phi (ϕ) angles and the strong HN(i) to Hβ(i+1) correlations observed (Shi et al. 2002). Another strong indication of helical structures is the presence of Hα(i) to HN(i+3) and Hβ (i+3) cross peaks. These peaks arise because the orientation of the Hα in a helix brings the amide and side chain protons at residue i+3 spatially close in alpha helices (i+2 in 3_{10}-helices). The presence of HN(i) to HN(i+2) distance restraints for other residues indicates the formation of a tighter structure than the α-helix predicted by the strong HN(i) to HN(i+1) NOE cross peaks. This would also be consistent with the characteristics of 3_{10}-helices, which maximize the hydrogen bonding of the helices and decrease the number of free main chain hydrogen bond donors and acceptors at terminating helices.

Although the presence of several NOE correlations to Hα indicates that the central glycine residue is found in a defined conformation, defining the local geometry of this is difficult due to the degenerate chemical shift of the glycine

Figure 3.1 Sequence alignment of NOE secondary structure patterns and $^3J_{HNHA}$ couplings constants consistent with helices (H) of YHGH in 100%, 90% and 80% TFE. NOE correlations are further characterized by the thickness of the line where thick lines represent strong correlations (~2.5-3.5 Å), intermediate lines represent medium correlations (up to 5 Å) and thin lines represent weak correlations (~5-6 Å).

α- protons and the short geminal distance relative to the range of the NOE distance restraints. Because the central glycine residue contains two α-protons, the absolute identification of local geometry is limited by the inability to stereospecifically measure the $^{3}J_{HNHA}$ coupling constant. This, in addition to the absence of side chain protons on the glycine residue for global NOE restraints in the glycine region and prediction of the backbone angle psi (ψ) of glycine, results in a large set of structures that fit the overall NOE and dihedral constraints.

The structures elucidated from NMR data and the NOE secondary structure patterns in 100% TFE identify YHGH as an N-terminal helix formed by residues YVAEAKTS (1-8), termination of this helix at S8, starting the break at the glycine-serine (9-10) residues, and a C-terminal helix of residues RVAEAK (11-15). Figure 3.2 shows the results of aligning the C-terminal and N-terminal helices separately in NMR-derived structures of the YHGH sequence in 100% TFE. The conformational ambiguity of the central GS sequence as described above allows variation in the relative orientations of the N and C-terminal helices with the two helical segments consistently identified as stable, regular helices. The calculated helicity from these structures is consistent with the estimates of helicity determined by CD spectroscopy. Formation of the two helices in this short of a fragment is surprising since calculations for the minimum number of residues under the most favorable conditions to form multiple helices suggest that at least 20 residues are required (Zimm and Bragg 1959; Werner et al. 2002) and that non-hydrogen bonded residues between helices should not be less than three (Zimm and Bragg 1959).

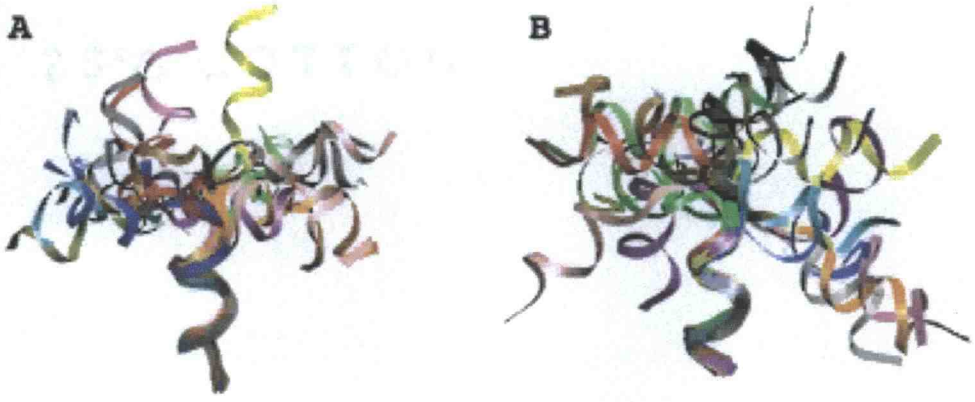


Figure 3.2 Alignment of 20 calculated structures of the YHGH sequence with A) N-terminal helices aligned and B) C-terminal helices aligned using the Globalcore/Globalmerge program (Karplus 2004). This figure was prepared with VMD (Humphrey et al. 1996). The large overhang observed in the C-terminal structure is a result of the short helix at this end of the peptide.

End fraying of helices (Rohl et al. 1992) cannot explain the decrease in helicity identified by comparing the CD data of the YHGH sequence vs. the YHAH sequence in 100% TFE nor can it explain the decrease in NOE patterns in the central SGS sequence as TFE concentrations are decreased. The strong helix formation properties of the VAEAK sequence in TFE and the ability to propagate through all 18 amino acid substitutions (except proline) of the central glutamic acid indicates that VAXAK is a helix initiating sequence. Similarly, a dynamic model of the α-helix being in rapid equilibrium between all residues would require a significant entropic event to overcome the less effective hydrogen bonding of the backbone amides to the solvent. There is no indication that helices become more dynamic once they fold and the ability of TFE to increase the effective intraelement hydrogen bonding within the two helices explains why non-alanine based studies report stabilization by and propagation in TFE (Nelson and Kallenbach 1986; Lu et al. 1997).

Whether the different conformations observed in the nonhelical segment are a result of conformational heterogeneity, mobility, or an artifact arising from the inability to stereospecifically assign the glycine α-protons is not known. However, the observed NOE pattern data clearly indicate that the region immediately adjacent to the glycine residue contains the only residues that display no characteristics of helical secondary structure at any of the TFE concentrations. Although glycine is a known helix breaker, the presence of glycine in helices found in proteins suggests that the propensity of glycine against helix formation can be overcome when placed in the correct combination of environment and

sequence context. Other sequences containing glycine have been shown to form helices in TFE and glycines have even been shown to participate in helix propagation (Krittanai and Johnson 2000; Lawrence and Johnson 2002), but in the present study the propensity of glycine in an indifferent sequence cannot be overcome even in 100% TFE.

Retention of residues in helical conformations in decreasing TFE percentage occurs in the two VAEAK sequences. The most convincing evidence of this is the maintained presence of NOE correlations from Hα of V2 and V12 to HN and Hβ of K6 and K16, respectively. The decrease in helicity concomitant with a change from 100% to 80% TFE corresponds to a two-residue shift from helix to random coil. As helices can propagate, the question of whether the two observed helices are formed independently or as a result of propagation from one helix and subsequent unfolding of the central residues arises. The retention of these two helices after TFE titration strongly suggests that the helices are formed independently and are not a result of a propagation of a single helix that subsequently breaks into two helices. The NMR data of YHGH in 80% TFE are identical regardless if the sample is diluted from 100% TFE, directly suspended in 80% TFE or titrated up to 80% TFE. Our results indicate that strategies for predicting helical structure require consideration for residue environment and sequence context.

3.4 Conclusion

If an understanding of protein folding is to be resolved, prediction of secondary structure from the primary sequence must be both accurate and precise. This is the first reported structure of a short peptide containing multiple nucleation sites. These results suggest that helix initiation, which requires more amino acid backbone dihedral angles to randomly form the first structure compared to a one amino acid propagation, may not be as thermodynamically unfavorable as previously proposed (Brook 1996; Doig 2002). The formation of only a two-residue break in the helix is unusual since, ideally, it would be predicted that the formation of a helix through this central sequences would require fewer random conformational events than a second nucleation. This isn't surprising considering that the terms for initiation and propagation in current models have largely been determined from systems containing large amounts of alanine. Alanine-based helices do not coincide with natural diversity in primary sequence that uniquely determines the tertiary structure and, therefore, cannot differentiate sequences containing more than one nucleation site, or sequences that are very strong breakers. This indicates that individually calculated propensity of each amino acid is insufficient to adequately describe the initiation of secondary structures.

3.5 Materials and Methods

Peptide synthesis and circular dichroism (CD) measurements were completed as previously described (Lawrence and Johnson 2002). NMR samples were prepared in 50mM sodium phosphate (pH 5). Samples were dissolved in trifluoroethanol-D_2 (Cambridge Isotope Labs).

All NMR data were collected on a Bruker DRX-600 at 282K. TOCSY spectra were collected using a mixing time of 35 msec. $^3J_{HNHA}$ coupling constants for determining phi (ϕ) backbone dihedral angles were measured from J-Resolved spectra. NOESY spectra were collected using a mixing time of 300 msec and psi (ψ) dihedral angles were estimated from $NOE_{\beta i\text{-}HNi}$:$NOE_{\beta i+1\text{-}Hni}$ ratios (Shi et al. 2002).

All data were processed using Bruker XWINNMR software. Data visualization and assignments were performed using the SPARKY assignment program developed at the University of California at San Francisco (Goddard and Kneller 2004). Simulated annealing-based structure calculations were performed using XPLOR (Brunger 1992). Structure visualization and alignment were done using the VMD program with the XPLOR (VMD-XPLOR) extensions (Humphrey et al. 1996) and globalmerge/globalcore programs (Karplus 2004), respectively. Secondary structures were identified using the XTLSSTR structure assignment program (King and Johnson 1999).

Chapter 4

Chemical Modification as a Method to Study a Mimic of Protein Function: Increase in the Fast Motions of *E. coli* Thioredoxin Backbone Amides

Joshua M. Hicks and Victor L. Hsu

Formatted for Submission

4.1 Summary

The backbone dynamics of *E. coli* thioredoxin chemically adducted with a cysteine-ethyl glutathione group at the active site N-terminal cysteine (C32) were determined to identify residue-specific changes resulting from the modification. This particular adduct was studied to suggest possible relationships to the transient state of thioredoxin's redox mechanism and how this state is accommodated by the rearrangements of the active site. Dynamical data for the modified protein are compared to the data for the oxidized and reduced forms as reported by Stone et al. (1993). 98% of all residues containing backbone amides were specifically identified and the individual longitudinal (T1) and transverse (T2) relaxation time constants and the heteronuclear $\{^1H\}$-^{15}N NOE values were measured. Model-free formalism analysis (Lipari and Szabo 1982) of the data was used to determine the generalized order parameter (S^2), contributions of ^{15}N exchange broadening (R_{EX}), the effective correlation time for internal motions (τ_E) and the overall molecular correlation time (τ_C). On the picosecond to nanosecond time scale, chemically modified thioredoxin exhibits similar trends to both the reduced and oxidized forms, but with an 8% and 7% decrease, respectively, in the average generalized order parameter indicating that the chemically modified structure is less rigid. Significant increases in the backbone motion occur in residues A67-I75 and V91-L94, which are the two loops that form the hydrophobic surface important for protein-protein interactions at the active site. Surprisingly, a significant amount of the dynamical changes in the

picosecond to nanosecond time scale range is found in residues that are spatially and sequentially distant from the active site and is often within well-defined secondary structure elements.

4.2 Introduction

Thioredoxin is a small (Mr ~11.7 kD), ubiquitous protein that catalyzes oxidation-reduction reactions via an active site thiol-disulfide mechanism utilizing a highly conserved C-X-X-C motif which is also found in homologs such as thioltransferases (aka glutaredoxins), protein disulphide isomerase (PDI), disulphide-bond formation facilitator (DsbA) and thioredoxin reductase (Holmgren 1985; Martin et al. 1993; Freedman et al. 1994; Yang et al. 1998). The solvent-exposed sulfur group of the N-terminal cysteine (Fig. 4.1) in the thioredoxin active site C-G-P-C motif (C32, pKa 6.7) has a lower pKa compared to the other active site cysteine (C35, pKa 9.0) and free cysteine (pKa 8.4) (Chivers and Raines 1997; Dyson et al. 1997). Although there is still some controversy in the literature as to the exact pKa of these residues (Dyson et al. 1997; Chivers et al. 1997b), the overall consensus is that the N-terminal cysteine sulfur group in reduced C-X-X-C motifs exists as a thiolate anion. The lower pKa results in a disulphide bond that is easier to reduce, and stabilization of the thiolate anion is a mechanism by which thioredoxin can modulate pKa and redox equilibrium (Holmgren 1985; Gane et al. 1995; Grauschopf et al. 1995).

