The Dissertation Committee for Suwei Wang Certifies that this is the approved version of the following dissertation:

ANALYSIS OF THE THERMODYNAMIC DETERMINANTS OF PROTEIN FOLD SPECIFICITY IN THE DENATURED STATE ENSEMBLE

Committee:

Vincent J. Hilser, Ph.D. , Supervisor

Robert O. Fox, Ph.D.

Bernard M. Pettitt, Ph.D.

Andres F. Oberhauser, Ph.D.

R. Bryan Sutton, Ph.D.

Dean, Graduate School

ANALYSIS OF THE THERMODYNAMIC DETERMINANTS OF PROTEIN FOLD SPECIFICITY IN THE DENATURED STATE ENSEMBLE

by

Suwei Wang

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas Medical Branch

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

THE UNIVERSITY OF TEXAS MEDICAL BRANCH

August, 2008

DEDICATION

This dissertation is gratefully dedicated to my family, especially …..

to Dad and Mom for instilling the importance of
hard working and love.

to my sister and brother-in-law for their
understanding and support.

ACKNOWLEDGMENTS

Over the course of my graduate career, I owe my gratitude to all the people who have made this dissertation possible. My graduate experience at UTMB has been one of the most valuable experiences that I will cherish forever.

My deepest gratitude is to my advisor, Dr. Vincent Hilser. It is my great fortune to have such a great advisor who gave me the inspiration of science, the freedom to explore on my own, and the valuable guidance for my research. In Dr. Hilser's research group, I have had the privilege to work with many talented individuals. I'd like to express my gratitude to Drs. Jenny Gu, Scott Larson, and Steven Whitten who have made contributions to my research experience. The harmonious atmosphere and friendly people in our lab will always be my beautiful memory in my life.

I owe my special thanks to Dr. Robert Fox, my committee chairman for his encouraging, critical advice for my research. I also would like to thank the members of my supervisory committee for their agreeing to serve on my committee and their thoughtful comments for my dissertation; Drs. Andres Oberhauser, Bryan Sutton and Montgomery Pettitt.

I also like to thank the BSCB program and BMB program directors, Drs. Wayne Bolen and Lillian Chan for their generous support and guidance.

Finally, I would like to thank the numerous individuals who have offered help for me during my educational experience at UTMB. Their generosity will always be in my heart and encourage me to overcome any difficulties in my life.

ANALYSIS OF THE THERMODYNAMIC DETERMINANTS OF PROTEIN FOLD SPECIFICITY IN THE DENATURED STATE ENSEMBLE

Publication No. _____*

Suwei Wang, Ph.D.

The University of Texas Medical Branch, 2008

Supervisor: Vincent J. Hilser

Although the thermodynamic control of protein folding has been known for decades, a complete understanding of the thermodynamic determinants that defining protein folds is still elusive. In this regard, it is becoming clear that focusing only on the native states of protein folds will be insufficient for deciphering the protein folding problem. Knowledge of the thermodynamics of the denatured state is also necessary. In this project, the thermodynamic determinants of the native fold, present in the denatured ensemble, were investigated and the critical role of the denatured state ensemble in controlling protein folding is discussed. In this work, the COREX algorithm, used until now to model the native state ensemble, was for the first time used to model the denatured state ensembles and investigate the relationship between denatured ensemble energetics and sequences, as well as between denatured ensemble energetics and secondary structures. Substantial thermodynamic differences were found between the denatured and the native states ensembles. The generality and robustness of our results were validated by performing fold-recognition experiments that matched sequences with their respective folds using only energetic information. The success of our study and the

unique energetic information found in denatured states suggest a wide range of strategies for developing novel algorithms for protein prediction and classification.

In addition, this work has particular medical relevance. Understanding the chemical and physical processes underlying thermodynamic determinants of protein folding specificity will enable the rational design of drugs to combat the rapidly expanding family of misfolding diseases. Some misfolding diseases are known to be related to non-specific β-sheet formation. The value of this project lies in the detailed analysis between denatured ensemble energetics and sequences, as well as between energetics and secondary structures. Correlation analysis between structure and energetic information revealed that denatured states have evolutionarily evolved to avoid early β-sheet formation, suggesting that the therapeutic strategies to combat misfolding diseases (especially for diseases related to non β-sheet formations) could be found in the energetics information of the denatured states rather than the native states.

TABLE OF CONTENTS

PAGE

LIST OF ILLUSTRATIONS

Figure **Page**

LIST OF TABLES

Table **Page**

LIST OF ABBREVIATIONS

The abbreviations used in this dissertation are as follows:

ASA	-Accessible Surface Area
FTIR	- Fourier Transform Infrared Spectroscopy
ID	-Intrinsically Disordered
NMR	-Nuclear Magnetic Resonance
PAM	-Partitioning Around Medoids
PDB	-Protein Data Bank
TE_D	-Denatured Thermodynamic Environment
TE_N	-Native Thermodynamic Environment
3D-1D	-Three dimensional one dimensional

CHAPTER 1

General Information

One of the most challenging problems in biology is the protein folding problem. In general, the protein folding problem refers to the question of how a chain of amino acids correctly and rapidly folds into the three-dimensional, functional form. Interest in the protein folding problem can be traced back to the early 1960s. Levinthal first noticed that amino acid sequences adopt a unique fold through a non-random search of the conformation space (Levinthal, 1968). Extensive interest in the protein folding problem has increased since the Nobel Prize winner Christian Anfinsen put forward his famous “thermodynamic hypothesis” for protein folding (Anfinsen, 1973). One of the most important implications of Anfinsen’s thermodynamic hypothesis is that all the information defining the final fold is contained in the primary sequence and that information is thermodynamic in nature. Although experimentalists and computational biologists have studied and attempted to understand the thermodynamic determinants of protein folding since Anfinsen’s hypothesis, most efforts were concentrated on the natively folded state of proteins. This state is favored because it is relatively stable and is biologically functional state that is experimentally easier to control. For this reason, denatured state of proteins has been ignored for years because of its relatively unstable and structurally heterogeneous characteristics. With the development of new spectroscopic techniques and computational simulation approaches, denatured state is gradually drawing the attention of biologists.

Results from nuclear magnetic resonance (NMR) studies have shown that the denatured state contains residual structures instead of adopting completely random coils (Hennig et al., 1999). Computational simulations also support the notion that non-random coil structures are found in the denatured state (Daura et al., 1999). Theoretical arguments also exist for the importance of the denatured state (Dill and Shortle, 1991). First, the denatured state represents the starting point of protein folding. Knowledge about the denatured state will help us understand the initiation of protein folding and the efficiency of folding process. Second, the denatured state is as equally important as the native state in determining the protein stability that is related to biological functions. In addition to the important role of the denatured state in protein folding and stability, this state has also been recognized as important for transport across membranes and protein turnover (Tompa, 2003).

More recently however, the denatured state has gained significant prominence with the recognition that many proteins are intrinsically disordered (ID) or contain ID regions under otherwise physiological conditions (Wright, 1999). This suggests that many proteins may have evolved to use the denatured state for functions previously associated with folded, native proteins. Indeed, disorder has been found to be a conserved feature (Chen et al., 2006; Romero et al., 2004; Ward et al., 2004), and its importance has already been established to processes such as catalysis (Gu et al., 2007; Kukreja et al., 2005;) and molecular recognition (Dunker et al., 2001; Iakoucheva et al., 2001; Meszaros et al., 2007), and the advantages of coupling allosteric control to the folding of ID regions has recently been developed (Hilser and Thompson, 2007).

Entire proteins or regions thereof are classified as intrinsically disordered if they lack a stable, well-defined, ordered structure as observed in natively folded proteins. The lack of ordered structure in these regions has usually been experimentally established by high temperature factors (B-factors), unresolved amino acids in X-ray crystallographic experiments, and/or spectroscopic techniques that do not detect regular structure formation (Blow et al., 1977; Ringe et al, 1986; Romero et al., 1997; 2004; Tompa, 2002; Uversky et al, 2002; Wright, 1999;). In spite of the observed disorder however, pre-formed elemental structures have been detected within several ID regions using both experimental and computational approaches (Fuxreiter et al., 2004; Hennig et al., 1999). For example, a fragment of the disordered N-terminal domain of glucocorticoid receptor can be induced to fold with the osmolyte TMAO (Kumar et al., 1999). This result is important because it indicates that although the conformational ensemble under native conditions is dominated by unstructured states, there nevertheless exists a relatively restricted conformational manifold of folded, compact structures that are important for functional interactions. As a consequence, a quantitative thermodynamic understanding of the denatured states of non-ID proteins may provide insight into the how ID proteins can be functionally and/or structurally characterized.

The current understandings of disordered regions have been largely gleaned from examining results of successful machine learning training efforts that detect recurring patterns within these regions (Dosztanyi et al., 2005; Gu et al., 2006; Jones and Ward, 2003; Linding et al., 2003; Liu and Rost, 2003; Romero et al., 1997; Yang et al., 2005). Though some disorder region predictors are successful in identifying disordered regions to help understand the identified sequences patterns and provide many useful applications, there is little

understanding regarding the thermodynamic properties of these regions. Perturbations of the denatured state ensemble have been shown to impact protein stability and folding kinetics (Mok et al., 2001; Shortle et al., 1992; Wrabl and Shortle, 1996; Wrabl and Shortle, 1999) since the free energy of stability is dependent on both the denatured and natively folded state of the protein. A better physical understanding of unfolded states of proteins may also be of particular importance due to the observation that many human diseases are associated with unfolded proteins (Ross et al., 2004) , as well as the existence of proteins that are biologically active in the intrinsically disordered state (Kukreja et al., 2005).

Although thermodynamic control of protein folding has been recognized for years, and the importance of the denatured state has drawn extensive attention recently, a detailed understanding of the thermodynamic role of denatured states in modulating protein folding remains elusive. Knowledge of protein fold specific thermodynamic determinants in the denatured state will provide new insights into deciphering the protein folding problem. Information on characteristics of thermodynamic determinants in denatured states will help understand why some proteins are misfolded while others retain their native folds. Therefore, deciphering the thermodynamic determinants that govern protein fold specificity in the denatured state will assist in developing therapeutic approaches to treat misfolded protein diseases.

In this work, the thermodynamic determinants of protein fold specificity in denatured states are characterized to understand the thermodynamic rules that relate sequence to fold as well as thermodynamic control of denatured states in protein folding.

Previous data showed that the native states of proteins share common thermodynamic properties that are independent of and even transcend structural similarities (Larson and Hilser, 2004; Wrabl et al. 2002;). This was done by developing a position-specific energetic description of each protein and determining the amino acid propensities for the different thermodynamic environment. The utility of our energetic representation was established by matching (with an 84% success rate) a protein's sequence to a one-dimensional representation of that protein's energy landscape. That result conclusively demonstrated that the organizing principles for native proteins can be represented in purely energetic terms, and that the specific thermodynamic descriptors developed in that work were sufficient to quantitatively characterize a diverse database of human protein structures.

For achieving the goals of characterizing the thermodynamic determinants in denatured states and understanding the role of denatured states in thermodynamic control of protein folding, we examine the thermodynamics of the denatured states across multiple proteins in human protein database in order to determine: 1) whether similar thermodynamic organizing principles exist as those in the native state, 2) what is the nature of these thermodynamic organizing principles 3) the relationship of this thermodynamic organizing schema with both sequence and structure, and 4) the quantitative similarity with the native state energetics.

CHAPTER 2

COREX Algorithm and the Denatured State Ensemble

Introduction

Large numbers of conformations interconvert at a fast rate in the denatured state ensemble thus making it difficult to characterize the dynamic nature of these states. Although the availability of advanced experimental techniques such as NMR and Fourier transform infrared spectroscopy (FTIR) can provide structural and dynamic information for denatured states, quantitative energetic description of the denatured states across multiple proteins has still proven elusive. To capture the thermodynamic characteristics of denatured states across a large protein database, an ensemble-based thermodynamic model, COREX, is used to generate the denatured ensemble of each protein in our *Homo sapiens* protein database containing nonhomologous proteins. COREX models the native state of a protein as an ensemble of partially unfolded conformational microstates instead of a single static state. The algorithm uses a high-resolution structure of a protein to generate a statistical thermodynamic ensemble of states by alternatively folding and unfolding a certain number of residue stretches within the sequence for all possible combinations. Denatured ensembles are generated by perturbing the full ensemble to favor unfolded states. Position-specific thermodynamics of each protein in the database were calculated and characterized. The unique features of those position-specific thermodynamics are discussed.

Results

Generation of the Ensemble of microstates by COREX

The statistical thermodynamic model COREX describes a native protein as an ensemble of states rather than as a single static structure (Hilser and Freire, 1996). A high resolution X-ray or NMR structure of protein is used as a template. The individual conformations are generated by unfolding the various regions of the protein in exhaustive combinations (Hilser and Freire, 1996; Wrabl et al., 2001). To systematically vary the folding units, the boundaries are advanced per residue in the sequence for each partition until the scheme repeats itself (Figure 2-1).

Probability of microstates:

The probability of any conformation (microstate) of a protein under equilibrium conditions is calculated by Equation 2.1:

$$P_i = \frac{K_i}{\sum_{i=1}^{N_{states}} K_i} = \frac{K_i}{Q} \qquad (2.1)$$

where $K_i = e^{(-\Delta Gi/RT)}$ is the statistical weight of each microstate and the summation in the denominator is the partition function, Q, for the system. The Gibbs free energy for each microstate, ΔG_i is calculated as:

$$\Delta G_i = \Delta H_{i,solvaiton} - T\,(\Delta S_{i,salvation} + W\Delta S_{i,conformational}) \qquad (2.2)$$

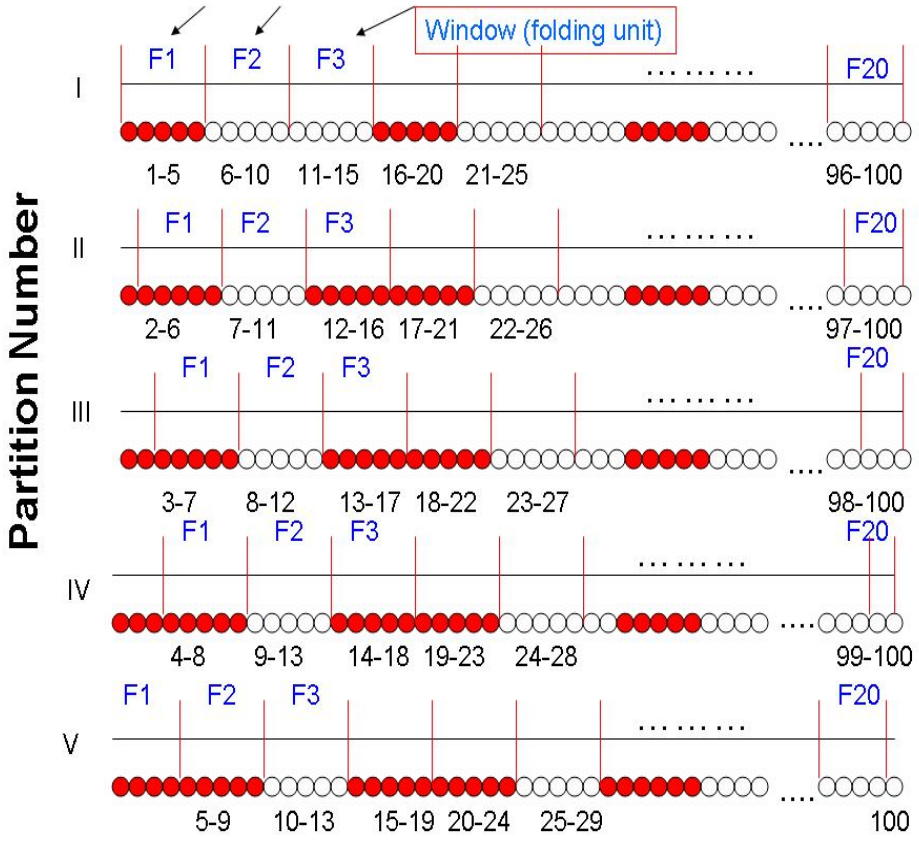


Figure 2-1: An illustration of the COREX algorithm to generate partially unfolded states

A protein of 100 amino acids is divided into 20 different sequential folding units (labeled F1-F20), each with 5 residues. Each circle represents one residue in the three-dimensional x-ray crystal structure of the native protein. The partitioning scheme is then overlaid on the high-resolution structure, the total combinations among different folding units and unfolded units generate in 2^N-2 (1028574) partially folded states, N is the number of folding units. An exhaustive enumeration of the partially unfolded states can be generated by systematically varying the folding units and sliding the unit ahead by one residue for each partition. For each partition, the procedure is repeated again (in this example 5 partitions needed until all possible partially folded conformations are generated).

The relative apolar and polar free energies of each state were calculated by accessible-surface-area based parameterization equations (Hilser and Freire, 1996; D'Aquino *et al.*, 1996; Xie and Freire, 1994):

$$\Delta G_{apolar,i}(T) = -8.44 * \Delta ASA_{apolar,i} + 0.45 * \Delta ASA_{apolar,i} * (T - 333) - T * \left(0.45 * \Delta ASA_{apolar,i} * \ln(T/385)\right) \quad (2.3)$$

$$\Delta G_{polar,i}(T) = 31.44 * \Delta ASA_{polar,i} - 0.26 * \Delta ASA_{polar,i} * (T - 333) - T * \left(-0.26 * \Delta ASA_{polar,i} * \ln(T/335)\right) \quad (2.4)$$

The conformational entropy ΔS_{conf} was determined by summarized three primary contributions: (1) $\Delta S_{bu \rightarrow ex}$ is the entropy change associated with transferring buried side chains to solvent exposure (2) $\Delta S_{ex \rightarrow u}$, the entropy change gained by a surface-exposed side-chain upon unfolding the peptide backbone; and (3) ΔS_{bb}, the backbone entropy change gained by unfolding itself (Hilser and Freire, 1996). W is an entropy weighting factor to control the contributions of completely unfolded states. COREX generated ensemble of states and the probabilities of each state is shown in (Figure 2-2) for G protein (PDB number 1KAO).Another important statistical descriptor of the equilibrium can be evaluated for each residue in the protein which is defined as the residue stability constant, $\kappa_{f,j}$. This quantity is the ratio of the summed probability of all states in the ensemble i in which a particular residue *j* is in a folded conformation ($\Sigma P_{f,j}$) to the summed probability of all states in which *j* is in an unfolded conformation ($\Sigma P_{nf,j}$):

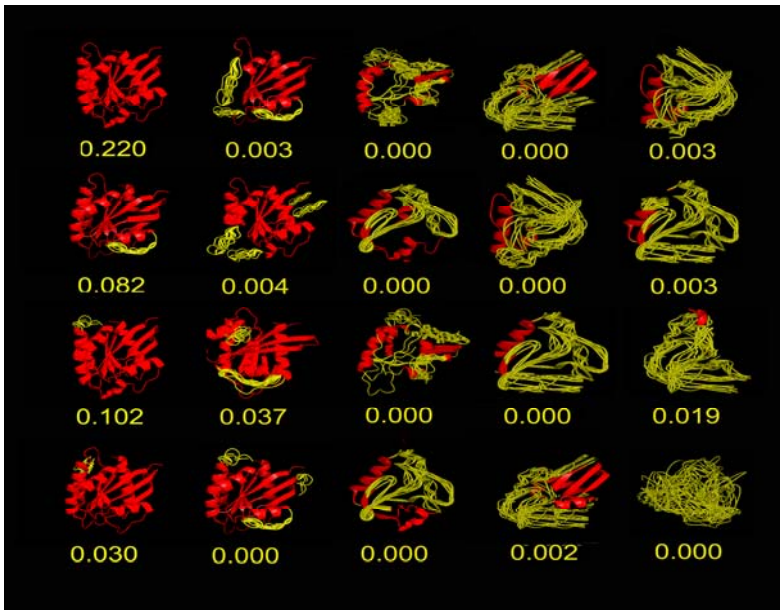


Figure 2-2: COREX generated ensemble of microstates and calculated probabilities of sample G protein (PDB: 1KAO)
Samples of partially folded microstates generated by COREX and the relative probabilities are shown. Each column represents microstates with different percentages of unfolding. Red regions in each state are treated as native-like. Yellow regions are treated are denatured-like (for schematic purposes only). Actual probabilities (numbers in yellow) of each state were shown below each state.

$$\kappa_{f,j} = \frac{\sum P_{f,j}}{\sum P_{nf,j}} \qquad (2.5)$$

From the stability constant, the position-specific free energy can be written as:

$$\Delta G_{f,j} = -RT \bullet ln\kappa_{f,j} \qquad (2.6)$$

The importance of the stability constant and its good agreement with experimental data has previously been shown with hydrogen deuterium exchange comparisons (Hilser and Freire, 1996).

Position-specific thermodynamic descriptors

Position-specific thermodynamic descriptors were calculated by taking the difference in folded and unfolded subensemble quantities.

$$[\Delta H]_{pol,j} = < \Delta H_{pol,f,j} > - < \Delta H_{pol,nf,j} > \qquad (2.7)$$

$$[\Delta H]_{apol,j} = < \Delta H_{apol,f,j} > - < \Delta H_{apol,nf,j} > \qquad (2.8)$$

$$[\Delta S]_{conf,j} = < \Delta S_{conf,f,j} > - < \Delta S_{conf,nf,j} > \qquad (2.9)$$

Quantities in folded and unfolded sub ensembles were calculated as:

$$\langle \Delta H \rangle = \sum_{i=1}^{N_{states}} P_i \bullet \Delta H_i = \sum_{i=1}^{N_{states}} \frac{K_i \bullet \Delta H_i}{Q}. \qquad (2.10)$$

$$\langle \Delta S \rangle = \sum_{i=1}^{N_{states}} P_i \bullet \Delta S_i = \sum_{i=1}^{N_{states}} \frac{K_i \bullet \Delta S_i}{Q} \qquad (2.11)$$

Generation of the Denatured State Ensemble by COREX

A database of 122 nonhomologous proteins (Table 2-1) was analyzed using the COREX algorithm. Because our previous analysis was determined under native conditions, the Boltzmann-weighted thermodynamic values reported at each position were dominated by contributions from structured states, as they have the highest probability under native conditions (Figure 2-2). To determine whether structure encoding information is also contained within any other subset of states in the full ensemble, the ensemble was systematically perturbed by increasing the stability of each state in a manner proportional to the amount of unfolded structure, as described in Materials and Methods. The net effect of such a perturbation strategy is to redistribute the ensemble to favor more unfolded states. These conditions can be referred to as denaturing because the ensemble probabilities are dominated by states where the folded regions account for less than 20% of the residues in any given state.

Table 2-1: *Homo sapiens* proteins analyzed by the COREX algorithm

PDB	Length	SCOP class	SCOP family
1A17	159	All alpha	Tetratricopeptide repeat (TPR)
1A3K	137	All beta	Galectin (animal S-lectin)
1AD6	185	All alpha	Retinoblastoma tumor suppressor
1ALY	146	All beta	TNF-like
1B56	133	All beta	Fatty acid binding protein-like
1B9O	123	Alpha and beta (a+b)	C-type lysozyme
1BD8	156	Alpha and beta (a+b)	Ankyrin repeat
1BIK	110	Small	Small Kunitz-type inhibitors
1BKF	107	Alpha and beta (a+b)	FKBP proline isomerase
1BKR	108	All alpha	Calponin-homology domain
1BR9	182	All beta	Metalloproteinases, TIMP
1BUO	121	Alpha and beta (a+b)	BTB/POZ domain
1BY2	111	Alpha and beta (a+b)	Scavenger receptor cysteine-rich
1BYQ	213	Alpha and beta (a+b)	Heat shock protein 90, N-terminal
1CBS	137	All beta	Fatty acid binding protein-like
1CDY	178	All beta	C2 set domains
1CLL	144	All alpha	Calmodulin-like
1CTQ	166	Alpha and beta (a/b)	G proteins
1CY5	92	All alpha	DEATH domain
1CZT	160	All beta	Coagulation factor C2 domain
1D7P	159	All beta	Coagulation factor C2 domain
1DG6	149	All beta	TNF-like
1DV8	128	Alpha and beta (a+b)	C-type lectin domain
1E21	119	Alpha and beta (a+b)	Ribonuclease A-like
1E87	117	Alpha and beta (a+b)	C-type lectin domain
1EAZ	103	All beta	Pleckstrin-homology domain
1ESR	75	Alpha and beta (a+b)	Interleukin 8-like chemokines

Table 2-1, cont.: *Homo sapiens* proteins analyzed by the COREX algorithm

PDB	Length	SCOP class	SCOP family
1FAO	100	All beta	Pleckstrin-homology domain
1FIL	139	Alpha and beta (a+b)	Profilin (actin-binding protein)
1FL0	163	All beta	Myf domain
1FNA	89	All beta	Fibronectin type III
1FNL	172	All beta	I set domains
1FP5	208	All beta	C1 set domains
1FW1	208	All alpha	Glutathione S-transferases
1G1T	157	Alpha and beta (a+b)	C-type lectin domain
1G96	111	Alpha and beta (a+b)	Cystatins
1GEN	200	All beta	Hemopexin-like domain
1GGZ	144	All alpha	Calmodulin-like
1GH2	107	Alpha and beta (a/b)	Thioltransferase
1GLO	217	Alpha and beta protein	Cathespin
1GNU	117	Alpha and beta (a+b)	GABARAP-like
1GP0	133	All beta	Mannose 6-phosphate receptor
1GQV	135	Alpha and beta (a+b)	Ribonuclease A-like
1GR3	132	All beta	TNF-like
1GSM	202	All beta	I set domains
1H6H	143	Alpha and beta (a+b)	PX domain
1HDO	205	Alpha and beta (a/b)	Oxidoreductases
1HDR	236	Alpha and beta (a/b)	Tyrosine oxidoreductases
1HMT	131	All beta	Fatty acid binding protein-like
1HNA	217	All alpha	Glutathione S-transferases
1HUP	141	Alpha and beta (a+b)	C-type lectin domain
1HZI	129	All alpha	Short-chain cytokines
1I1N	223	Alpha and beta (a/b)	Protein-L-isoaspartyl
1I27	69	All alpha	C-terminal rap74 subunit
1I2T	61	All alpha	PABC (PABP) domain

Table 2-1, cont.: *Homo sapiens* proteins analyzed by the COREX algorithm

PDB	Length	SCOP class	SCOP family
1I4M	108	Alpha and beta (a+b)	Prion-like
1I71	83	Small	Kringle modules
1I76	163	Alpha and beta (a+b)	Matrix metalloproteases
1IAM	185	All beta	I set domains
1IAP	190	All alpha	Regulator of G-protein signaling
1IFR	110	All beta	Lamin A/C globular tail domain
1IHK	157	All beta	Fibroblast growth factors (FGF)
1IJR	103	Alpha and beta (a+b)	SH2 domain
1IJT	128	All beta	Fibroblast growth factors (FGF)
1IKT	115	Alpha and beta (a+b)	Sterol carrier protein, SCP
1IMJ	208	Alpha and beta (a/b)	Ccg1/TafII250-interacting factor B
1J74	139	Alpha and beta (a+b)	Ubiquitin conjugating enzyme
1JHJ	160	All beta	Anaphase-promoting complex
1JK3	158	Alpha and beta (a+b)	Matrix metalloproteases
1JSF	130	Alpha and beta (a+b)	C-type lysozyme
1JSG	111	All beta	Oncogene products
1JWF	139	All alpha	VHS domain
1JWO	97	Alpha and beta (a+b)	SH2 domain
1K04	142	All alpha	Focal adhesion kinase
1K1B	228	Alpha and beta (a+b)	Ankyrin repeat
1K59	122	Alpha and beta (a+b)	Ribonuclease A-like
1KAO	167	Alpha and beta (a/b)	G proteins
1KCQ	103	Alpha and beta (a+b)	Gelsolin-like
1KEX	155	All Beta	B1 Domain of Neuropilin - 1
1KMV	186	Alpha and beta (a/b)	Dihydrofolate reductases
1KPF	111	Alpha and beta (a+b)	HIT protein kinase-interacting

Table 2-1, cont.: *Homo sapiens* proteins analyzed by the COREX algorithm

PDB	Length	SCOP class	SCOP family
1KTH	58	Small	BPTI-like
1L8J	170	Alpha and beta (a+b)	MHC antigen-recognition domain
1L9L	74	All alpha	NKL-like
1LCL	141	All beta	Galectin (animal S-lectin)
1LDS	96	All beta	C1 set domains
1LN1	203	Alpha and beta (a+b)	STAR domain
1LPJ	133	Alpha and beta	Human Crbp IV
1LSL	113	All beta	Thrombospondin-1
1M7B	179	Alpha and beta (a/b)	G proteins
1M9Z	104	Small	Extracellular domain, cell surface
1MFM	153	All beta	Cu,Zn superoxide dismutase-like
1MH1	180	Alpha and beta (a/b)	G proteins
1MH9	194	Alpha and beta	Deoxyribonucleotidase
1MJ4	79	Alpha and beta	Cytocrome B5 Sulfite Oxidase
1MWP	96	Alpha and beta (a+b)	A heparin-binding domain
1N6H	167	Alpha and beta	Rab5A
1NKR	195	All beta	I set domains
1PBK	116	Alpha and beta (a+b)	FKBP immunophilin/proline
1PBV	195	All alpha	Sec7 domain
1PHT	83	All beta	SH3-domain

Table 2-1, cont.: *Homo sapiens* proteins analyzed by the COREX algorithm

PDB	Length	SCOP class	SCOP family
1POD	124	All alpha	Vertebrate phospholipase A2
1QB0	177	Alpha and beta (a/b)	Cell cycle control phosphatase
1QDD	144	Alpha and beta (a+b)	C-type lectin domain
1QKT	248	All alpha	Nuclear receptor ligand-binding
1QUU	245	All alpha	Spectrin repeat
1RBP	174	All beta	Retinol binding protein-like
1RLW	124	All beta	PLC-like (P variant)
1SRA	151	All alpha	Osteonectin
1TEN	89	All beta	Fibronectin type III
1TN3	137	Alpha and beta (a+b)	C-type lectin domain
1ZON	181	Alpha and beta (a/b)	Integrin A (or I) domain
1ZXQ	192	All beta	C2 set domains
2ABL	162	Alpha and beta (a+b)	SH2 domain
2CPL	164	All beta	Cyclophilin
2FCB	173	All beta	I set domains
2FHA	172	All alpha	Ferritin
2ILK	155	All alpha	Interferons/interleukin-10 (IL-10)
2PSR	96	All alpha	S100 proteins
2TGI	112	Small	Transforming growth factor
3FIB	249	Alpha and beta (a+b)	Fibrinogen C-terminal domain
3IL8	68	Alpha and beta (a+b)	Interleukin 8-like chemokines
5PNT	157	Alpha and beta (a/b)	Phosphotyrosine phosphatase

Position-specific Energetics in the Denatured State Ensemble is the Unique Property of Ensemble

As indicated previously (Larson and Hilser, 2004), position-specific energetics in the native ensemble are the ensemble averaged thermodynamic reporters and do not represent the energetic contribution of an amino acid to the stability of the molecule. In this chapter, COREX generates denatured ensembles by perturbing the ensemble to favor unfolded states. To investigate whether position-specific energetics in the denatured ensemble still represent the ensemble averaged thermodynamic property after perturbation, correlation analysis between position-specific descriptors and the contribution of the amino acid at the same position to the accessible surface area (ASA) in protein is shown in Figure 2-3. The absence of a correlation between the position-specific descriptors in denatured ensemble and the energetic contributions for all proteins within the database reveals (Table 2-2) that the position-specific energetics in denatured ensemble are a property of the ensemble as a whole and can characterize protein fold in an effective way independent from the amino acid sequence at the position.

