2011

Mechanistic insights on important biomolecules derived using simple dynamics models from extending the reach of elastic network modeling

Michael Thomas Zimmermann

Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Bioinformatics Commons

Recommended Citation

Zimmermann, Michael Thomas, "Mechanistic insights on important biomolecules derived using simple dynamics models from extending the reach of elastic network modeling" (2011). Graduate Theses and Dissertations. 12546.

https://lib.dr.iastate.edu/etd/12546

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Mechanistic insights on important biomolecules derived using simple dynamics models from extending the reach of elastic network modeling

by

Michael Thomas Zimmermann

A dissertation submitted to the graduate faculty in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Robert L. Jernigan, Co-Major Professor
Edward Yu, Co-Major Professor
Guang Song
Zhijun Wu
Amy Andreotti

Iowa State University
Ames, Iowa
2011

Copyright © Michael Thomas Zimmermann, 2011. All rights reserved.
# Table of Contents

LIST OF TABLES .................................................................................................................. vi
LIST OF FIGURES ................................................................................................................ viii
ACKNOWLEDGEMENTS ..................................................................................................... xiii
ABSTRACT ............................................................................................................................ xiv

CHAPTER 1. INTRODUCTION ............................................................................................. 1

1.1. Overview of Protein Dynamics .................................................................................... 1

1.2. Significant contributions of this dissertation ............................................................ 4

CHAPTER 2. Elastic Network Modes Capture the Motions Apparent in Multiple tRNA Structures ....... 6

2.1. Abstract ....................................................................................................................... 6

2.2. Introduction ................................................................................................................ 6

2.3. Methods ..................................................................................................................... 8

2.3.1. Dataset (Ensemble of Structures) ........................................................................... 8

2.3.2. Elastic Network Model ......................................................................................... 8

2.3.3. Principal Component Analysis (PCA) ................................................................. 11

2.3.4. Structure Alignment ............................................................................................. 13

2.3.5. Structure Representation ....................................................................................... 13

2.3.6. Comparing PCs and Modes .................................................................................. 13

2.4. Results ....................................................................................................................... 14

2.5. Discussion ................................................................................................................ 22

2.6. Conclusions .............................................................................................................. 25

CHAPTER 3. Heavy metal ion transport complex CusCBA assembly, structure, and mechanism ........ 27

3.1. CusA Structure Summary ........................................................................................... 27

3.2. CusBA Structure Summary ....................................................................................... 27

3.3. CusBA mechanistic insights derived from structural studies ....................................... 29

3.4. Modeling of the interaction of CusC with the CusA+CusB co-crystal and the injection mechanism .................................................................................................................. 30

3.5. Extracting the ion-exchange mechanism .................................................................. 32

CHAPTER 4. Structural control over telomerase mechanism ............................................. 38

4.1. Abstract ....................................................................................................................... 38
4.2. Introduction .................................................................................................................. 38
4.3. Results and discussion ............................................................................................... 42
4.4. Conclusions ................................................................................................................. 50
4.5. Methods ....................................................................................................................... 50

CHAPTER 5. Ribosome mechanism explored using coarse-grained models .................... 53
5.1. Background .................................................................................................................. 53
5.2. Extracting the Ribosomal Mechanism from Simulations ........................................... 53
5.3. The ribosome is like a clock-work .............................................................................. 60
5.4. Assembly Order of Ribosomal Proteins Relates to Their Computed Protein Dynamics .... 63
  5.4.1. Abstract .................................................................................................................... 63
  5.4.2. Introduction ............................................................................................................. 63
  5.4.3. Methodology ........................................................................................................... 67
     Analysis of Contacts in the Assembled 30S subunits ..................................................... 67
     Molecular Dynamics Simulations .................................................................................. 67
     Elastic Network Modeling ............................................................................................. 69
     Statistical Analysis of Contact Residue Mobility ......................................................... 69
  5.4.4. Results and Discussion .......................................................................................... 70
     Ribosomal proteins are enriched with positively charged amino acids ....................... 70
     Structures and contact residues are more conserved than sequences ....................... 72
     Dynamics and conformational Changes of S15 ........................................................... 74
     Dynamics and conformational Changes of S17 ........................................................... 76
     Dynamics and conformational change of S20 ............................................................. 79
     General trends based on Elastic Network Modeling ..................................................... 81
  5.4.5. Conclusion ............................................................................................................. 84