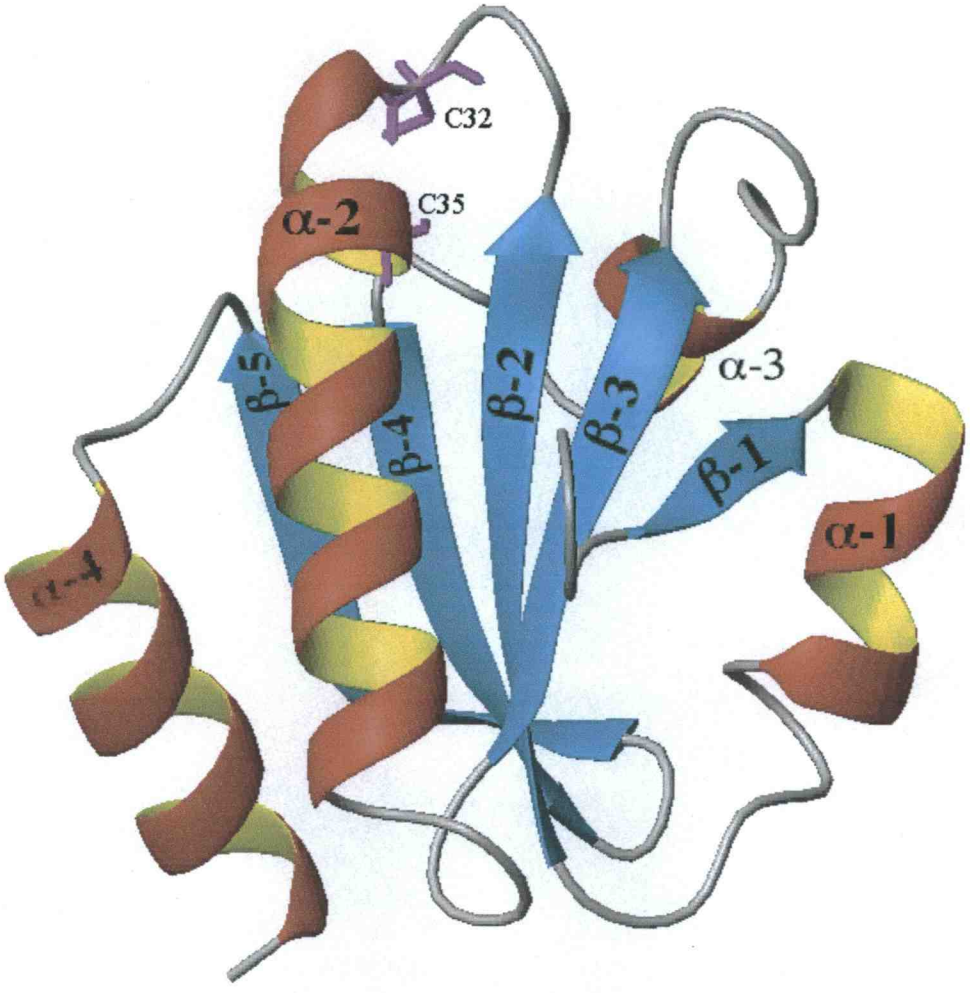


Figure 4.1 Ribbon model of reduced *E. coli* thioredoxin (1XOB). The active site motif C-G-P-C is found at the N-terminus of helix α-2, here pictured at the top of the structure as indicated by the side chains of C32 and C35 (violet). This figure was prepared using MOLMOL (Koradi et al. 1996)

Thioredoxin was first identified as the electron donor to ribonucleotide reductase in the conversion of nucleoside diphosphates to deoxynucleoside diphosphates (Laurent et al. 1964). Since then thioredoxin has been identified in a myriad of systems including its use as a structural unit of the T7 DNA polymerase complex (Modrich and Richardson 1975; Huber et al. 1986), as the inducer of *bis*-phosphatases in chloroplasts (Nishizawa and Buchanan 1981), and in interactions with several proteins necessary for cell division and energy transduction (Kumar et al. 2004). Thioredoxin is 1,000 times more efficient than dithiol threitol (DTT) at reducing insulin (Holmgren 1985). Figure 4.2a diagrams the general mechanism for disulphide reduction by thioredoxin as proposed by Jeng et al. (1995). Briefly, the transient state of thioredoxin function includes the formation of a mixed disulphide intermediate involving nucleophilic attack of the C32 thiolate on the disulphide located in the target protein (Jeng et al. 1995).

Chemical modification of expressed proteins can be used to mimic or model the modifications that proteins are subjected to *in vivo*. Adduction of a protein can perturb the conformation, change the thermal stability and/or alter folding and unfolding pathways (Radford et al. 1991; Eyles et al. 1994). Chemical modification of side-chains provides a method for studying the effects on structural/functional elements of a protein without altering the primary sequence. For example, 1,2-Dichloroethane (DCE) is commonly manufactured for the synthesis of vinyl chloride and due to its high reactivity it is toxic. DCE is a causative agent in renal failure and chronic exposure to DCE is known to cause liver damage (Williams and Diwan 1994). The cellular detoxification of DCE is

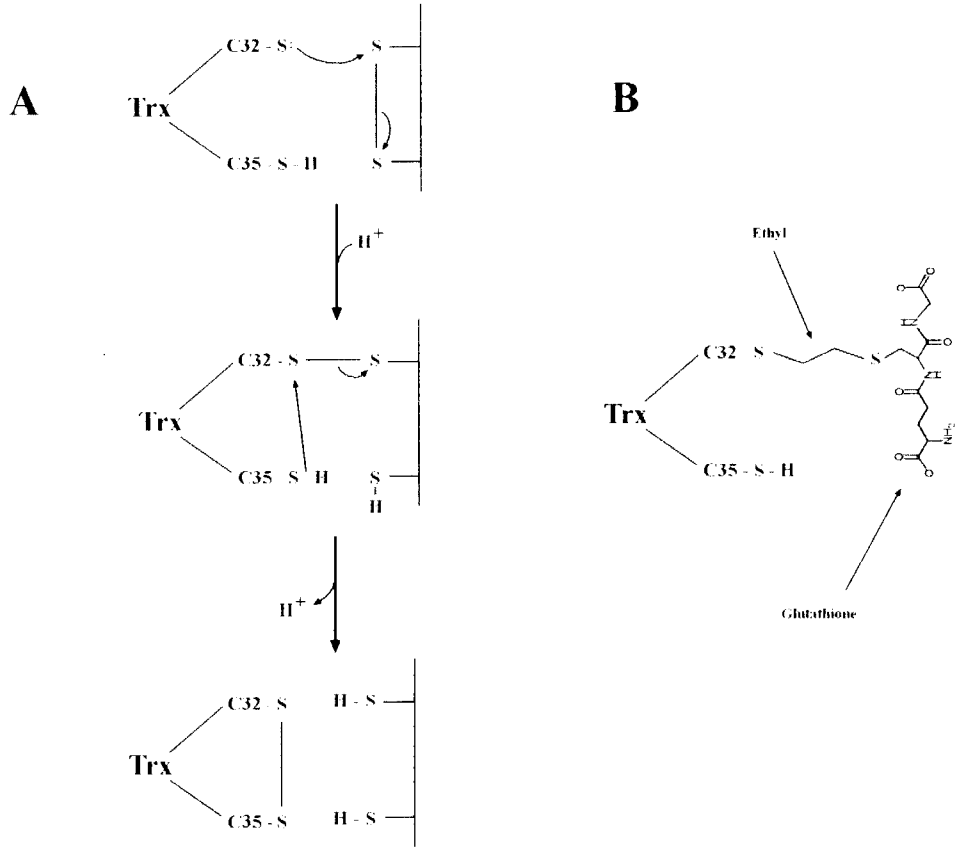


Figure 4.2 Schematic representation of the mechanism for thioredoxin function and the cysteine-ethyl glutathione adducted structure. (A) The mechanism involves recognition of the target protein containing a disulphide bond. This is followed by nucleophilic attack by the C32 thiolate anion to one of the oxidized cysteines of the substrate, thus forming the mixed disulphide intermediate. Subsequent reduction of the target protein results in oxidation of the active site disulphide of thioredoxin (Jeng et al. 1995). (B) Glutathione is irreversibly adducted to C32-sulfur of thioredoxin by an ethyl group attached the glutathione cysteine sulfur (Kim et al. 2001).

predominately mediated via glutathione S-transferase producing a cysteine-ethyl-glutathione (CEG) conjugate. CEG can form an electrophilic episulfonium ion that has been shown to be mutagenic via the alkylation of N7 of guanine nucleotides (Humphreys et al. 1990) and can also target functional/structural protein residues (Pumford and Halmes 1997). Proteins such as hemoglobin (Erve et al. 1996) have been demonstrated to be a cellular target of CEG and adduction of thioredoxin at the N-terminal active site cysteine (C32) by CEG has been identified (Erve et al. 1995; Kim et al. 2002). The presence of the tripeptide CEG-moiety as a result of the irreversible adduction of C32 could be a good mimic of the mixed disulphide intermediate in thioredoxin function (Fig. 4.2b).

To map the orientation and interactions of the glutathione adduct to thioredoxin, a previous study compared the mass spectrometrically-monitored backbone amide proton H/D exchange rates of the glutathione adduct to the H/D exchange rates of a cysteine-ethyl-cysteine (CEC) adduct, thereby eliminating the γ-glutamic acid and glycine residues of glutathione (Kim et al. 2002). The data show that glutathione led to additional protection of the thioredoxin backbone amide protons of I75 and either G92 or A93, but the lack of a cleavage site prevented distinguishing between G92 and A93. To account for these identified protecting groups, the resulting hypothesized structure included hydrogen bonding between the backbone amide proton of I75 and the side chain carbonyl of γ-glutamic acid and from the backbone amide proton of either G92 or A93 to the backbone carbonyl of the glutathione glycine residue. This orientation was

modeled from the NMR derived structure of the glutathione-thioltransferase mixed disulphide intermediate (Yang et al. 1998). In the current NMR study, nuclear Overhauser (NOE) spectroscopy experiments indicate that the orientation of glutathione in the active site cleft of thioredoxin is similar to the previously predicted structure (Kim et al. 2002).

NMR can provide additional information on the changes in backbone dynamics as a result of the chemical adduction of the functionally important C32. ^{15}N nuclear spin relaxation measurements are used to characterize the motions observed in backbone conformational dynamics unique to all amino acids containing backbone amide protons. Quantitation of internal protein motion is necessary for identifying the residual entropy important for the energetics of protein function (Wand 2001). The distribution and magnitude of residual conformational entropy in proteins is often associated with changes in protein activity and functional states (Stock 1999; Lee et al. 2000; Mayer et al. 2003). Residue-specific dynamics of the reduced and oxidized forms of thioredoxin have previously been reported (Stone et al. 1993). By comparing the motions between the CEG adduct and the motions of the native reduced and oxidized forms, changes in the dynamical profile of each amino acid provide a residue-specific map of backbone dynamics possibly analogous to those that might be seen in the transient state of the thioredoxin redox mechanism.

This study compares the model-free analysis of the CEG-C32 modified thioredoxin (CEG-thioredoxin) with the dynamical information previously obtained for the reduced and oxidized forms of thioredoxin (Stone et al. 1993).

Characterization of the backbone conformational dynamics in CEG-thioredoxin exhibits similar trends to the dynamics of unmodified reduced and oxidized thioredoxin, but faster motions are observed in most of the amide containing residues in the adducted form. CEG-thioredoxin shows a decrease in the generalized order parameter of residues in the active site and the two loops that make up the hydrophobic face at the active site. Increases in backbone dynamics were also identified in the secondary structures that lead in and out of these functionally important structures. Thus, the backbone dynamics of thioredoxin are examined to resolve residue-specific motions in the hope of learning about thioredoxin function.

4.3 Results and Discussion

4.3.1 Formation of Aggregates

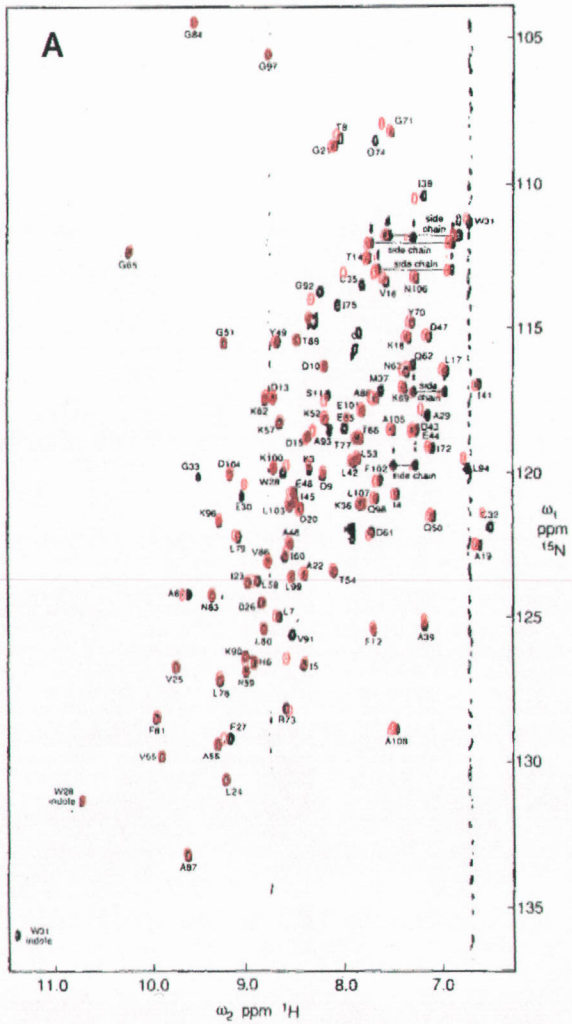
Thermodynamically, thioredoxin is a very stable protein, it resists denaturation up to 67 C and never completely unfolds in the presence of detergents (Bhutani and Udgaonkar 2003). Due to precipitation, active site adduction prevented the use of high concentrations of CEG-thioredoxin in this NMR investigation. NMR spectra of the CEG-thioredoxin were collected using identical buffer and temperature conditions used in previous studies so that results can be compared (Stone et al. 1993). However, the maximum sample concentration in this study was limited to 100 μM compared to previous studies of 1-4 mM (Chandrasekhar et al. 1991; Stone et al. 1993). To ensure that there is no

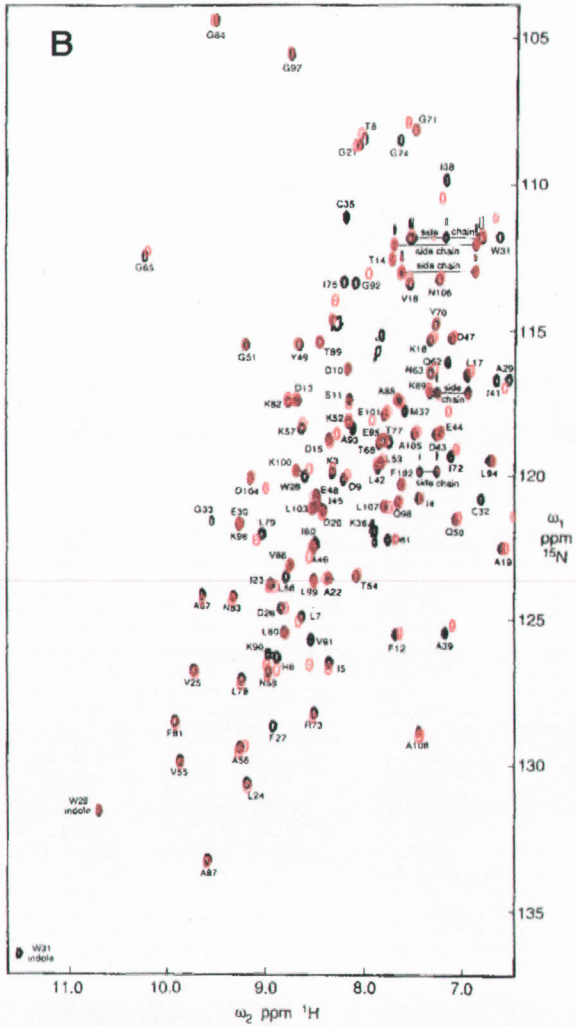
effect on the relaxation measurements, all data were collected at sub-aggregate concentrations (< 80 μM). Insoluble aggregates form rapidly even at concentrations as low as 100 μM. Previously reported mass spectrometry studies of CEG-thioredoxin (Kim et al. 2001; Kim et al. 2002) did not require the high concentrations typically used for NMR spectroscopy and thus aggregate formation was never observed. The adduction has resulted in a structure more susceptible to form aggregates *in vitro,* suggesting that, if the CEG-adduct is a good mimic, the formation of any mixed disulphide could convert thioredoxin to a state that is susceptible to aggregation *in vivo*.