Stability Constants in Denatured Ensemble – Agreement between COREX Calculation and Experimental Data

Previous work has demonstrated that denatured proteins contain residual structures (Dill and Shortle, 1991; Dobson et al., 1992; Shortle et al., 2001). Those relatively stable, persistent individual residual structures are hypothesized to play a significant role in the energetics of protein folding. Most residual structures were detected by NMR studies which experimentally support the important roles of denatured states involved in protein folding.

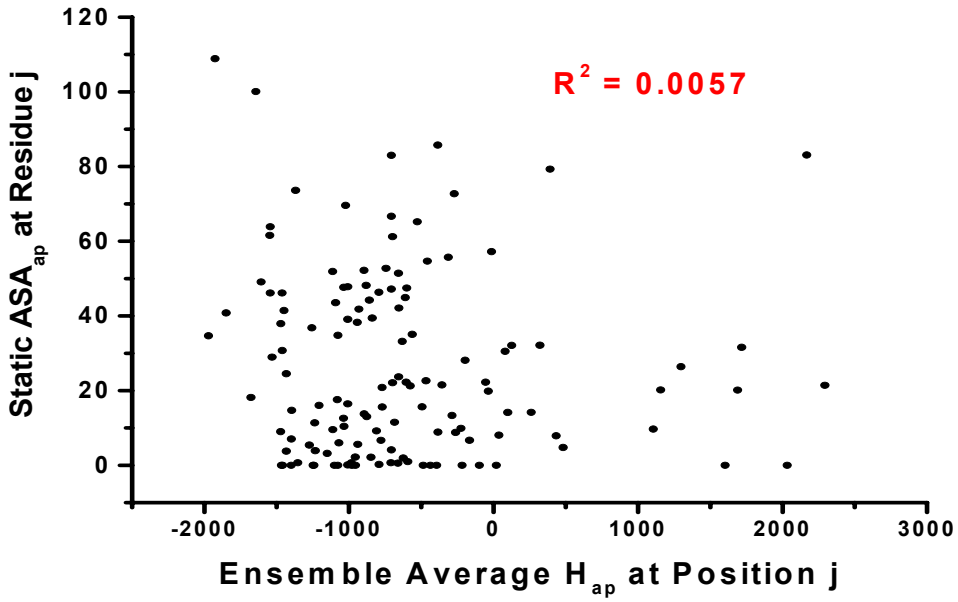


Figure 2-3: Residue-specific accessible surface area (ASA) vs position-specific thermodynamics in human lysozyme protein (PDB: 1JSF)

Each point in the plot represents one residue in protein. The ordinate is the ASA for each residue of the protein which represents the residue-specific energetic contribution to the thermodynamics of the protein. The abscissa is the COREX algorithm calculated thermodynamic descriptor (Apolar enthalpy). The correlation coefficient (R^2) for ASA vs position-specific thermodynamic descriptors is 0.0057, suggesting no correlation. Correlation statistics for the entire database are summarized in Table 2-2.

Table 2-2: Correlation (R^2) table of accessible surface area contributors versus denatured ensemble-averaged thermodynamic descriptors

R^2	ΔG	$\ln k_f$	ΔH_{apol}	ΔH_{pol}	$T\Delta S_{conf}$
ASA_{apol}	0.015826	0.015826	0.000246	0.018176	0.000169
ASA_{pol}	0.001971	0.001971	0.002843	0.005849	0.001363
ASA_{sc}	0.017912	0.017912	0.003285	0.012893	0.000137

As discussed previously, the stability constants calculated by COREX represent regional differences in stability within the protein at the resolution of each residue position. For the native ensemble, previous experiments have demonstrated good agreement between COREX calculated stability constants and experimental protection factors (Hilser & Freire, 1996). To demonstrate that the COREX calculated stability constants for the denatured ensemble still serve to represent true characteristics of the denatured ensemble, COREX calculated stability constants and experimental data are compared. As indicated by Figure 2-4, the relatively stable region (residues 83-86) of staphylococcal nuclease reported experimentally under denatured condition is also captured by COREX (Alexandrescu and Shortle, 1994; Wang and Shortle, 1995). Furthermore, good agreement was also found between COREX calculations and cold denaturation experiments (Babu et al., 2004).

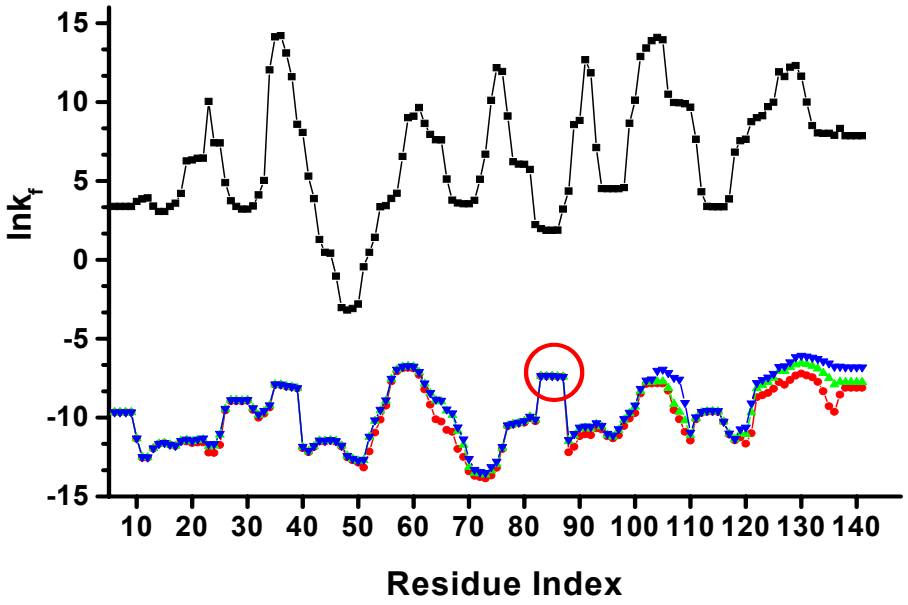


Figure 2-4: COREX calculated stability constants of staphylococcal nuclease (PDB: 1STN)

Stability constants of staphylococcal nuclease under native and denatured conditions were calculated. The X-axis indicates the residue numbers and Y-axis is the stability constants under native condition (Top black line) and denatured condition (Bottom colored lines). Under denatured conditions, different numbers of allowable folded segments (1, 2, and 3) were used to calculate the stability constants (Blue line represents the result from a maximum of one folded units, green line represents the result from a maximum two folded, units and red line represents the result from maximum three folded units). Resides (83-86) in the red circle are relatively stable residues found under denatured conditions identified by COREX as well as reported by experimental data.

Materials and Methods

Nonredundant database of *Homo sapiens* proteins

A dataset of nonredundant *Homo sapiens* proteins (Table 2.1) with protein structures in the Protein Data Bank (PDB) was used for this analysis. This dataset contains 122 proteins with a total of 17802 residues. The selection criteria for this dataset are 1) Proteins containing 50-250 amino acids with a maximum of 50% sequence identity within the set. 2) Only X-ray structures having a resolution better than 2.5 Å were selected. These criteria were set with consideration for computational demands and structure quality.

Computational Details of COREX algorithm

The COREX algorithm is a statistical thermodynamic model in which a native protein is depicted as an ensemble of states rather than as a single static structure. The thermodynamic energetics of each 122 proteins in the *Homo Sapiens dataset* was calculated using COREX algorithm. Monte Carlo sampling was used to generate the protein ensembles with the following parameters: (1) 50,000 state/partition for proteins larger than 80 residues, all states were selected for proteins less than 80 residues. For the native ensemble, entropy weighting factor W=0.5; for denatured ensemble, W=1.5. Simulation temperature is set at 25^{o}C and the window size for local unfolding is 5 residues with a

minimum window size set at 4 residues. For generating denatured ensemble, the maximum number of folded windows in each partition is set at 2. Because the size of the proteins in the analyzed database ranged form 50 to 250 residues, the maximum number of residues folded range from 10 to 50.

CHAPTER 3

Energetic information in Native and Denatured State Ensembles

Introduction

As mentioned, both native and denatured states are involved in protein folding and contribute to protein stability. Consequently, characterization of the energetic information in both native and denatured states becomes important for a complete understanding of the protein folding process as well as the thermodynamic control of protein stability and functions. Although much attention has been focused on the energetic landscape of the native states, investigating the native state alone will be insufficient for deciphering the protein folding problem (Bowler, 2007). Quantitative thermodynamic description of the denatured state will be critical for a detailed understanding the thermodynamic control of protein folding, stability, and function.

We previously showed that the energetic information derived from the native state ensemble can be used to determine the regional differences in stability for a database of multiple proteins (Larson and Hilser, 2004). With this distinction, proteins can be categorized in terms of the thermodynamics of the energy landscape of the native states, rather than in terms of the secondary or tertiary structure of the folded native conformation (Larson and Hilser, 2004; Wrabl et al., 2002). In the previous chapter, denatured ensembles were generated using COREX for a database of *Homo sapiens* proteins and position-specific thermodynamics were calculated. In this chapter, the energetic

landscape of proteins under denaturing conditions was examined and the energetic information from native and denatured ensembles was compared. The goal of this chapter is to describe the common energetic properties in denatured states across multiple proteins in the human protein database and quantitatively compare the energetic information between the native state ensemble and the denatured state ensemble.

Results

Defining Thermodynamic Environments

Previous work has shown that three thermodynamic environments (low stability, medium stability and high stability) (Wrabl et al., 2001) or eight thermodynamic environments (Larson and Hilser 2004) could be defined based on ensemble-based energetics. With a combination of eight thermodynamic environments, 90% of the energetic variability within a *Homo sapiens* structural database can be captured (Larson and Hilser, 2004). In a similar way, thermodynamic environments for both native ensembles and denatured ensembles were defined in this chapter.

Native ensembles and denatured ensembles were generated by COREX as shown in previous chapter. Position-specific thermodynamic descriptors were calculated for each protein in the database. Thermodynamic environments within proteins were defined through the use of the partitioning around medoids (PAM) clustering applied to the position-specific thermodynamic descriptors. Four descriptors used in defining thermodynamic environments were: stability

constants (lnk_f), apolar enthalpy (H_{apol}), polar enthalpy (H_{pol}) and conformational entropy (TS_{conf}) Figure 3-1 is the outline of how eight thermodynamic environments were defined.

Characterization of Thermodynamic Environments in Native and Denatured State Ensembles

To characterize the energetic information in each thermodynamic environment, mean energetic values were calculated for eight environments in the native ensemble (Table 3-1) and the denatured ensemble (Table 3-2). The normalized mean energetic values for each thermodynamic environment are shown in Figure 3-2 for native ensemble and Figure 3-3 for denatured ensemble. As shown in Figure 3-2 and Figure 3-3, thermodynamic environments (TE_N and TE_D) represent a systematic partitioning of the space, where the different environments corresponds to a different stabilities, as well as different thermodynamic mechanisms for achieving that stability. Comparisons between normalized energetic values of eight environments in native ensemble and denatured ensemble show significant difference.

To compare with a conventional view based on secondary structure, one of proteins in the database (1KAO) was structurally colored coded with eight thermodynamic environments in native ensemble and eight thermodynamic environments in the denatured ensemble as shown in Figure 3-4. Eight thermodynamic environments represented by eight different colors. The primary sequence has been colored according to the relative thermodynamic environment segments.

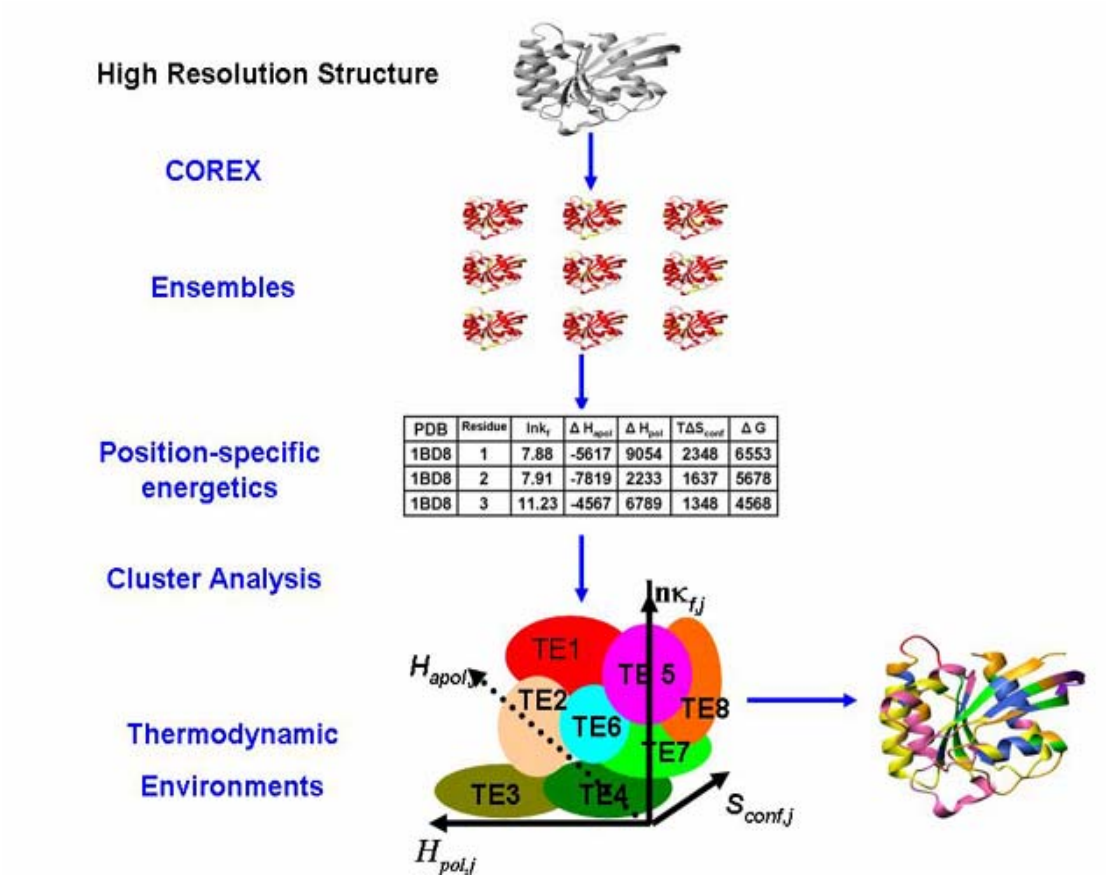


Figure 3-1: Schematic illustration of defining eight thermodynamic environments for G protein (PDB ID: 1KAO)
COREX algorithm was used to generate ensembles based on crystal structure. Position-specific thermodynamic descriptors were calculated. The whole database of total 17802 residues with position-specific thermodynamic descriptors was clustered based on four dimensions (lnk_f, H_{apol}, H_{pol} and TS_{conf}) Eight clusters with different colors represent eight different thermodynamic environments based on four dimensional clustering. The whole structure of G protein was color coded based on eight thermodynamic environments.

Table 3-1: Mean energetic properties of eight thermodynamic environments in native ensemble

TE Native	ΔG (cal/mol)	ΔH_{pol} (cal/mol)	ΔH_{apol} (cal/mol)	ΔTS_{conf} (cal/mol)
1	-3527	-6083	4859	-3124
2	-4938	-8649	6292	-4101
3	-8548	-12388	8756	-4356
4	-9822	-15059	9078	-5225
5	-12425	-16275	14167	-5563
6	-10527	-12487	14045	-4573
7	-7473	-9443	10902	-4065
8	-6385	-11496	6494	-4885

Table 3-2: Mean energetic properties of eight thermodynamic environments in denatured ensemble

TE Denatured	ΔG (cal/mol)	ΔH_{apol} (cal/mol)	ΔH_{pol} (cal/mol)	ΔTS_{conf} (cal/mol)
1	7500	-1083	-915	-8323
2	9082	-996	44	-9312
3	8427	-1115	1085	-7704
4	7611	-303	-2168	-9790
5	9100	662	-400	-10327
6	9859	-1028	-1075	-10902
7	10178	-1522	2426	-8363
8	10696	-1538	1094	-9939

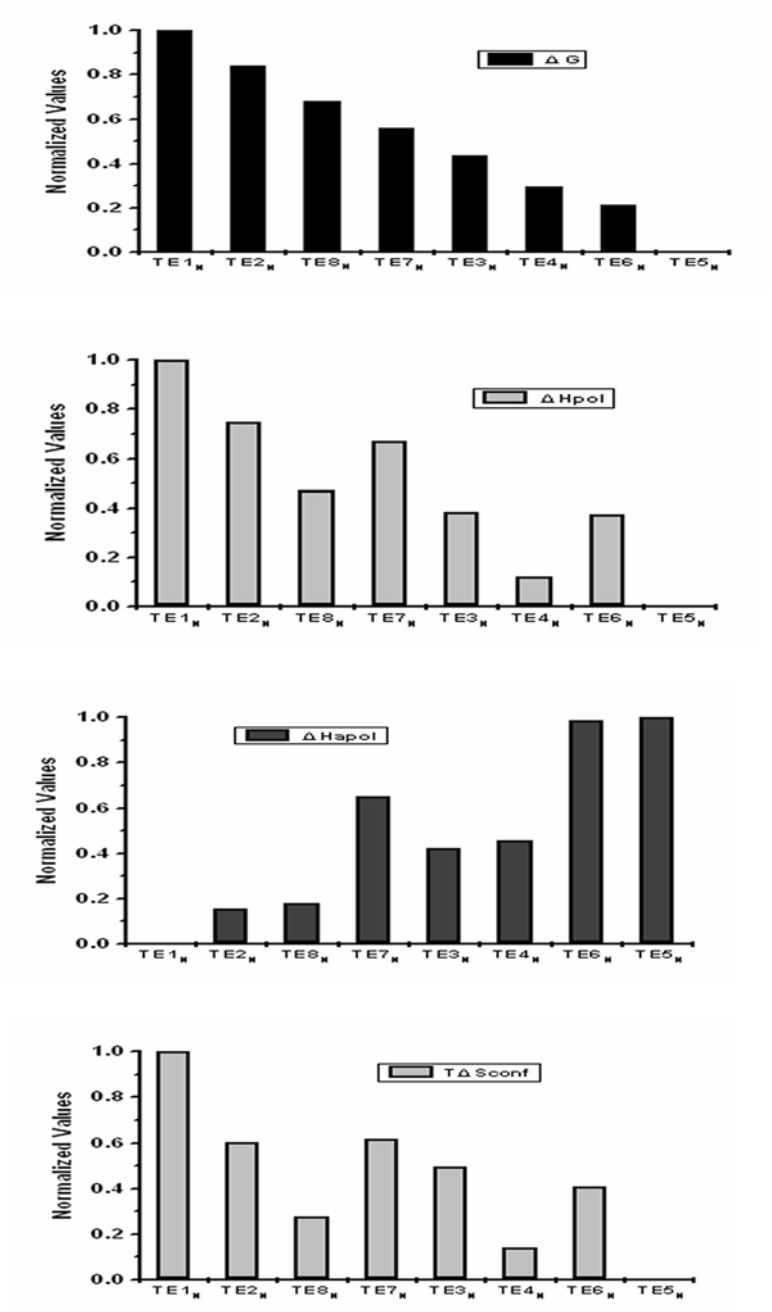


Figure 3-2: Comparison of the eight thermodynamic environments in native state ensemble

Each thermodynamic environment is defined based on the clustering of these four thermodynamic descriptors. Plotted are the mean values for the four thermodynamic descriptors with in each cluster: free energy ΔG, apolar enthalpy ΔH_{apol}, polar enthalpy ΔH_{pol} and conformational entropy $T\Delta S_{conf.}$ The X-axis is the thermodynamic environments (TE_D) listed in order of increasing stability. The Y-axis is the normalized mean values of the corresponding thermodynamic descriptors.

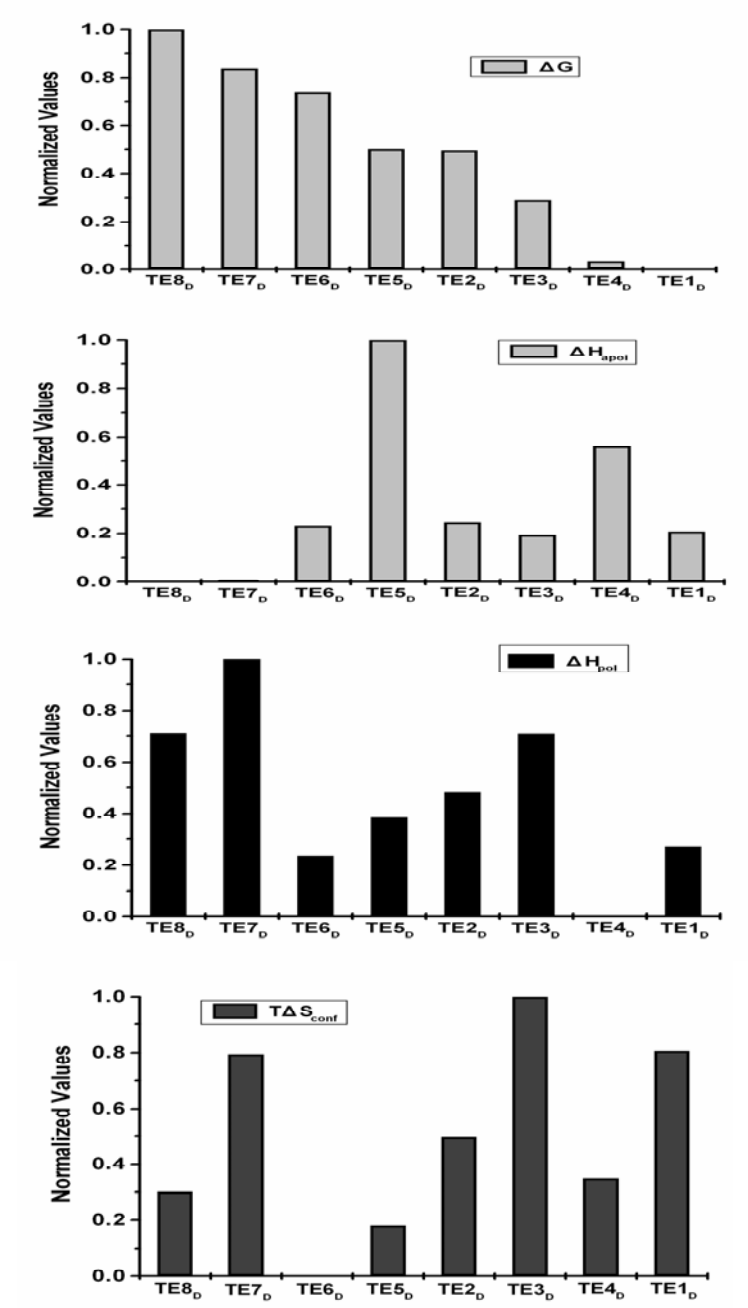


Figure 3-3: Comparison of the eight thermodynamic environments in denatured state ensemble

Each thermodynamic environment is defined based on the clustering of these four thermodynamic descriptors. Plotted are the mean values for the four thermodynamic descriptors with in each cluster: free energy ΔG, polar enthalpy ΔH_{pol}, apolar enthalpy ΔH_{apol} and conformational entropy $T\Delta S_{conf.}$ The X-axis is the thermodynamic environments (TE_D) listed in order of increasing stability. The Y-axis is the normalized mean values of the corresponding thermodynamic descriptors.

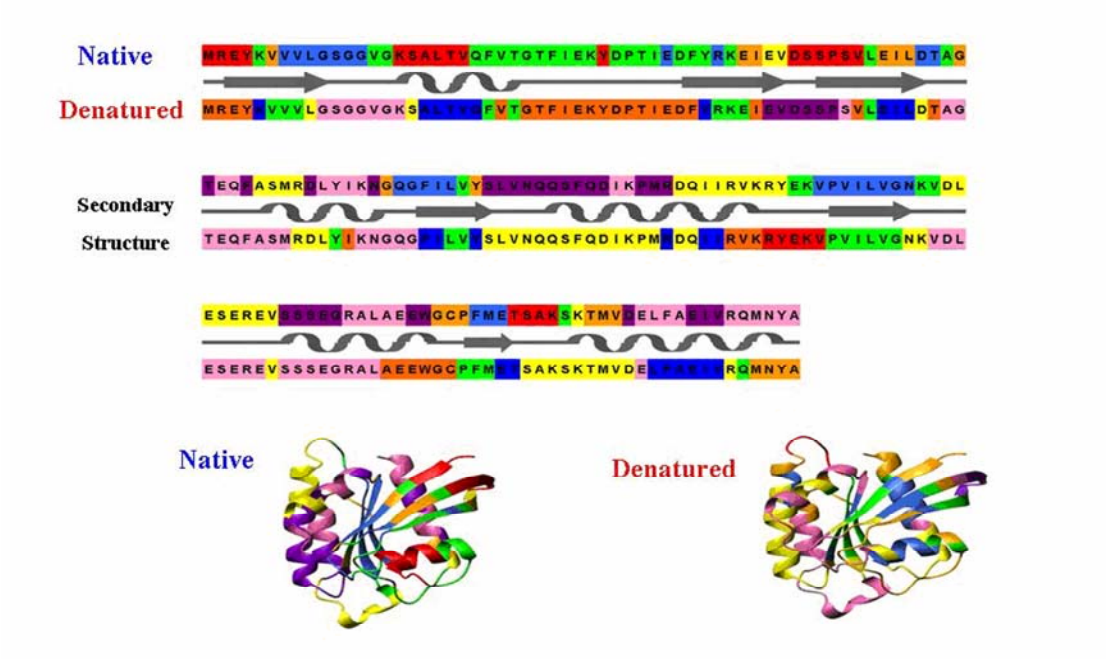


Figure 3-4: Thermodynamic environments characterization for the GTP binding protein (PDB: 1KAO) in native and denatured state ensembles
Eight thermodynamic environments represented by eight different colors. The primary sequence has been colored according to the relative thermodynamic environment segments. Secondary structures are shown for comparison with thermodynamic environment segments. High-resolution structures have been mapped with native ensemble-based energetics and denatured ensemble-based energetics to show the differences.