CHAPTER 6. Functional insights into Immunoglobin-G from structural dynamics ............. 86
6.1. Abstract ....................................................................................................................... 86
6.2. Introduction .................................................................................................................. 86
6.3. Materials and methods .............................................................................................. 88
  6.3.1. Kinematics of Proteins .......................................................................................... 88
  6.3.3. Computing Changes in Internal Distances ............................................................ 88
  6.3.4. Fractal and Spectral Dimension ............................................................................ 89
6.4. Results and Discussion ................................................................. 90
6.5. Conclusion ................................................................................. 99

CHAPTER 7. Combining 4-body energy with entropies calculated from elastic networks .......... 101
7.1. Incorporating vibrational entropy into decoy discrimination ............................................ 101
    7.1.1 Abstract 101
    7.1.2 Introduction ........................................................................... 101
    7.1.3. Four-body contact potentials .................................................. 103
    7.1.4. Geometric construction of four-body contacts ........................................... 103
    7.1.5. Four-Body Contact Potential Energy Function ........................................... 105
    7.1.6. Performances of different individual potential functions for model ranking .......... 106
    7.1.7. Obtaining an optimized potential ................................................. 107
    7.1.8. Performance of the optimized potential .............................................. 107
    7.1.9. Elastic network model (ENM) entropy ............................................. 108
    7.1.10. Energy and entropy changes for different structure pairs of the same protein ... 110
7.2. Incorporating mean square fluctuation as entropy ......................................................... 119
7.3. Discriminating docking poses with computed free-energies ........................................... 120

CHAPTER 8. MAVENs: Motion analysis and visualization of elastic networks and structural ensembles ................................................................................................................. 123
8.1. Abstract ......................................................................................... 123
8.2. Introduction ................................................................................... 124
8.3. Implementation ............................................................................... 125
8.4. Elastic Network Model ................................................................. 127
8.5. Comparing ENMs to Ensembles of Structures ......................................................... 127
8.6. Model Generation ........................................................................... 127
8.7. Discussion ...................................................................................... 129
8.8. Conclusions ................................................................................... 130

CHAPTER 9. Exploring protein fold space .............................................................................. 131
9.1. Discrete versus continuous ........................................................................ 131
9.2. Modularity of genes and its effect on protein structure ............................................. 132
9.3. Introduction ..................................................................................... 133
9.4. Results ......................................................................................... 136
9.5. Discussion ........................................................................................................ 142
9.6. Methods ........................................................................................................ 143
  9.6.1. Datasets ................................................................................................. 143
  9.6.2. Protein Structure Evaluation Metrics ..................................................... 143
  9.6.3. Analysis of Structural Classification ....................................................... 144
  9.6.4. Graph Construction ................................................................................ 144
  9.6.5. Cluster Generation and Comparision ..................................................... 144

CHAPTER 10. Conclusions ...................................................................................... 146

APPENDIX A. Computational free energy evaluation to explore energy minima ........... 148
  A1. Sampling multiple minima from a starting structure .................................... 149
  A2. Structure refinement .................................................................................. 150

APPENDIX B. Developmental differences in IL1B expression .................................. 152
  B.1. Abstract .................................................................................................... 152
  B.2. Introduction ............................................................................................... 152
  B.3. Methods .................................................................................................... 153
  B.3. Results ...................................................................................................... 154
  B.4. Discussion and conclusions ...................................................................... 156

APPENDIX C. Structural regulation of kinase activity .......................................... 158

APPENDIX D. RNA nanostructure assembly and verification .................................. 161

APPENDIX E. Animating cellular and sub-cellular images and tomographic maps ........ 164

APPENDIX F. Characterization of YAP and YBHB by Mass Spectrometry .............. 167
  F.1. Abstract .................................................................................................... 167
  F.2. Manuscript Excerpts ................................................................................ 167
    F.2.1. Collision cross section measurements .................................................. 168
    F.2.2. Collision cross section calculations ..................................................... 169
  F.3. Conclusion ................................................................................................. 170

BIBLIOGRAPHY .................................................................................................... 171
LIST OF TABLES

Table 2.1: Percent of tRNA ensemble variance captured by the first 3 PCs using each alignment method. .......................................................... 24