4.3.2 Affect of Adduction on Structure

Heteronuclear single quantum correlation (HSQC) overlays of the CEG-thioredoxin spectra onto the reduced and oxidized spectra (Chandrasekhar et al. 1991) are shown in Figure 4.3. Chemical shifts for most backbone amide groups of CEG-thioredoxin were identical to chemical shifts in the oxidized (Fig. 4.3a) and reduced (Fig. 4.3b) spectra. Due to the sensitivity of chemical shifts they are commonly used in secondary structure prediction, and spectral comparisons can be used to accelerate residue assignment and determine secondary and tertiary structures among homologs (Case et al. 1994; Wishart and Case 2001). Here, previously reported ^{1}H-^{15}N chemical shifts (Chandrasekhar et al. 1991) are used to assign 84% of the residues containing backbone amide protons. Those residues not identified by chemical shift were assigned via individual spin systems from

Figure 4.3 Heteronuclear single quantum correlation (HSQC) spectra overlay of CEG-thioredoxin (red) onto the HSQC spectra of (A) oxidized thioredoxin and (B) reduced thioredoxin (Chandrasekhar et al. 1991).


A
105
110
115
120
125
130
135
ω_1 ppm ^{15}N
11.0
10.0
9.0
8.0
7.0
ω_2 ppm ^{1}H


B
105
110
115
120
125
130
135
ω_1 ppm ^{15}N
11.0
10.0
9.0
8.0
7.0
ω_2 ppm ^{1}H

HSQC-Total Correlation Spectroscopy (HSQC-TOCSY) and sequential NOE assignments (Wuthrich 1986).

Amide proton chemical shifts that vary from both the oxidized and reduced forms are located in the active site region (F27, A29, C32, G33, C35, M37 and I38) and the two loops (G74-I75 and V91-L94) that form the hydrophobic surface believed to be important in protein-protein interactions required for the redox mechanism of thioredoxin function. CEG-thioredoxin is structurally closer to the oxidized form. Besides the active site and the two hydrophobic loops, only backbone amides E30, E85 and L94 differed between CEG-thioredoxin and the oxidized form while chemical shifts of H6, W28, W31, K36, A39, Q62, L79, K90 and E101 differed between the CEG-thioredoxin and the reduced form. Interestingly the side-chain indole proton of W28 does not shift in any of the forms indicating that the packing of strand β-3 is not altered by the presence of the ethyl-glutathione group or by the redox state of thioredoxin. Just as in the MS and CD structural analysis (Kim et al. 2001; Kim et al. 2002), this NMR analysis confirms that the tertiary structure is largely unaffected by the presence of the adduct.

4.3.3 Orientation of the Glutathione Group in the Active Site

NMR nuclear Overhauser spectroscopy (NOESY) experiments conducted in this study confirm the close approach of the side chain of the γ-glutamic acid of glutathione to the amide proton of I75, and of the backbone amide proton of G92

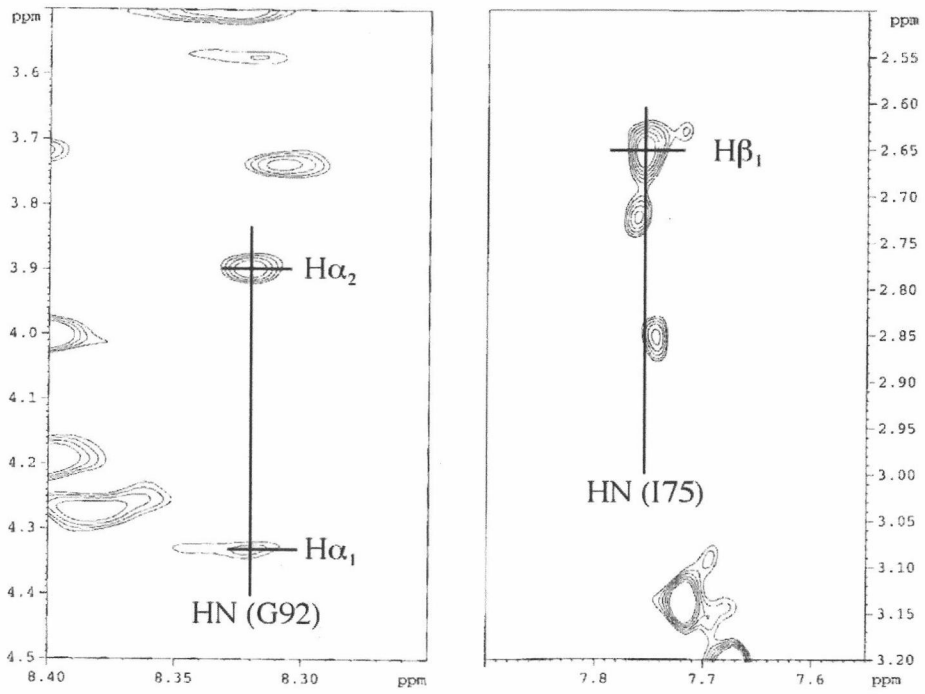


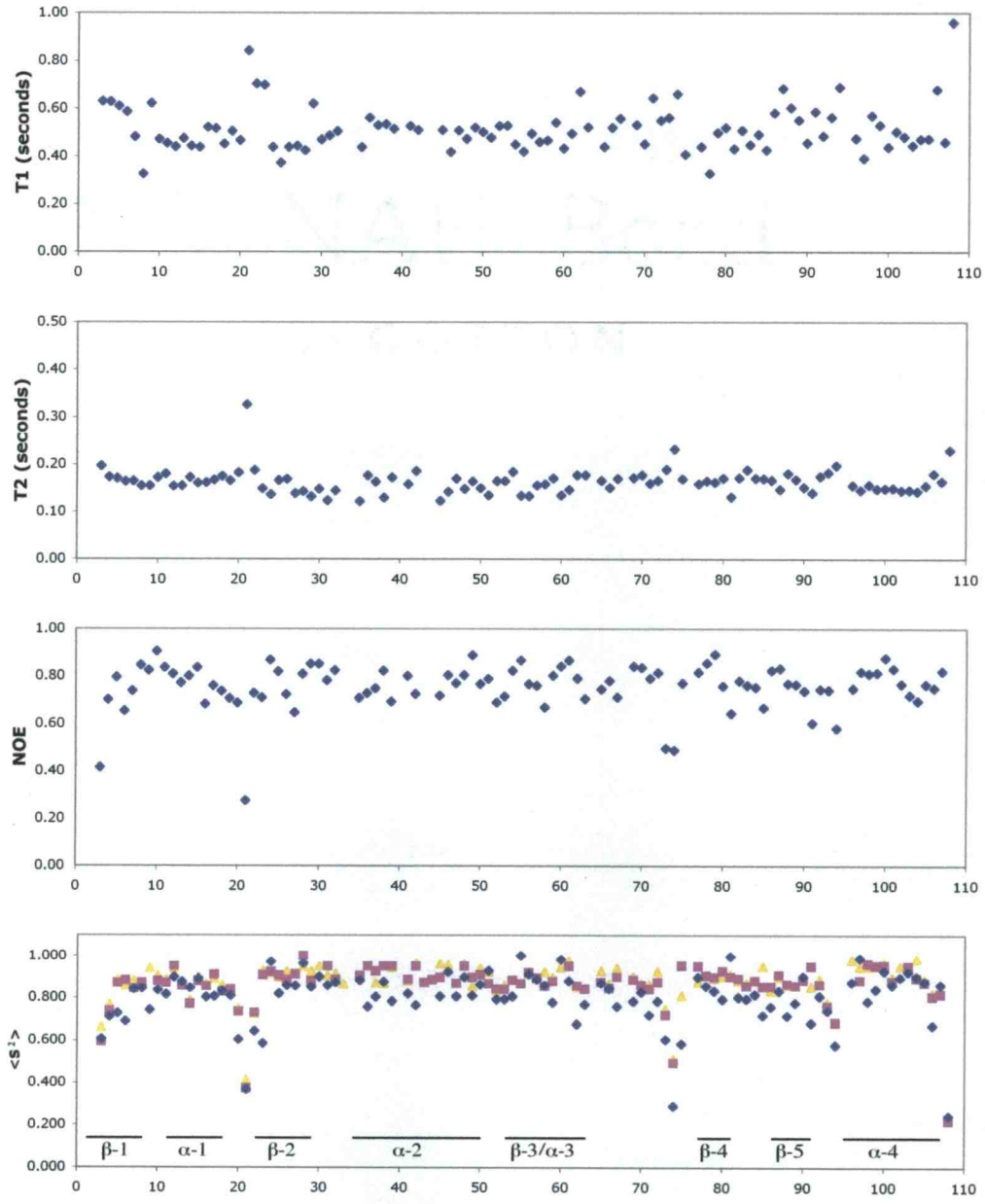
Figure 4.4 Nuclear Overhauser spectroscopy assignments that define the orientation of the CEG-moiety relative to the thioredoxin active site. (A) Correlation between HN(G92) and the CEG-glycine $H\alpha_1$ and $H\alpha_2$. (B) Correlation between HN(I75) and the CEG-γ-glutamic acid side chain $H\beta_1$.

is close enough to the glutathione-glycine to participate in a hydrogen bond with the backbone carbonyl of the glutathione-glycine residue (Fig. 4.4). These distance restraints are very similar to the modeled orientation predicted by Kim et al. (2002). However, NMR has described the protection of backbone amide protons from solvent exchange to G92 instead of the previously predicted A93. Although glutathione is not a substrate for thioredoxin, the resulting interaction definitively determines the orientation of the CEG adduct as being similar to other protein family members solved in a mixed disulphide with glutathione that represents a true intermediate in their catalytic cycle (Bushweller et al. 1994; Yang et al. 1998).

4.3.4 Affect of Adduction on Backbone Dynamics

Figure 4.5 plots sequentially all of the raw ^{15}N longitudinal (T1) and transverse relaxation (T2) rates and the $\{^{1}H\}$-^{15}N NOE exchange rate for CEG-thioredoxin. Because the chemical modification has little effect on the tertiary structure of thioredoxin, model-free analysis of CEG-thioredoxin was performed as in previous studies assuming isotropic tumbling (Chandrasekhar et al. 1991). Initial estimates of the apparent correlation time can be calculated for a spherical molecule from the T1/T2 ratio if certain precautions are taken (Kay et al. 1989; Clore et al. 1990). Residues whose T1/T2 value fall one standard deviation outside the average T1/T2 ratio for all the residues were excluded from initial calculation of the apparent correlation time. This eliminates residues with shortened T2 values as a result of slower conformation averaging compared to the

Figure 4.5 Sequence alignment with measured relaxation rates in CEG-thioredoxin. The x-axis indicates the residue number with the secondary structures of thioredoxin indicated in panel D. Panel A shows the T1 (± 1.8%) spin-lattice relaxation rates. Panel B shows T2 (± 2.2%) spin-spin relaxation rates and panel C shows the steady-state $\{^{1}H\}$-^{15}N heteronuclear NOE (± 5%) enhancements. Panel D shows the generalized order parameter (S^2 ± 0.03) where blue diamonds are the values for CEG-thioredoxin and the reduced and oxidized forms are yellow triangles and pink squares, respectively. Generalized order parameters for the reduced and oxidized forms of thioredoxin were reproduced from the supplemental material (Stone et al. 1993).


T1 (seconds)
T2 (seconds)
NOE
<s²>
β-1
α-1
β-2
α-2
β-3/α-3
β-4
β-5
α-4

difference in ^{15}N chemical shifts of the conformers. For example, residues 21-23, 94, 87 and 108 have noticeably higher T1 values, but these high values can be ignored because these rapid motions lengthen both T1 and T2 by the same fraction. However, residues T8 and L78 deviate by more than two standard deviations below the average T1 range with no comparable decrease in T2; this also accounts for their inability to be fit to any of the models (Mandel et al. 1995). Next, the assumption that the motions on the τ_E time scale do not contribute to T1 is not valid when $\{^1H\}$-^{15}N NOE < 0.6 (Clore et al. 1990), as reported for residues 3, 21, 73, 74, 94 and 108. With these considerations, the initial value of τ_C was estimated to be 4.50 ns and τ_C was optimized to 4.54 ns in the final model-free analysis.

Panel D in Figure 4.5 plots the generalized order parameter (S^2) as a function of the primary sequence. The generalized order parameter specifies the degree of spatial restriction of the bond vector, and motions represented by the generalized order parameter will be referred to as dynamics on the picosecond to nanosecond timescale. The values of the generalized order parameter range from zero for isotropic internal motion to one for completely restricted motion (Lipari and Szabo 1982). Overall, CEG-thioredoxin dynamics when compared to the reduced and oxidized forms show similar trends in motions in the picosecond to nanosecond time scale, but residue to residue comparison shows that the chemical modification decreases the generalized order parameter for most of the residues and for the first time identifies several unobvious regions possibly associated with the transient function of thioredoxin. On average, the reported generalized order

parameter for the reduced form is 0.87 and for the oxidized form is 0.86 indicating little variation in the overall rigidity of thioredoxin between redox states. CEG-thioredoxin has an average generalized order parameter of 0.80 indicating an overall increase in mobility in the picosecond to nanosecond time scale as a result of the chemical adduction.

The generalized order parameter for most residues in CEG-thioredoxin is within the expected values for a compact, folded protein (> 0.7) and sequence alignment exhibits trends consistent with the secondary and tertiary structures (Fig. 4.5d). Residues L24, W28, G51, I60, F81 and G97 all have S^2 values greater than 0.95 indicating limited mobility in the picosecond to nanosecond time scale comparatively described as also being less mobile in the reduced and oxidized forms (Stone et al. 1993). A108 has the lowest generalized order parameter (0.24) and the absence of resonances for the first two residues, most likely due to conformational exchange broadening, and the low generalized order parameter of K3 (0.61) indicates that the N-terminus and C-terminus of CEG-thioredoxin are just as disordered as in the reduced and oxidized forms.