Stability Constant and Energetic Correlation between Native and Denatured Ensembles

As previously established protein stability is defined by both native and denatured states. One of the most important parameters calculated by COREX is the residue stability constant as shown in the previous chapter. The COREX calculated stability constant has good agreement with experimentally verifiable hydrogen exchange protection factors (Hilser et al. 1998) and it can provides a meaningful characterization of regional stability within the protein at the level of each residue position (Larson et al. 2004). Quantitative comparison of the position specific stability constant under native and denatured conditions provides a better understanding for protein stability. As shown in Figure 3-5, stability constants for G protein (1KAO) were calculated under native and denatured conditions. It is clear that regions of high stability under native conditions often correspond to regions of low stability under denaturing conditions and vice versa. However, correlation analysis of all residues stability constants within whole database under native and denatured conditions shows that there is no correlation between stability constants under native and denatured conditions (Figure 3-6). No correlation between native stability constants and denatured stability constants suggests that native ensembles and denatured ensembles contribute differently to protein stability. In fact, all four position-specific thermodynamic variables (i.e., the free energy of unfolding, [ΔG], entropic [$T\Delta S$], and enthalpic contribution from apolar [ΔH_{ap}] and polar [ΔH_{p}] residues) show no correlation between native and denatured ensembles (Table 3-3).

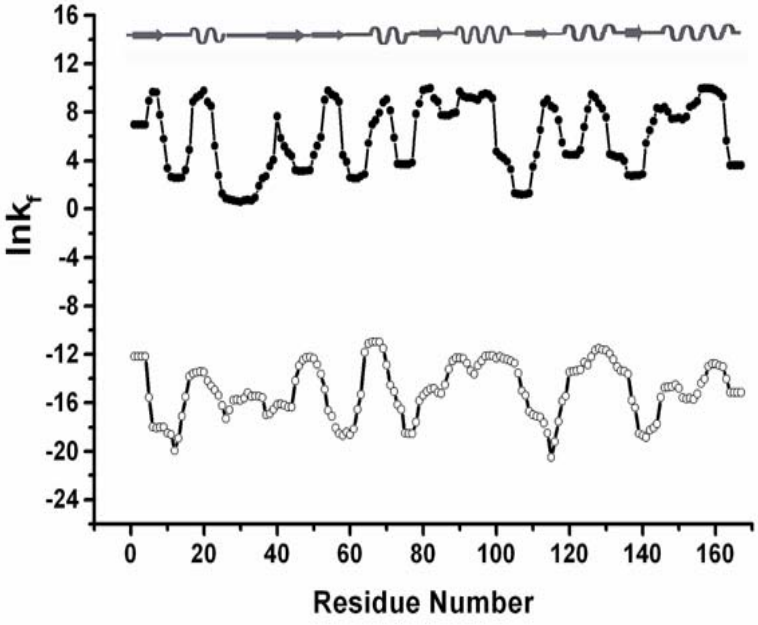


Figure 3-5: Stability constants for the GTP binding protein (PDB: 1KAO) under native and denatured conditions

Stability constants under native conditions are shown as close circles; stability constants under denatured conditions are shown as open circles. Secondary structures are shown on top. The X-axis is the residue number. The Y-axis is stability constants (lnk_f).

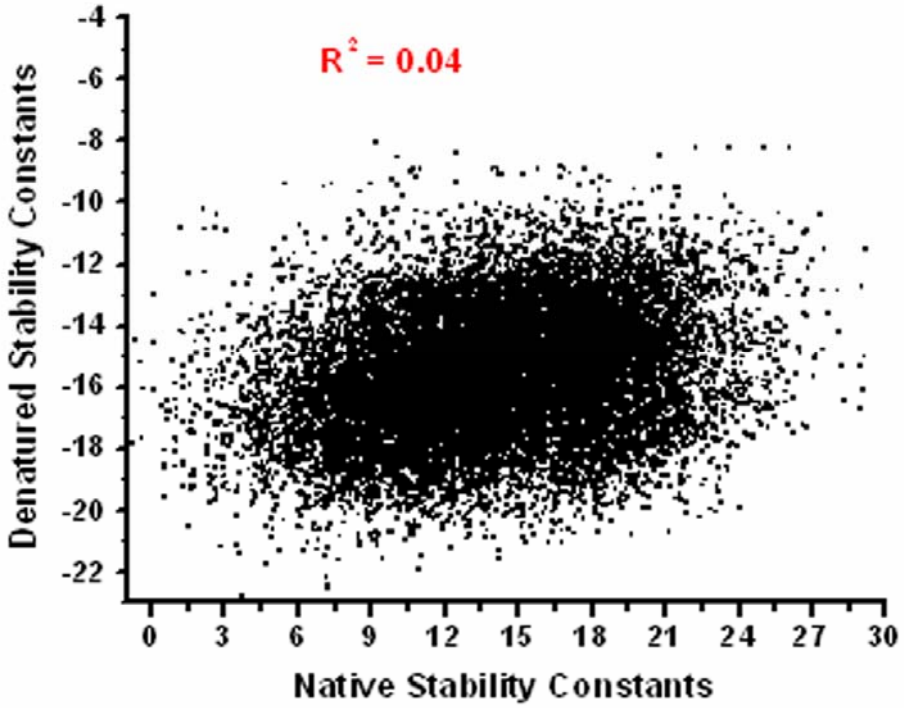


Figure 3-6: Calculated Stability Constants of all residues in the database using native and denatured ensembles show no correlation
Each point of the scatter plot is a residue position for each protein in a nonredundant *Homo sapiens* database. The ordinate is stability constants ($\ln k_f$) calculated from denatured ensembles using COREX. The abscissa is the stability constants ($\ln k_f$) calculated from native ensembles. Denatured ensembles contain largely unfolded proteins with up to 20% residual native structure whereas native ensembles consist mainly of natively folded regions in different combinations. The correlation coefficient (R^2) between the observed $\ln k_f$ values for each ensemble is 0.04, indicating no correlation.

Table 3-3: Calculated Energetics Using Native and Denatured Ensembles Show no Correlation

Energetics	$\langle \Delta G \rangle$	$\langle \ln K_f \rangle$	$\langle \Delta H_{ap} \rangle$	$\langle \Delta H_{pol} \rangle$	$\langle T\Delta S_{conf} \rangle$
R^2	0.0405	0.0405	0.0479	0.0385	0.0772

Discussion

Although there are some reports that characterize the energetics of denatured proteins (Bowler, 2007, Carra and Privalov, 1995, Godbole et al, 2000), identifying common thermodynamic properties of denatured states across multiple proteins is very limited. In this chapter, position-specific thermodynamic descriptors for each protein in the database of non-homologous *Homo sapiens* protein structures were calculated under native and denatured conditions. Energetic information under native and denatured conditions was analyzed to identify the charactering thermodynamic environments and the contributing factors to the observed relative stabilities in different states. The Boltzmann-weighted energetic information determined under native conditions differs considerably from those values determined under denaturing conditions. Quantitative comparison of energetic information under native and denatured conditions shows no correlation. The significance of this result is two-fold. First, it shows that the calculations on the native and denatured ensembles are monitoring different physical properties. Whereas under denaturing conditions the energetics are largely reporting on the stability of isolated pieces of the native structure in the absence of the stabilization effects of neighboring segments, the energetics under native conditions are reporting on the stability of each region in the context of those stabilizing interactions of neighboring segments. Second, the results indicate that there is no correlation between the stability of the individual pieces of structure that make up a protein and the stability of that region in the final fold. For instance, the loop at position 105 to 111 in G protein (PDBID: 1KAO) is one of the most stable elements of structure under denaturing

conditions, yet under native conditions it is among the least stable regions (Figure 3-5). This is because most of the stability of this region stems from local interactions; the folding of the remainder of the molecule adds comparatively little to the stability of this region. Conversely, many positions involved in the beta sheet (e.g., residue 55-57, 78-81) have relatively low stability in the denatured ensemble, owing to the dearth of short-range stabilizing interactions, but acquire significant stability in the native ensemble.

Materials and Methods

Clustering analysis algorithm

Partitioning Around Medoids (PAM) clustering method was used to cluster all 17802 residues in the *Homo sapiens* dataset based on four position-specific thermodynamic descriptors (ΔG, ΔH_{apol}, ΔH_{pol} and $T\Delta S_{conf}$) to identify 2, 4, 6, 8, 10, 12, 14, 16, and 18 medoids. Thermodynamic environments are labeled according to clustering medoids numbers. Manhattan distance was used to measure dissimilarity between medoids. The clustering analyses were performed using the S-Plus 6.0 professional software.

Stability constants calculation

Residue stability constant is calculated as:

$$\kappa_{f,j} = \frac{\sum P_{f,j}}{\sum P_{nf,j}} \qquad (3.1)$$

Where $P_{f,j}$ and $P_{nf,j,}$ are the probabilities of all states in the ensemble in which residue j is either folded or unfolded, respectively. Under equilibrium conditions, the probability of any given conformational microstate, *i*, in the ensemble is given by

$$P_i = \frac{K_i}{\sum_{i=1}^{N_{states}} K_i} = \frac{K_i}{Q} \qquad (3.2)$$

where $K_i = e^{(-\Delta Gi/RT)}$ is the statistical weight of each microstate and the summation in the denominator is the partition function, Q, for the system. The Gibbs free energy for each microstate, ΔG_i is calculated as:

$$\Delta G_i = \Delta H_{i,\ solvation} - T(\Delta S_{i,\ solvation} + W\Delta S_{i,\ conformational}) \qquad (3.3)$$

The relative apolar and polar free energies of each state was calculated by accessible-surface-area based parameterization equations in Chapter 2.

CHAPTER 4

Investigation of Fold Information in the Native and the Denatured State Ensembles

Introduction

According to Anfinsten's thermodynamic hypothesis, the information defining the final fold is contained in the primary sequence and is thermodynamic in nature (Anfinsen, 1973). Analysis of the thermodynamic determinants of protein fold specificity will provide a better understanding of the energetics that drive protein folding. Previous work has shown that a database of proteins can be represented with thermodynamic building blocks, and the thermodynamic environments can be used to match the sequences with their native folds successfully (Wrabl et al. 2002). Also, as shown previously, eight thermodynamic environments derived from native ensembles captured 90% of the energetic variability across the *Homo sapiens* protein database (Larson et al. 2004).

Thermodynamic environments under native and denatured conditions were defined and characterized in Chapter 3. In this chapter, the native ensemble thermodynamic environments (TE_N) and denatured ensemble thermodynamic environments (TE_D) are used to determine the fold information content through fold recognition experiments. The strategy is to establish amino acid propensity scales for different energetic environments (similar to the propensity scales for secondary structural environments) and to determine the

generality of these preferences by successfully matching sequences to their respective folds (defined in energetic terms). Also, the full ensembles under native and denatured conditions were partitioned into different sub-ensembles to investigate the fold contributions of sub-ensembles.

Results

Fold Recognition Experiments based on Energetic Information of the Native State Ensemble

Figure 4-1 shows a schematic outline of the energetic-based fold recognition experiments. Thermodynamic environments within proteins were detected through the use of clustering methods applied to the position-specific thermodynamic descriptors. This was followed by an indirect determination of the information content through fold recognition experiments. Fold recognition success was defined as the case where the target sequence scored higher than 99% of the decoy library. The decoy library utilized here contained 431 sequences and therefore the target sequence had to score amongst the top 4 sequences. The full ensemble under native conditions is partitioned into five sub ensembles with different percentages of folding (0-20%, 20-40%, 40-60%, 60-80%, 80-100%). Figure 4-2 shows fold recognition results obtained by threading 20 amino acids into different thermodynamic environments. Fold recognition success was defined as the case where the target sequence scored higher than 99% of the decoy library. The decoy library utilized here contained 431 sequences (Larson and Hilser, 2004) and therefore the target sequence had to

score amongst the top 4 sequences. As Figure 4-2 indicates, using the 8 thermodynamic environments in the database of native ensemble energetics (TE_N), 83.6% of the folds were successfully recognized by their sequence. To confirm that the residue specific information at each position originates from native like states, thermodynamic environments derived by clustering only the sub-ensembles containing 80-100% folded structure showed no depreciable effect. It should be noted that almost no fold recognition success was achieved for the control calculations wherein the sub-ensembles containing only 60-80%, 40-60%, or 20-40% of the native fold were used. This important finding indicates that the energy landscape of intermediately folded states (i.e. states with between 20-80% of the residues folded) do not determine the fold that a particular sequence will adopt.

Since the propensities of 20 amino acids in different thermodynamic environments were used in fold recognition experiments, it would be interesting to know the distributions of amino acids in each thermodynamic environment. Figure 4-3 shows the hierarchical clustering of 20 amino acids in eight thermodynamic environments of native ensemble. As shown previously (Larson et al., 2004), the contribution of each amino acid was not correlated to the thermodynamics of the environment to which it belongs. Consequently, the hierarchical grouping did not completely reflect a traditional chemical property selection mechanism such as hydrophobicity. Within the identified six groups, for example, aromatic amino acids (Trp, Phe,Tyr) make up one group while proline occupies a different group, similar to traditional chemical property selection.

However, it is also interesting to see that chemically and structural similar amino acids, for example, lysine and arginine were in different groups and showed different thermodynamic environment propensities. In contrast to the traditional way of grouping of 20 amino acids based on chemical properties, grouping amino acids based on propensities of amino acids in thermodynamic environments enable the description of protein folds in thermodynamic terms. Figure 4-2 shows eight thermodynamic environments, obtained using the full ensemble under native condition, reaches 83% success in fold recognition experiments. Further increases in the number of thermodynamic environments didn't improve the fold recognition success. To determine the energetic information based on hierarchical clustering, fold recognition experiments were carried out based on grouping information of 20 amino acids within eight thermodynamic environments. Figure 4-4 shows the fold recognition success as a function of amino acids clustering (Larson and Hilser, 2004). From Figure 4-4 we can see that fold recognition curve saturates at six clusters of amino acids with ~80% success and six amino acids clusters are enough to capture the majority of energetic information in our database (Larson and Hilser, 2004).

Fold Recognition Experiments based on Energetic Information in the Denatured State Ensemble

We previously showed that proteins can be categorized in terms of the thermodynamics of the energy landscape of the native protein, rather than in terms of the structural attributes of the folded native conformation. We demonstrated the utility of such a categorization scheme by successfully

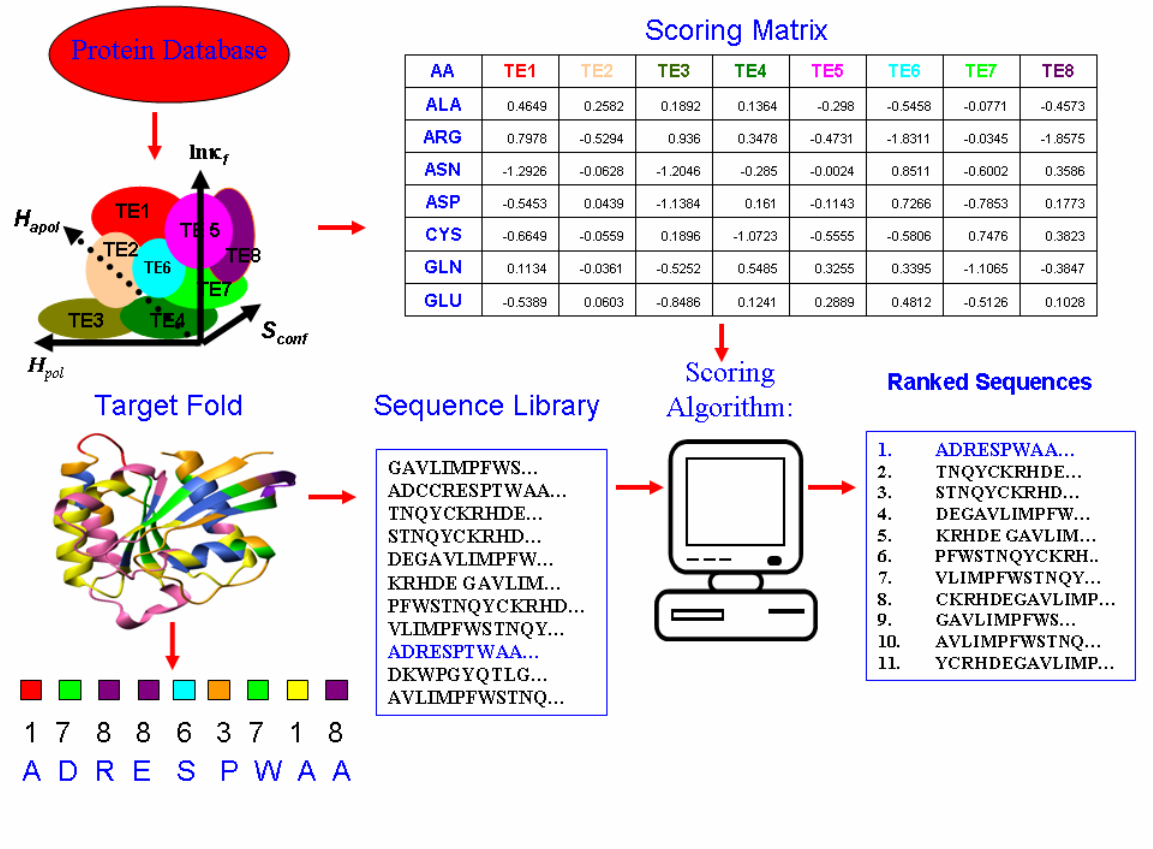


Figure 4-1: Schematic illustration of energetic-based fold recognition experiment

Eight thermodynamic environments were detected through clustering position-specific thermodynamic descriptors calculated by COREX for each residue (17802) in the database. Using these identified clusters, the amino acid propensities for each thermodynamic environment were calculated and used to formulate the scoring matrix in subsequent fold recognition experiments to match sequences to the target fold. For each protein (target fold) in the database, the three-dimensional structure can be represented by a one dimensional string of thermodynamic environment numbers (1 to 8) identified during the clustering analysis. Each target fold will be aligned to sequences in a library containing 431 sequences. Alignments are scored based on the Smith-Waterman local alignment using a scoring matrix of 20 amino acids to 8 thermodynamic environments. Fold recognition success was defined as the case where the target sequence scored higher than 99% of the decoy library (scored in the top four).

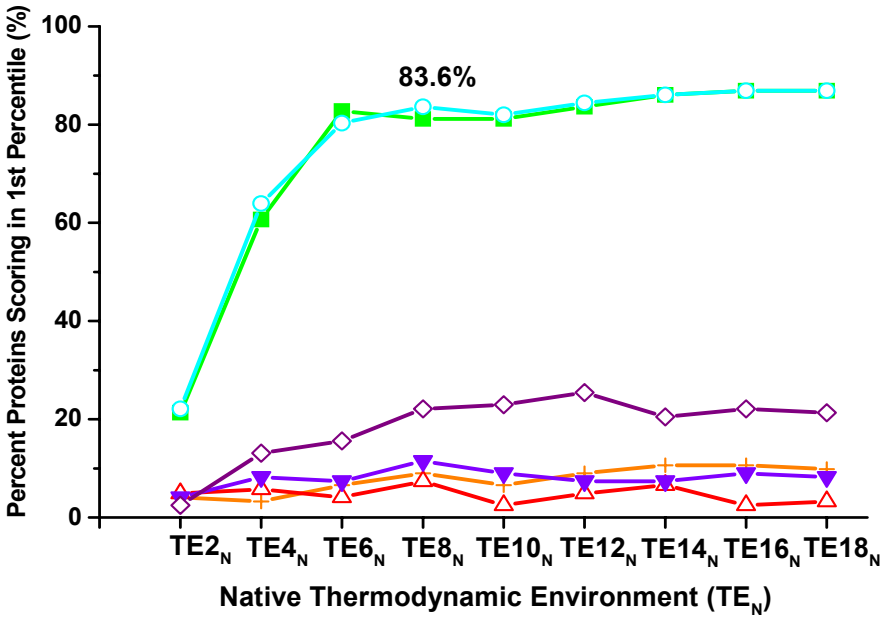


Figure 4-2: Fold recognition successes as a function of native state thermodynamic environments (TE_N)
Fold recognition experiments use scoring matrices composed of the log-odds probability of the 20 amino acids observed in each thermodynamic environments. A successful fold recognition experiments is defined as scoring the target protein among the top four proteins (1%) out of 431 sequences. Native ensembles were divided into five sub-ensembles (0-20% folded, 20-40% folded, 40-60% folded, 60-80% folded and 80-100% folded) to determine the sub-ensemble contribution to fold recognition. Each line represents a different sub-ensemble (from top to bottom): full ensemble (open circle in cyan), 80-100% folded sub-ensemble (closed square in green), 60-80% folded sub-ensemble (open diamond in purple), 0-20% folded sub-ensemble (cross in orange), 40-60% folded sub-ensemble (closed downtriangle in blue), 20-40% folded sub-ensemble (open uptriangle in red). With eight $TE_{N,}$ success rate was achieved at 83.6 %.

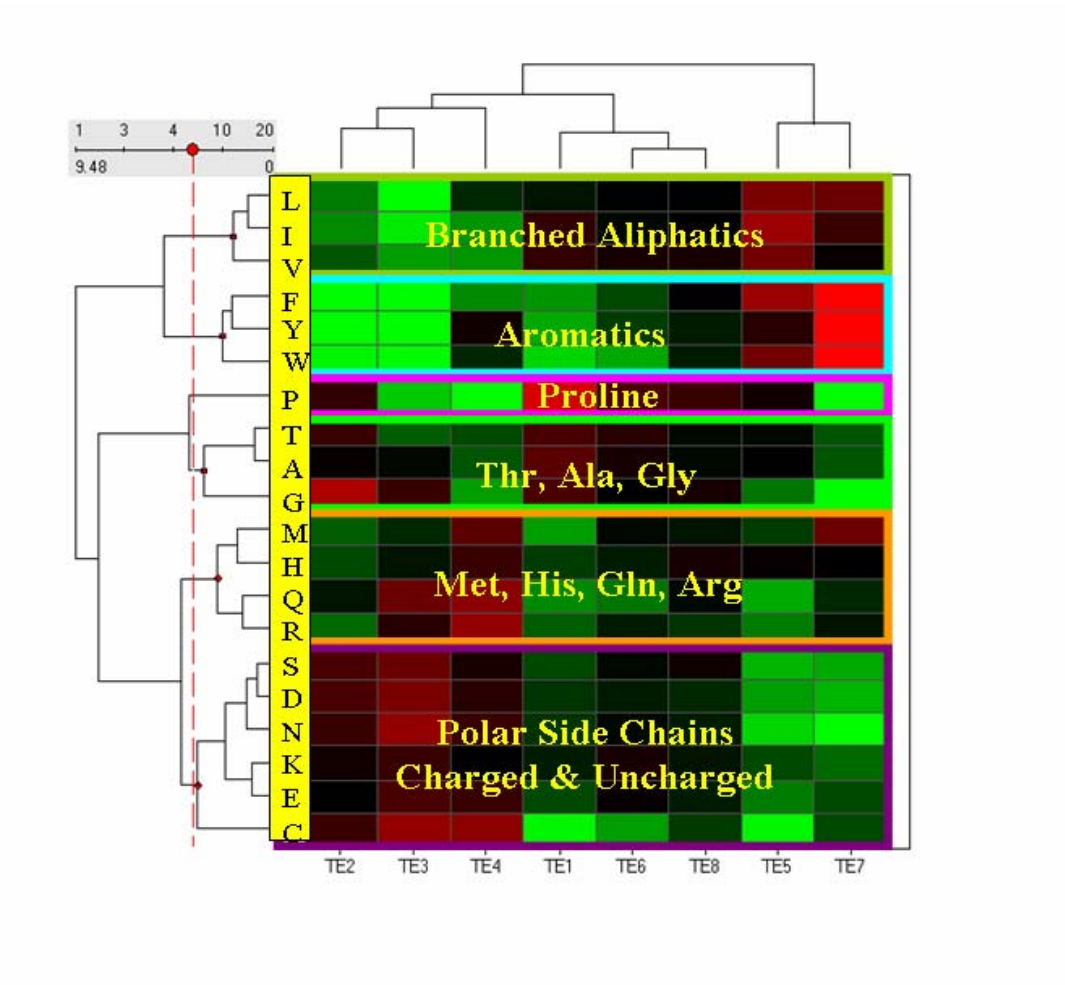


Figure 4-3: Double hierarchical cluster analysis of amino acid propensities for eight native state thermodynamic environments (TE_N)

Cluster results are shown in heat map in which rows are twenty amino acids and columns are eight native thermodynamic environments. Negative propensities are green, propensities near zero are black, and positive propensities are red. The color intensity reflects the magnitude of the propensities. The row dendrogram shows groupings of amino acids with similar log-odds probabilities for the thermodynamic environments. The gray scale above the amino acid dendrogram is the cluster slider. The numbers below the scale are the calculated dissimilarity measures. The red dotted line is positioned at the level of six amino acid clusters. Each of the six amino acid cluster nodes is indicated by a red dot and is highlighted.