Table 2.2: Percent of variance apparent in the PCs captured by the first 6 normal modes shows that combining modes from ANMs with different power dependences is most efficient. .......................................................... 25

Table 5.1: Pulling on the ribosome bound mRNA induces functional motions that are similar to the dominant normal modes, especially the 30S-50S ratchet and the mRNA tunnel gating. ........................................................................................................ 61

Table 5.2: Net charge of r-proteins from sequences .......................................................................................................................... 71

Table 5.3: Breakdown of total contacts in terms of contacts made by positively charged residues. .......................................................................................................................... 72

Table 5.4: Average MD RMSF values (Å), (standard deviation) and enrichment factors ............ 76

Table 5.5: ANM enrichment factors and significance for 30S proteins ........................................ 84

Table 7.1. The eight classes of residue types used in the 4-body potential, chosen to reduce the amino acid alphabet. .......................................................... 103

Table 7.2. Five structure pairs from the molmovedb database (239) used in this study. ........ 111

Table 7.3. Total change in entropy and energy upon deformation by the first three normal modes for the five structure pairs. .......................................................... 111

Table 7.4. The mean difference in computed energy (V) by ENM, entropy (S), and free energy (G) for the five structure pairs. .......................................................................................................................... 113

Table 7.5. Free energy changes by combining the optimized four body energies with the entropies estimated from the elastic network model................................. 114

Table 7.6: Decoy discrimination is significantly improved compared to 4-body energy alone by incorporating fluctuation-based entropy. .......................................................... 119

Table 9.1: Clustering metrics for graphs generated from common CATH or SCOP annotations across both datasets .......................................................................................................................... 141

Table 9.2: Top 10 domains by CATH Topology Usage ........................................................................ 141

Table A.1: RMSD change from the initial adenylate kinase structure to members of the structural ensemble generated by iterative ANM .......................................................................................................................... 150
Table E.1: Correlation of mean square fluctuation for 5 types of ENMs applied to the membrane leaflet using a coarse-graining threshold of 9 units................................................................. 164

Table F.1: Calculated cross sections for YbhB protein from E. coli using crystal structure data. The values include the sizes of the hydrogen atoms................................................................. 168
LIST OF FIGURES

Figure 1.1: Examples of structure representations at different resolutions and using different data sources.................................................................................................................................................. 1

Figure 1.2: We illustrate the ENM model building procedure using the HIV-1 protease structure, 1T3R. .................................................................................................................................................................................................................................................. 2

Figure 1.3: Graphical example of residue level coarse-graining for use in mixed-resolution ENM modeling and application of the interaction cutoff radius $r_c$. ................................................................................................. 3

Figure 2.1: PCs of the tRNA ensemble constructed from all common atoms and highlighting the physical limits of deformation which are not the same in each direction. ................................................................. 12

Figure 2.2: Normal mode space and dominant motions from atomic ANM visualized similarly to Figure 2.1 .................................................................................................................................................................................................................................... 15

Figure 2.3: Comparison of multiple ANM representations with the first two PCs using A) cutoff and B) distance dependent ANMs. ........................................................................................................................... 18

Figure 2.4: Comparing the Very Lowest Frequency Normal Modes with the PCs. ................................................ 20

Figure 2.5: Correlation between crystallographic B-Factors and computed fluctuations for A) cutoff based ANM and B) ANM with distance dependent springs where all atoms are considered to be interacting through connected springs with a weight of $d_{ij}^{-x}$ where $x$ is the power dependence............................................................................................................................................................................................................................... 22

Figure 3.1: The swinging motion of PC2 coupled to TM8 is the dominant structural difference between the apo and copper bound forms and is captured by ANM................................................................. 28

Figure 3.2: The neighbor dependent opening and closing motions of the CusA structure using ENMs shows an asymmetric mechanism.............................................................................................................. 29

Figure 3.3: Docking of CusC to CusBA to assemble the full CusCBA system with CusC either on top of the CusB ring or injected into it. ................................................................................................................ 30

Figure 3.4: Structural model showing one candidate pose for the CusCBA system, emphasizing one point of evidence for CusC to be “injected” into the CusB ring. ................................................................. 31

Figure 3.5: The top portion of the CusBA crystal structure is shown, displaying the directions with which the CusB annulus is computationally opened................................................................................... 32