Two loops made up of residues A67-I75 (P68 contains no amide proton) and V91-L95, previously identified as being significantly more mobile than other residues (Stone et al. 1993), here display a substantial increase in backbone dynamics over the corresponding residues in the reduced and oxidized forms (Table 4.1). In the oxidized structure, residues A67-I75 are found packed against the active site where the side chain of I75 is within van der Waals distance of the disulphide bond (Jeng et al. 1994). The change in S^2 of G74 (decreased by 0.22

Table 4.1 Residue-specific generalized order parameters of loops A67-I75 and V91-L94.

	Residue	Reduced	Oxidized	CEG
Active site loop A67-I75	A67	0.94	0.90	0.76
	K69	0.90	0.89	0.79
	Y70	0.87	0.86	0.82
	G71	0.86	0.84	0.72
	I72	0.93	0.87	0.78
	R73	0.75	0.72	0.60
	G74	0.51	0.50	0.29
	I75	0.81	0.95	0.58
Average		0.82	0.82	0.67
Active site loop V91-L94	V91	0.85	0.95	0.68
	G92	0.89	0.86	0.81
	A93	0.78	0.76	0.73
	L94	0.68	0.68	0.57
Average		0.80	0.81	0.70

compared to the reduced form and 0.21 compared to the oxidized form) and I75 (decreased by 0.23 compared to the reduced form and decreased by 0.37 compared to the oxidized form) show the largest decrease in the generalized order parameter of residues in the A67- I75 sequence, where the average decrease in S^2 is 0.15 (Table 4.1). The second loop is made up of V91-L94, a series of hydrophobic residues that complete the hydrophobic face of the active site. S95 does not have resolved resonances in CEG-thioredoxin or in either of the two previous structures, indicating that this residue is highly disordered in all three thioredoxin forms. The average decrease in S^2 in both loops of CEG-thioredoxin is 0.13, indicating that the plasticity of the two loops that make up the hydrophobic surface surrounding the active site are more effected by the chemical modification than most residues.

The comparison of other individual residue generalized order parameters can be used to identify those residues and/or regions that have a larger than average ($\Delta S^2 > 0.07$) increase in their backbone dynamics from the two native forms. The helix α-2 contains a region (residues A39 and I41-I45) that has a significant decrease in S^2 (Table 4.2). Inspection of the region around A39 and I41-I45 (residue 40 is proline and therefore contains no amide proton) in helix α-2 reveals correlating increases in the neighboring sequences (Fig. 4.6). Molecular modeling of the thioredoxin structure shows that the methyl group of A39 is part of a hydrophobic face that includes the side chains of I41, L42, I45, L94 and L99, which are packed against the central β-sheet. Residues I41, L42, I45 and L99 are shielded from the solvent with the side chain of K96 crossing the between the

Table 4.2 Generalized order parameters of residues with a greater than 0.07 decrease in S^2 compared to the reduced and oxidized forms. The first group corresponds to the residues discussed in Figure 4.6, the second group corresponds to the residues discussed in Figure 4.7 (G21 is included for the continuity of the discussion), and the third grouping is the other identified residues.

Location	Residue	Reduced	Oxidized	CEG
Helix α-2	A39	0.94	0.95	0.78
Helix α-2	I41	0.88	0.89	0.81
Helix α-2	L42	0.96	0.95	0.77
Helix α-2	D43	-	0.87	0.80
Helix α-2	E44	-	0.89	0.80
Helix α-2	I45	0.96	0.90	0.81
Loop	L94	0.68	0.68	0.57
Helix α-4	K96	0.98	-	0.79
Helix α-4	Q98	0.95	0.96	0.84
Helix α-4	L99	0.95	0.95	0.84
Avcrage		0.91	0.89	0.78
Loop	D20	0.75	0.74	0.60
Loop	G21	0.41	0.37	0.37
Strand β-2	A22	0.73	0.73	0.64
Strand β-2	I23	0.93	0.91	0.59
Strand β-4	L80	0.90	0.93	0.79
Strand β-4	K82	0.88	0.89	0.80
Strand β-5	E85	0.95	0.85	0.72
Strand β-5	V86	0.83	0.85	0.76
Strand β-5	A87	0.90	0.91	0.83
Strand β-5	A88	0.86	0.86	0.72
Strand β-5	T89	0.86	0.86	0.77
Helix α-4	N106	0.82	0.81	0.67
Average		0.82	0.81	0.69
Strand β-1	I5	0.89	0.87	0.73
Strand β-1	H6	0.86	0.88	0.69
Loop	D9	0.94	-	0.74
Loop	L17	0.89	0.91	0.80
Loop	N59	0.90	0.88	0.78
Average		0.90	0.89	0.75

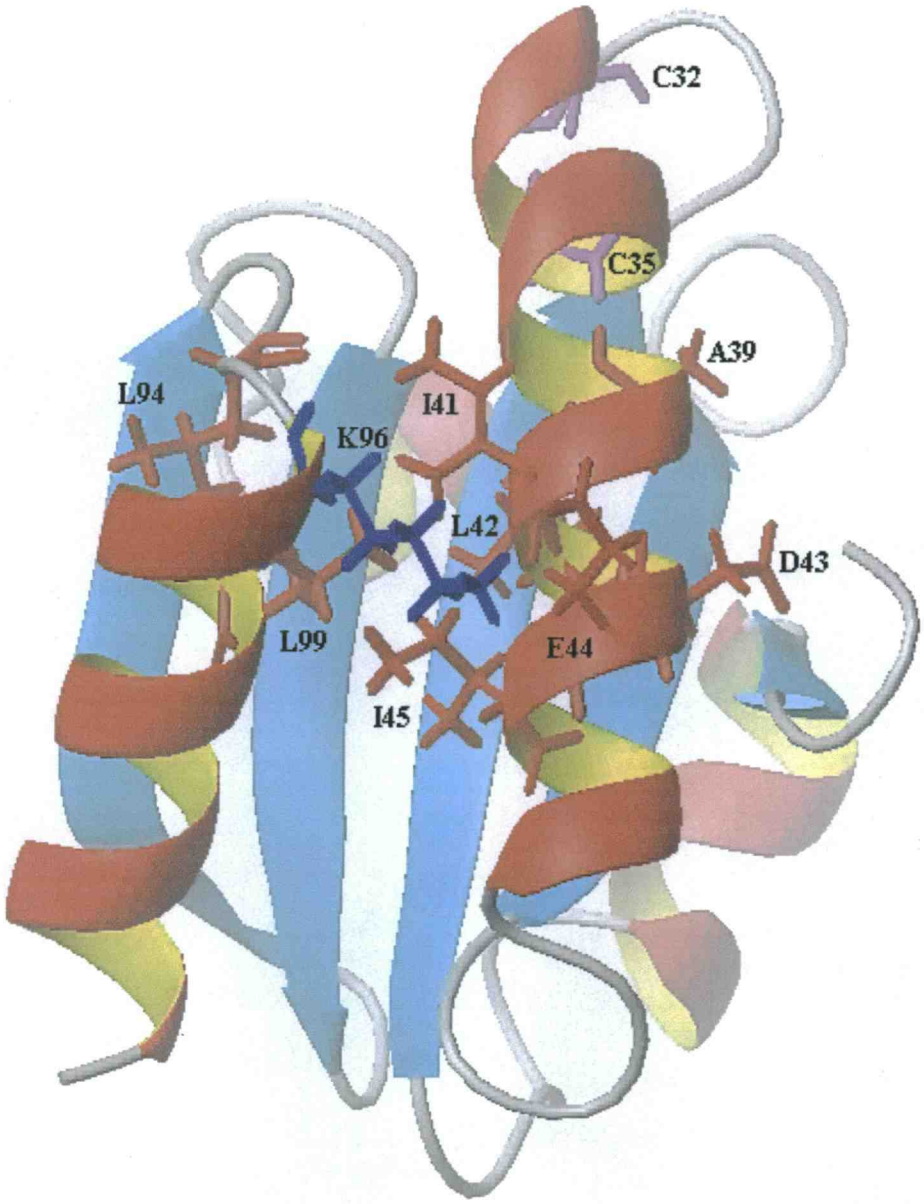


Figure 4.6 Model of the hydrophobic interactions associated with the residues in helix α-2. Residue side-chains of A39, I41, L42, D43, E44, I45, L94, K96 and L99, which all have a decrease in S^2 greater than 0.07, are indicated. Lysine 96 (K96) is indicated in blue for contrast. Side-chains of C32 and C35 are included for orientation. This figure was prepared with MOLMOL (Koradi et al. 1996).

solvent and I41, L42 and L99 with the K96- Nξ group forming a hydrogen bond to the side chain carbonyl oxygen of E44. Through this series of hydrophobic interactions, the significance of loop K90-L94 in the function of thioredoxin is that its accommodating motions can induce dynamical changes in the stabilizing structures.

Interestingly, many of the residues with a significant decrease in the generalized order parameter are not found at the active site (Table 4.2). Residues D20-G21 are located in the loop between helix α-1 and strand β-2 with residues A22 and I23 as the first two residues at the N-terminus of strand β-2. Both the reduced and oxidized forms show a similar generalized order parameter in residue G21, but these low values of G21 were believed to be a result of the conformational freedom of glycine (Stone et al. 1993). Here, chemical modification results in an increase in fast motional times in D20 and the first two residues of strand β-2, A22 and I23. Residues D20, A22 and I23 compose the sequence neighboring G21 indicating that this turn and secondary structure are effected by the conformation of the active site. Structurally, strand β-2 is the secondary structure that precedes the active site. An increase in the dynamics at this part of the β-sheet is complemented by an increase in the dynamics of strand β-4 residues L80 and K82 and strand β-5 residues E85-T89 (Fig. 4.7). In this group of amino acids A87 and A88 both pack next to N106 and L80, and the side chain amide of K82 is aligned with the carbonyl oxygen of N106. There is no correlation of residue packing to the active site, but because this region is

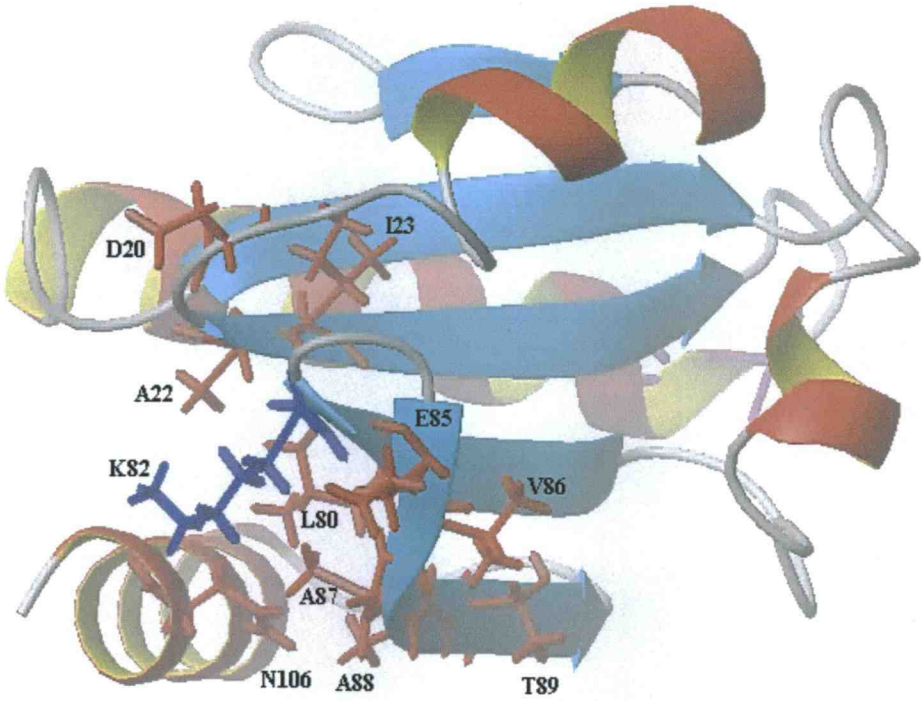


Figure 4.7 Model of structurally/functionally stabilizing elements of residues that have a decrease in the generalized order parameter (S^2) greater than 0.07. Amino acids D20, A22, I23, L80, E85, V86, A87, A88, T89, K82 and N106 are represented by side-chain groups. The K82 side-chain group is colored blue for contrast. This figure was prepared with MOLMOL (Koradi et al. 1996).

organized around the secondary structures leading to the active site and loops G71-I75 and V91-L94, together these residues make up the structures that pack the functional elements against one side of the core β-sheet.

Residues I5, H6, D9, L17 and N59 are more mobile in CEG-thioredoxin, but could not be associated with any of the fast regions described above. Although I5 and H6 are highly mobile, the backbone amide of H6 is solvent exposed and the lack of main chain hydrogen bonding would make this sequence an ideal "breaking point" for terminating strand β-1 of the β-sheet. The other three residues are all solvent exposed with D9 found in the loop between strand β-1 and helix α-1, L17, which is the N-terminal residue of helix α-1, and N59, a linker between strand β-3 and helix α-3. Although these sparse residues have not been correlated to thioredoxin structure/function in this study, they represent how distributive the effects of chemical modification on functionally important residues can be (Fig. 4.8).