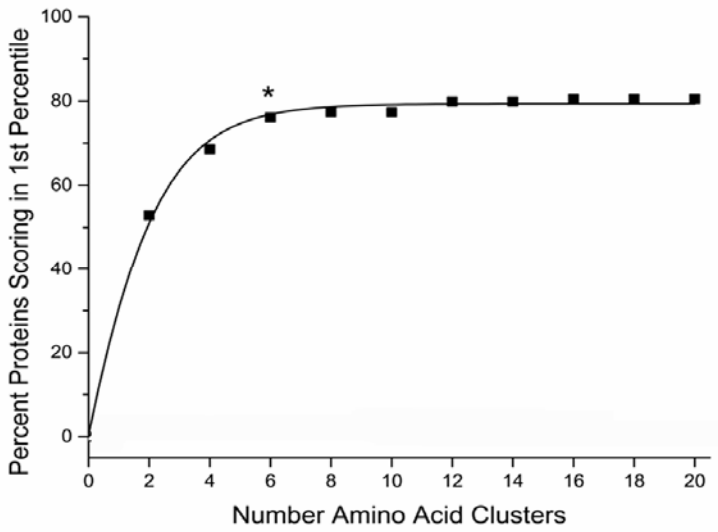


Figure 4-4: Fold recognition success as a function of amino acid cluster number in native state ensemble

The solid squares represent fold recognition experiments using scoring matrices composed of the propensities of a series of amino acid clusters for the eight native thermodynamic environments. A successful fold recognition experiment is one in which the native amino acid sequence, of the target protein, scores higher than 99% of the sequences (scoring in the top four) in the sequence library. The x-axis indicates the number of amino acid clusters used to generate the scoring matrix used in the associated fold recognition experiment. The asterisk denotes six amino acid groups are necessary to encode the eight thermodynamic environments of the proteins in our database.

matching (with an 84% success rate) a protein's sequence to a one-dimensional representation of that protein's energy landscape. That result conclusively demonstrated that the thermodynamics of the native state ensemble are important determinants of what fold a sequence will adopt, and that the specific thermodynamic descriptors developed in that work were sufficient to quantitatively characterize a diverse database of human protein structures. Here energetic information in the denatured ensemble and its sub ensembles are analyzed in a similar way based on fold recognition experiments. As Figure 4-5 indicates, in the case of the denaturing conditions, fold recognition success was found to plateau at 98.3%, using eight environments from full ensemble, an improvement over the same calculation performed native conditions. Partitioning the ensemble to investigate the contribution of the different sub-ensembles, reveals that under denaturing conditions the information content of the sub-ensemble containing 0-20% structure (80-100% unfolded regions) achieved exactly the same fold recognition success as full ensemble. 80-100% structure (i.e. 0-20% unfolded regions) was sufficient to produce fold recognition success at a rate of only 35%. Fold recognition results from other sub-ensembles (20-40% folded, 40-60% folded, 60-80% folded) were less than 20% success. Hierarchical clustering result (Figure 4-6) based on propensities of 20 amino acids in eight thermodynamic environments of denatured ensemble reveals different groups when comparing results to those obtained using the native ensemble (Figure 4-3). For example, among the six groups illustrated, Gly and Arg each occupies one group, while Ser and Glu make up one group, aromatic

amino acids(Trp, Tyr and Phe) and Asn, Glu, Asp are in the same group. The observed differences in amino acid groups between results obtained using hierarchical clustering of native and denatured ensemble suggest that different mechanisms of thermodynamic control are involved. To further illustrate the significance of the observed differences, simple fold recognition experiments were performed by using the hierarchical clustering results obtained with the denatured ensemble. Figure 4-7 is the fold recognition result based on amino acid cluster numbers. As Figure 4-7 indicates, clustering the database to be represented by four clusters of amino acids in the denatured ensemble captures 80% success of fold recognition within the database compared to 6 clusters in native ensembles capturing the same success rate. Figure 4-8 is a control experiment to show the folding information in native ensemble and denatured ensemble is not random. As Figure 4-8 reveals, the probability to obtain fold recognition success higher than 15% if energetic information was just randomly selected is very small.

Identifying the Source of Denatured State Fold Recognition Success

Fold recognition experiment results based on native ensemble energetics and denatured ensemble energetics clearly indicated that denatured ensemble achieves a higher success rate when matching sequence to fold. The experiments were performed using on PROFILESEARCH (Bowie et al, 1991) which basically follows Smith-Waterman local alignment algorithm with success was defined by the overall local alignment scores.

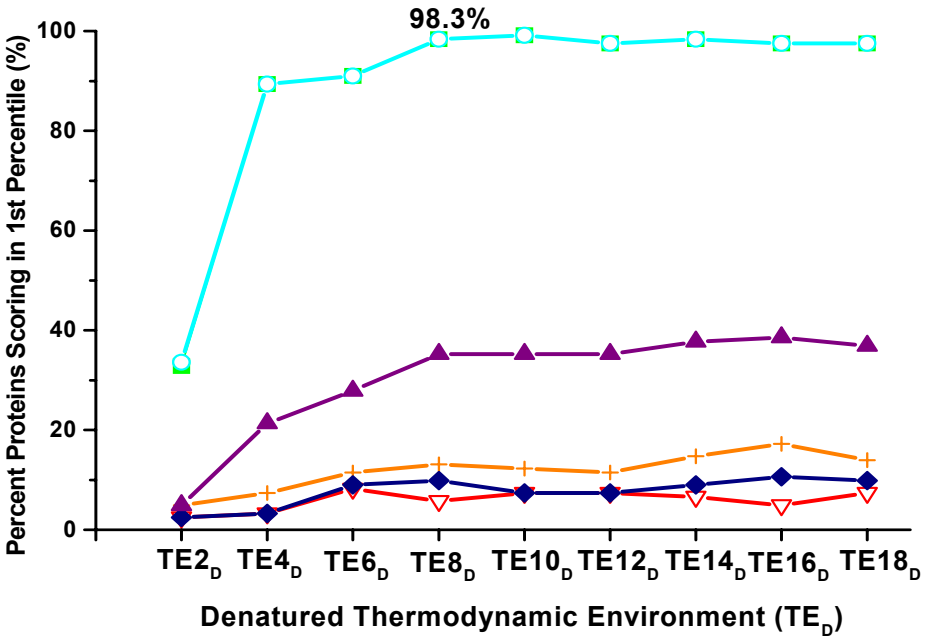


Figure 4-5: Fold recognition successes as a function of denatured state thermodynamic environments (TE_D)
Fold recognition experiments using scoring matrices composed of the log-odds probability of 20 amino acids for each thermodynamic environment. A successful fold recognition experiments was defined as scoring the target protein among the top four proteins (1%) out of 431 sequences. Denatured ensembles were divided into five sub-ensembles (0-20% folded, 20-40% folded, 40-60% folded, 60-80% folded and 80-100% folded) to determine the sub-ensemble contribution to fold recognition. Full ensemble (open circle in cyan), 0-20% folded sub-ensemble (closed square in green), 80-100% folded sub-ensemble (closed uptriangle in purple), 20-40% folded sub-ensemble (cross in orange), 60-80% folded sub-ensemble (closed diamond in blue) and 40-60% folded sub-ensemble(open down triangle in red). With eight $TE_{D,}$ 98.3 % fold recognition is achieved.

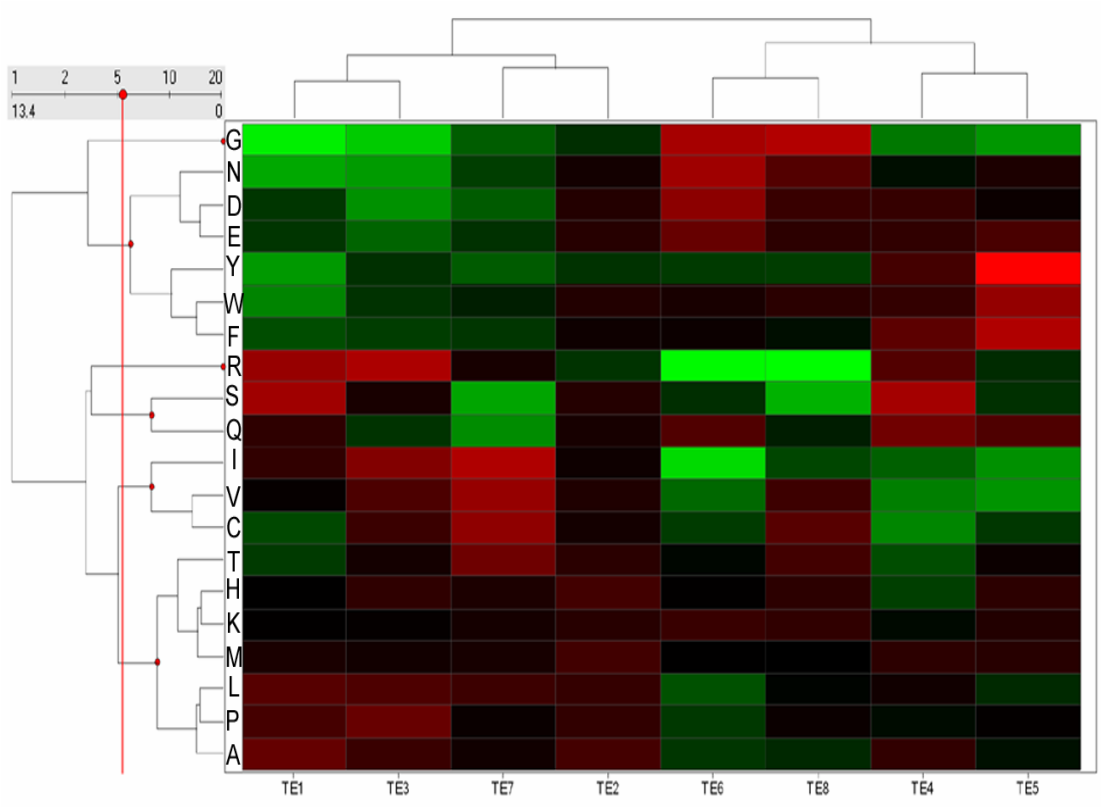


Figure 4-6: Double hierarchical cluster analysis of amino acid propensities for eight denatured state thermodynamic environments (TE_D)

Cluster results are shown in heat map in which rows are twenty amino acids and columns are eight denatured thermodynamic environments. Negative propensities are green, propensities near zero are black, and positive propensities are red. The color intensity reflects the magnitude of the propensities. The row dendrogram shows groupings of amino acids with similar log-odds probabilities for the thermodynamic environments. The gray scale above the amino acid dendrogram is the cluster slider. The numbers below the scale are the calculated dissimilarity measures. The red dotted line is positioned at the level of six amino acid clusters. Each of the six amino acid cluster nodes is indicated by a red dot.

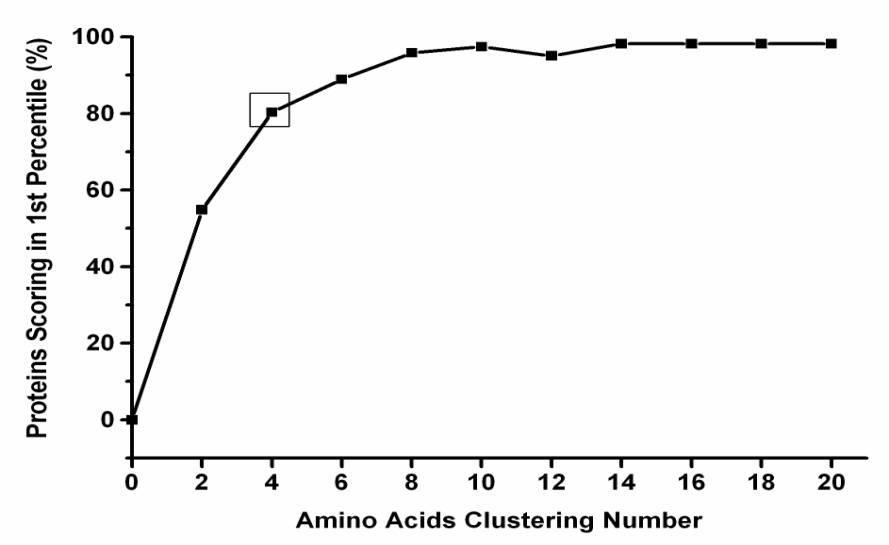


Figure 4-7: Fold recognition success as a function of amino acid cluster number in denatured state ensemble

The solid squares represent fold recognition experiments using scoring matrices composed of the propensities of a series of amino acid clusters for the eight denatured thermodynamic environments. A successful fold recognition experiment is one in which the native amino acid sequence, of the target protein, scores higher than 99% of the sequences (scoring in the top four) in the sequence library. The x-axis indicates the number of amino acid clusters used to generate the scoring matrix used in the associated fold recognition experiment. The box denotes that four amino acid groups are necessary to encode the eight denatured thermodynamic environments of the proteins in our database.

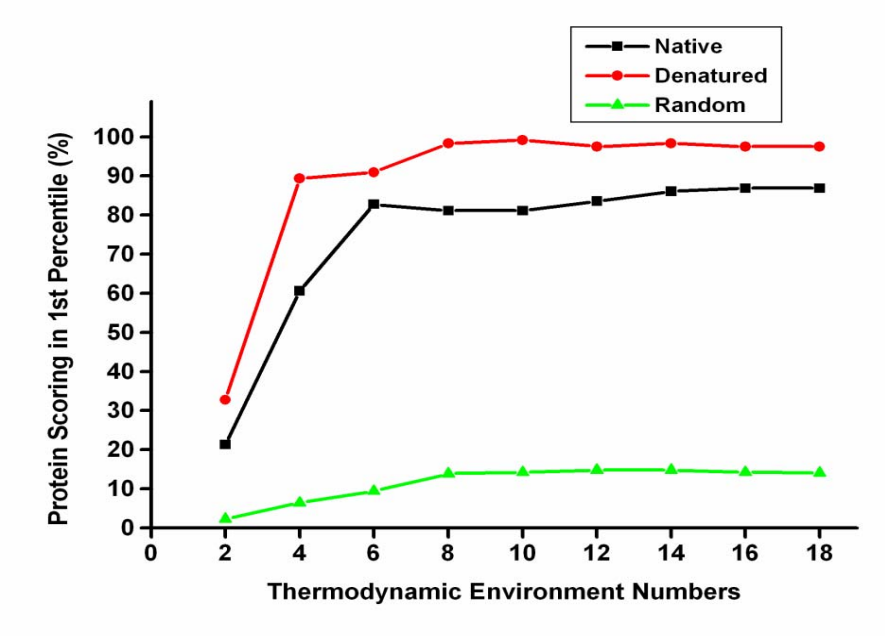


Figure 4-8: Fold recognition performance using thermodynamic environments identified with native, denatured and randomly generated ensemble

Fold recognition experiments use scoring matrices composed of the log-odds probability of 20 amino acids for each thermodynamic environment. A successful fold recognition experiments is defined as scoring the target protein among the top four proteins (1%) out of 431 sequences. X-axis is the number of thermodynamic environments. Y-axis is the percentage of proteins that ranked in the top four in fold recognition experiments. Red line is fold recognition results as a function of denatured thermodynamic environments. Black line is fold recognition results as function of native thermodynamic environments. Green line is the control experiment in which states were randomly picked up in COREX calculation and the random states were used to calculate position-specific thermodynamic descriptors and do cluster to define thermodynamic environments.

To identify the source of the improved fold recognition success using the denatured state thermodynamics, several analysis were made. First, fold recognition scoring matrix based on the propensities of 20 amino acids in each thermodynamic environments was investigated. Table 4-1 shows the propensities of 20 amino acids to eight native thermodynamic environments (TE_N). Table 4-2 shows the propensities of 20 amino acids to eight denatured thermodynamic environments (TE_D). Comparison of these propensities of amino acids in TE_N and TE_D reveals that the propensities ranges (from the lowest value to the highest value) for 12 amino acid types are broader for TE_D than those observed for TE_N. A broader range of propensities range suggests more variability in the propensities of amino acids to specific thermodynamic environments, thus imparts a distinguishing feature using TE_D for different amino acids. Second, residue-specific scores were calculated based on the propensity of amino acid to thermodynamic environments. Figure 4-9 displays the normalized residue-specific score distribution for all residues in the native ensemble and denatured ensemble. As Figure 4-9 indicates, residues in the denatured ensemble have a higher average position-specific score. Figure 4-10 shows the normalized residue-specific score for a fatty acid binding protein (PDB ID: 1CBS) in database. For each residue position, using the propensities for the thermodynamic environments in the denatured ensemble yields a higher normalized position-specific score which possibly accounts the observed improvement in the overall fold recognition score. Third, the quality of alignments between the target sequence and fold were investigated. Figure 4-11 shows the

sequence-thermodynamic environment alignments for sample proteins. Improved alignments were observed; resulting in a 15% increased in success rates in the fold recognition experiments. To quantitatively assess the improvements in the alignments obtained from the denatured state energetics, we compared the average identities for structural, energetic and sequence information obtained from native and denatured fold recognition experiments (Figure 4-12). For instance, the mean identity of thermodynamic environments between the actual and the aligned structures is 69.5% using denatured state energetics for fold recognition compared to just 56.6% (P= 0.02) when using native thermodynamic environments. Similarly, secondary structure identities also display a statistically significant improvement (+8%) when using information from the denatured rather than native state energetics are used as the basis for alignment. As the statistics revealed, alignments from denatured state information were more successful at matching both secondary structure and thermodynamic information than the alignments using native state thermodynamic information, demonstrating that both the length and the quality of the alignment are increased when derived from denatured state energetics.

Table 4-1: Propensities of 20 amino acids in eight native state thermodynamic environments

AA	TE_N1	TE_N2	TE_N3	TE_N4	TE_N5	TE_N6	TE_N7	TE_N8
ALA	0.452	0.0451	0.3908	-0.4474	-0.9298	0.146	0.0219	-0.287
ARG	-0.5305	-0.4176	0.2835	0.3992	0.6595	-0.4422	-0.2893	-0.5501
ASN	-0.2671	0.101	0.0691	0.4479	-0.2559	-1.1134	-0.4657	0.5433
ASP	-0.2186	0.1131	0.1469	0.3189	-0.4081	-0.6068	-0.4246	0.4669
CYS	-0.286	-0.1818	-0.0896	0.6559	0.3862	-0.5575	-1.6171	0.3574
GLN	-0.1899	-0.2493	0.2229	0.4402	0.4385	-0.595	-0.8924	0.1481
GLU	-0.3687	-0.0684	-0.1196	0.3897	0.1349	-0.4693	-0.3297	0.4107
GLY	0.3714	0.6621	-0.2742	-0.3651	-1.2817	-1.0835	-0.2451	0.655
HIS	-0.1836	-0.2404	0.0444	0.0761	0.4483	0.1015	-0.1903	-0.2558
ILE	0.1229	-0.2366	-0.2888	-0.745	-0.2082	0.5674	0.5648	-0.7713
LEU	-0.2289	-0.1989	-0.0341	-0.4416	-0.0871	0.5504	0.392	-0.8531
LYS	-0.1054	0.0522	0.0088	0.2026	-0.2481	-0.4288	-0.1209	0.3891
MET	-0.2796	-0.3912	0.0723	0.1778	0.5658	0.1152	-0.5707	-0.1148
PHE	-1.5335	-1.0511	-0.5042	-0.6605	0.5836	0.9802	0.1756	-1.0493
PRO	0.9902	0.6253	-0.1364	-1.3557	-2.0027	-0.4897	0.3847	-0.4242
SER	-0.0349	0.0469	0.3706	0.397	-0.3678	-0.5806	-0.3935	0.0411
THR	0.1525	0.1661	-0.1591	-0.1067	-0.5085	-0.1616	0.238	0.1897
TRP	-0.9032	-0.8229	-1.0371	-0.7906	1.0874	0.7166	-0.0256	-0.8265
TYR	-0.9283	-1.0909	-0.5537	-0.0852	1.0604	0.4242	0.0466	-0.9461
VAL	0.2	-0.0298	-0.0987	-0.6274	-0.4392	0.3847	0.5053	-0.6003

Table 4-2: Propensities of 20 amino acids in eight denatured state thermodynamic environments

AA	TE_D1	TE_D2	TE_D3	TE_D4	TE_D5	TE_D6	TE_D7	TE_D8
ALA	0.4649	0.2582	0.1892	0.1364	-0.298	-0.5458	-0.0771	-0.4573
ARG	0.7978	-0.5294	0.936	0.3478	-0.4731	-1.8311	-0.0345	-1.8575
ASN	-1.2926	-0.0628	-1.2046	-0.285	-0.0024	0.8511	-0.6002	0.3586
ASP	-0.5453	0.0439	-1.1384	0.161	-0.1143	0.7266	-0.7853	0.1773
CYS	-0.6649	-0.0559	0.1896	-1.0723	-0.5555	-0.5806	0.7476	0.3823
GLN	0.1134	-0.0361	-0.5252	0.5485	0.3255	0.3395	-1.1065	-0.3847
GLU	-0.5389	0.0603	-0.8486	0.1241	0.2889	0.4812	-0.5126	0.1028
GLY	-1.7588	-0.4908	-1.5225	-0.9766	-1.179	0.8965	-0.8042	0.9762
HIS	-0.183	0.2478	0.1207	-0.607	0.098	-0.1758	-0.0008	0.1018
ILE	0.1348	-0.0976	0.665	-0.8211	-1.1361	-1.6186	0.9568	-0.6523
LEU	0.3759	0.1566	0.3141	-0.0763	-0.4603	-0.7243	0.2045	-0.2181
LYS	-0.176	0.0674	-0.1519	-0.2516	0.038	0.1796	-0.0465	0.1252
MET	-0.0142	0.2408	-0.072	0.1026	0.0679	-0.1809	-0.0374	-0.1929
PHE	-0.6994	-0.094	-0.5922	0.4158	0.9647	-0.106	-0.5456	-0.2853
PRO	0.2657	0.1363	0.4901	-0.2655	-0.1623	-0.5578	-0.1204	-0.116
SER	0.8622	0.0501	-0.031	0.886	-0.506	-0.4946	-1.2752	-1.3582
THR	-0.5699	0.086	-0.0548	-0.6999	-0.109	-0.2337	0.5298	0.2458
TRP	-1.059	0.0346	-0.519	0.1492	0.7841	-0.0305	-0.3891	0.0854
TYR	-1.196	-0.5211	-0.5093	0.2577	1.4776	-0.5757	-0.7906	-0.5903
VAL	-0.1493	0.0191	0.3099	-1.0219	-1.1597	-0.8711	0.7973	0.2095

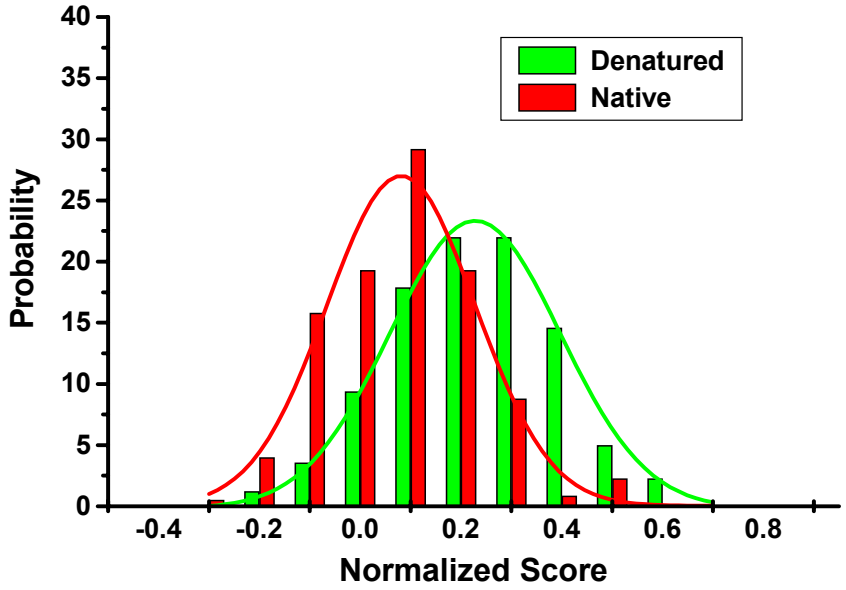


Figure 4-9: Normalized position-specific score distributions for native and denatured state ensembles

Scores are calculated based on the propensities of twenty amino acids to eight thermodynamic environments in the native and denatured ensembles. Scores are normalized based on averaging the scores of five neighboring residues. The x-axis represents normalized score ranges. The y-axis represents the distribution probabilities) of normalized scores in each range. Position-specific scores follow a normal distribution and normalized scores in denatured ensemble have a higher mean, thus suggesting that average position-specific scores are higher for denatured ensembles when matching sequence to folds.

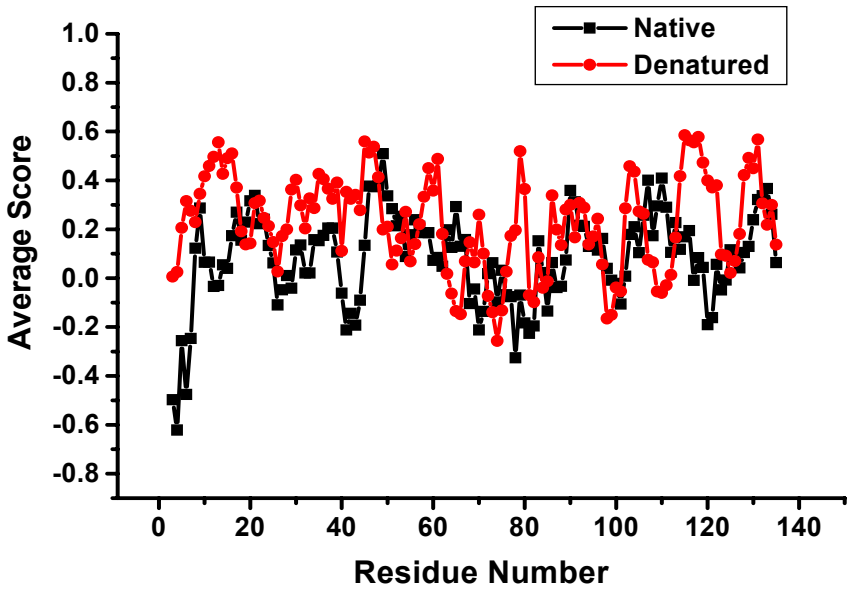


Figure 4-10: Position-specific alignment scores for a fatty acid binding protein (PDB: 1CBS) calculated from native and denatured state ensembles

The x-axis represents the residue number and y-axis is the average score for each position. Average scores were calculated from the native (black line) and denatured (red) ensemble. Average scores were calculated by averaging scores within a window of five neighboring residues. As shown in the figure, scores calculated using the denatured ensemble are higher than scores calculated from native ensemble in most positions.

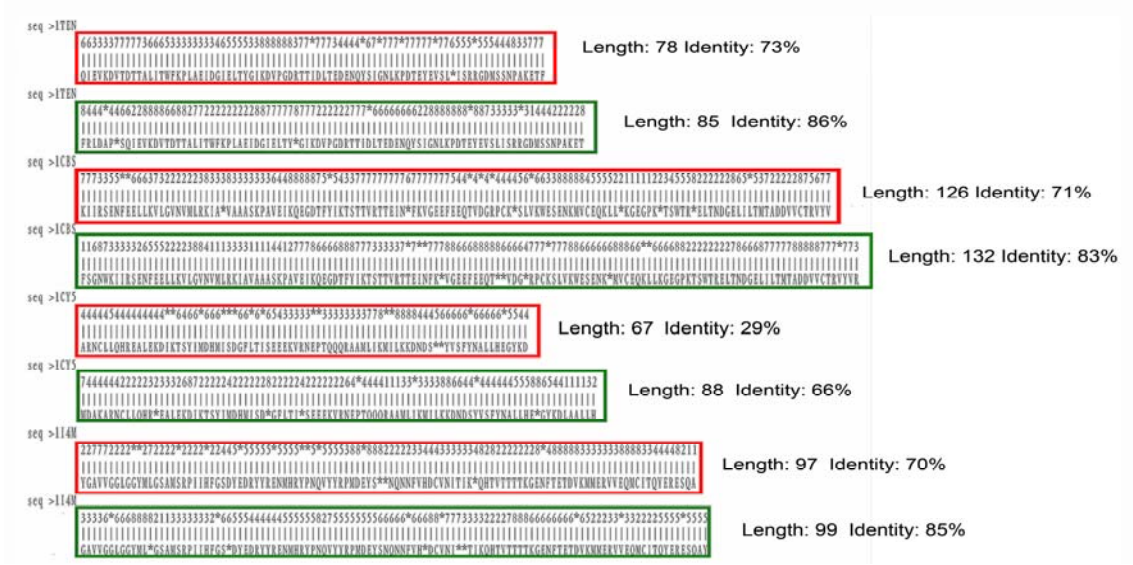


Figure. 4-11: Comparison of alignments generated from fold recognition experiments using native and denatured state thermodynamic environments

Alignments were generated using the Smith-Waterman local alignment algorithm to score proteins for fold recognition based on the identified thermodynamic descriptors. Alignments using TE_N are boxed in red while alignments using TE_D in boxed in green (gaps are represented as asterisks *). Local alignment length and identity are shown next to the alignment and clearly shows that alignments using denatured ensemble thermodynamic environments are longer matched with higher identities.