Figure 3.6: Depiction of the MD environment used to refine the CusC-CusBA post-injection complex......................................................................................................................................................................................................................... 32
Figure 3.7: Comparison between standard ANM modes and mode-motions where virtual bond lengths and approximated Van Der Waals exclusion are taken into account. .......................33

Figure 3.8: Transition between the Cu$^{+1}$ bound and apo states, emphasizing the dynamic motion between PC2 and the ion-exchange. .................................................................34

Figure 3.9: Directionality Imparted to the Cu$^{+}$ Ion by Normal Modes 1-3 Suggests a Specific Structural Mechanism. .........................................................................................35

Figure 3.10: The motion from Figure 3.9 is shown from a wider view, showing the details included in our simulation. ..........................................................................................36

Figure 3.11: Distance changes between paired methionine residues upon normal mode deformation relates to the ion-shuttling mechanism. .................................................................37

Figure 4.1: Partial model of human telomerase ribonucleoprotein complex generated computationally in this study. .................................................................42

Figure 4.2: Structural model for the processive motion of human telomerase. ..........................48

Figure 5.1: The ribosome showing the mRNA and the nascent peptide channel (NC) tunnels. ...54

Figure 5.2: Viewing the 70S ribosome with mRNA and nascent chain tunnel volumes. ..........55

Figure 5.3: Two examples of matrices of vector overlaps using the most collective mode of motion; A) L3 and L22 and B) S5 and the mRNA tunnel. .................................................................56

Figure 5.4: Visualizations of the first mode of motion – a ratchet motion with opposite rotation of the 50S and 30S subunits.................................................................57

Figure 5.5: The 5$^{th}$ mode of motion – a rocking hinge motion which alternately opens and closes the two ends of the mRNA exit tunnel. .................................................................58

Figure 5.6: Distribution of eigenvalues. Jumps in eigenvalues (squares of frequencies) indicate a change in the time scale of these motions. .................................................................60

Figure 5.7: External force imposed to pull the mRNA ahead. We pull the ribosome bound mRNA in the direction $F_{app}$, which induces two major types of motion. .........................61

Figure 5.8: Two major types of motion within the 30S shown by our models are displayed – the 30S head-swivel as the mRNA tunnel gates. .................................................................62

Figure 5.9. Previous work in the Jernigan lab (Figure 7 in Kurkcuoglu et. al. (65) reproduced with permission) using ANM showed that S3 is the most mobile of the mRNA entrance proteins, followed by S4, which is in agreement with our observed dynamics (Figure 5.8). .................................................................62
Figure 5.10: The 30S *T. thermophilus* subunit (1J5E), interface side colored by domain.............65
Figure 5.11: Crystal structure comparisons for A) S15, B) S17, and C) S20. .................................73
Figure 5.12: RMSD calculated from the X-ray crystal structure for A) S15, B) S17, and C) S20 across MD trajectories and for both species. .................................................................74
Figure 5.13: RMSF values for A) S15, B) S17, and C) S20 calculated from MD trajectories............77
Figure 5.14: Structural variations during MD simulations for all six proteins. .................................81
Figure 5.15: Sequence conservation shown by using ConSurf for the 30S r-proteins, displaying S15, S17, and S20 specifically with rRNA contacts. .................................................................83
Figure 6.1: The structure of IgG (1IGT) highlighting the domains and hinges. .................................87
Figure 6.2: The first six modes of motion capture most of the total motion in IgG and the CDR. 91
Figure 6.3: Hinges and the structure behavior identified by ANM computations. ..........................92
Figure 6.4: The effect of sampling hinge dihedral angles on the position of the CDR. ......................93
Figure 6.5: Mean-square internal distance changes within the entire immunoglobulin. .................94
Figure 6.6: Intermolecular crystal packing in 1IGT shows that substrate free structures are still bound CDR bound........................................................................................................95
Figure 6.7: The computed positional mean-square fluctuations of Fab and CDR residues shows increased mobility of the sequence-variable loops.........................................................96
Figure 6.8: Correlation of experimental and computed temperature factors of the CDRs is stronger when the full IgG structure is considered than when parts of the structure are considered. .................................................................................................................97
Figure 6.9: Inclusion of the full IgG structure diminishes magnitude of internal distance changes. .................................................................................................................................98
Figure 6.10: Spectral and Fractal dimension of IgG and its domains shows the full structure to have the most efficient communication..............................................................99
Figure 7.1: Identification of residue points for use in the four-body contacts.................................104
Figure 7.2: Comparison of the change in energy and entropy upon deformation by the first three normal modes. ........................................................................................................112
Figure 7.3: Performance of classifying decoy structures by Local Character of ΔS.......................115
Figure 7.4: Result of classifying CASP9 monomer decoys using vibrational entropy, energy, and a free energy................................................................................................................116
Figure 7.5: RMSD comparison of the best decoy for each target and the best decoy as judged by \( \Delta S, \Delta E \), and a combination of the two shows the efficacy of fluctuation-based entropy. ................................................................. 120