4.3.5 Comparison of Backbone Dynamics to H/D Exchange Rates

Comparing NMR-derived dynamical data to CEG-thioredoxin backbone amide proton H/D exchange rates measured by mass spectroscopy (Kim et al. 2001) provides no additional information as to how the motions on the picosecond to nanosecond timescale might affect the conformational stability of thioredoxin. For example, comparing CEG-thioredoxin to the reduced and oxidized forms shows that Q62 has an increase in S^2 with a small change in the MS derived H/D exchange rate while N63 has a small change in S^2 and a large change in the H/D

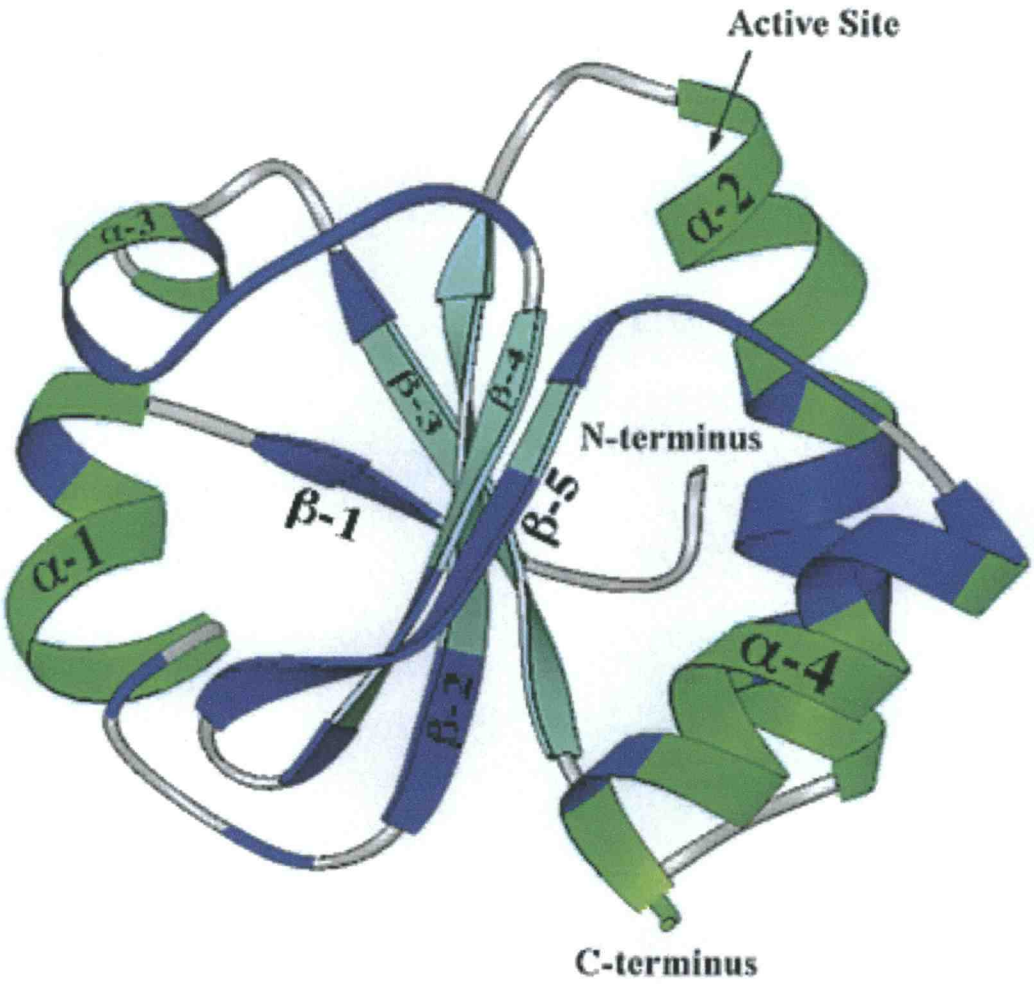


Figure 4.8 Ribbon model of CEG-thioredoxin showing residues found to have significant increases in backbone mobility. Helices (green), strands (light blue) and loops (grey) are represented. Residues with a decrease in the generalized order parameter (S^2) greater than 0.07 are shown in blue. This figure was prepared with Ribbons (Carson and Bugg 1986).

exchange rate. The lack of correlation is not surprising since only changes in the order of a segment or changes in the conformational heterogeneity can lead to an increase or decrease in the H/D exchange rate of amide protons. Therefore, H/D exchange rates measured by mass spectroscopy are not used in the analysis of the NMR-derived changes in fast motions observed in CEG-thioredoxin.

4.4 Conclusion

The purpose of this study was to map residue-specific changes in the structure and the internal motions as a result of chemical adduction of the functionally important C32 in thioredoxin. The results indicate that initial mass spectroscopy H/D exchange-based prediction of the interactions of the glutathione with thioredoxin is very close to the structure as seen in this NMR analysis and similar to those identified in thioltransferase. Relaxation measurements show that the active site and the hydrophobic loops G71-I75 and V91-L94 are predictably affected, which likely result from a rearrangement to accommodate the cysteine-ethyl glutathione group. Interestingly, larger than average increases in backbone dynamics also occur in the secondary structures that lead in and out of the active site and the two loops. There were also a series of residues, which unexpectedly had a larger than average increase in S^2, but were not correlated to any series of interactions or structural regions of thioredoxin. Although the mechanism of these changes in mobility is not known, the inability of this adduct to disrupt the secondary or tertiary structure and the ability of it to alter the dynamical profile

suggest that rearrangements at the active site are accommodated by the plasticity of the protein.

Acknowledgements

We would like to thank P. Andrew Karplus for critical reading of this manuscript. This publication was made possible by grant number P01 ES00040 and P30 ES00210 from the National Institute of Environmental Health Sciences (NIEHS), NIH. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS, NIH. We also acknowledge the Nucleic Acids and Proteins Facility Services Core of the Environmental Health Sciences Center at Oregon State University.

4.5 Materials and Methods

Thioredoxin was expressed in the pBR325/pBHK8 vector with the Ampicillin resistance marker from Lunn et al. (Lunn et al. 1984) and transformed into the *E coli* K-12 MG1655 strain acquired from the *E. coli* genome project at the University of Wisconsin (Blattner et al. 1997). The MG1655 strain was chosen due to its properties as a cell line capable of growth in minimal media. Cells were grown overnight by adding a 50 ml-inoculum to 1 L KD medium. KD medium was made by autoclaving 5 g yeast extract, 10 g peptone from casein, tryptic digest in 900 ml deionized water. Once cooled, 100 ml 10x KD salts (1.0 g $MgSO_4$, 9.99 g citric acid, 5.0 g K_2HPO_4 and 1.76 g Na/NH_4HPO_4), 20 ml 20%

glycerol (sterile), 10 ml Basal vitamins (or 1 mg thiamine) and 300 μl ampicillin stock (150 mg/ml) was added. Thioredoxin was purified from lysed cells as previously described (Dyson et al. 1997). Thioredoxin modification with the dichloroethane glutathione (CEG) conjugate and subsequent purification were done as described (Erve et al. 1995). Fractions were resuspended in 3.5 M urea (0.4 M ammonium bicarbonate, pH 7.7) and dialyzed extensively in 0.1 M potassium phosphate (pH 5.7) to match the final buffer conditions as reported by Stone et al. (1993). The final molecular weight and purity of CEG-thioredoxin were confirmed using MALDI-TOF mass spectrometry analysis.

Samples were prepared in 0.5 M potassium phosphate buffer (pH 5.7) as previously described (Chandrasekhar et al. 1991; Stone et al. 1993). Relaxation spectra were processed using two different protocols for resolution enhancement of various cross-peaks in the ^{1}H-^{15}N heteronuclear correlation spectra. Most peaks were resolved using the first protocol with an 8 Hz exponential line broadening in ω1. Protocol 2 was used to quantify overlapped peaks by using a sine function applied to ω1. T1 and T2 spin relaxation rates were quantized using the two-parameter non-linear Levenberg-Marquardt fitting algorithm and Monte Carlo simulation procedure. Improvement of the two parameter fitting function was accomplished by removing time points determined to be comparable to noise in order to prevent a bias in how the function approaches zero (Viles et al. 2001). The heteronuclear $\{^{1}H\}$-^{15}N NOE intensities were determined using the cross-peaks of ^{1}H-^{15}N correlation spectra obtained with and without proton saturation. Generalized order parameters were determined using the model-free method

described by Lipari and Szabo (Lipari and Szabo 1981) and were quantitated with ModelFree 4.0 (Palmer 1991; Mandel et al. 1995) using the FastModelFree (Cole and Loria 2003) for model selection described by Mandel et al. (Mandel et al. 1995).

For ^{15}N, relaxation is predominantly governed by the dipolar interaction with directly bound protons and can be interpreted in terms of dynamic molecular properties without a detailed understanding of the structure. Non-linear Levenberg-Marquardt curve analysis of T1 and T2 relaxation rates were optimized by eliminating time points comparable to noise to prevent a bias in the function as it approaches zero (Viles et al. 2001). Depending on model selection (Mandel et al. 1995), fitting of the relaxation data to the Lipari and Szabo model-free formalism provides up to four parameters to describe amide motions on the picosecond to nanosecond and microsecond to millisecond time scales: S_S^2, S_F^2, R_{EX} and τ_E, where S_S^2 is the slow component of the generalized order parameter (S^2), S_F^2 is the fast component and $S^2=S_S^2+S_F^2$. For most models S^2 is described by S_S^2 (e.g. $S_F^2=0$). Motions on the order of the picosecond to nanosecond time scales are characterized by the generalized order parameter (S^2) and the effective internal correlation time (τ_E) of these motions. The value of S^2 ranges from zero for isotropic internal motion, except in specific anisotropic motional models (Lipari and Szabo 1982), to one for completely restricted motion. Motions on the order of microseconds to milliseconds are described by the R_{EX} term that represents the slow component of amide motions such as conformational changes.

Correlation spectra of ^{15}N T1 and T2 relaxation time measurements and $\{^{1}H\}$-^{15}N NOEs were recorded using standard pulse sequences (Kordel et al. 1992). A recycle delay of 1.7 seconds was used in T1 and T2 measurements to ensure recovery of ^{1}H magnetization. A recycle delay of 5.0 seconds, or approximately five times the longest ^{15}N T1, was used between scans to ensure that maximal NOEs developed before acquisition. Each T1 and T2 experiment was measured by setting the relaxation delay to several different values similar to those found in Stone et al. (1993) T1 experiments, where the delay represents the longitudinal relaxation period, used the seven delays: 42(x2), 125(x2), 242(x2), 500(x2), 1014, 2026 and 3037 ms. The duration of the CPMG pulse sequence in T2 measurements was set to 8 delays of: 18 (x2), 53 (x2), 100 (x2), 200 (x2), 350, 500, 800 and 1200 ms.

All NMR data were collected on a Bruker DRX-600 at 308K. For all data, 4K complex points were collected in the ω_2 domain (sw = 14ppm) with the ^{1}H carrier frequency set to water. Data processing for all T1, T2 and $\{^{1}H\}$-^{15}N NOE spectra was done with NMRPipe, all other data were processed using XWINNMR from Bruker. NMR spectral visualization and chemical shift assignments were performed using the SPARKY assignment program developed at the University of California at San Francisco (Goddard and Kneller 2004). Visualization was accomplished using the MOLMOL (Koradi et al. 1996) and RIBBONS (Carson and Bugg 1986) molecular modeling programs as indicated in the figure captions.

Chapter 5

Conclusion

Each of the studies that comprise this dissertation describes several physical properties important for understanding protein structure and function. Chapter 2 combines the large number of structures in the PDB repository (Berman et al. 2000) with the XTLSSTR secondary structure algorithm (King and Johnson 1999) to identify the commonly overlooked 3_2-helix as a significant secondary structure important for NA recognition. The results presented in Chapter 3 represent one further step in the understanding of helix forming tendencies of residues and how helix initiation and propagation vary with environment (Chakrabartty et al. 1994; Munoz and Serrano 1994; Waterhous and Johnson 1994; Doig and Baldwin 1995; Celinski and Scholtz 2002; Doig 2002; Lawrence and Johnson 2002; Werner et al. 2002). Chapter 4 quantitates an increase in the fast motions of a mimic of a thioredoxin catalytic intermediate; these motions have recently gained interest due to their identification in several functional mechanisms of proteins (Stock 1999; Ishima and Torchia 2000; Spyracopoulos and Sykes 2001; Wand 2001; Mayer et al. 2003).

The 3_2-helix is an Important Secondary Structure

Chapter 2 documents that 5% of the residues in 258 nucleic acid binding proteins are in the 3_2-conformation. This is not surprising as there is recurring evidence of 3_2-helices as a common secondary structure (Woody 1992; Adzhubei and Sternberg 1993; Sreerama and Woody 1994; Stapley and Creamer 1999). However, this study was also able to specifically identify three instances where the 3_2-helices existed within the unassigned regions of reported structures, and that these 3_2-helices were responsible for nucleic acid recognition. In each case the 3_2-motif was capable of forming both specific recognition of the nucleobases and non-specific recognition of the sugar-phosphate backbone.

Although there is recurring evidence for 3_2-helix content in protein structure and 3_2-helices are well represented in protein-protein interactions (Ren et al. 1993; Jardetzky et al. 1996; Mahoney et al. 1997; Reinherz et al. 1999; Kanyalkar et al. 2001), the reporting of this secondary structure has not been included in the reporting of protein structures to the PDB. The DSSP algorithm (Kabsch and Sander 1983) has been instrumental in the understanding of secondary structures in proteins, but it is clearly limited by its identification of secondary structures via hydrogen bonding patterns. In contrast, the XTLSSTR algorithm (King and Johnson 1999) uses intercarbonyl vectors to give a better overall description of chain conformation and XTLSSTR assignments are not limited to common secondary structures like those identified by the DSSP algorithm. With these considerations, it would benefit all structural studies to adopt a hierarchical reporting of secondary structure assignments using

XTLSSTR assignments followed by analysis of the hydrogen bonding patterns identified by the DSSP algorithm. By carefully examining chain conformation and the hydrogen bonding patterns they form, the resulting secondary structure assignment will provide a more detailed annotation of the elements that make up protein structure.

α-Helix Initiation and Propagation

It is well known that the conformation of an amino acid in the final protein structure is dependent on its neighboring sequence and the local environment. Studies involving alanine-based peptides, while informative, are limited due to the lack of variation observed in the primary sequences of those peptides. Helices designed with large numbers of alanine residues ignore the natural diversity in the amino acid sequence that is used to encode different secondary and tertiary structures and are, therefore, may not give reliable insight into the variation in initiation and propagation that occurs in real proteins.