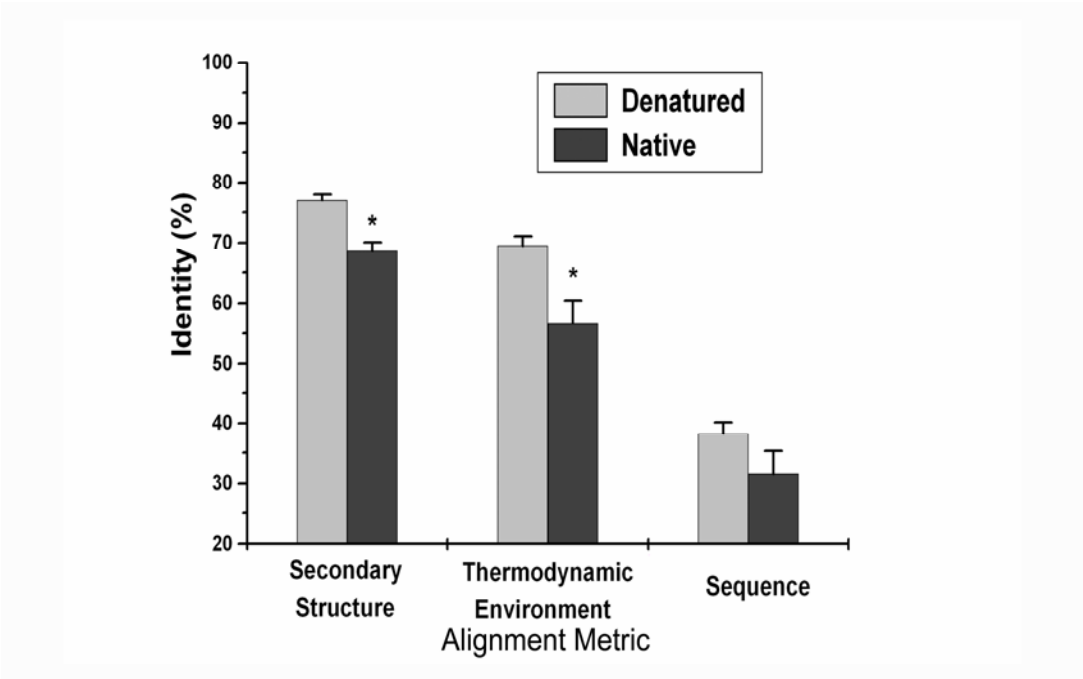


Figure. 4-12: Alignments identity calculated based on the fold recognition experiments using native and denatured state thermodynamic environments

Alignments were generated using the Smith-Waterman local alignment algorithm to score proteins for fold recognition based on the identified thermodynamic descriptors. Alignments identity in alignments obtained using TE_N (black bar) and TE_D (grey bar) using different metrics (secondary structure, thermodynamic environments and sequence as shown on the x-axis) and the percentage of exact match on the y-axis.

Discussion

In the previous chapter, we demonstrated that the energetic information contained within the native and denatured ensembles is different. The results presented here revealed that there is different fold information content contained in the native and denatured ensemble as revealed by the different success in fold recognition experiments matching sequences to their respective folds. Energetic information in the denatured ensemble achieved a higher success rate compared to native ensemble; which suggests that there should be some unique folding information in the denatured ensemble. Close investigation of these results revealed that higher fold recognition success originates from the improved quality of alignment between sequence and thermodynamic environments using the information content of the denatured ensemble. Although energetic determinants within the native ensemble have been shown to effectively recognize folds (Larson and Hilser 2004; Wrabl et al. 2002), the results presented here show that the energetic determinants within the denatured ensemble are more discriminating. Combined with the energetic determinants of the native ensemble energetic important for fold specificity (Larson and Hilser 2004), the success of this study and the unique nature of the energetic determinants in the denatured ensemble will provide a better thermodynamic description of protein folds and open new opportunities to realize a new energetic classification of protein folds.

Methods

Propensities of amino acids in each thermodynamic environment

Propensities of amino acids in each thermodynamic environment are the double-normalized log-odd probabilities of each amino acid in thermodynamic environments calculated as double normalized log-odd probabilities:

$$\text{Log Odds}_{(AA,TE)} = \ln \frac{\frac{AA_{TE}}{Total_{AA}}}{\frac{Total_{TE}}{Total_{Residues}}} \qquad (4.1)$$

AA_{TE} is the total number of one type amino acid in one thermodynamic environment. $Total_{AA}$ is the total number of one type of amino acid in the whole database. $Total_{TE}$ is the total number of amino acids in one thermodynamic environment. $Total_{Residues}$ is the total number of residues in the whole database.

Double hierarchical cluster analysis

Double hierarchical cluster analysis was shown on heat map generated by SpotFire DecisionSite Statistics 7.2 software. Agglomerative hierarchical approach was used with complete linkage (measure the maximum distance between two clusters) clustering method and city block distance (Manhattan distance) to measure the dissimilarity.

Propensities of twenty amino acids in eight thermodynamic environments was clustered and visualized by different colors in heat map. The color range is set to continuous coloring and spans from green, to black, to red. The range is set so that propensities equal to zero (no propensity) are colored black, lower propensities are colored green and higher propensities are colored red. The relative intensity of the colors reflects their distance from zero propensities.

Fold recognition experiments based on amino acid propensities for thermodynamic environments

Fold-recognition experiments were based on PROFILESEARCH of Eisenberg and coworkers (Bowie et al., 1991) as described previously (Wrabl et al 2001,2002). Based on clustering results, each protein (profile) in database was represented by one-dimensional string with each residue assigned to a thermodynamic environment. There are 431 decoy sequences including the 122 native sequences in our dataset from which correct fold recognition is tested. PROFILESEARCH implements the Smith-Waterman local alignment algorithm (Smith & Waterman, 1981) that is used to align each profile with each sequence in this search database. Log-odds probabilities of amino acids in thermodynamic environments (Equation 5.2) were used to construct scoring matrix for alignment. All other parameters in PROFILESEARCH, specifically the gap open and extension penalties, were defaults. A successful fold recognition experiment is one in which the native sequence had an alignment cumulative score among the top four (1%) scores out of the total 431 sequences that has been scored.

Fold recognition experiments based on random information

Fold recognition experiments based on random information were performed by randomly selecting microstates from the full ensemble. The randomly selected microstates were different states containing different percentage of folded and unfolded structure generated by COREX. The ensemble made up of randomly selected microstates was then used to calculate position-specific thermodynamics and subsequent clustering results were then used to perform fold recognition.

Alignment identity calculation

Alignment program based on PROFILESERACH was written in Perl. Alignment output is generated using thermodynamic environments as the profile that was used to match the sequences (target). Because each sequence can be represented as an amino acid sequence, string of thermodynamic environment numbers or secondary structure types, identities based on each representation can be calculated. Identity is calculated as the percentage of matched positions divided by the total length of alignment. Secondary structure assignments for residues in the target sequences, the template structure used for fold recognition, were assigned using STRIDE (Frishman and Argo, 1995). Simple statistical t-test was used to compare the mean identity between alignments using denatured and native state energetics. Statistical tests were performed using the open statistics software, R (www.r-project.org).

CHAPTER 5

The Relationship Between Denatured State Energetics and Secondary Structures

Introduction

The protein architecture is defined hierarchically with four classes starting with the primary amino acid sequence followed by the secondary, tertiary and quaternary structure. Secondary structures are building blocks of three-dimensional, functional native folds. Information from secondary structure can be used to correctly predict functional sites and assign sequences to folds (Bowie et al 1991, Jones et al. 1999, Rose et al. 2006). In the previous chapter, we demonstrated the success of using denatured ensemble energetics to match sequences to folds. The high alignment identity for secondary structures using denatured state energetics suggests that the algorithm may be capturing local energetics that are specific to different secondary structure types. Therefore, it is interesting to know whether there is some correlation between energetics in the denatured state and the building blocks (secondary structures) of the final native fold. In this chapter, propensities for different secondary structure types were calculated under denatured conditions and the relationship between denatured ensemble energetics and secondary structures were investigated.

Results

Propensities of Secondary Structure in Thermodynamic Environments

To address the question of whether a correlation exists between energetics in denatured ensemble and secondary structure, the propensity of each secondary structure type was calculated for each thermodynamic environment in native (Table 5-1) and denatured (Table 5-2) conditions. Propensities were calculated as the log-odds probabilities of each secondary structure in each thermodynamic environment. As Table 5-1 and Table 5-2 reveal, the same secondary structure types in native and denatured conditions show different propensities. For regular structures (alpha helix and beta sheet), propensities in denatured thermodynamic environments (Table 5-2) showed greater variability; while propensities of irregular structures (coil and turn) showed greater variability in native thermodynamic environments (Table 5-1). For better illustration, propensities for each secondary structure type were plotted as a function of thermodynamic environments in different conditions (native and denatured). Several observations can be made. First, the propensity of each secondary structure for different thermodynamic environments is non-random in both the native and the denatured states, resulting in “thermodynamic signatures” for different secondary structural elements. Second, within the native state, there appear to be only two general signatures, one that is shared by regular secondary structures (i.e. alpha helix and beta sheet) (Figure 5-1) and one that is shared by irregular structures (coil and turn) (Figure 5-2). Positions that adopt

both alpha helices and beta strands have positive propensities for environments TE_N 5, 6 and negative propensities for TE_N 1, 2, and 8; whereas positions that adopt both coil and turn have positive propensities for environments TE_N 1, 2, and 8 and negative propensities for TE_N 4, 5, and 6. Third, unlike the native state signatures, within the denatured state thermodynamic environments, the thermodynamic signatures for regular secondary structures (alpha helix and beta sheet) show clear differences (Figure 5-3). Positions that adopt alpha helix in the folded protein show preferences to be in TE_D 1, 2, 3, 4, 5 whereas positions that adopt beta strands prefer TE_D 7 and 8. Fourth, although the signatures for helix and sheet (and to a lesser extent turn) contain strong propensities, there are no significant propensities for coil (Figure 5-4) when compared to the magnitude of the propensities in the native thermodynamic environments (Figure 5-2). These propensities show clearly that while the thermodynamics of the native state can discriminate between regular and non-regular structure, the denatured state thermodynamics can discriminate between different types of regular structure.

Variability in Denatured State Ensemble Energetic Propensities - Implication for Secondary Structure

The clear separation of thermodynamic propensities (Figure 5-3) based on the secondary structure adopted by that position in the final fold suggests that the TE_D may be useful in making inferences about secondary structure. To challenge this hypothesis, the thermodynamic environment of each position (from TE_D) was assigned to a secondary structure based on the propensity of observing the structure for that environment (Figure 5-3, 5-4). For example, because alpha

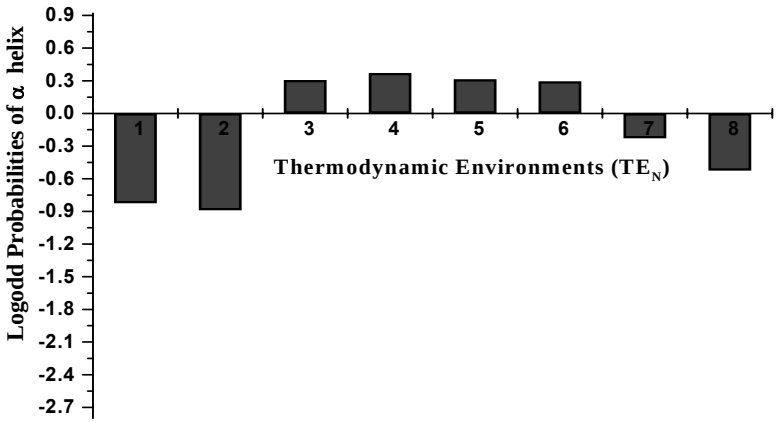
Table 5-1: Propensities of secondary structures in native state thermodynamic environments

Native TE	Alpha Helix	Beta Sheet	Coil	Turn
1	-0.904	-0.646	0.807	0.344
2	-0.938	-0.448	0.441	0.620
3	0.298	-0.338	-0.108	-0.071
4	0.394	-0.153	-0.389	-0.196
5	0.378	0.453	-1.295	-1.069
6	0.357	0.427	-0.797	-1.261
7	-0.288	0.255	0.095	-0.185
8	-0.734	-0.300	0.251	0.496
Variability	0.606	0.419	0.683	0.691

Table 5-2: Propensities of secondary structures in denatured state thermodynamic environments

Denatured TE	Alpha Helix	Beta Sheet	Coil	Turn
1	0.581	-1.302	0.141	-0.356
2	0.283	-0.326	-0.081	0.016
3	0.197	-0.028	0.233	-0.416
4	0.663	-1.454	-0.496	-0.048
5	0.296	-0.160	-0.279	-0.138
6	-0.52	-0.489	0.0317	0.648
7	-2.685	0.919	-0.066	-0.879
8	-1.724	0.450	0.219	0.202
Variability	1.2181	0.8033	0.253	0.4535

A)



B)

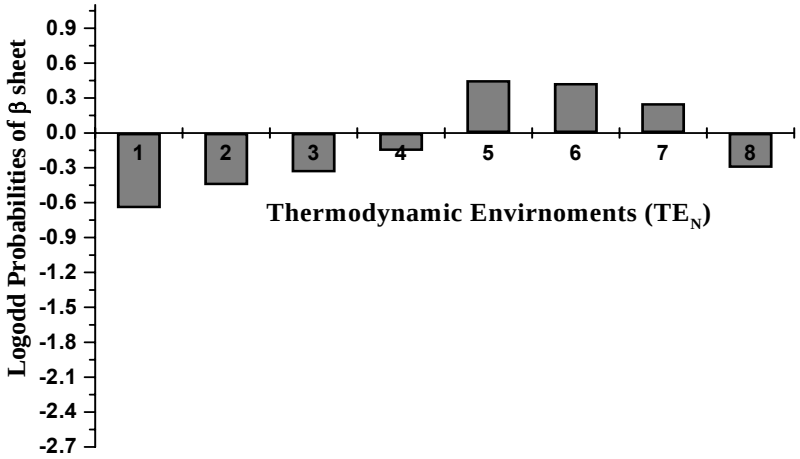
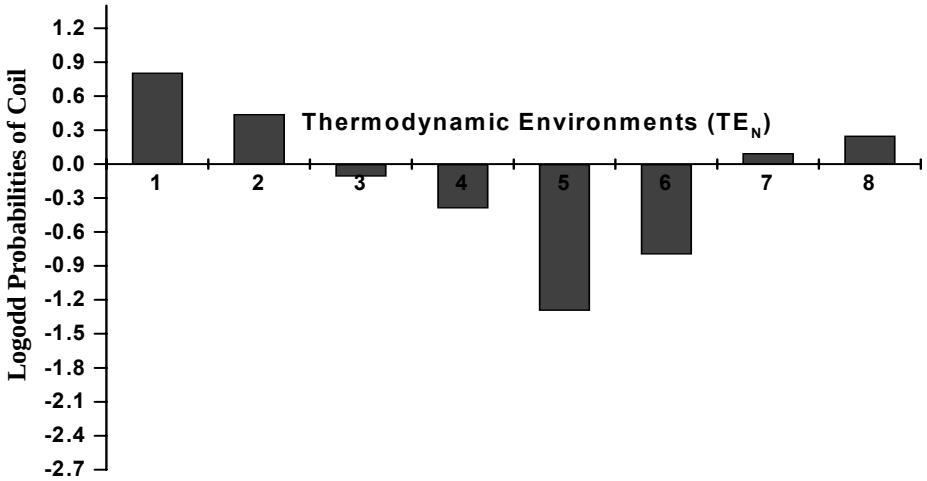


Figure 5-1: Regular secondary structure propensities for native state thermodynamic environments (TE_N)

In each plot, the eight environments are aligned on the ordinate and the log-odds probabilities of the secondary structure are plotted against the abscissa. The log-odds probabilities of (A) alpha helices and (B) beta sheets are shown.

A)



B)

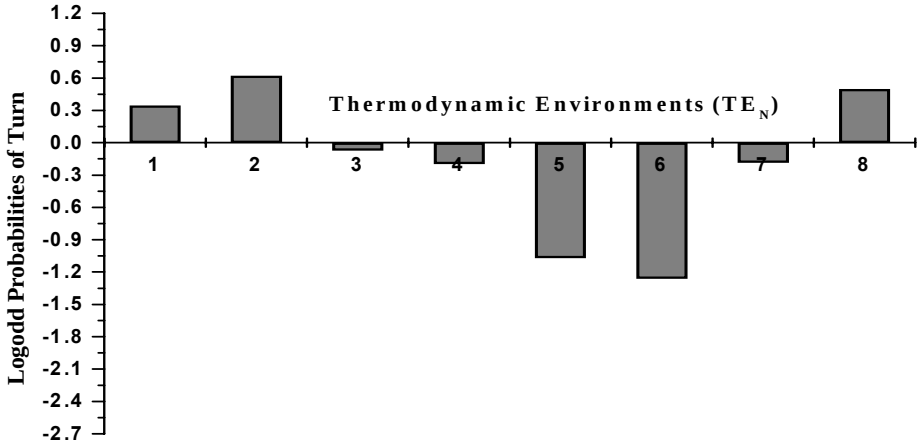
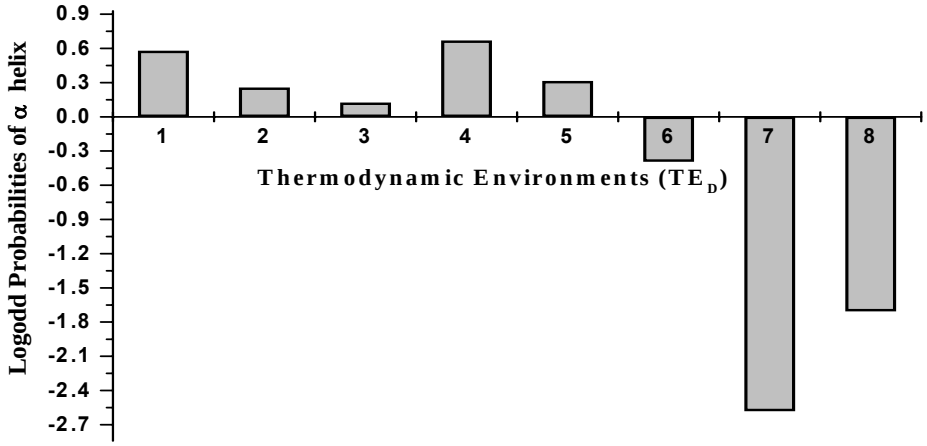


Figure 5-2: Irregular secondary structure propensities for native state thermodynamic environments (TE_N)

In each plot, the eight environments are aligned on the ordinate and the log-odds probabilities of the secondary structure are plotted against the abscissa. The log-odds probabilities of (A) coil and (B) turn are shown.

A)



B)

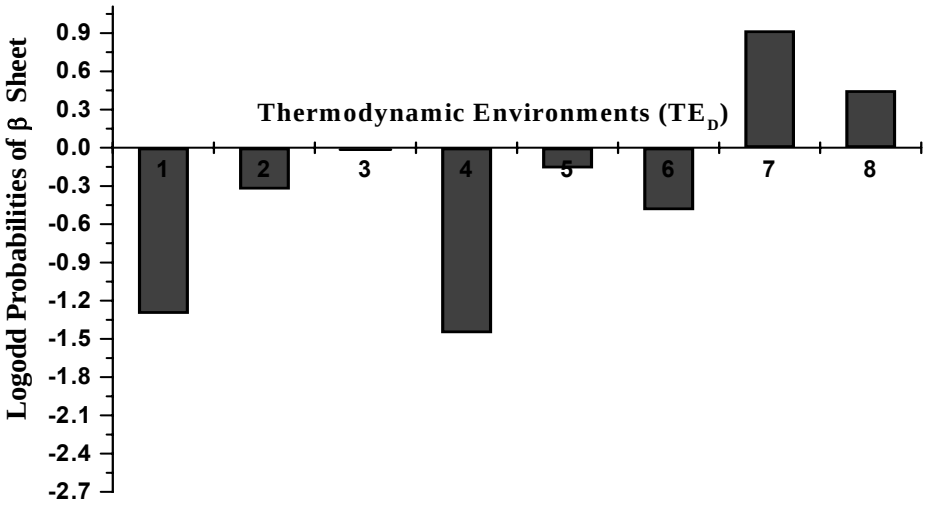
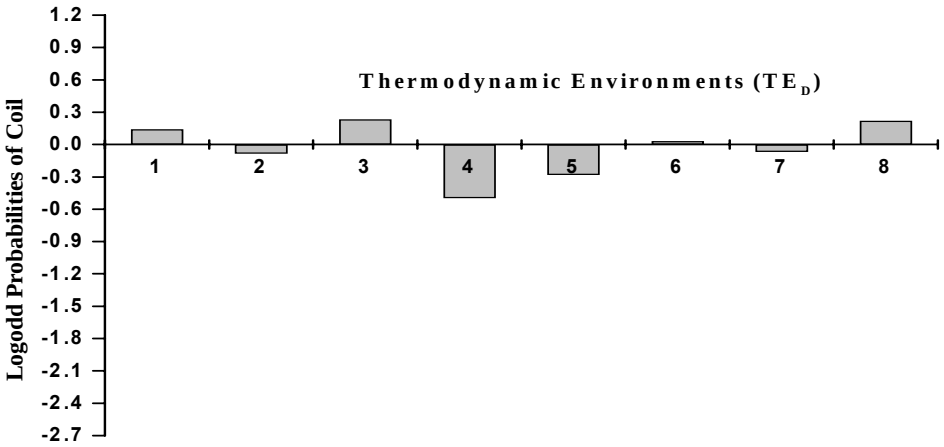


Figure 5-3: Regular secondary structure propensities for denatured state thermodynamic environments (TE_D)

In each plot, the eight environments are aligned on the ordinate and the log-odds probabilities of the secondary structure are plotted against the abscissa. The log-odds probabilities of (A) alpha helices and (B) beta sheets are shown.

A)



B)

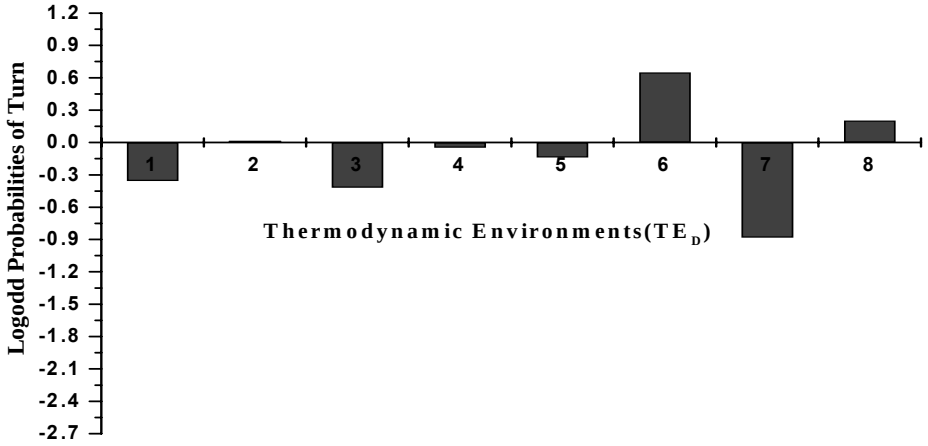


Figure 5-4: Irregular secondary structures propensities for denatured state thermodynamic environments (TE_D)

In each plot, the eight environments are aligned on the ordinate and the log-odd probabilities of the secondary structure are plotted against the abscissa. The log-odds probabilities of (A) coil and (B) turn are shown.

helix has a high propensity for TE_D 1, 2, 4 and 5, any positions with TE_D 1, 2, 4 or 5 were assigned to alpha helix. That assignment was then compared to the secondary structure observed in the native fold. The fraction of matches for each secondary structure is shown in Figure 5-5 and Figure 5-6. For comparison, the number of matches obtained by randomly assigning secondary structure from a fixed number of counts for each secondary structure type (i.e. controlling for the composition of each secondary structure: Figure 5-5) reveals that the predictions are significant in all cases, but especially for helix and sheet. Similarly, when the results are compared to the number of matches obtained by randomly assigning entire elements to consecutive stretches of positions (i.e. controlling for composition and continuity of each secondary structural element: Figure 5-6), it is clear that in successfully matching sequence to fold, the denatured state thermodynamic information performs disproportionately well with regular secondary structure, and only marginally well in turns and coils.

The clear separation of thermodynamic propensities of secondary structures also opens the opportunity of predicting secondary structures based on this simple assignment approach. Figure 5-7 shows a comparison of secondary structure predication results between our approach and currently popular secondary structure prediction algorithms. As Figure 5-7 reveals, although admittedly a crude method for making an assignment, it nonetheless allows us to gauge our secondary structure encoding information with more

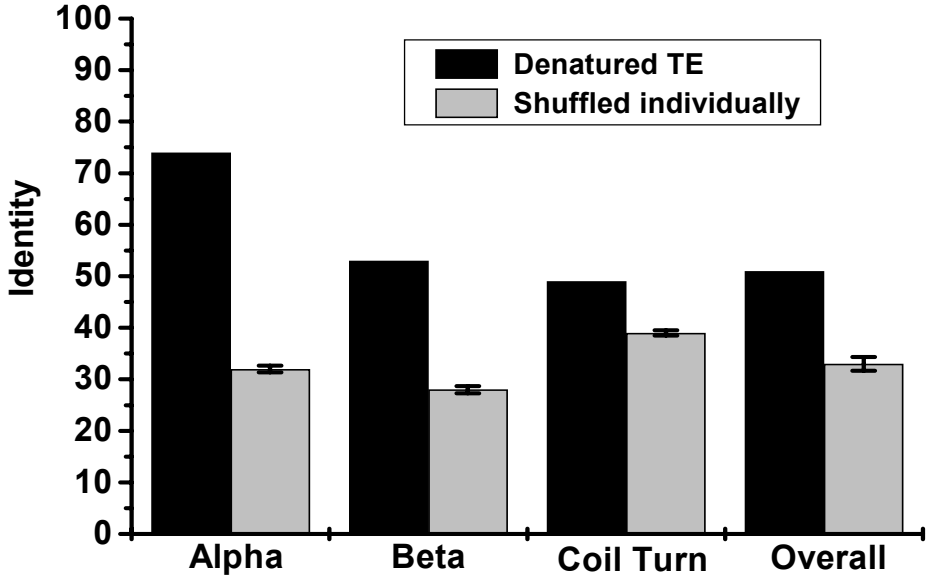


Figure 5-5: Comparison of secondary structure assignment using thermodynamic environment information to the random assignment of secondary structure

The overall identity and those reported for each subcategories (alpha, beta, and coil) using thermodynamic environment information (black bar) is compared to identities calculated using the random assignment of secondary structure (grey bar). Secondary structures were randomly shuffled individually and reassign to each position. Irregular structures (including coil and turn) were categorized as coil. The identities calculated with the random assignment are the average of 100 repetitions.

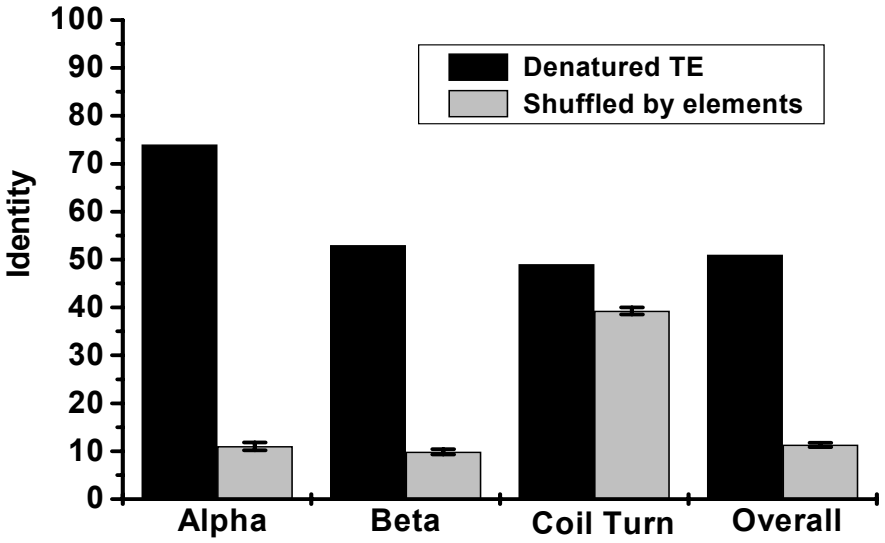


Figure 5-6: Comparison of secondary structure assignment using thermodynamic environment information to the random assignment of secondary structure elements.

The overall identity and those reported for each subcategories (alpha, beta, and coil) using thermodynamic environment information (black bar) is compared to identities calculated using the random assignment of secondary structure (grey bar).Secondary structure segments within the database were randomly assigned. Irregular structures (including coil and turn) were categorized as coil. The identities calculated with randomly assignment are the average of 100 repetitions.

sophisticated approaches that use more input features and advanced technologies. While the predictions did not outperform more recent developments in secondary structure predictions; our aim was not to construct a new secondary structure predictor. Instead, it was to show that TE_D does capture energetics that are associated with the energetics of the secondary structures that each position adopted in the native fold for each position. The assignment is achieved using only the propensities of each residue rather than incorporating information as implemented by other secondary structure predictors.