Figure 7.6: Selecting docked forms using computed free energies with mean square fluctuation based entropy improves selection of docking poses in unconstrained docking, showing advantages compared to energy alone. ................................................................. 121

Figure 7.7: Using mean square fluctuation based entropy improves selection of docking poses in constrained docking. ........................................................................ 122

Figure 8.1: MAVEN interface overview and examples of model generation features. ........... 126

Figure 8.2: Analysis features of MAVEN. .............................................................................. 128

Figure 9.1: The relationship between average shortest path computed using the Watts-Strogatz approximation and the TM-score threshold for retaining edges in the graph............. 132

Figure 9.2: Metrics for Analysis of MCL clustering on TM-score graphs. ............................. 137

Figure 9.3: Structural comparison of the reverse transcriptase fold and conformational variation – a test of state-of-the-art structure superposition. ................................................. 139

Figure 9.4: Frequency of CATH topology usage in three datasets........................................ 142

Figure A.1: Energy funnel analogy depicting the two aims – searching for alternate conformations of a structure and refinement of a modeled structure......................... 148

Figure A.2: Graphical depiction of the two aims: structure refinement and conformational change................................................................................................................. 149

Figure A.3: Preliminary simulations to refine a predicted structure using ANM and 4-body potentials, without using the target structure. .................................................. 151

Figure B.1: Comparison of the predicted structures of the two porcine IL1β structures. ....... 155

Figure B.2: We superimpose onto 1ITBA the 5 best ITASSER homology models of each of the two pig IL1-β pre-cleavage protein sequences using the CE algorithm.......................... 157

Figure B.3: Structure analysis to determine if a differential proline residue induces changes that could prevent cleavage by CASP1................................................................. 157

Figure C.1: Domain structure of Itk kinase reproduced from (307). ........................................ 158

Figure C.2: Itk structure with regulatory and functional regions highlighted – reproduced from (308).................................................................................................................. 159
Figure C.3: Snapshots from a preliminary REMD trajectory on the Itk free-no-linker structure showing diversity in hydrophobic spine alignment and loop melting. ................................................. 160

Figure D.1: Schematic of possible uses for RNA nanorings in delivering therapeutic small interfering (si) RNA - reproduced from (310). ......................................................................................... 162

Figure D.2: The technosquare test case ANM showing ring closure using the dominant normal mode. ......................................................................................................................... 162

Figure D.3: Example of all-atom models of pentameric nanorings where ANM again shows ring closure, but is insensitive to alterations in the monomer junction interactions............. 163

Figure D.4: Unlike globular proteins, individual modes of motion for the technosquare are not significantly affected by removal of extended termini................................................................. 163

Figure E.1: We display three views of the 3D model generated from cellular images.............. 164

Figure E.2: The coarse graining threshold, T, is varied and the resulting 3D models are displayed to show the relative density of each region. ................................................................. 165

Figure E.3: Different modes of motion of the image. We plot the structure colored by \( \log_{10}(\text{Temperature Factor}) \) where blue are small temperature factors and red large. .. 165

Figure E.4: ANM motion of the structure shows the opening and closing of pores along the membrane leaflet. ........................................................................................................... 165

Figure E.5: The next phase of this project will focus on sub-cellular structures, animating Golgi apparatus and endoplasmic reticulum tomographic models.............................................. 166