The results from Chapter 3 provide the first ever documentation of a short (< 20 amino acids) peptide that forms two distinct helical segments. The designed YHGH peptide has identical strong helix forming segments at the amino and carboxy terminal ends, and a Thr-Ser-Gly-Ser-Arg sequence separating them. NMR-based structure determination showed that the two termini are helical and the central Gly-Ser residues are not. This result is not consistent with current models of helix initiation and propagation. Specifically, propagation of a helix is considered more favorable than initiation, and it is thought that the end separation

between helix termini cannot be less than three residues in order to enable sufficient solvation of the termini. It is ambiguous what aspects of the initiation-propagation model need to be modified to account for this result, but clearly more than a single helical propensity is required for predicting secondary structure from protein sequence.

Structural Studies of an Adducted Thioredoxin

A thorough understanding of protein stability and binding affinity require dynamical measurements for evaluating the residual entropy in proteins. Synchronous, fast internal motions associated with protein function may lead to the concerted motions necessary for overcoming activation barriers encountered by the transition states (Mayer et al. 2003). These correlated motions have been shown to influence the kinetics of ligand binding and catalysis (Bruice and Benkovic 2000; Osborne et al. 2001). Similarly, quenching of these fast motion in enzymes can be exploited as a mechanism of inhibition (Matko et al. 1980).

The results reported in Chapter 4 show the effects that chemical adduction had on the structure and internal dynamics thioredoxin. First, the position of the cysteine-ethyl glutathione adduct is defined by observed NOEs, and is shown to be consistent with the position predicted from mass spectrometry-derived H/D exchange data (Kim et al. 2001). Second, it was shown that, although the modification doesn't affect the tertiary structure, it does disrupt the backbone dynamics of the entire structure. Many of the increases in backbone dynamics in the picosecond to nanosecond time scale were identified in regions known to be

important in thioredoxin function. Interestingly, many other regions and several isolated residues were also identified as having a larger than average increase in the backbone dynamics. Likely these changes in mobility result from the internal mechanisms that accommodate the rearrangement of the active site due to the cysteine-ethyl glutathione adduct. Although the origin of these changes in the dynamics of a protein is not known, quantitating variations in the dynamical profile could potentially give insight into the sources of residual entropy important for thioredoxin function.

BIBLIOGRAPHY

Abkevich, V. I., Gutin, A. M. and Shakhnovich, E. I. (1994). "Specific nucleus as the transition state for protein folding: evidence from the lattice model." *Biochemistry* **33**(33): 10026-10036.

Abragam, A. (1961). *The principles of nuclear magnetism.* London, England, Oxford University Press.

Adzhubei, A. A., Eisenmenger, F., Tumanyan, V. G., Zinke, M., Brodzinski, S. and Esipova, N. G. (1987). "Approaching a complete classification of protein secondary structure." *J Biomol Struct Dyn* **5**(3): 689-704.

Adzhubei, A. A., Eisenmenger, F., Tumanyan, V. G., Zinke, M., Brodzinski, S. and Esipova, N. G. (1987). "Third type of secondary structure: noncooperative mobile conformation. Protein Data Bank analysis." *Biochem Biophys Res Commun* **146**(3): 934-938.

Adzhubei, A. A., Eisenmenger, F., Tumanyan, V. G., Zinke, M., Brodzinski, S. and Esipova, N. G. (1987b). "Approaching a complete classification of protein secondary structure." *J Biomol Struct Dyn* **5**(3): 689-704.

Adzhubei, A. A. and Sternberg, M. J. (1993). "Left-handed polyproline II helices commonly occur in globular proteins." *J Mol Biol* **229**(2): 472-493.

Akke, M. (2002). "NMR methods for characterizing microsecond to millisecond dynamics in recognition and catalysis." *Curr Opin Struct Biol* **12**(5): 642-647.

Allison, L. A., Moyle, M., Shales, M. and Ingles, C. J. (1985). "Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases." *Cell* **42**(2): 599-610.

Altieri, A. S., Hinton, D. P. and Byrd, R. A. (1995). "Association of biomolecular systems via pulsed field gradient NMR self-diffusion measurements." *Journal of the American Chemical Society* **117**: 7566-7567.

Anfinsen, C. B. (1973). "Principles that govern the folding of protein chains." *Science* **181**(96): 223-230.

Anfinsen, C. B., Haber, E., Sela, M. and White, F. H., Jr. (1961). "The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain." *Proc Natl Acad Sci U S A* **47**: 1309-1314.

Aqvist, J., Luecke, H., Quiocho, F. A. and Warshel, A. (1991). "Dipoles localized at helix termini of proteins stabilize charges." *Proc Natl Acad Sci U S A* **88**(5): 2026-2030.

Artymiuk, P. J., Blake, C. C., Grace, D. E., Oatley, S. J., Phillips, D. C. and Sternberg, M. J. (1979). "Crystallographic studies of the dynamic properties of lysozyme." *Nature* **280**(5723): 563-568.

Aue, W. P., Karhan, J. and Ernst, R. R. (1976). "Two-dimensional spectroscopy. Application to nuclear magnetic resonance." *The journal of chemical physics* **64**: 4226.

Avbelj, F. and Fele, L. (1998). "Role of main-chain electrostatics, hydrophobic effect and side-chain conformational entropy in determining the secondary structure of proteins." *J Mol Biol* **279**(3): 665-684.

Bakk, A., Hoye, J. S. and Hansen, A. (2001). "Heat capacity of protein folding." *Biophys J* **81**(2): 710-714.

Baldwin, R. L. (2002). "Making a network of hydrophobic clusters." *Science* **295**(5560): 1657-1658.

Baldwin, R. L. and Rose, G. D. (1999). "Is protein folding hierarchic? I. Local structure and peptide folding." *Trends Biochem Sci* **24**(1): 26-33.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. (2000). "The Protein Data Bank." *Nucleic Acids Res* **28**(1): 235-242.

Bhutani, N. and Udgaonkar, J. B. (2003). "Folding subdomains of thioredoxin characterized by native-state hydrogen exchange." *Protein Sci* **12**(8): 1719-1731.

Bienkiewicz, E. A., Moon Woody, A. and Woody, R. W. (2000). "Conformation of the RNA polymerase II C-terminal domain: circular dichroism of long and short fragments." *J Mol Biol* **297**(1): 119-133.

Bierzynski, A. and Pawlowski, K. (1997). "Helix-coil transition theories. Are they correct?" *Acta Biochim Pol* **44**(3): 423-432.

Biosym/MSI (1994). *Insight II Users Guide.* San Diego.

Blanch, E. W., Morozova-Roche, L. A., Cochran, D. A., Doig, A. J., Hecht, L. and Barron, L. D. (2000). "Is polyproline II helix the killer conformation? A Raman optical activity study of the amyloidogenic prefibrillar intermediate of human lysozyme." *J Mol Biol* **301**(2): 553-563.

Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. and Shao, Y. (1997). "The complete genome sequence of Escherichia coli K-12." *Science* **277**(5331): 1453-1474.

Bochkarev, A., Bochkareva, E., Frappier, L. and Edwards, A. M. (1998). "The 2.2 A structure of a permanganate-sensitive DNA site bound by the Epstein-Barr virus origin binding protein, EBNA1." *J Mol Biol* **284**(5): 1273-1278.

Brook, C. L. r. (1996). "Helix-coil kinetics: Folding time scales for helical peptides from a sequencial kinetic modl." *Journal of Physical Chemistry* **100**: 2546-2549.

Brook, C. r. (1996). "Helix-coil kinetics: Folding time scales for helical peptides from a sequencial kinetic modl." *Journal of Physical Chemistry* **100**: 2546-2549.

Bruice, T. C. and Benkovic, S. J. (2000). "Chemical basis for enzyme catalysis." *Biochemistry* **39**(21): 6267-6274.

Brunger, A. T. (1992). *X-PLOR*. New Haven, Yale University.

Bushweller, J. H., Billeter, M., Holmgren, A. and Wuthrich, K. (1994). "The nuclear magnetic resonance solution structure of the mixed disulfide between Escherichia coli glutaredoxin(C14S) and glutathione." *J Mol Biol* **235**(5): 1585-1597.

Carson, M. and Bugg, C. E. (1986). "Algorithm for Ribbon Models of Proteins." *Journal of Molecular Graphics* **4**: 121-122.

Carter, P., Andersen, C. A. and Rost, B. (2003). "DSSPcont: Continuous secondary structure assignments for proteins." *Nucleic Acids Res* **31**(13): 3293-3295.

Case, D. A., Dyson, H. J. and Wright, P. E. (1994). "Use of chemical shifts and coupling constants in nuclear magnetic resonance structural studies on peptides and proteins." *Methods Enzymol* **239**: 392-416.

Celinski, S. A. and Scholtz, J. M. (2002). "Osmolyte effects on helix formation in peptides and the stability of coiled-coils." *Protein Sci* **11**(8): 2048-2051.

Chakrabartty, A., Kortemme, T. and Baldwin, R. L. (1994). "Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions." *Protein Sci* **3**(5): 843-852.

Chamberlain, A. K., Handel, T. M. and Marqusee, S. (1996). "Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH." *Nat Struct Biol* **3**(9): 782-787.

Chandrasekhar, K., Krause, G., Holmgren, A. and Dyson, H. J. (1991). "Assignment of the 15N NMR spectra of reduced and oxidized Escherichia coli thioredoxin." *FEBS Lett* **284**(2): 178-183.

Chivers, P. T., Prehoda, K. E., Volkman, B. F., Kim, B. M., Markley, J. L. and Raines, R. T. (1997b). "Microscopic pKa values of Escherichia coli thioredoxin." *Biochemistry* **36**(48): 14985-14991.

Chivers, P. T. and Raines, R. T. (1997). "General acid/base catalysis in the active site of Escherichia coli thioredoxin." *Biochemistry* **36**(50): 15810-15816.

Chou, P. Y. and Fasman, G. D. (1974). "Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins." *Biochemistry* **13**(2): 211-222.

Clore, G. M., Driscoll, P. C., Wingfield, P. T. and Gronenborn, A. M. (1990). "Analysis of the backbone dynamics of interleukin-1 beta using two-dimensional inverse detected heteronuclear 15N-1H NMR spectroscopy." *Biochemistry* **29**(32): 7387-7401.

Clore, G. M. and Gronenborn, A. M. (1998). "Determining the structures of large proteins and protein complexes by NMR." *Trends Biotechnol* **16**(1): 22-34.

Cole, R. and Loria, J. P. (2003). "FAST-Modelfree: a program for rapid automated analysis of solution NMR spin-relaxation data." *J Biomol NMR* **26**(3): 203-213.

Creamer, T. P. (1998). "Left-handed polyproline II helix formation is (very) locally driven." *Proteins* **33**(2): 218-226.

Danielsson, J., Jarvet, J., Damber, P. and Graslund, A. (2002). "Translational diffusion measured by PFG-NMR on full length and fragments of the Alzheimer AB(1-40) peptide. Determination of hydrodynamic radii of random coil peptides of varying length." *Magnetic resonance in chemistry* **40**: S89-S97.

Dekant, W., Vamvakas, S. and Anders, M. W. (1989). "Bioactivation of nephrotoxic haloalkenes by glutathione conjugation: formation of toxic and mutagenic intermediates by cysteine conjugate beta-lyase." *Drug Metab Rev* **20**(1): 43-83.

Dillet, V., Dyson, H. J. and Bashford, D. (1998). "Calculations of electrostatic interactions and pKas in the active site of Escherichia coli thioredoxin." *Biochemistry* **37**(28): 10298-10306.

Dinner, A. R., Sali, A., Smith, L. J., Dobson, C. M. and Karplus, M. (2000). "Understanding protein folding via free-energy surfaces from theory and experiment." *Trends Biochem Sci* **25**(7): 331-339.

Dobson, C. M., Sali, A. and Karplus, M. (1998). "Protein folding: a perspective from theory and experiment." *Angew. Chem. Int. Ed.* **37**(7): 868-893.

Doig, A. J. (2002). "Recent advances in helix-coil theory." *Biophys Chem* **101-102**: 281-293.

Doig, A. J. and Baldwin, R. L. (1995). "N- and C-capping preferences for all 20 amino acids in alpha-helical peptides." *Protein Sci* **4**(7): 1325-1336.

Dyson, H. J., Jeng, M. F., Tennant, L. L., Slaby, I., Lindell, M., Cui, D. S., Kuprin, S. and Holmgren, A. (1997). "Effects of buried charged groups on cysteine thiol ionization and reactivity in Escherichia coli thioredoxin: structural and functional characterization of mutants of Asp 26 and Lys 57." *Biochemistry* **36**(9): 2622-2636.

Englander, S. W. and Mayne, L. (1992). "Protein folding studied using hydrogen-exchange labeling and two-dimensional NMR." *Annu Rev Biophys Biomol Struct* **21**: 243-265.

Ernst, R. R. and Anderson, W. A. (1966). "Application of Fourier transform spectroscopy to magnetic resonance." *Review of scientific instrumentation* **37**.

Erve, J. C., Barofsky, E., Barofsky, D. F., Deinzer, M. L. and Reed, D. J. (1995). "Alkylation of Escherichia coli thioredoxin by S-(2-chloroethyl)glutathione and identification of the adduct on the active site cysteine-32 by mass spectrometry." *Chem Res Toxicol* **8**(7): 934-941.

Erve, J. C., Deinzer, M. L. and Reed, D. J. (1996). "Reaction of human hemoglobin toward the alkylating agent S-(2-chloroethyl)glutathione." *J Toxicol Environ Health* **49**(2): 127-143.

Eyles, S. J., Radford, S. E., Robinson, C. V. and Dobson, C. M. (1994). "Kinetic consequences of the removal of a disulfide bridge on the folding of hen lysozyme." *Biochemistry* **33**(44): 13038-13048.

Feher, V. A. and Cavanagh, J. (1999). "Millisecond-timescale motions contribute to the function of the bacterial response regulator protein Spo0F." *Nature* **400**(6741): 289-293.

Feng, S., Chen, J. K., Yu, H., Simon, J. A. and Schreiber, S. L. (1994). "Two binding orientations for peptides to the Src SH3 domain: development of a general model for SH3-ligand interactions." *Science* **266**(5188): 1241-1247.

Fersht, A. R. (1995). "Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications." *Proc Natl Acad Sci U S A* **92**(24): 10869-10873.