Discussion

The results in this chapter indicated that the denatured state thermodynamic signature contains significant fold encoding information, and that the majority of correctly aligned positions are in regions containing alpha helices and beta strands. Distinct preferences for particular thermodynamic environments based on regular secondary structures (alpha helices and beta sheets) show that there are some correlation between energetics in the denatured ensembles and secondary structures. Because the energetics in the denatured ensembles can be successfully correlated back to the observed structural features found in the natively folded protein, it is important to emphasize that these identified energetics could potentially be related back to specific functional features.

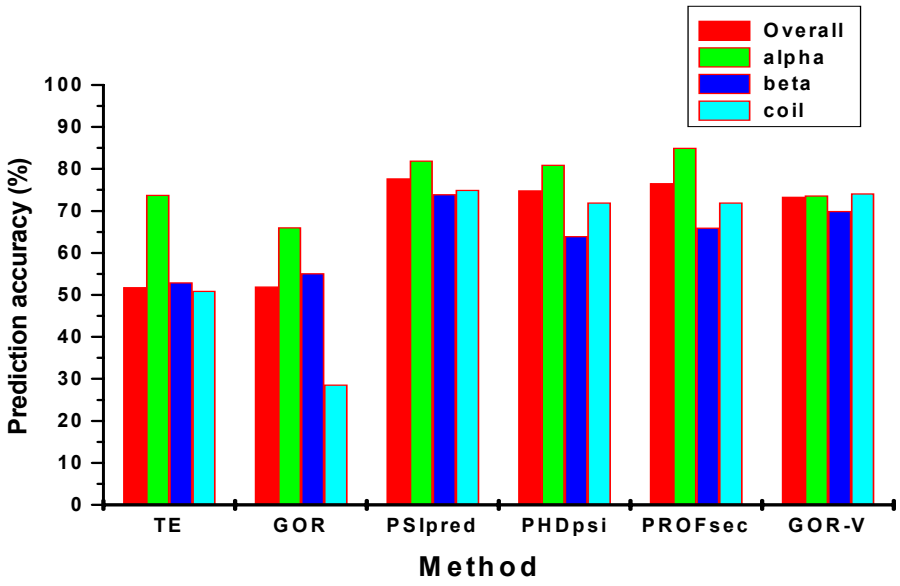


Figure 5-7: Comparison of secondary structure prediction performance using thermodynamic environments (TE) with other predictors

The accuracies for overall and subcategories (alpha, beta, and coil, turn) is compared to 5 other predictors: GOR (Garnier et al, 1978), PSIpred (Jones, 1999), PHDpsi (Przybylski, 2002), PROFsec (Rost, unpublished data), and GOR-V (Sen et al., 2005). Predication accuracy for overall (red bar), alpha helix (green bar), beta sheet (blue bar) and coil turn (cyan bar).

Currently, the three popular protein secondary structure prediction algorithms are the Chou-Fasman/GOR methods, neural network models and nearest-neighbor methods. Developed in the early 1970's, the Chou-Fasman method was an empirical method based on the analysis of the relative frequency of amino acids in each secondary structure types to make these structural predictions (Chou and Fasman 1978). The success of GOR, a method introduced in 1978, assumes that the amino acids flanking the central amino acid influence the secondary structure adopted by the central residue is likely to adopt. Neural network models are sophisticated machine learning techniques which can be trained to predict secondary structures. Nearest-neighbor methods predict secondary structure by identifying similar sequences with known structures.

Although secondary structure predictions based on the presented thermodynamic environment approach did not outperform leading secondary structure prediction algorithms, we found that the performance of assignment was comparable. The current leading technologies for secondary structure predictors utilize additional input features such as evolutionary information or multiple alignments to improve these predictions. Using the TE_D based assignment did not require the use of machine learning techniques such as neural networks because the TE_D are biophysical descriptors that describe properties of secondary structural elements in the denatured ensembles. These thermodynamic environments represent the composite enthalpic and entropic

contributions that are preferred, if not required, for the formation of the observed secondary structural element.

The significance of the results we wish to emphasize is the relationship between denatured ensemble energetics and secondary structures. The presented correlation provides us with a better understanding about how thermodynamics under denatured conditions are related to the building blocks (secondary structures) of native folds. An improved understanding of the correlations between denatured ensemble energetics and secondary structure will allow us to achieve the goal of obtaining better secondary structure prediction results.

Methods

Calculation of propensities of secondary structures in thermodynamic environment

Secondary structures were assigned to each residue in database by secondary structure assignment program STRIDE (Frishman et al. 1995). Log-odds probabilities of four secondary structure categories (alpha, beta, coil and turn) in thermodynamic environments were calculated as:

$$Log\ Odds_{(SS,TE)} = \ln \frac{\frac{AA_{SS}}{Total_{SS}}}{\frac{Total_{TE}}{Total_{\text{Residues}}}} \qquad (5\text{-}1)$$

AA_{ss} is the total number of one type secondary structure in one thermodynamic environment. $Total_{ss}$ is the total number of one type of secondary

structure in the whole database. $Total_{TE}$ is the total number of amino acids in one thermodynamic environment. $Total_{Residues}$ is the total number of residues in the whole database.

Secondary structure prediction identity calculation

Positions of each residue (represented by thermodynamic environment) were assigned a secondary structure according to the log-odd probabilities of secondary structures in thermodynamic environments. For example, if α-helices have a high log-odds probability in thermodynamic environment 1, residues in this environment are classified as α-helices. Secondary structure assignment for residues in original structure was assigned using STRIDE (Frishman and Argo, 1995). Identity is calculated as the percentage of matched positions divided by total number of residues in original structure. Coil and turn were assigned individually, but categorized in one group in the identity calculations as is usually reported by other secondary structure prediction algorithms.

CHAPTER 6

Investigating the Roles of Structure and Sequence in the Local Energetics in the Denatured State Ensemble

Introduction

The increasing evidence of residual structure (native-like or non native like) present in the denatured state (Dill and Shortle, 1991, Mok et al., 1998) suggest that they may be guiding points for the folding process. To investigate how the denatured state controls the protein folding process thermodynamically, the relationship between structure in the denatured state and the local energetics must be established. The contributions of inherent sequence properties to local energetics of the denatured ensemble should also be explored since the fold defining information is contained within the primary sequence. Identifying the relationship between sequence and the energetics of the denatured ensemble will help us understand the role of the denatured state in folding process. The results from chapter 4 and 5 indicate that the denatured state ensemble can be viewed as a sort of thermodynamic signature containing significant fold encoding information, and that the majority of positions that are correctly aligned are in regions of containing α-helices and β-strands. To investigate whether this result was simply due to our model of the denatured state (i.e. the denatured ensemble is comprised of states with isolated segments of native-like structure), we

systematically explored the importance of alternative denatured state conformations.

Results

Role of Non-native Structure in the Local Energetics of the Denatured State

To investigate the importance of the local structure features (i.e. α-helix, β–sheet or turn) to the energetics at each position, we utilized a previously reported algorithm which generates self-avoiding conformations by randomly sampling backbone and torsion angles (Whitten et al. In Press). For this analysis, a subset of 12 proteins (DATASET1) (Table 6.1) was selected from the entire database so that each structural class (i.e., all alpha (all-α), all beta (all-β), alpha and beta (α+β), and small proteins (small)) had 3 representative members. Multiple random, self-avoiding chains were generated for each sequence and the stability of each local segment of structure (relative to an ensemble of disordered conformations) was calculated. The averages and standard deviations over all conformations for each sequence are shown for each structural class (Figure 6-1 is for all alpha (all-α), Figure 6-2 is for all beta (all-β), Figure 6-3 is for alpha and beta (α+β), and Figure 6-4 is for small proteins (small)). Interestingly, the stability profiles from denatured ensembles (assuming only native-like conformations in structured regions) are similar to those determined from the cases where the denatured state is generated from actual random conformations. Inspection of equation 6.1 (See Material and Methods) reveals the origin of this behavior. Because the thermodynamics of the unfolded sub-ensemble for each residue j,

$<\Delta G_{nf,j}>$, is dominated by the probability of the completely unfolded state, P_U, it is determined from the additive contribution of the individual unfolded state values for each amino acid (Tables 6.2A & 6.2B)(D'Aquino et al., 1996, Hilser et al. 1996, Hilser et al., 2006, Lee et al., 1994). As such the unfolded sub-ensemble for each amino acid is identical. The similarities in the energy profiles calculated from different denatured state structures indicates that the differences in stability between the different regions of the sequence are far greater than the stability differences between each alternative conformation of a specific region. In other words, the peaks and valleys that are visible in each sequence provide an ensemble-averaged 'foldability' metric (i.e. the probability of finding that residue in the context of its sequence neighbors in a unique conformation, relative to being disordered).

Although the negative sign for $\ln\kappa_f$ in figures 6-1 – 6-4 indicates that the denatured ensemble is dominated by a broad conformational repertoire at every position, there are nonetheless, significant position-specific differences in the foldability metric. For example, the difference between the denatured state stability at positions 33 and 40 of the protein 1KTH (Figure 6-5) reveals that regardless of the specific conformation, position 40 is highly unlikely to adopt a single structure as compared to position 33, and will instead populate a broader ensemble. In other words, the combined probability for the ensemble of alternative conformations at position 40 is far greater than at position 33, and ensures that the ensemble will be distributed among many states.

Table 6-1: *Homo sapiens* proteins in DATASET1

PDB	Length	SCOP class	SCOP family
1I27	69	All alpha	C-terminal rap74 subunit
1I2T	61	All alpha	PABC (PABP) domain
1L9L	74	All alpha	NKL-like
1FNA	89	All beta	Fibronectin type III
1LDS	96	All beta	C1 set domains
1TEN	89	All beta	Fibronectin type III
1ESR	75	Alpha and beta	Interleukin 8-like chemokines
1MJ4	79	Alpha and beta	Cytocrome B5 Sulfite Oxidase
1MWP	96	Alpha and beta	A heparin-binding domain
1KTH	58	Small	BPTI-like
1I71	83	Small	Kringle modules
1M9Z	104	Small	Extracellular domain, cell surface

Table 6.2A Amino acid denatured state properties

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE
$ASA_{ex,apol}$ ($Å^2$)[1]	70.0	87.1	38.1	42.1	30.3	65.0	71.1	26.2	90.0	110.7
$ASA_{ex,pol}$($Å^2$)[1]	36.1	126.1	104.0	95.0	75.05	121.6	94.3	43.12	68.0	10.9
S_{sc}[2] ($cal \cdot mol^{-1} \cdot K^{-1}$)	0	-0.84	2.24	2.16	0.61	2.12	2.27	0	0.79	0.67
S_{bb}[3] ($cal \cdot mol^{-1} \cdot K^{-1}$)	4.1	3.4	3.4	3.4	3.4	3.4	3.4	6.5	3.4	2.18

Table 6.2B Amino acid denatured state properties

	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYP	VAL
$ASA_{ex,apol}$($Å^2$)[1]	122.3	101.3	104.6	186.8	100.8	55.1	79.5	184.5	175.8	88.7
$ASA_{ex,pol}$($Å^2$)[1]	27.5	79.0	64.0	36.5	15.6	81.9	41.1	52.3	71.1	17.8
S_{sc}[2] ($cal \cdot mol^{-1} \cdot K^{-1}$)	0.25	1.02	0.58	1.51	0.0	0.55	0.48	1.15	1.74	1.29
S_{bb}[3] ($cal \cdot mol^{-1} \cdot K^{-1}$)	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4	3.4	2.18

(1) Solvent accessible apolar ($ASA_{ex,\ apol}$) and polar ($ASA_{ex,\ pol}$) surface area for each amino acid in the denatured state (Lee et al., 1994, Hilser and Feire, 1996, Hilser et al. 2006. Murphy et al., 1992). (2) Side chain conformational entropy differences (ΔS_{sc}) between the completely unfolded state and the state in which each residue is folded. This corresponds to the $\Delta S_{ex\text{-}u}$) previously determined and applied as described before (Hilser and Feire, 1996). (3) Backbone conformational entropy differences (ΔS_{bb}) between the completely unfolded state and the state in which each residue is folded.

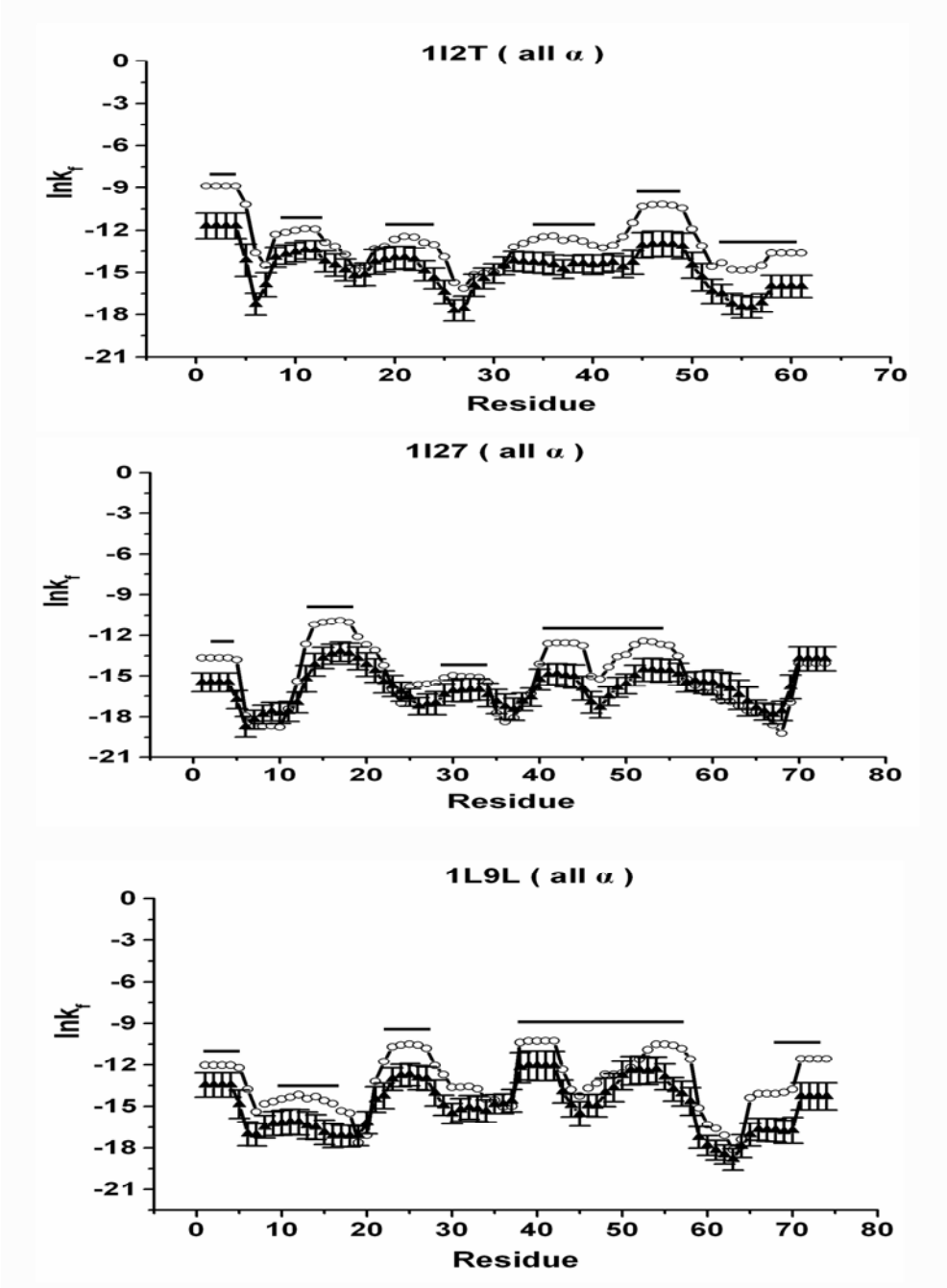


Figure 6-1: Examining the structural effects to calculated stability constants (lnk_f) using denatured state ensemble for alpha proteins

Three α proteins (1I2T, 1I27 and 1L9L) were selected and calculated lnk_f (open circles). These values were compared to the null model where structures were randomly generated for subsequent calculation of the stability constant (RAND_3D, closed triangles with error bars). Regions of α helices are highlighted with a black bar.

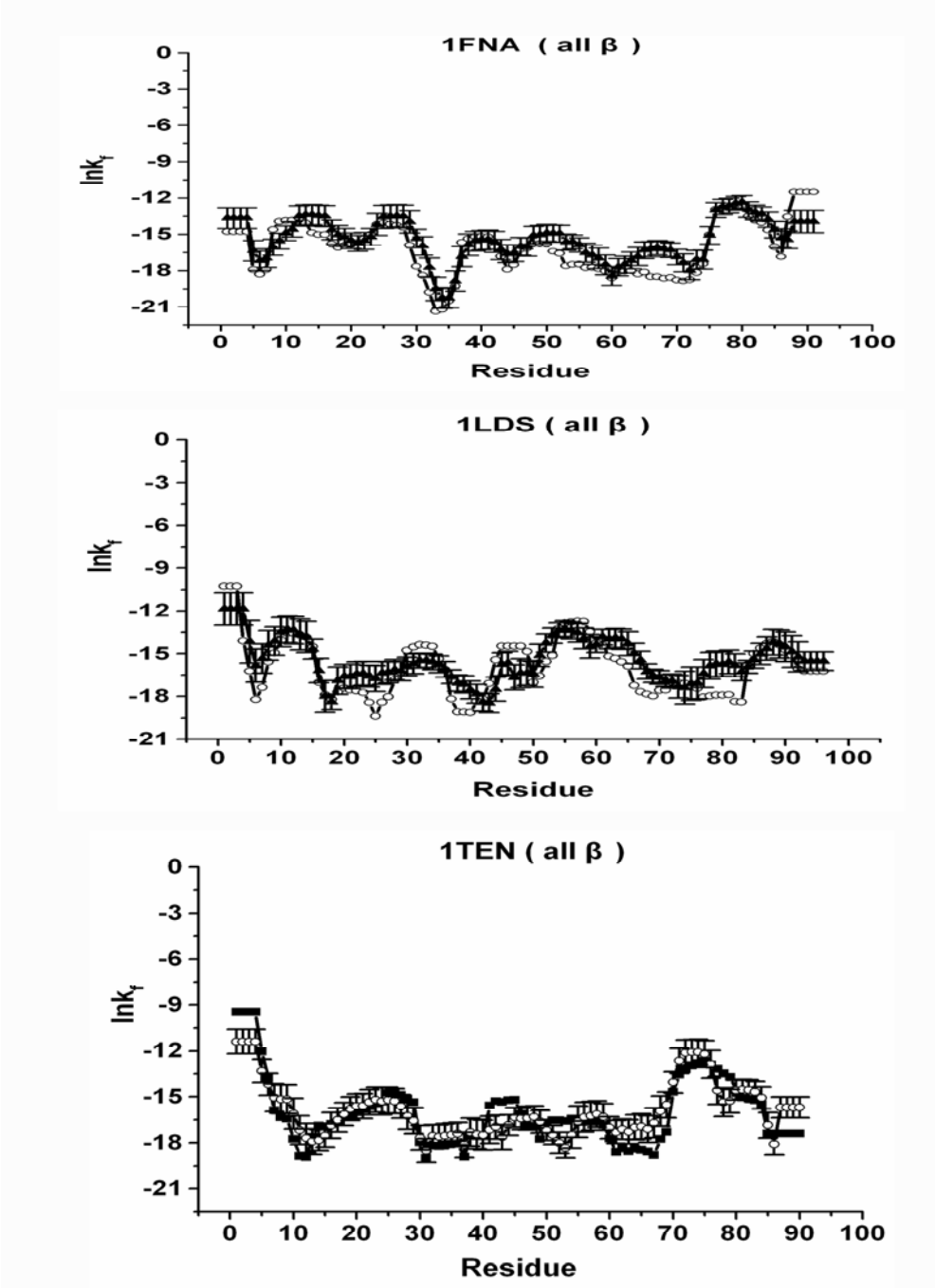


Figure 6-2: Examining the structural effects to calculated stability constants (lnk_f) using denatured state ensemble for beta proteins

Three β proteins (1FNA, 1LDS, and 1TEN) were selected and calculated lnk_f (open circles). These values were compared to the null model where structures were randomly generated for subsequent calculation of the stability constant (RAND_3D, closed triangles with error bars).

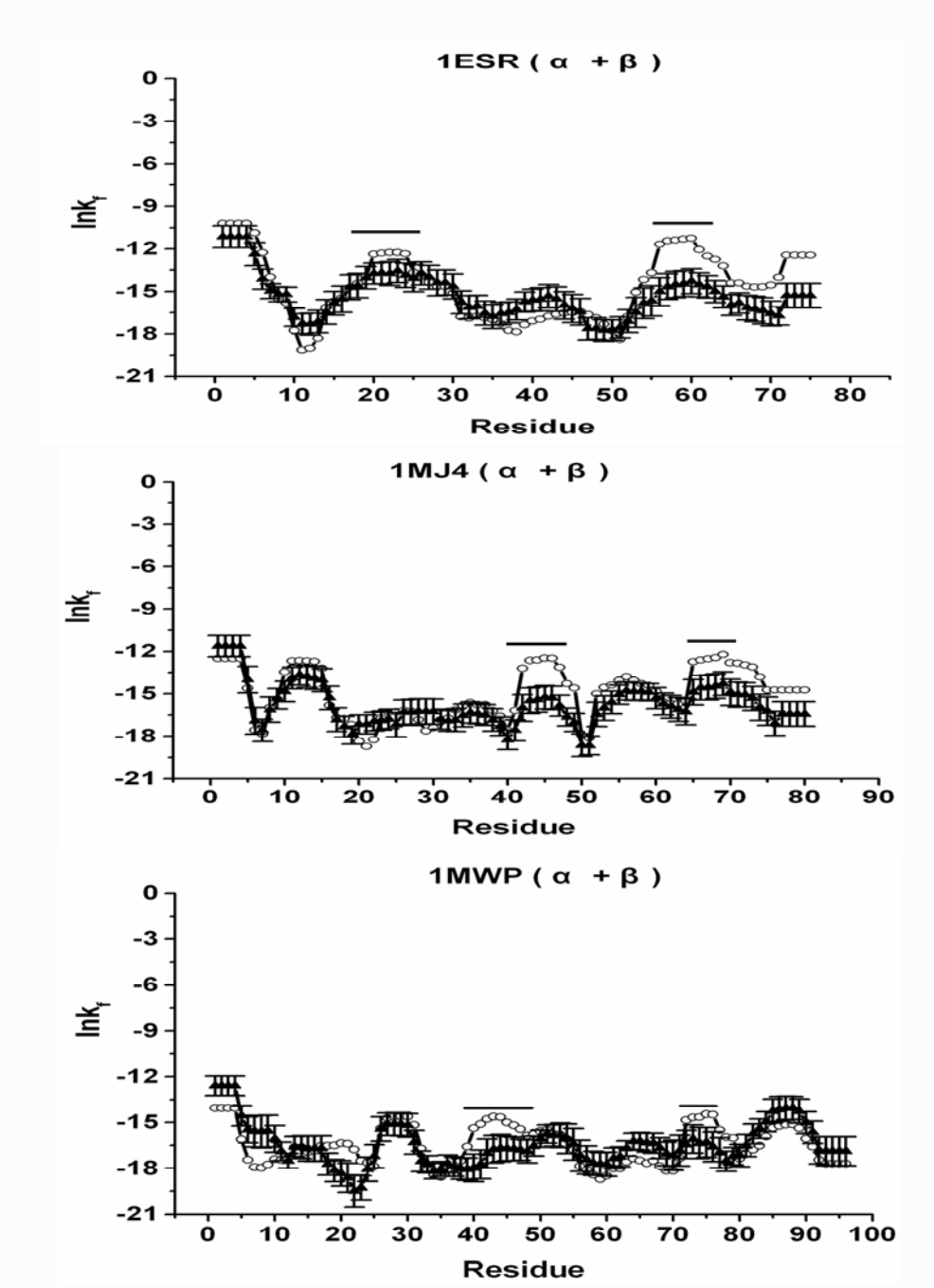


Figure 6-3: Examining the structural effects to calculated stability constants (lnk_f) using denatured state ensemble for alpha + beta proteins

Three $\alpha + \beta$ proteins (1ESR, 1MJ4, and 1MWP) were selected and calculated lnk_f (open circles). These values were compared to the null model where structures were randomly generated for subsequent calculation of the stability constant (RAND_3D, closed triangles with error bars). Regions of α helices are highlighted with a black bar.

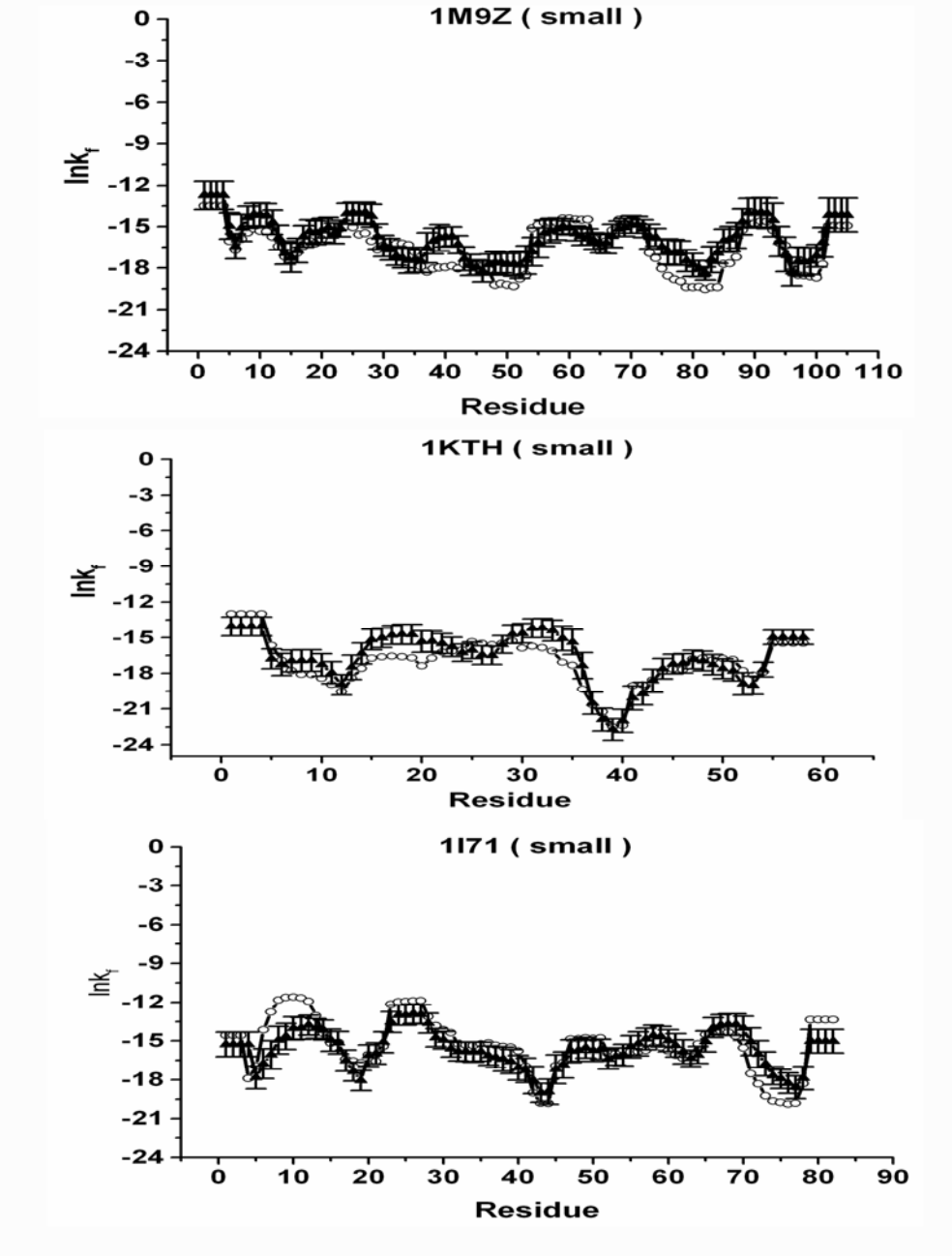


Figure 6-4: Examining the structural effects to calculated stability constants (lnk_f) using denatured state ensemble for small proteins

There small proteins (1M9Z, 1KTH, and 1I71) were selected and calculated lnk_f (open circles). These values were compared to the null model where structures were randomly generated for subsequent calculation of the stability constant (RAND_3D, closed triangles with error bars).