Figure F.1: Measured cross sections for YbhB protein monomer and dimer from *E. Coli* .......... 169
ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and writing this dissertation. First, Dr. Robert L. Jernigan has been my primary advisor at Iowa State University. His guidance, insight, and thoughtful nature have played a crucial role in my professional development and, without which, this body of work would not have developed. The members of my Program of Study committee, Dr. Edward Yu, Dr. Guang Song, Dr. Amy Andreotti, and Dr. Zhijun Wu, have each positively influenced my research directly through individual conversations, but also through their impact on campus: the expertise and activities of their lab members, talks, meetings, and department activities. Many other faculty and staff have been meaningful to my graduate experience including Dr. Drena Dobbs, Dr. Christopher Tuggle, Dr. Volker Brendel, Dr. Andrzej Kloczkowski, Dr. Xuefeng Zhao and Dr. Taner Sen. I would also like to thank Trish Stauble and Mary Jane McCunn for their advice and coordination during my time as a student in this interdepartmental program.

The support of family has been a true asset and source of comfort. My loving wife, Elizabeth Zimmermann, has been understanding, patient, and a joy. The firm foundation laid by my parents, Rev. Fred and Carol Zimmermann, has been paramount for my development and studies.

Finally, for the immeasurable value of many conversations and opportunities to learn, I am grateful to my peer collaborators: Dr. Fadi Towfic, Dr. Yves Sucaet, Ataur Katebi, Sumudu Leelananda, Dr. John Van Hemert, Dr. Aris Skliros, Dr. Haitao Cheng, Usha Muppirala, Deepak Reyon, Tonia Schwartz, Dr. Lei Yang, Dr. Yaping Feng, Erin Boggess, Scott Boyken, Jonathan Hurst, Nicholas Larson, and Ben Lewis.

Financial support provided by National Institutes of Health grants R01GM081680, R01GM072014, and R01GM073095 supported most of this research and is gratefully acknowledged.
ABSTRACT

The dynamics of biomolecules are important for carrying out their biologic functions, but these remain difficult to probe in detail experimentally, so that their accurate computational evaluation is an important field of ongoing study. Critical questions remain open such as what are the importance of individual interactions within a structure, the composition of denatured states and equilibrium native ensembles, as well as the role and conservation of flexibility in functional dynamics. The tools of Molecular Dynamics, Monte Carlo simulation, and Normal Mode Analysis coupled with knowledge-based approaches represent the mainstay of computational approaches used in this field.

The primary focus of this dissertation is to explore the functional dynamics of important biomolecules while extending the utility of Normal Mode Analysis using Elastic Network Models through the application of novel analysis methods. Many of these techniques have been made available to the scientific community through the software tool MAVEN which integrates and automates many of the steps in model building and analysis. By utilizing these tools, we have discerned structural dynamics characteristics and mechanistic behaviors of antibodies, ribosomes, telomerase, and efflux systems. Modes from multiple Anisotropic Network Models capture collective as well as local motions which accurately describe a large set of experimental tRNA structures. Mechanistic understanding of biomolecular motion can aid in the understanding of physiology, disease states, and our ability to engineer new structures with novel functions.

The ability to distinguish native-like structures from a set of computational predictions is important not only in structure prediction, but also in molecular docking and for predicting conformational changes. We propose a new algorithm for evaluating the entropy of motion of biomolecules, showing that it leads to enhanced discrimination between native-like and non-native-like models in both structure predictions and protein-protein docking. Our findings indicate that the shape of a protein or complex contains sufficient information to distinguish it from poorer quality predictions. Graph theoretical approaches have also been employed to investigate the connectedness of the protein structure universe, showing that the modularity of protein domain architecture is of fundamental importance for future improvements in structure matching. All of the studies herein impact our understanding of protein domain evolution and modification.
CHAPTER 1. INTRODUCTION

1.1. Overview of Protein Dynamics

Figure 1.1: Examples of structure representations at different resolutions and using different data sources.

The Cα representation is the most common level of coarse-graining for proteins where only these single points per residue are used to represent the structure. For detailed questions, retaining all heavy atoms may also be appropriate, or mixed-resolution may be used (Figure 1.3). In cases where only low resolution data is available (such as in electron microscopy), the molecular surface of a globular protein may be filled with lattice points.