Fischer, D. and Eisenberg, D. (1996). "Protein fold recognition using sequence-derived predictions." *Protein Sci* **5**(5): 947-955.

Fox, D. G., Cary, P. D. and Kneale, G. G. (1999). "Conformational studies of the C-terminal domain of bacteriophage Pf1 gene 5 protein." *Biochim Biophys Acta* **1435**(1-2): 138-146.

Frauenfelder, H., Petsko, G. A. and Tsernoglou, D. (1979). "Temperature-dependent X-ray diffraction as a probe of protein structural dynamics." *Nature* **280**(5723): 558-563.

Freedman, R. B., Hirst, T. R. and Tuite, M. F. (1994). "Protein disulphide isomerase: building bridges in protein folding." *Trends Biochem Sci* **19**(8): 331-336.

Frishman, D. and Argos, P. (1995). "Knowledge-based protein secondary structure assignment." *Proteins* **23**(4): 566-579.

Fuentes, E. J. and Wand, A. J. (1998). "Local dynamics and stability of apocytochrome b562 examined by hydrogen exchange." *Biochemistry* **37**(11): 3687-3698.

Gane, P. J., Freedman, R. B. and Warwicker, J. (1995). "A molecular model for the redox potential difference between thioredoxin and DsbA, based on electrostatics calculations." *J Mol Biol* **249**(2): 376-387.

Garcia, C., Nishimura, C., Cavagnero, S., Dyson, H. J. and Wright, P. E. (2000). "Changes in the apomyoglobin folding pathway caused by mutation of the distal histidine residue." *Biochemistry* **39**(37): 11227-11237.

Goch, G., Maciejczyk, M., Oleszczuk, M., Stachowiak, D., Malicka, J. and Bierzynski, A. (2003). "Experimental investigation of initial steps of helix propagation in model peptides." *Biochemistry* **42**(22): 6840-6847.

Goddard, T. D. and Kneller, D. G. (2004). Sparky. University of California, San Francisco.

Goldman, M. and Porneuf, M. (1994). *NMR and more, in honro of Anatole Abragam*. Les Ulis, France.

Grauschopf, U., Winther, J. R., Korber, P., Zander, T., Dallinger, P. and Bardwell, J. C. (1995). "Why is DsbA such an oxidizing disulfide catalyst?" *Cell* **83**(6): 947-955.

Gross, J. D., Gelev, V. M. and Wagner, G. (2003). "A sensitive and robust method for obtaining intermolecular NOEs between side chains in large protein complexes." *J Biomol NMR* **25**(3): 235-242.

Hodsdon, M. E. and Frieden, C. (2001). "Intestinal fatty acid binding protein: the folding mechanism as determined by NMR studies." *Biochemistry* **40**(3): 732-742.

Holley, L. H. and Karplus, M. (1989). "Protein secondary structure prediction with a neural network." *Proc Natl Acad Sci U S A* **86**(1): 152-156.

Holmgren, A. (1985). "Thioredoxin." *Annu Rev Biochem* **54**: 237-271.

Holmgren, A. (1989). "Thioredoxin and glutaredoxin systems." *J Biol Chem* **264**(24): 13963-13966.

Huber, H. E., Russel, M., Model, P. and Richardson, C. C. (1986). "Interaction of mutant thioredoxins of Escherichia coli with the gene 5 protein of phage T7. The redox capacity of thioredoxin is not required for stimulation of DNA polymerase activity." *J Biol Chem* **261**(32): 15006-15012.

Humphrey, W., Dalke, A. and Schulten, K. (1996). "VMD: visual molecular dynamics." *J Mol Graph* **14**(1): 33-38, 27-38.

Humphreys, W. G., Kim, D. H., Cmarik, J. L., Shimada, T. and Guengerich, F. P. (1990). "Comparison of the DNA-alkylating properties and mutagenic responses of a series of S-(2-haloethyl)-substituted cysteine and glutathione derivatives." *Biochemistry* **29**(45): 10342-10350.

Ishima, R. and Torchia, D. A. (2000). "Protein dynamics from NMR." *Nat Struct Biol* **7**(9): 740-743.

Jackson, S. E. and Fersht, A. R. (1991). "Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition." *Biochemistry* **30**(43): 10428-10435.

Jackson, S. E. and Fersht, A. R. (1991). "Folding of chymotrypsin inhibitor 2. 2. Influence of proline isomerization on the folding kinetics and thermodynamic characterization of the transition state of folding." *Biochemistry* **30**(43): 10436-10443.

Jaravine, V. A., Alexandrescu, A. T. and Grzesiek, S. (2001). "Observation of the closing of individual hydrogen bonds during TFE-induced helix formation in a peptide." *Protein Sci* **10**(5): 943-950.

Jardetzky, T. S., Brown, J. H., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L. and Wiley, D. C. (1996). "Crystallographic analysis of endogenous peptides associated with HLA-DR1 suggests a common, polyproline II-like conformation for bound peptides." *Proc Natl Acad Sci U S A* **93**(2): 734-738.

Jeener, J., Meier, B. H., Bachmann, P. and Ernst, R. R. (1979). "Invesitgation of exchange processes by two-dimensional NMR spectroscopy." *The journal of chemical physics* **71**: 4546-4553.

Jeng, M. F., Campbell, A. P., Begley, T., Holmgren, A., Case, D. A., Wright, P. E. and Dyson, H. J. (1994). "High-resolution solution structures of oxidized and reduced Escherichia coli thioredoxin." *Structure* **2**(9): 853-868.

Jeng, M. F., Holmgren, A. and Dyson, H. J. (1995). "Proton sharing between cysteine thiols in Escherichia coli thioredoxin: implications for the mechanism of protein disulfide reduction." *Biochemistry* **34**(32): 10101-10105.

Jennings, A. J., Edge, C. M. and Sternberg, M. J. (2001). "An approach to improving multiple alignments of protein sequences using predicted secondary structure." *Protein Eng* **14**(4): 227-231.

Jennings, P. A. and Wright, P. E. (1993). "Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin." *Science* **262**(5135): 892-896.

Johnson, C. S. J. (1999). "Diffusion ordered nuclear magnetic resonance spectroscopy: principles and applications." *Progress in nuclear magnetic resonance spectroscopy* **34**: 203-256.

Kabsch, W. and Sander, C. (1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." *Biopolymers* **22**(12): 2577-2637.

Kanyalkar, M., Srivastava, S. and Coutinho, E. (2001). "Conformation of N-terminal HIV-1 Tat (fragment 1-9) peptide by NMR and MD simulations." *J Pept Sci* **7**(11): 579-587.

Karplus, M. and McCammon, J. A. (2002). "Molecular dynamics simulations of biomolecules." *Nat Struct Biol* **9**(9): 646-652.

Karplus, P. A. (2004). Globalcore/Globalmerge. Corvallis, OR.

Kay, L. E., Torchia, D. A. and Bax, A. (1989). "Backbone dynamics of proteins as studied by 15N inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease." *Biochemistry* **28**(23): 8972-8979.

Kelly, M. A., Chellgren, B. W., Rucker, A. L., Troutman, J. M., Fried, M. G., Miller, A. F. and Creamer, T. P. (2001). "Host-guest study of left-handed polyproline II helix formation." *Biochemistry* **40**(48): 14376-14383.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H. and Phillips, D. C. (1958). "A three-dimensional model of the myoglobin molecule obtained by x-ray analysis." *Nature* **181**(4610): 662-666.

Kim, M. Y., Maier, C. S., Reed, D. J. and Deinzer, M. L. (2002). "Conformational changes in chemically modified Escherichia coli thioredoxin monitored by H/D exchange and electrospray ionization mass spectrometry." *Protein Sci* **11**(6): 1320-1329.

Kim, M. Y., Maier, C. S., Reed, D. J., Ho, P. S. and Deinzer, M. L. (2001). "Intramolecular interactions in chemically modified Escherichia coli thioredoxin monitored by hydrogen/deuterium exchange and electrospray ionization mass spectrometry." *Biochemistry* **40**(48): 14413-14421.

King, S. M. and Johnson, W. C. (1999). "Assigning secondary structure from protein coordinate data." *Proteins* **35**(3): 313-320.

Klein-Seetharaman, J., Oikawa, M., Grimshaw, S. B., Wirmer, J., Duchardt, E., Ueda, T., Imoto, T., Smith, L. J., Dobson, C. M. and Schwalbe, H. (2002). "Long-range interactions within a nonnative protein." *Science* **295**(5560): 1719-1722.

Klemm, J. D., Rould, M. A., Aurora, R., Herr, W. and Pabo, C. O. (1994). "Crystal structure of the Oct-1 POU domain bound to an octamer site: DNA recognition with tethered DNA-binding modules." *Cell* **77**(1): 21-32.

Koradi, R., Billeter, M. and Wuthrich, K. (1996). "MOLMOL: a program for display and analysis of macromolecular structures." *J Mol Graph* **14**(1): 51-55, 29-32.

Kordel, J., Skelton, N. J., Akke, M., Palmer, A. G., 3rd and Chazin, W. J. (1992). "Backbone dynamics of calcium-loaded calbindin D9k studied by two-dimensional proton-detected 15N NMR spectroscopy." *Biochemistry* **31**(20): 4856-4866.

Krittanai, C. and Johnson, W. C., Jr. (2000). "The relative order of helical propensity of amino acids changes with solvent environment." *Proteins* **39**(2): 132-141.

Kumar, J. K., Tabor, S. and Richardson, C. C. (2004). "Proteomic analysis of thioredoxin-targeted proteins in Escherichia coli." *Proc Natl Acad Sci U S A* **101**(11): 3759-3764.

Kundu, S., Melton, J. S., Sorensen, D. C. and Phillips, G. N., Jr. (2002). "Dynamics of proteins in crystals: comparison of experiment with simple models." *Biophys J* **83**(2): 723-732.

Lam, S. L. and Hsu, V. L. (2003). "NMR identification of left-handed polyproline type II helices." *Biopolymers* **69**(2): 270-281.

Laurent, T. C., Moore, E. C. and Reichard, P. (1964). "Enzymatic synthesis of deoxyribonucleotides IV." *Journal of biological chemistry* **239**: 3436-3444.

LaVoie, M. J. and Hastings, T. G. (1999). "Dopamine quinone formation and protein modification associated with the striatal neurotoxicity of methamphetamine: evidence against a role for extracellular dopamine." *J Neurosci* **19**(4): 1484-1491.

Lawrence, J. R. and Johnson, W. C. (2002). "Lifson-Roig nucleation for alpha-helices in trifluoroethanol: context has a strong effect on the helical propensity of amino acids." *Biophys Chem* **101-102**: 375-385.

Lee, A. L., Kinnear, S. A. and Wand, A. J. (2000). "Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex." *Nat Struct Biol* **7**(1): 72-77.

Leopold, P. E., Montal, M. and Onuchic, J. N. (1992). "Protein folding funnels: a kinetic approach to the sequence-structure relationship." *Proc Natl Acad Sci U S A* **89**(18): 8721-8725.

Levinthal, C. (1968). "Are there pathways for protein folding?" *J chim phys phys-chim biol* **65**: 44-45.

Levitt, M. and Warshel, A. (1975). "Computer simulation of protein folding." *Nature* **253**(5494): 694-698.

Lewis, H. A., Chen, H., Edo, C., Buckanovich, R. J., Yang, Y. Y., Musunuru, K., Zhong, R., Darnell, R. B. and Burley, S. K. (1999). "Crystal structures of Nova-1 and Nova-2 K-homology RNA-binding domains." *Structure Fold Des* **7**(2): 191-203.

Lewis, H. A., Musunuru, K., Jensen, K. B., Edo, C., Chen, H., Darnell, R. B. and Burley, S. K. (2000). "Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome." *Cell* **100**(3): 323-332.

Lifson, S. and Roig, A. (1961). "On the theory of helix-coil transition in polypeptides." *The journal of chemical physics* **34**(6): 1963-1974.

Lipari, G. and Szabo, A. (1981). "Nuclear magnetic resonance relaxation in nucleic acid fragments: models for internal motion." *Biochemistry* **20**(21): 6250-6256.

Lipari, G. and Szabo, A. (1982). "Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity." *Journal of the American Chemical Society* **104**(17): 4546-4559.

Lu, H., Buck, M., Radford, S. E. and Dobson, C. M. (1997). "Acceleration of the folding of hen lysozyme by trifluoroethanol." *J Mol Biol* **265**(2): 112-117.

Lunn, C. A., Kathju, S., Wallace, B. J., Kushner, S. R. and Pigiet, V. (1984). "Amplification and purification of plasmid-encoded thioredoxin from Escherichia coli K12." *J Biol Chem* **259**(16): 10469-10474.

Mahoney, N. M., Janmey, P. A. and Almo, S. C. (1997). "Structure of the profilin-poly-L-proline complex involved in morphogenesis and cytoskeletal regulation." *Nat Struct Biol* **4**(11): 953-960.

Maity, H., Maity, M. and Englander, S. W. (2004). "How cytochrome c folds, and why: submolecular foldon units and their stepwise sequential stabilization." *J Mol Biol* **343**(1): 223-233.

Mandel, A. M., Akke, M. and Palmer, A. G., 3rd (1995). "Backbone dynamics of Escherichia coli ribonuclease HI: correlations with structure and function in an active enzyme." *J Mol Biol* **246**(1): 144-163.

Marqusee, S., Robbins, V. H. and Baldwin, R. L. (1989). "Unusually stable helix formation in short alanine-based peptides." *Proc Natl Acad Sci U S A* **86**(14): 5286-5290.

Martin, J. L., Bardwell, J. C. and Kuriyan, J. (1993). "Crystal structure of the DsbA protein required for disulphide bond formation in vivo." *Nature* **365**(6445): 464-468.

Matko, J., Tron, L., Balazs, M., Hevessy, J., Somogyi, B. and Damjanovich, S. (1980). "Correlation between activity and dynamics of the protein matrix of phosphorylase b." *Biochemistry* **19**(25): 5782-5786.

Mayer, K. L., Earley, M. R., Gupta, S., Pichumani, K., Regan, L. and Stone, M. J. (2003). "Covariation of backbone motion throughout a small protein domain." *Nat Struct Biol* **10**(11): 962-965.