Role of Sequence in Local Energetics of the Denatured State

The sequence contribution to the observed residue stabilities for each secondary structural class was investigated by comparing to the thermodynamic signature from the actual sequence with signatures from sequences that have been randomized. Stability constants calculated from actual sequence and randomized sequences are shown for each structural class (Figure 6-6 is for all alpha (all-α), Figure 6-7 is for all beta (all-β), Figure 6-8 is for alpha and beta (α+β), and Figure 6-9 is for small proteins (small)). Several observations can be made from this comparison. First, the sequence composition determines the mean stability for each protein with little deviation from this mean, even when the sequence has been shuffled several times (see Materials and Methods) (Figure 6-10). The difference between the mean value of the cases where just the conformation was shuffled and the case where both the sequence and conformation were shuffled is not statistically significant (p=0.7683). Second, the sequence order impacts the variance of residue stabilities within the protein sequence (p=0.03) indicating that neighboring residues have significant stabilizing and destabilizing contributions(Figure 6-11) and that the thermodynamics that are calculated at each position are not simply reporting on the properties of the individual amino acids.

Discussion

The roles of non-native structures and sequences in local energetics of denatured states are investigated. Though different conformations show similar stability profiles, there are still significant differences in the position-specific probabilities of adopting a unique conformation. Also the mean stability for each protein is independent of their structural class (Figure 6-10). Sequences regions

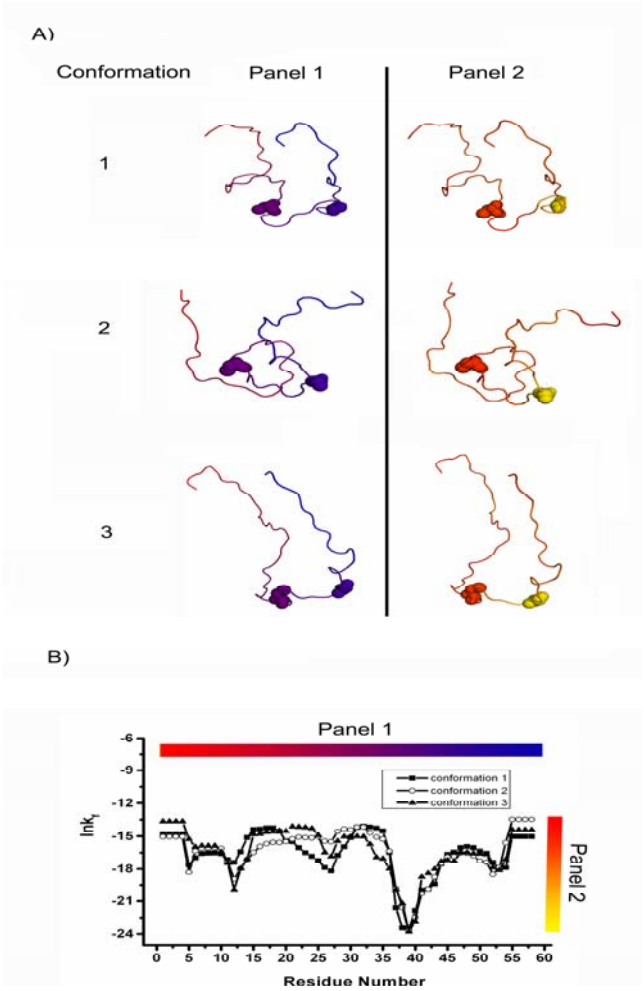


Figure 6-5: Random generated three structures of small Kunitz-type inhibitors protein (PDB: 1KTH) and the stability constants under denatured conditions

(A) Three random generated structures. Panel1 shows conformations colored by residues. Panel 2 show conformations colored by stability constants values. Clearly, three randomly generated structures show different conformations. (B) Stability constants for three conformations under denatured conditions. Clearly, three randomly generated structures show a similar stability profile.

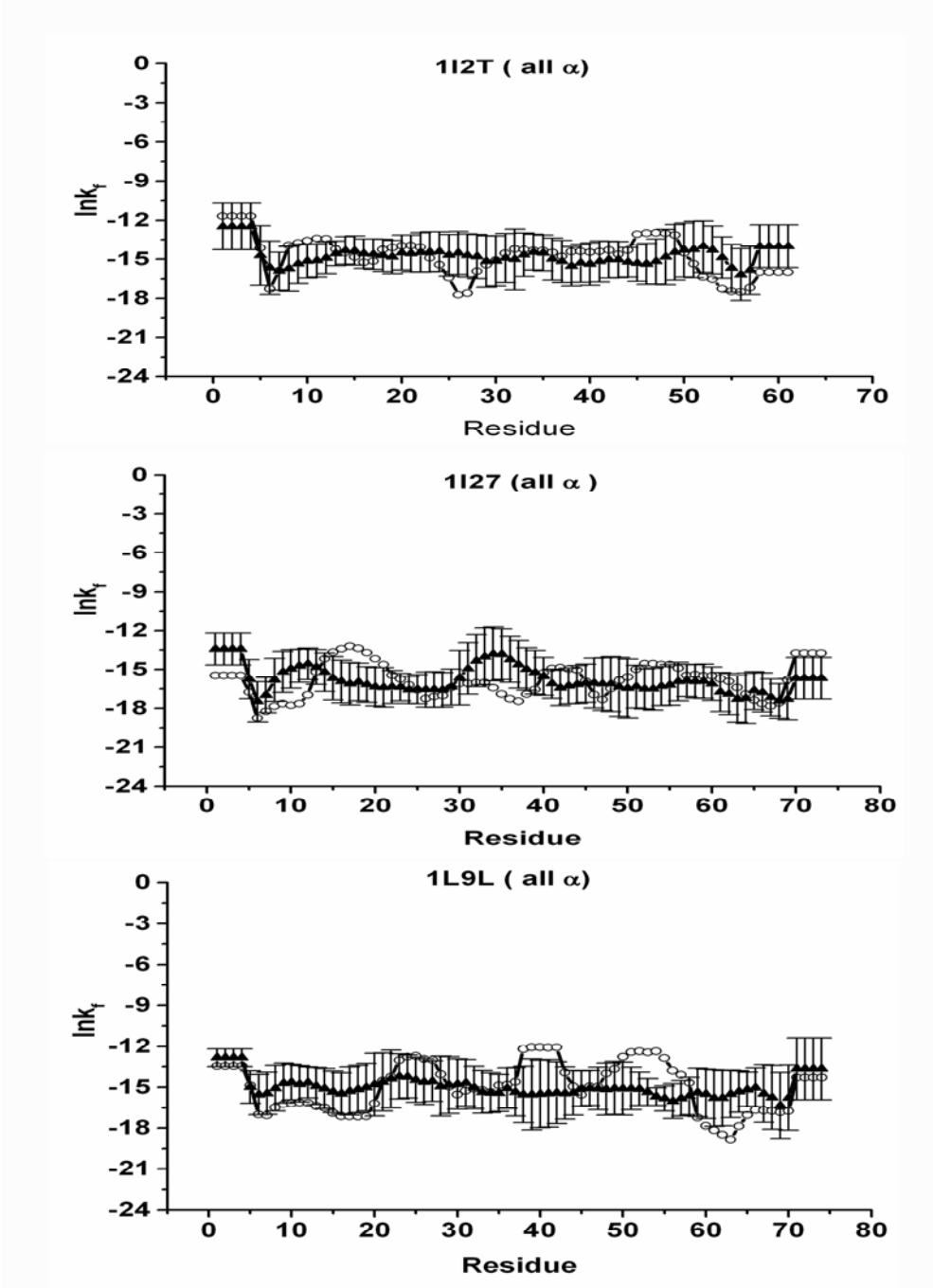


Figure 6-6: Examination of sequence contribution to the stability constant (lnk_f) in alpha proteins

Three alpha proteins (1I2T, 1I27 and 1L9L) were selected. Changes in stability constant was investigated between proteins with randomly generated structures (RAND_3D, open circles) and randomly generated structures + randomly shuffled sequences serving as the second null model (RAND_3DSEQ, closed triangles with error bars). The same proteins from DATASET1 were used.

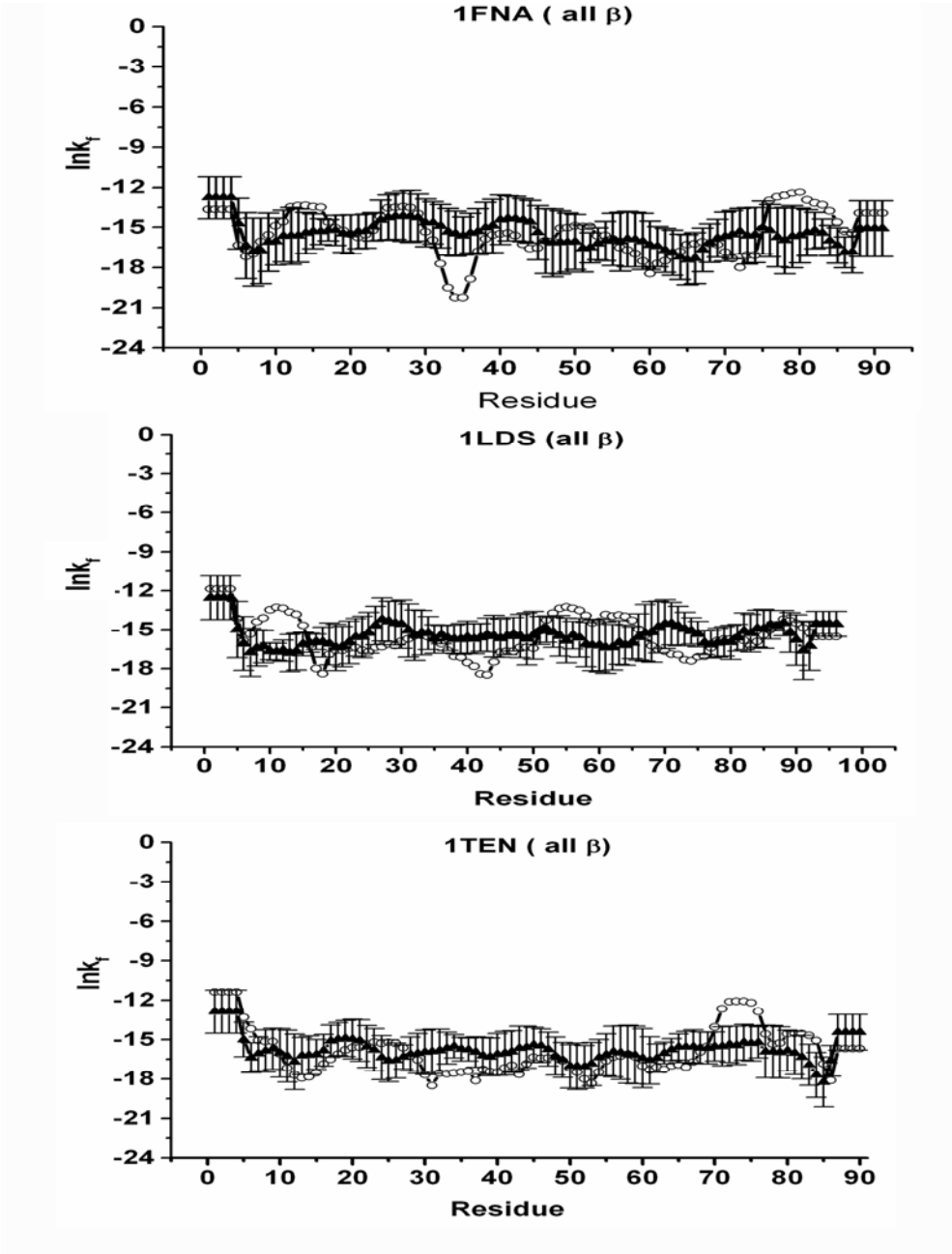


Figure 6-7: Examination of sequence contribution to the stability constant (lnk_f) in beta proteins

Three beta proteins (1FNA, 1LDS and 1TEN) were selected. Changes in stability constant was investigated between proteins with randomly generated structures (RAND_3D, open circles) and randomly generated structures + randomly shuffled sequences serving as the second null model (RAND_3DSEQ, closed triangles with error bars). The same proteins from DATASET1 were used.

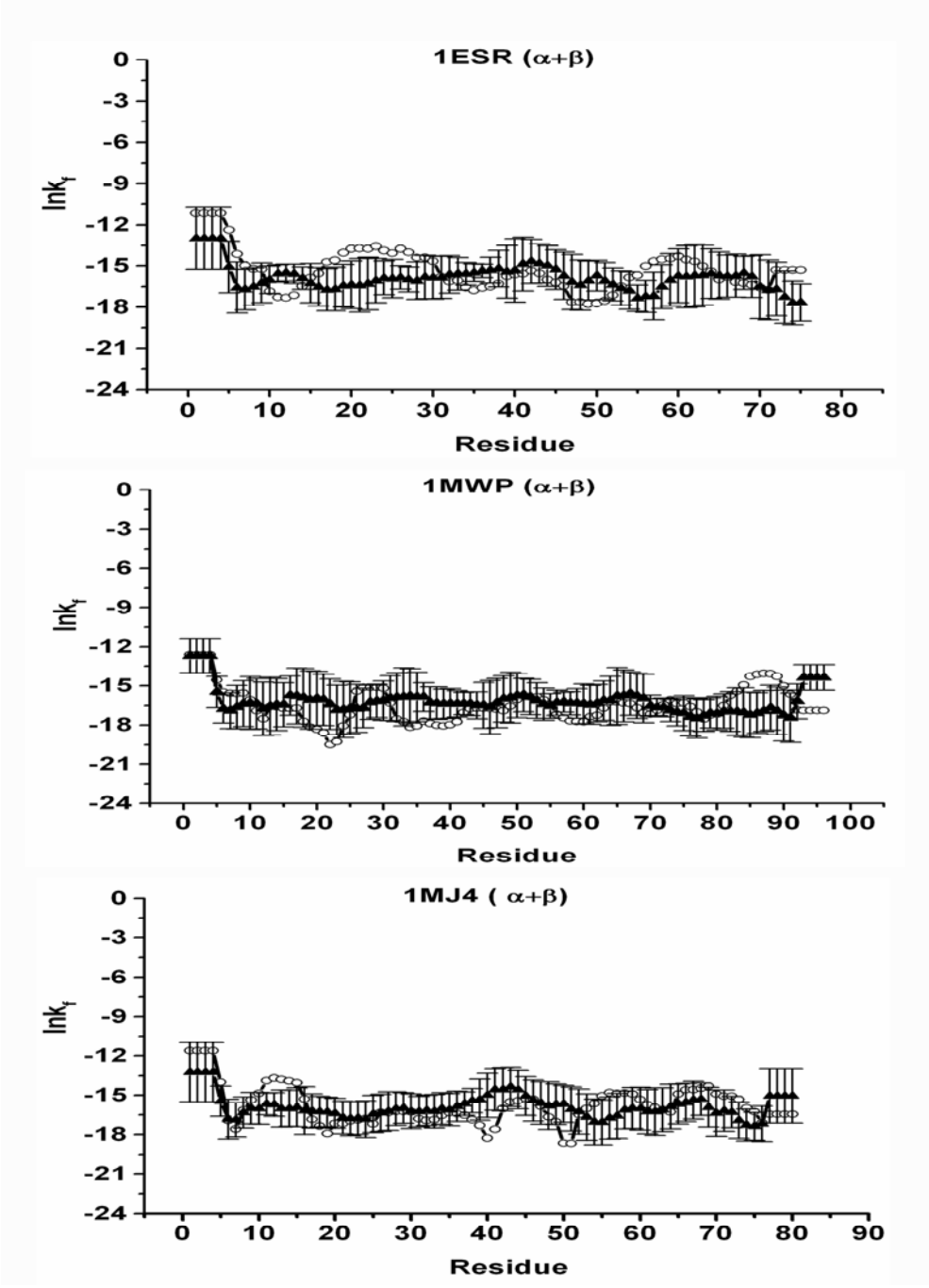


Figure 6-8: Examination of sequence contribution to the stability constant (lnk_f) in alpha + beta proteins

Three alpha + beta proteins (1ESR, 1MWP and 1MJ4) were selected. Changes in stability constant was investigated between proteins with randomly generated structures (RAND_3D, open circles) and randomly generated structures + randomly shuffled sequences serving as the second null model (RAND_3DSEQ, closed triangles with error bars). The same proteins from DATASET1 were used.

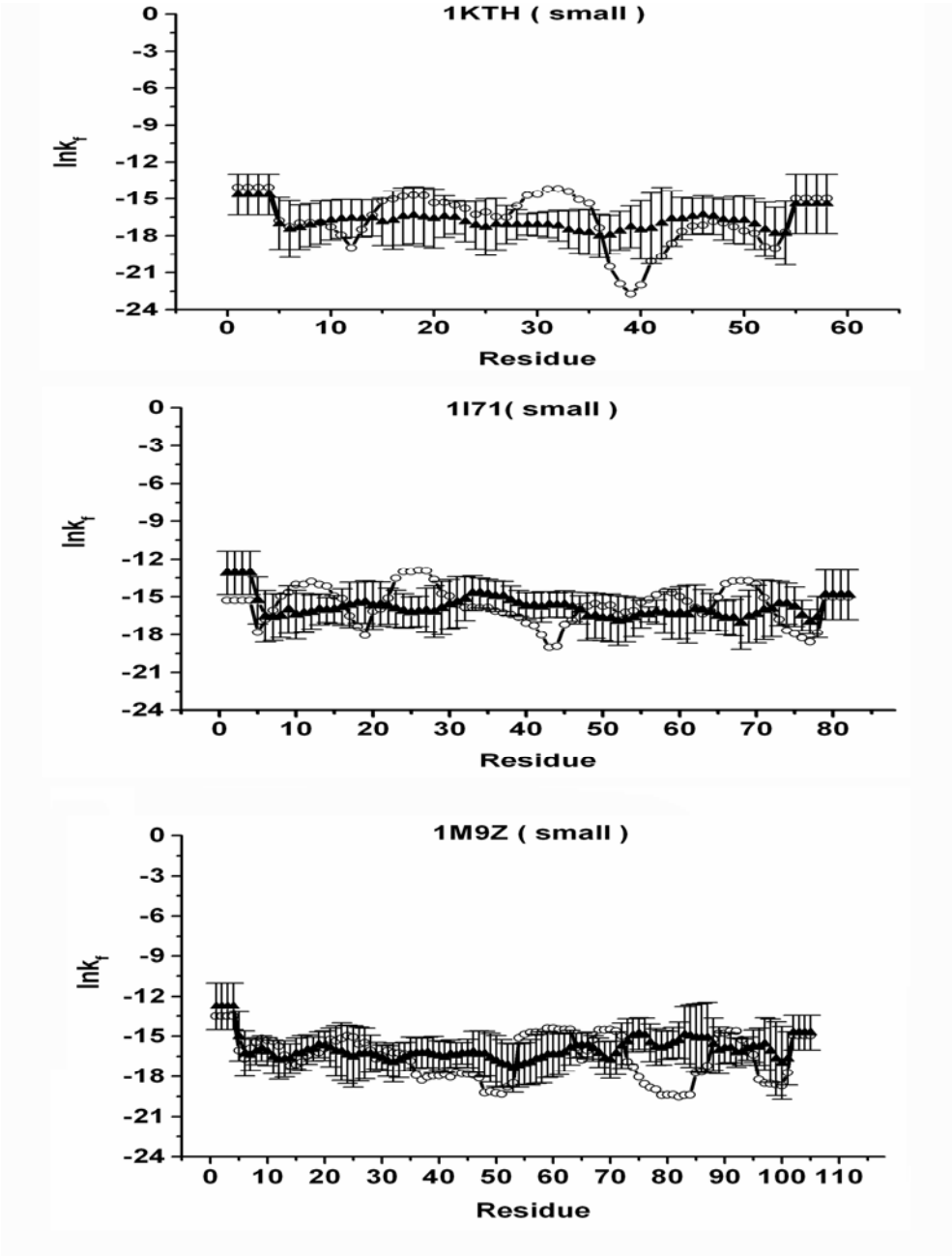


Figure 6-9: Examination of sequence contribution to the stability constant (lnk_f) in small proteins

Three small proteins (1KTH, 1I27 and 1M9Z) were selected. Changes in stability constant was investigated between proteins with randomly generated structures (RAND_3D, open circles) and randomly generated structures + randomly shuffled sequences serving as the second null model (RAND_3DSEQ, closed triangles with error bars). The same proteins from DATASET1 were used.

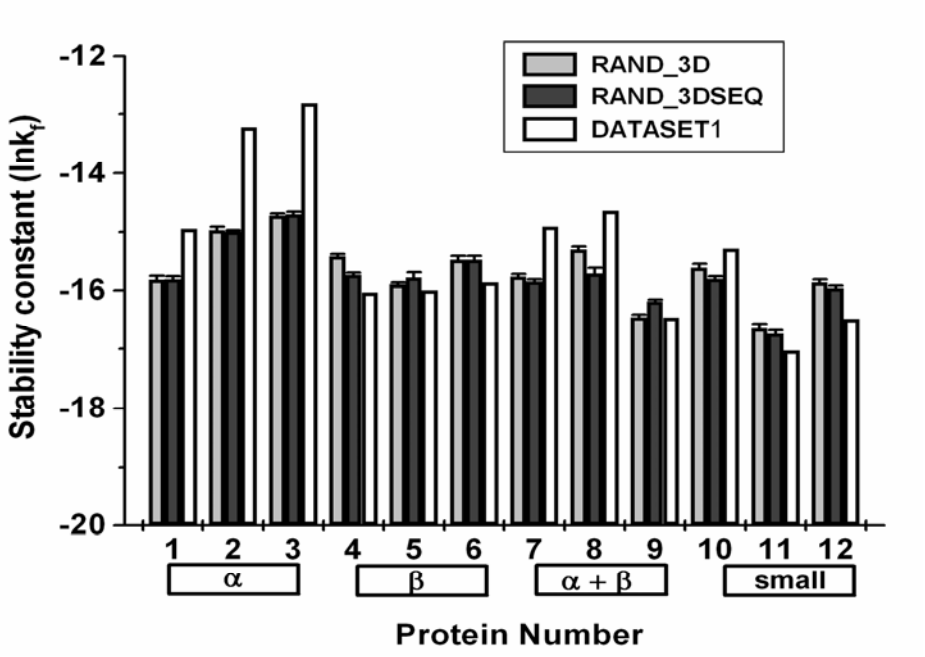


Fig. 6-10: Sequence composition affects the mean of stability constants

The mean stability constant (lnk_f) of each protein and DATASET1 (white), RAND_3D (grey) and RAND_3DSEQ (black) is calculated. X-axis is the 12 proteins in DATASET1 labeled from 1 to 12. Y-axis is the calculated mean stability constant for each protein.

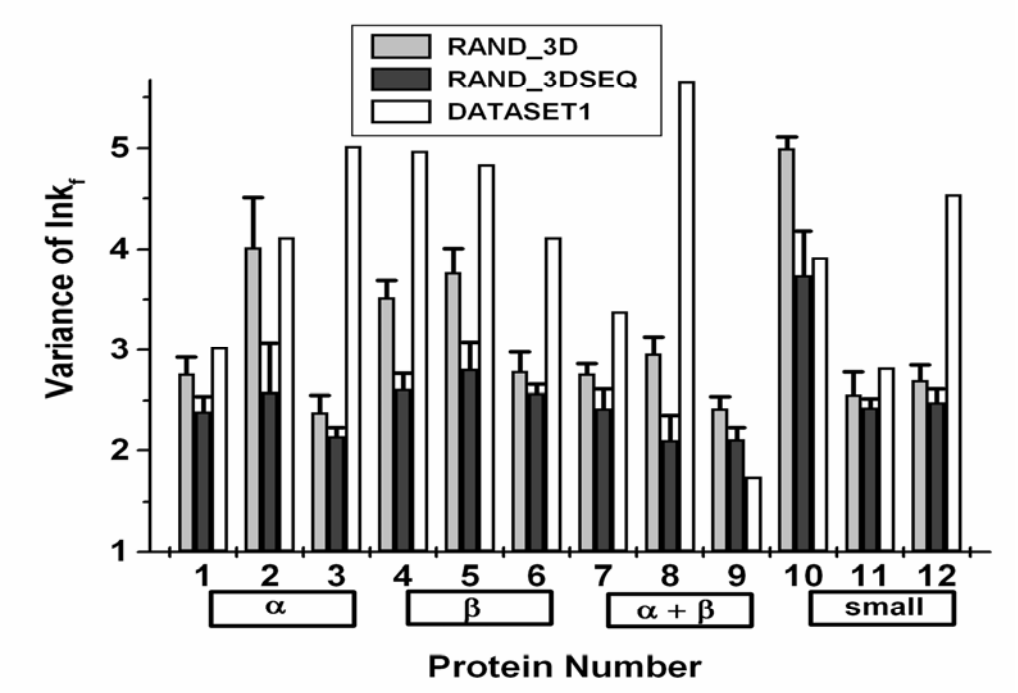


Fig. 6-11: Sequence order affects the variance of stability constants

The variance of stability constants (lnk_f) of each protein and DATASET1 (white), RAND_3D (grey) and RAND_3DSEQ (black) is calculated. X-axis is the 12 proteins in DATASET1 labeled from 1 to 12. Y-axis is the calculated variance of stability constants for each protein.

with alpha helix forming propensities show higher stabilities (more stable), but this observation pertains to the protein's denatured state and does not necessarily suggest that all-α proteins have a higher mean stability over other structural classes. Thus the stabilizing contribution from alpha helical formation applies locally and does not necessarily lead to an overall improvement in global stability. Local energetics in denatured states also shows substantial sequence dependence. The dependence of the mean stability on the composition of the sequence shows that proteins in the denatured state are not thermodynamically equal. Regional stabilities are also locally modified by the sequence order of amino acids in the local context, another example supporting that Flory's isolated-pair hypothesis does not hold true (Pappu et al., 2000) and that residue conformations are not independent of each other. Understanding these effects will help facilitate our understanding of how mutations impact the denatured state of the protein that is also intrinsically linked with the natively folded proteins.

Materials and Methods

Selection of proteins used in dataset

1) DATASET1

A sub dataset contains 12 selected proteins from the entire database described in Chapter 2 to allow for a more in depth analysis of sequence and structural contributions to the observed stability. Three proteins were selected for each of four common SCOP class category (all α, all β, α + β, and small) to construct a representative dataset.

2) RAND_3D (NULL MODEL1)

The null model used to investigate the structural contribution to the calculated energetics was generated for DATASET1 was called RAND_3D. Random conformations were generated by a program called MPMOD for each protein in DATASET1.

3) RAND_3DSEQ (NULL MODEL2)

The null model used to investigate the sequence contributions to the calculated energetics was generated for DATASET1 was called RAND_3DSEQ. Each protein in DATASET1 have randomly shuffled sequence and randomly generated conformers, also using MPMOD. Denatured ensemble energetics are calculated for each generated protein in RAND_3DSEQ using COREX, assuming the randomly generated conformer as the "native state x-ray structure". The energetics of RAND_3D was compared to this second null model to investigate the sequence contribution to the stabilization of the region. Sequences were randomly shuffled 10 times and 1 random conformer was generated for each randomized sequences.

Generated Random Conformations

Random conformations were generated by a program called MPMOD (Whitten et al., In press) for each protein in DATASET1. As shown in Figure 6.12, MPMOD generates random conformations that do not violate van der Waals radii

limits. First, the backbone φ, ψ, and ω angles are randomly selected, then the algorithm checks to see if the chain is self-avoiding. If there are no violations, rotamers are randomly selected from a rotamer library (Lovell et al., 2000) for each residue type to build the side chain. The atomic positions of the side chains are then checked for violations with limits of allowable van der Waals interactions. Denatured ensemble energetics were then calculated for these generated conformations in this dataset with COREX. 50 random conformers were generated for each protein.