Globular proteins are essential components of all living organisms, responsible for a wide range of biological functions which exhibit high specificity or execution on a myriad of targets. These multifold functions include enzymes which catalyze biochemical reactions, structural scaffolds to build the structural parts of organisms, muscle proteins that produce mechanical work from chemical reactions, transporters such as hemoglobin that shuttle essential metabolites, antibodies that bind to foreign materials to signal the immune response, and transcription factors that modulate gene expression. The polypeptide chain is folded compactly into a characteristic three-dimensional structure, often referred to as the protein fold. It is commonly observed that the packing density of the protein components is high and residues with hydrophilic character tend to remain near the protein surface (1) leaving a primarily hydrophobic core (2).
Figure 1.2: We illustrate the ENM model building procedure using the HIV-1 protease structure, 1T3R.

Atoms that are being considered are colored red, those yet to be considered, gray, and the interaction cutoff radius shown as a blue sphere. A) The distance from each atom in the structure to other atoms is calculated. B) Harmonic springs are placed between any pairs that are within the interaction cutoff distance. C) This is carried out for all atoms in the structure – here we have shown Cα atoms only. D) After all pairs are considered, the structure is represented by a network of harmonic interactions – an elastic network.

Protein structures are not static, but sample a conformational ensemble that is influenced by its environment. This ensemble is often described by an energy landscape analogy where the energy of a given conformation determines its relative population, deeper minima corresponding to lower energy states, and the energy barriers between local minima determining important interconversion rates. It is only very recently that protein folding has been able to be studied in detail in vivo experimentally (3). However, there is a rich literature concerning the study of equilibrium dynamics in vivo and in vitro using tagging techniques like Fluorescence Resonance Energy Transfer. Computational approaches have been extremely useful to fill in gaps in understanding. Molecular dynamics simulations are increasingly common and able to handle larger and more complex systems. Also, extensive study has shown that for certain properties the many types of software and force fields used are coming to a consensus (4). However, large systems remain very costly to simulate and intractable for small laboratories. Elastic Network Models (ENMs), in particular the Anisotropic Network Model (ANM), have proven useful for obtaining the dominant functional motions of biomolecules through many studies (5-12). It has been shown that the motions computed using ENMs correspond well to the principal components of MD
trajectories (13-15), can aid in molecular structure determination (16), refining of structural models (17), and in flexible docking (18, 19). It has become common to coarse-grain a structure for use in ENMs. This is the practice of picking out a representative set of spatial points to represent the system (see Figure 1.1). These points are then connected with harmonic springs as described in Figure 1.2. Sometimes the details remain important, such as the effect of mutation on an active site. In this case, a mixed coarse-graining approach (20) can be implemented, as shown in Figure 1.3.

![Figure 1.3: Graphical example of residue level coarse-graining for use in mixed-resolution ENM modeling and application of the interaction cutoff radius $r_c$.](image)

The activity of proteins depends not only on their structure, but also the fluctuations or dynamics of the structure. Experimental studies early in protein science showed that substantial structural fluctuations occur in proteins, and that these fluctuations are essential to protein function (21, 22). Theoretical advances have helped to clarify the nature and function of protein dynamics. Such motions are more than local fluctuations as large-scale collective motions have been experimentally observed in nucleic acid and protein structures (23-27). Allosteric mechanisms were identified early on in protein science (28, 29) and represent the general case of collective motion. In the words of J. A. McCammon in 1984, “of special importance is the strong coupling observed between local and collective displacements; this coupling governs the character of many ligand-binding processes and structural transformations that are essential to biological function” (30). Function mediated by dynamics even plays a role at the tissue specificity level. For example, Engler et. al. (31) analyzed adhesion molecules and tissue specific effects mediated by soluble factors and mechanical signals.
1.2. Significant contributions of this dissertation