McCammon, J. A., Gelin, B. R. and Karplus, M. (1977). "Dynamics of folded proteins." *Nature* **267**(5612): 585-590.

Mertens, J. J., Sterck, J. G., Lau, S. S., Monks, T. J., van Bladeren, P. J. and Temmink, J. H. (1990). "Cytotoxicity of nephrotoxic glutathione-conjugated halohydroquinones." *Toxicol Lett* **53**(1-2): 147-149.

Metcalfe, E. E., Zamoon, J., Thomas, D. D. and Veglia, G. (2004). "(1)H/(15)N heteronuclear NMR spectroscopy shows four dynamic domains for phospholamban reconstituted in dodecylphosphocholine micelles." *Biophys J* **87**(2): 1205-1214.

Modrich, P. and Richardson, C. C. (1975). "Bacteriophage T7 Deoxyribonucleic acid replication in vitro. A protein of Escherichia coli required for bacteriophage T7 DNA polymerase activity." *J Biol Chem* **250**(14): 5508-5514.

Munoz, V. and Serrano, L. (1994). "Elucidating the folding problem of helical peptides using empirical parameters." *Nat Struct Biol* **1**(6): 399-409.

Munoz, V. and Serrano, L. (1995). "Elucidating the folding problem of helical peptides using empirical parameters. II. Helix macrodipole effects and rational modification of the helical content of natural peptides." *J Mol Biol* **245**(3): 275-296.

Munoz, V., Thompson, P. A., Hofrichter, J. and Eaton, W. A. (1997). "Folding dynamics and mechanism of beta-hairpin formation." *Nature* **390**(6656): 196-199.

Myers, J. K., Pace, C. N. and Scholtz, J. M. (1997). "Helix propensities are identical in proteins and peptides." *Biochemistry* **36**(36): 10923-10929.

Myers, J. K., Pace, C. N. and Scholtz, J. M. (1998). "Trifluoroethanol effects on helix propensity and electrostatic interactions in the helical peptide from ribonuclease T1." *Protein Sci* **7**(2): 383-388.

Nelson, J. W. and Kallenbach, N. R. (1986). "Stabilization of the ribonuclease S-peptide alpha-helix by trifluoroethanol." *Proteins* **1**(3): 211-217.

Nishizawa, A. N. and Buchanan, B. B. (1981). "Enzyme regulation in C4 photosynthesis: purification and properties of thioredoxin-linked fructose bisphosphatase and sedoheptulose bisphosphatase from corn leaves." *J Biol Chem* **256**: 6119-6126.

O'Neil, K. T. and DeGrado, W. F. (1990). "A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids." *Science* **250**(4981): 646-651.

Osborne, M. J., Schnell, J., Benkovic, S. J., Dyson, H. J. and Wright, P. E. (2001). "Backbone dynamics in dihydrofolate reductase complexes: role of loop flexibility in the catalytic mechanism." *Biochemistry* **40**(33): 9846-9859.

Ostrander, D. B., Ernst, E. G., Lavoie, T. B. and Gorman, J. A. (1999). "Polyproline binding is an essential function of human profilin in yeast." *Eur J Biochem* **262**(1): 26-35.

Palmer, A. G., 3rd (1997). "Probing molecular motion by NMR." *Curr Opin Struct Biol* **7**(5): 732-737.

Palmer, A. G., 3rd, Rance, M. and Wright, P.E. (1991). "Intramolecular mothions of a zinc finger DNA-binding domain from xfin characterized by proton detected natural abundance (13)C heteronuclear NMR spectroscoopy." *Journal of the American Chemical Society* **113**: 4371-4380.

Pauling, L. and Corey, R. B. (1951c). "The structure of fibrous proteins of the collagen-gelatin group." *Proc Natl Acad Sci U S A* **37**: 272-281.

Pervushin, K., Riek, R., Wider, G. and Wuthrich, K. (1997). "Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution." *Proc Natl Acad Sci U S A* **94**(23): 12366-12371.

Pumford, N. R. and Halmes, N. C. (1997). "Protein targets of xenobiotic reactive intermediates." *Annu Rev Pharmacol Toxicol* **37**: 91-117.

Qian, H. and Chan, S. I. (1996). "Interactions between a helical residue and tertiary structures: helix propensities in small peptides and in native proteins." *J Mol Biol* **261**(2): 279-288.

Radford, S. E., Woolfson, D. N., Martin, S. R., Lowe, G. and Dobson, C. M. (1991). "A three-disulphide derivative of hen lysozyme. Structure, dynamics and stability." *Biochem J* **273(Pt 1)**: 211-217.

Ramachandran, G. N. and Sasisekharan, V. (1968). "Conformation of polypeptides and proteins." *Adv Protein Chem* **23**: 283-438.

Reinherz, E. L., Tan, K., Tang, L., Kern, P., Liu, J., Xiong, Y., Hussey, R. E., Smolyar, A., Hare, B., Zhang, R., Joachimiak, A., Chang, H. C., Wagner, G. and Wang, J. (1999). "The crystal structure of a T cell receptor in complex with peptide and MHC class II." *Science* **286**(5446): 1913-1921.

Ren, R., Mayer, B. J., Cicchetti, P. and Baltimore, D. (1993). "Identification of a ten-amino acid proline-rich SH3 binding site." *Science* **259**(5098): 1157-1161.

Revington, M., Holder, T. M. and Zuiderweg, E. R. (2004). "NMR study of nucleotide-induced changes in the nucleotide binding domain of Thermus thermophilus Hsp70 chaperone DnaK: implications for the allosteric mechanism." *J Biol Chem* **279**(32): 33958-33967.

Richards, F. M. and Kundrot, C. E. (1988). "Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure." *Proteins* **3**(2): 71-84.

Richardson, F. C., Horn, D. M. and Anderson, N. L. (1994). "Dose-responses in rat hepatic protein modification and expression following exposure to the rat hepatocarcinogen methapyrilene." *Carcinogenesis* **15**(2): 325-329.

Richardson, J. S. (1981). "The anatomy and taxonomy of protein structure." *Adv Protein Chem* **34**: 167-339.

Roder, H. and Wuthrich, K. (1986). "Protein folding kinetics by combined use of rapid mixing techniques and NMR observation of individual amide protons." *Proteins* **1**(1): 34-42.

Rohl, C. A., Chakrabartty, A. and Baldwin, R. L. (1996). "Helix propagation and N-cap propensities of the amino acids measured in alanine-based peptides in 40 volume percent trifluoroethanol." *Protein Sci* **5**(12): 2623-2637.

Rohl, C. A., Scholtz, J. M., York, E. J., Stewart, J. M. and Baldwin, R. L. (1992). "Kinetics of amide proton exchange in helical peptides of varying chain lengths. Interpretation by the Lifson-Roig equation." *Biochemistry* **31**(5): 1263-1269.

Rost, B. and Sander, C. (1993). "Prediction of protein secondary structure at better than 70% accuracy." *J Mol Biol* **232**(2): 584-599.

Rost, B., Schneider, R. and Sander, C. (1997). "Protein fold recognition by prediction-based threading." *J Mol Biol* **270**(3): 471-480.

Sadqi, M., Lapidus, L. J. and Munoz, V. (2003). "How fast is protein hydrophobic collapse?" *Proc Natl Acad Sci U S A* **100**(21): 12117-12122.

Shi, Z., Olson, C. A., Rose, G. D., Baldwin, R. L. and Kallenbach, N. R. (2002). "Polyproline II structure in a sequence of seven alanine residues." *PNAS* **99**(14): 9190-9195.

Sklenar, H., Etchebest, C. and Lavery, R. (1989). "Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis." *Proteins* **6**(1): 46-60.

Spyracopoulos, L. and Sykes, B. D. (2001). "Thermodynamic insights into proteins from NMR spin relaxation studies." *Curr Opin Struct Biol* **11**(5): 555-559.

Sreerama, N. and Woody, R. W. (1994). "Poly(pro)II helices in globular proteins: identification and circular dichroic analysis." *Biochemistry* **33**(33): 10022-10025.

Stapley, B. J. and Creamer, T. P. (1999). "A survey of left-handed polyproline II helices." *Protein Sci* **8**(3): 587-595.

Stock, A. (1999). "Biophysics. Relating dynamics to function." *Nature* **400**(6741): 221-222.

Stolorz, P., Lapedes, A. and Xia, Y. (1992). "Predicting protein secondary structure using neural net and statistical methods." *J Mol Biol* **225**(2): 363-377.

Stone, M. J., Chandrasekhar, K., Holmgren, A., Wright, P. E. and Dyson, H. J. (1993). "Comparison of backbone and tryptophan side-chain dynamics of reduced and oxidized Escherichia coli thioredoxin using 15N NMR relaxation measurements." *Biochemistry* **32**(2): 426-435.

Storrs, R. W., Truckses, D. and Wemmer, D. E. (1992). "Helix propagation in trifluoroethanol solutions." *Biopolymers* **32**(12): 1695-1702.

Temmink, J. H., Bruggeman, I. M. and van Bladeren, P. J. (1986). "Cytomorphological changes in liver cells exposed to allyl and benzyl isothiocyanate and their cysteine and glutathione conjugates." *Arch Toxicol* **59**(2): 103-110.

Tiffany, M. L. and Krimm, S. (1968). "New chain conformations of poly(glutamic acid) and polylysine." *Biopolymers* **6**(9): 1379-1382.

Udgaonkar, J. B. and Baldwin, R. L. (1988). "NMR evidence for an early framework intermediate on the folding pathway of ribonuclease A." *Nature* **335**(6192): 694-699.

Van Holde, K. E., Johnson, W. C. and Ho, P. S. (1998). *Principles of physical biochemistry*. Upper Saddle River, N.J., Prentice Hall.

Viles, J. H., Duggan, B. M., Zaborowski, E., Schwarzinger, S., Huntley, J. J., Kroon, G. J., Dyson, H. J. and Wright, P. E. (2001). "Potential bias in

NMR relaxation data introduced by peak intensity analysis and curve fitting methods." *J Biomol NMR* **21**(1): 1-9.

Vuister, G. W. and Bax, A. (1994). "Measurement of four-bond HN-H alpha J-couplings in staphylococcal nuclease." *J Biomol NMR* **4**(2): 193-200.

Wand, A. J. (2001). "Dynamic activation of protein function: a view emerging from NMR spectroscopy." *Nat Struct Biol* **8**(11): 926-931.

Wang, A. C. and Bax, A. (1996). "Determination of the backbone dihedral angles in human ubiquitin from reparametrized emperical karplus equations." *Journal of the American Chemical Society* **118**: 2483-2494.

Wang, Y. and Shortle, D. (1995). "The equilibrium folding pathway of staphylococcal nuclease: identification of the most stable chain-chain interactions by NMR and CD spectroscopy." *Biochemistry* **34**(49): 15895-15905.

Waterhous, D. V. and Johnson, W. C., Jr. (1994). "Importance of environment in determining secondary structure in proteins." *Biochemistry* **33**(8): 2121-2128.

Werner, J., Dyers, R., Fesinmeyerm, R. and Andersen, N. (2002). "Dynamics of the primary processes of protein folding: helix nucleation." *Journal of chemical physics* **106**: 487-494.

Wetlaufer, D. B. (1973). "Nucleation, rapid folding, and globular intrachain regions in proteins." *Proc Natl Acad Sci U S A* **70**(3): 697-701.

Williams, M. and Diwan, S. (1994). Toxilogical profile for 1,2-dichloroethane. Atlanta, GA, US department of health and human services, agency for toxic substances and disease registry.

Williams, S., Causgrove, T. P., Gilmanshin, R., Fang, K. S., Callender, R. H., Woodruff, W. H. and Dyer, R. B. (1996). "Fast events in protein folding: helix melting and formation in a small peptide." *Biochemistry* **35**(3): 691-697.

Williamson, M. P. (1994). "The structure and function of proline-rich regions in proteins." *Biochem J* **297** (**Pt 2**): 249-260.

Wilson, D. S., Guenther, B., Desplan, C. and Kuriyan, J. (1995). "High resolution crystal structure of a paired (Pax) class cooperative homeodomain dimer on DNA." *Cell* **82**(5): 709-719.

Wishart, D. S. and Case, D. A. (2001). "Use of chemical shifts in macromolecular structure determination." *Methods Enzymol* **338**: 3-34.

Woody, R. W. (1992). "Circular dichroism and conformation of unordered peptides." *Adv Biophys Chem* **2**: 37-79.

Wuthrich, K. (1986). *NMR of proteins and nucleic acids*, John Wiley and Sons.

Yancey, P. H., Clark, M. E., Hand, S. C., Bowlus, R. D. and Somero, G. N. (1982). "Living with water stress: evolution of osmolyte systems." *Science* **217**(4566): 1214-1222.

Yancey, P. H. and Somero, G. N. (1979). "Counteraction of urea destabilization of protein structure by methylamine osmoregulatory compounds of elasmobranch fishes." *Biochem J* **183**(2): 317-323.

Yang, Y., Jao, S., Nanduri, S., Starke, D. W., Mieyal, J. J. and Qin, J. (1998). "Reactivity of the human thioltransferase (glutaredoxin) C7S, C25S, C78S, C82S mutant and NMR solution structure of its glutathionyl mixed disulfide intermediate reflect catalytic specificity." *Biochemistry* **37**(49): 17145-17156.

Yu, H., Chen, J. K., Feng, S., Dalgarno, D. C., Brauer, A. W. and Schreiber, S. L. (1994). "Structural basis for the binding of proline-rich peptides to SH3 domains." *Cell* **76**(5): 933-945.

Zdanowski, K. and Dadlez, M. (1999). "Stability of the residual structure in unfolded BPTI in different conditions of temperature and solvent composition measured by disulphide kinetics and double mutant cycle analysis." *J Mol Biol* **287**(2): 433-445.

Zhou, N. E., Kay, C. M., Sykes, B. D. and Hodges, R. S. (1993). "A single-stranded amphipathic alpha-helix in aqueous solution: design, structural characterization, and its application for determining alpha-helical propensities of amino acids." *Biochemistry* **32**(24): 6190-6197.

Zimm, B. and Bragg, J. (1959). "Theory of the phase transistion between helix and random coil in polypeptide chains." *The Journal of Chemical Physics* **31**(2): 526-535.