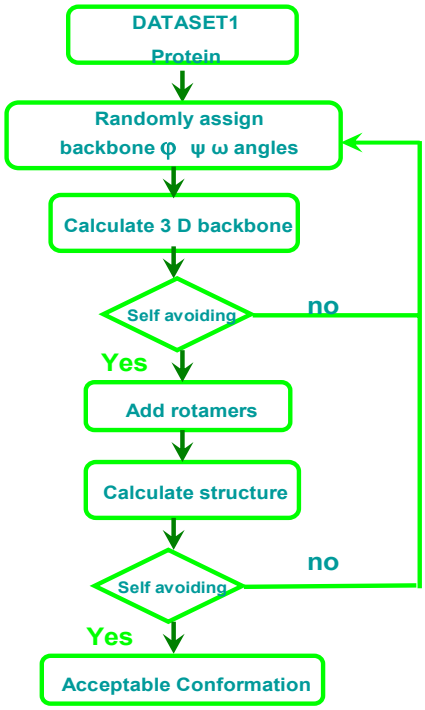


Figure 6-12: Schematic representation of MPMOD procedure on generating random conformations

Sequence of each protein in DATASET1 was used as input for MPMOD. Backbone angles were randomly selected and were calculated no violation of self avoiding. Rotamers are randomly selected from a rotamer library for each residue type to build the side chain. Final conformation was generated after no violation of self-avoiding was constraint was achieved.

CHAPTER 7

Probing the Role of the Denatured State in Protein Folding

Introduction

Although the protein folding problem has been recognized for decades, a detailed understanding of the actual folding process does not yet exist. Different protein folding models have been proposed to reflect the different insights on the folding process. The first protein folding model, the framework model, was proposed by Ptitsyn in 1973 and was later further refined (Ptitsyn, 1995). This model implies that individual amino acid residues and their local interactions with neighbors define the folding code. The model states that the formation of secondary structures is the cause for the assembly of the final structure, thus emphasizing the importance of the surrounding environment of each amino acid as determinants of the final fold. However, the competing hydrophobic collapse model argues instead that the non-local hydrophobic interactions are the driving force for the formation of final fold (Dill et al., 1995). Partially in agreement with the framework model, the nucleation model emphasizes the importance of strong and localized nucleus (native-like elements of secondary structure) in the formation of final structure (Wetlaufer, 1990). Particular models aside, the protein folding process involves ensembles from both fully folded and unfolded states; therefore, investigations focused on only the folded state will incompletely decipher the protein folding problem. Therefore, unveiling of the role of denatured

states in folding process will provide novel insights that will help complete the current understanding of the protein folding process.

In the previous chapters, the energetic information in denatured states were characterized and identified. The roles of denatured states in protein folding will then be investigated in this chapter and the relevant implications about the denatured states are discussed.

Results

Propensity of Secondary Structure in Thermodynamic Environments of the Denatured State Ensemble and Structural Forming Capacity

Figure 7.1 is a summary figure of structural effects on the stability of the denatured state ensemble. A significant observation from Figure 7.1 is that α-helix segments (labeled with black bar) are more stabilized in the protein when the model of the denatured ensemble accounts for only native-like conformations. The origin of this difference is that the structure of the α-helix is significantly more compact and stable than the randomly selected conformations, and it represents a very narrow region of the conformational space accessed by the random sampling of states. Nonetheless, it is noteworthy that in spite of this built-in bias, sequence segments destined for α-helix show peaks in the structure forming propensity, even when randomly generated structures are used. This result indicates that sequence segments which are destined for α-helix have a comparatively low energetic cost associated with constraining the ensemble to a unique structure. To investigate the role of denatured state in protein folding, propensities of secondary structures in each thermodynamic environment were re-inspected. Figure 7-2 is the average stabilities of eight thermodynamic

environments in denatured ensemble ordered from high to low and propensities of secondary structures in those eight environments were also shown. As Figure 7-2 indicates, the only positive propensities that are found in the most stable environments (i.e., $TE1_D$ and $TE4_D$) are those segments destined for α-helix. Sequences destined for all other secondary structure have low intrinsic structure forming capability (i.e., they are represented by troughs in their $\ln k_f$ values; Figure 7-2). Interestingly, sequence segments with high intrinsic structure forming capability (i.e. peaks in their $\ln k_f$ values; Figure 7-2) have among the highest negative propensities for β–sheet. In other words, relative to all other secondary structure, β–forming sequences characteristically favor high conformational degeneracy when in isolation.

Is the Denatured State Poised to Minimize Unfavorable Folding?

The framework model emphasizes the roles of local environment of amino acid and the secondary structure elements in determining the final fold. Interestingly, what we found here is that local structural forming capacities of amino acid sequence correlated well with the surrounding thermodynamic environment for each amino acid residue in the primary sequences. This correlation between the structural and thermodynamic characteristics of the denatured state reveals interesting and previously unreported trends for the denatured state, and these trends further support a framework model of protein folding (Udgaonkar et al., 1988, White et al., 2005). The role of the denatured state in protein folding is proposed as shown in Figure 7-3. As indicated in Figure 7-3, the denatured state is predicted to be macroscopically heterogeneous, with the propensity for any single structure being highly improbable across the entire

sequence. Within this background many regions, particularly those destined for α-helix or coil, will flicker (in the context of small isolated segments) into the folded conformation far more often than those regions destined for other secondary structures (Figure 7-3). Most important, however, is that our model provides a statistical picture which reveals that regions destined for β–sheet will form unique local structure much less often than random. In effect, the denatured state thermodynamics (particularly with regard to β and α- helical structures) are dominated by strong negative propensities (Figure 7-1).

Discussion

The role of denatured states in protein folding is investigated in this chapter. Although our studies do not establish the underlying reasons why protein denatured states have evolved with these propensities, there is at least one plausible hypothesis. Because β–strands interact with other β–strands through backbone hydrogen bonding, the potential for partnering with incorrect strand formation is relatively high. α-helices, however, presumably exclude potential non-specific backbone interactions through the formation of local i to i+3 hydrogen bonds of the helix. As a consequence, most of the favorable (and unfavorable) interactions between helical regions can in principle be controlled or modulated through individual site mutations, as they will involve mostly side chain interactions. Controlling for incorrect β-strand pairing, on the other hand, will be less amenable to modulation through single site mutation, and would presumably require a more global solution. Our results provide insight into such a solution. The thermodynamic architecture of the denatured state indicates that the denatured ensemble is biased in a way that minimizes the probability of

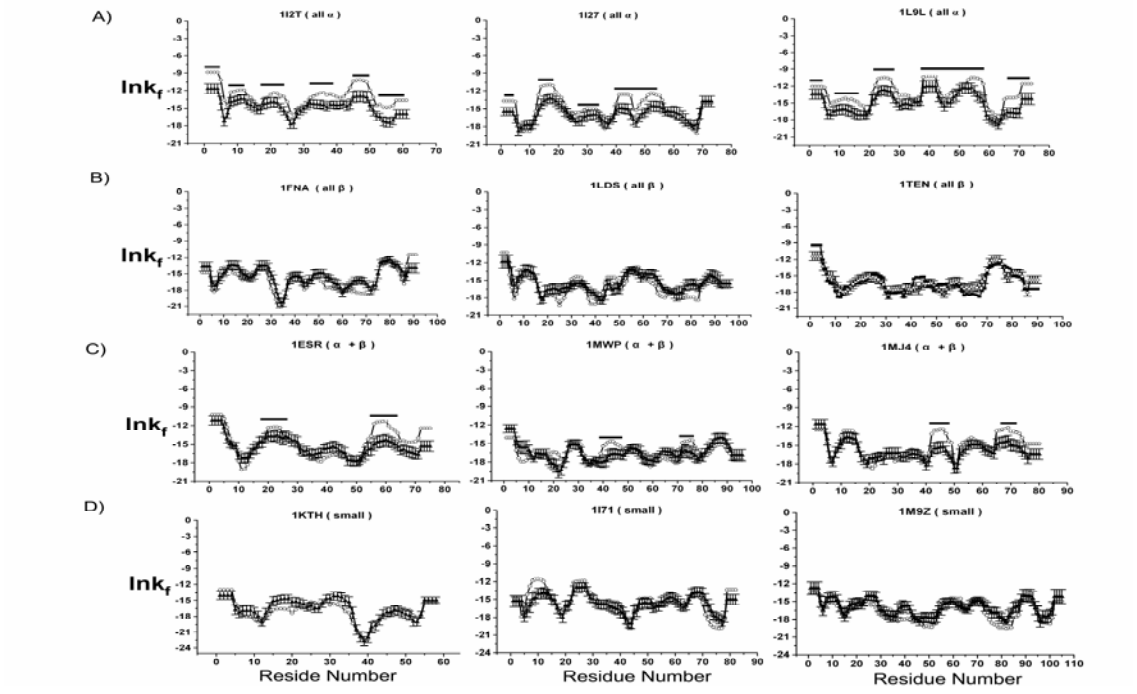


Figure 7-1: Examining the structural effects to calculated stability constants (lnk_f) using denatured state ensemble
Twelve proteins, three from each structural class ((A) all alpha (B) all Beta (C) alpha + beta (D) small) were randomly selected (DATASET1, open circles). These values were compared to the null model where structures were randomly generated for subsequent calculation of the stability constant (RAND_3D, closed triangles with error bars). Regions of alpha helices are highlighted with a black bar.

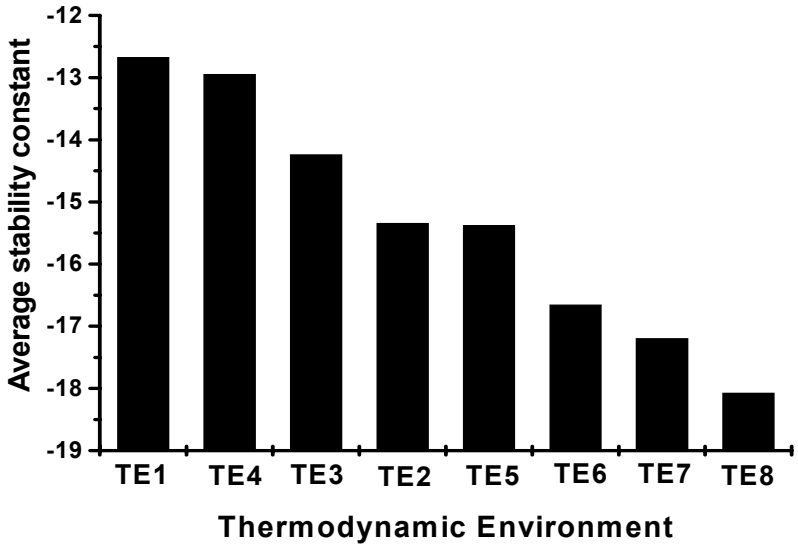


Figure 7-2: Average stabilities of eight thermodynamic environments under denatured conditions

Eight thermodynamic environments (TEs) were ranked from high stability to low stability on X-axis. Stability of each thermodynamic environment was represented by black bar. TE1 and TE4 are two thermodynamic environments with high stabilities, TE7 and TE8 are two thermodynamic environments with low stabilities.

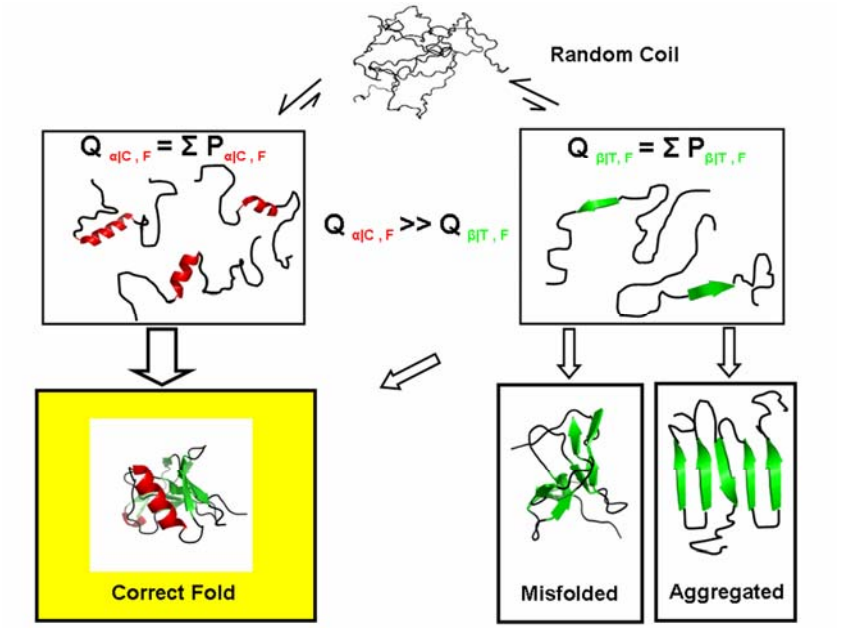


Figure 7-3: Schematic representation of the denatured state energy landscape

Shown is a hypothetical unfolded protein (top), which is depicted as having no structural propensity. The strong negative bias for α-structure formation coupled to the modest propensity for a-helix and turn structure formation, suggests the that the sub-partition function for states involving isolated folded segments of helix and coil ($Q_{\alpha|C}$; Left) is significantly higher than the sub-partition function for states where isolated segments of β-strand ($Q_{\beta|T}$; Right) are folded. This pre-collapse equilibrium does not obligatorily signify that nucleation between different parts of the structure subsequent folding occurs through helix and coil, only that those segments in isolation have high folding probabilities.

equilibrium states that could promote folding to non-productive end states (Fig. 12).It is interesting that protein mis-folding into amyloid fibrils has been associated with β structure formation, indicating that non-specific β structure is indeed a potential problem. The fact that our results suggest that the denatured states evolved to minimize this problem raises the possibility that the determinants of amyloid propensity for a sequence may be found in the denatured state thermodynamics, rather than in the properties of the native state. Whether this is indeed the case awaits further study.

CHAPTER 8

Concluding Remarks

A number of original and significant contributions have been made by the presented work. The contributions can be summarized as the first application of COREX towards the analysis of the denatured state ensembles of multiple proteins in a human protein database revealing significant thermodynamic differences between sequence segments within the denatured ensemble, as well as between denatured and native states. The correlation between denatured states and secondary structures and the crucial roles of denatured states in the protein folding process has also been identified. The contributions from this work has opened new opportunities for future study of the protein folding problem with potential applications to uncover novel strategies for the treatment of amyloid fibrils associated with protein misfolding diseases.

Although the importance of the protein denatured state has long been recognized, the thermodynamic role of the denatured states in protein folding remains elusive. In chapter 1, the importance of investigating the denatured state was emphasized and the specific aims of the study in this dissertation were proposed to capture the common characteristics of the denatured state across multiple proteins in *Homo sapiens* database.

Simulation of the denatured state and characterization of associated energetic features across multiple proteins was not an easy task. Chapter 2 described the initial effort to develop a strategy for simulating and characterizing

the denatured state using the statistical thermodynamic model COREX. The good correlation between experimental data and COREX calculations demonstrated the robustness and efficiency of this algorithm in simulating the denatured state and calculating position-specific thermodynamic descriptors, thus providing a solid foundation for the studies conducted in this dissertation.

In Chapter 3, energetic information in the denatured state was characterized and then compared with those of the native state. Quantitative analysis of the energetics between denatured and native states revealed that there is no correlation between these states. This observation further supported the point that a study focused on the native state will insufficiently address the protein folding problem.

Energetic information contained within the denatured ensemble was investigated in chapter 4 by conducting fold recognition experiments to match sequence to fold. Surprisingly, denatured states showed a significant improvement in the ability to match sequences to folds compared to native states. This finding suggested that there was unique information in denatured ensembles and the thermodynamic determinants contributing to protein fold specificity was different from that found in the native ensemble. More importantly, the significant observations in this chapter open a wide range of future possibilities for developing new approaches of protein classification and prediction by using information in denatured states.

The primary goal of chapter 5 was to investigate the relationship between denatured state energetics and secondary structure. Propensities of secondary structures in thermodynamic environments of the denatured state across multiple proteins were analyzed and revealed that the denatured state energetics can be

related to structural features of the native state. This observation was intriguing as it suggests that there are thermodynamic signatures in denatured states for structural and possibly functional properties. The correlation between energetics in the denatured state and native state secondary structures also opens opportunities to develop simple but efficient algorithms for protein secondary structure prediction using simple techniques.

The roles of structure and sequence in denatured states energetics were investigated in chapter 6. The finding that local energetics in denatured states reflected the structural forming abilities of local sequence was interesting. Also inherent properties of sequences were found to play an important role in local energetics of denatured states. The observations in this chapter provided a better understanding for the relationships between sequences, denatured states and secondary structures.

What exactly is the role of the denatured state in protein folding? Based on the results presented in the previous six chapters, an intriguing mechanism of protein folding involving denatured states was put forward in chapter 7. The findings suggested that the energetics of the denatured state avoided early β-sheet formation, opting instead of for α-helix and turn to serve as potential nucleation sites for protein folding. Early incorrect β-sheet formation has always been found in amyloid fibrils associated with misfolding diseases. Our results suggest that energetics in the denatured state evolved to minimize the possibility of the formation of incorrect β-sheet. Knowledge about thermodynamic determinants of protein folding specificity in the denatured state opens the possibility of detecting determinants of amyloid formation and providing capability to control amyloid related misfolding diseases.

REFERENCES

Alexandrescu, A. T., Shortle, D. (1994). Backbone dynamics of a highly disordered 131 residue fragment of staphylococcal nuclease. *J Mol Biol* **242**, 527-546.

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

Babu, C.R., Hilser, V.J., Wand, J. (2004). Direct access to the cooperative substructure of proteins and the protein ensemble via cold denaturation. *Nat Struct Mol Biol* **11**, 352 – 357.

Blow, D. M. (1977). Flexibility and rigidity in protein crystals. *Ciba Found Symp*, 55-61.

Bowler, B.E. (2007) Thermodynamics of protein denatured states. *Mol Biosyst* **2**, 88-99.

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

Carra, J.H, Privalov, P.L (1995). Energetics of denaturation and m values of staphylococcal nuclease mutants. *Biochemistry* **34** (6):2034-2041.

Chen, J. W., Romero, P., Uversky, V. N., Dunker, A. K. (2006). Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. *J Proteome Res* **5**, 888-898.

Daura, X, van Gunsteren, Mark AE. (1999). Folding-unfolding thermodynamics of a beta-heptapeptide from equilibrium simulations. *Proteins* **34** (3): 269-280.

Dill, K.A., Shortle, D (1991). Denatured states of proteins. *Annu Rev Biochem* **60**, 795-825.

Dill, K.A. Bromberg, S, Yue, K., Fiebig, K.M ,Yee, D.P , Thomasm P.D. , Chan, H.S. (1995) Principles of protein folding—a perspective from simple exact models, *Protein Sci* **4** , 561–602.

Dosztanyi, Z., Csizmok, V., Tompa, P., Simon, I. (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* **347**, 827-839.

Dunker, A. K., Obradovic, Z. (2001). The protein trinity - linking function and disorder. *Nat Biot* **19**, 805-806.

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

Fuxreiter, M., Simon, I., Friedrich, P., Tompa, P. (2004). Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J Mol Biol* **338**, 1015-1026.

Garnier, J., Osguthorpe, D. J.,Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* **120**, 97-120.

Godbole, S, Hammack B, Bowler B.E. (2000). Measuring denatured state energetics: deviations from random coil behavior and implications for the folding of iso-1-cytochrome c. *J Mol Biol* **296**, 217-228.

Gu, J., Gribskov, M., Bourne, P. E. (2006). Wiggle-predicting functionally flexible regions from primary sequence. *PLoS Comput Biol* **2**, e90.

Gu, J., Bourne, P. E. (2007). Identifying allosteric fluctuation transitions between different protein conformational states as applied to Cyclin Dependent Kinase 2. *BMC Bioinformatics* **8**, 45.

Hennig**,** M., Bermel, W., Spencer, A., Dobson, C. M., Smith, L. J. , Schwalbe, H.(1999) Side-chain conformations in an unfolded protein: chi1 distributions in denatured hen lysozyme determined by heteronuclear 13C, 15N NMR spectroscopy. *J Mol Biol* **288**, 705-723

Hilser, V. J. , Freire, E. (1996). Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. *J Mol Biol* **262**, 756-772.

Hilser, V. J., Townsend, B. D. , Freire, E. (1997). Structure-based statistical thermodynamic analysis of T4 lysozyme mutants: structural mapping of cooperative interactions. *Biophys Chem* **64**, 69-79.

Hilser, V. J. , Freire, E. (1997). Predicting the equilibrium protein folding pathway: structure-based analysis of staphylococcal nuclease. *Proteins* **27**, 171-183.

Hilser, V. J., Garcia-Moreno, E. B., Oas, T. G., Kapp, G., Whitten, S. T. (2006). A statistical thermodynamic model of the protein ensemble. *Chem Rev* **106**, 1545-1558.

Hilser, V. J., Thompson, E. B. (2007). Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proc Natl Acad Sci U S A* **104**, 8311-8315.

Iakoucheva, L. M., Kimzey, A. L., Masselon, C. D., Bruce, J. E., Garner, E. C., Brown, C. J., Dunker, A. K., Smith, R. D. , Ackerman, E. J. (2001). Identification of intrinsic order and disorder in the DNA repair protein XPA. *Protein Sci* **10**, 560-571.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* **292**, 195-202.

Jones, D. T., Ward, J. J. (2003). Prediction of disordered regions in proteins from position specific score matrices. *Proteins* **53** Suppl 6, 573-578.

Kukreja, R.,Singh, B. (2005). Biologically active novel conformational state of botulinum, the most poisonous poison. *J Biol Chem* **280**, 39346-39352.

Kumar, R., Baskakov, I. V., Srinivasan, G., Bolen, D. W., Lee, J. C. ,Thompson, E. B. (1999). Interdomain signaling in a two-domain fragment of the human glucocorticoid receptor. *J Biol Chem* **274**, 24737-24741.

Larson, S. A., Hilser, V. J. (2004). Analysis of the "thermodynamic information content" of a Homo sapiens structural database reveals hierarchical thermodynamic organization. *Protein Sci* **13**, 1787-1801.

Lee, K. H., Xie, D., Freire, E., Amzel, L. M. (1994). Estimation of changes in side chain configurational entropy in binding and folding: general methods and application to helix formation. *Proteins* **20**, 68-84.

Levinthal, C. (1968). Are there Pathways for Protein Folding. *J Chem Phys* **65**:44-45.

Li, F., Gangal, M., Juliano, C., Gorfain, E., Taylor, S. S., Johnson, D. A. (2002). Evidence for an internal entropy contribution to phosphoryl transfer: a study of domain closure, backbone flexibility, and the catalytic cycle of cAMP-dependent protein kinase. *J Mol Biol* **315**, 459-469.

Linding, R., Russell, R. B., Neduva, V.,Gibson, T. J. (2003). GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* **31**, 3701-3708.

Liu, J., Rost, B. (2003). NORSp: Predictions of long regions without regular secondary structure. *Nucleic Acids Res* **31**, 3833-3835.

Liu, T., Whitten, S. T. , Hilser, V. J. (2007). Functional residues serve a dominant role in mediating the cooperativity of the protein ensemble. *Proc Natl Acad Sci U S A* **104**, 4347-4352.

Lovell, S. C., Word, J. M., Richardson, J. S. , Richardson, D. C. (2000). The penultimate rotamer library. *Proteins* **40**, 389-408.

Meszaros, B., Tompa, P., Simon, I.,Dosztanyi, Z. (2007). Molecular principles of the interactions of disordered proteins. *J Mol Biol* **372**, 549-561.

Mok, Y. K., Elisseeva, E. L., Davidson, A. R., Forman-Kay, J. D. (2001). Dramatic stabilization of an SH3 domain by a single substitution: roles of the folded and unfolded states. *J Mol Biol* **307**, 913-928.

Pappu, R. V., Srinivasan, R., Rose, G. D. (2000). The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc Natl Acad Sci U S A* **97**, 12565-12570.

Przybylski, D., Rost, B. (2002). Alignments grow, secondary structure prediction improves. *Proteins* **46**, 197-205.

Ptitsyn, O.B. (1995). Structures of folding intermediates. *Curr Opin Struct Biol* **5,** 74-8. *Curr Opin Struct Biol* **5**(1): 74-78.

Ringe, D., Petsko, G. A. (1986). Study of protein dynamics by X-ray diffraction. *Methods Enzymol* **131**, 389-433.

Romero, Obradovic, Kissinger, C., Villafranca, J. E., Dunker, A. K. (1997). Identifying Disordered Regions in Proteins from Amino Acid Sequences. *Proc. I.E.E.E. International Conference on Neural Networks* **1**, 90-95.

Ross, C. A. & Poirier, M. A. (2004). Protein aggregation and neurodegenerative disease. *Nat Med* **10 Suppl**, S10-7.

Romero, P., Obradovic, Z., Dunker, A. K. (2004). Natively disordered proteins: functions and predictions. *Appl Bioinformatics* **3**, 105-113.

Sen, T. Z., Jernigan, R. L., Garnier, J. ,Kloczkowski, A. (2005). GOR V server for protein secondary structure prediction. *Bioinformatics* **21**, 2787-8.

Shortle, D., Chan, H. S. , Dill, K. A. (1992). Modeling the effects of mutations on the denatured states of proteins. *Protein Sci* **1**, 201-215.

Smith, T. F., Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol* **147**, 195-7.

Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem Sci* **27**, 527-533.

Tompa, P. (2003). The functional benefits of protein disorder. *J Mol Stru Theoc* **666**, 361-371.

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

Uversky, V. N. (2002). Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* **11**, 739-756.

Wang Yi and Shortle, D. (1995). Equilibrium folding pathway of staphylococcal nuclease: identification of the most stable chain-chain interactions by NMR and CD spectroscopy. *Biochemistry*, **34**, 15985 – 15950.

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. , Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* **337**, 635-645.

Wetlaufer, D.B. (1990).Nucleation in protein folding — confusion of structure and process. *Trends Biochem Sci* **15**, 414–415.

White, G. W., Gianni, S., Grossmann, J. G., Jemth, P., Fersht, A. R., Daggett, V.(2005). Simulation and experiment conspire to reveal cryptic intermediates and a slide from the nucleation-condensation to framework mechanism of folding. *J Mol Biol* **350**, 757-775.

Whitten, S. T., Kurtz, A. J., Pometun, M. S., Wand, A. J., Hilser, V. J. (2006). Revealing the nature of the native state ensemble through cold denaturation. *Biochemistry* **45**, 10163-10174.

Wrabl, J. O. , Shortle, D. (1996). Perturbations of the denatured state ensemble: modeling their effects on protein stability and folding kinetics. *Protein Sci* **5**, 2343-2352.

Wrabl, J. , Shortle, D. (1999). A model of the changes in denatured state structure underlying m value effects in staphylococcal nuclease. *Nat Struct Biol* **6**, 876-883.

Wrabl, J. O., Larson, S. A., Hilser, V. J. (2001). Thermodynamic propensities of amino acids in the native state ensemble: implications for fold recognition. *Protein Sci* **10**, 1032-1045

Wright, P. E., Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* **293**, 321-331.

Yang, Z. R., Thomson, R., McNeil, P., Esnouf, R. M. (2005). RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*. **21**(16):3369-3376

VITA

Suwei Wang was born in Qiqihaer, Heilongjiang province in China on December, 18, 1972. In 1989, he entered Northeast Agricultural University in Harbin, China where he received the degree of Bachelor of Science in Agronomy. In 1993, Suwei was accepted into Plant Biochemistry & Physiology program in the Graduate School at Northeast Agricultural University and received his Master of Science in 1996. In 2001, Suwei entered University of Texas at Dallas and received his Master of Science in Computer Science in 2002. With his background in both biology and computer science, Suwei was accepted into Biophysical, Structural & Computational Biology program at The University of Texas Medical Branch in 2003. He joined Dr. Vincent J. Hilser's lab in 2004 and began work towards the Doctor of Philosophy degree.

Permanent Address: 7019 Lasker Dr. Apt 1223
Galveston, Texas 77551

This dissertation was typed by Suwei Wang