This dissertation focuses on applications of the Elastic Network Model (ENM) for determining the functional dynamics of important biomolecules, often requiring the ENMs to be used in new ways that further their reach by including additional aspects of interactions. Chapter 2 includes both of these aspects. The ability of ENMs to describe multiple experimental structures of tRNA is investigated, enabling us to learn how effective these models are for RNA and what the minimal coarse-grained models are for nucleotide systems and how these differ from those for proteins. Also, we demonstrate the ability of multiple dissimilar ANM models to better capture the ensemble compared to the use of a single model. Chapter 3 focuses on the mechanism of a new protein structure, CusCBA, which is a drug efflux system specialized to export copper and silver ions from the cytosol. This is a membrane bound protein, and we have used a novel treatment of the membrane with the ANM in order to compute the efflux mechanism. The mechanisms of telomerase (Chapter 4), the ribosome and ribosomal proteins (Chapter 5), and antibodies (Chapter 6) are also investigated, showing that structure and dynamics play essential roles for each of these biomolecules. Chapter 7 takes ENM studies further by using them to calculate vibrational or fluctuation-based entropies. Entropy is usually ignored in most structural studies due to the difficulties of its estimation and the myriad possible environmental contributions. Our entropy measure is then combined with an energy estimate using the newly developed 4-body potential of Feng, et. al. (32), to improve the recognition of a native (low energy) structure within a large set of predicted structures or the native binding pose within a collection of docked structures. Our preliminary results are positive and show significantly improved accuracy compared to the most accurate energy based classifications. Chapter 8 introduces the newly developed and freely available software MAVEN which has been constructed to allow users to more easily implement many types of ENMs as well as aiding in model preparation, generation, and analysis. The major part of this work concerns protein structures, which are often described by fold class, where a fold is defined as a specific arrangement of secondary structure elements. Chapter 9 discusses our work that computationally examines the relationships between protein folds. The goal there is to address the question of whether fold space is continuous or discrete, and how well state-of-the-art structure matching algorithms distinguish between differences in fold and the conformers of one fold.

Other projects have been placed into appendices, including an algorithm which combines many of the methods described in Chapters 1-8 into a pipeline that would computationally explore protein conformation space at a coarse level (Appendix A). The two major contributions of this algorithm, if
successful, would be the prediction of new low energy conformations of a protein's structure from a starting conformation, and the refinement of predicted or modeled structures. Each of the other appendices represents collaborative work. In Appendix B, we worked with Dr. Christopher Tuggle of Iowa State University to computationally analyze two copies of the IL1B gene that are expressed at different developmental stages in the pig (*Sus scrofa*) and in different tissues looking for insights into possible functional differences between them. Molecular Dynamics studies of the Tec and Src kinases are underway in collaboration with the lab of Dr. Amy Andreotti of Iowa State University (Appendix C). We also collaborated with Dr. Bruce Shapiro at the National Institute of Health to compute the dynamics of newly designed RNA structures (Appendix D). Preliminary work on animating sub-cellular images is presented in Appendix E. Future work on this project will use data provided by professors at the University of Florida and the University of Colorado at Boulder to animate images of the Golgi apparatus. Work with Dr. Robert Houk of Iowa State University is described in Appendix F where the highly conserved YbhB protein, which is believed to regulate cellular proliferation but remains poorly characterized, is studied by mass spectroscopy, determining that different organisms utilize different oligomeric states.
CHAPTER 2. Elastic Network Modes Capture the Motions Apparent in Multiple tRNA Structures

 manusipt to be submitted

2.1. Abstract

The structural variability within an ensemble of 75 superimposed experimentally determined tRNA structures indicates a limited set of motions that are captured by a small set of principal components (PCs). These limited motions closely resemble the motions computed with a small number of low frequency normal modes from Elastic Network Models (ENMs), either at atomic or coarse-grained resolution. By utilizing only six dissimilar modes from multiple ENMs, 92% of the structural ensemble variance in the first 4 PCs is captured. Improving the efficiency with which computational models of structural motions capture the dominant functional motions of biomolecules will increase the utility and applicability of the models. ENMs have previously been tested for proteins against atomic simulations and experimental ensembles with similar strong agreement. Various ENM model types, parameters, and structure representations are tested, exposing differences between optimal models for protein and for RNA. The loss of dynamical information upon coarse-graining is somewhat larger for tRNA than for globular proteins, indicating perhaps less cooperativity within the structures reflective of the relative packing densities within the structures. If structures are superimposed based on the dominant normal mode there is a modest gain in quality of the alignment. These findings further show the utility of ENMs and demonstrate their applicability to RNA structures and ribonucleoproteins such as the ribosome.

2.2. Introduction

In the Protein Databank (PDB) (33), many structures have been determined either under different conditions, for mutants, with different ligands bound or simply reported by different investigators. These static X-ray structures may not individually inform us about the dynamics of proteins, but collectively they constitute an ensemble of structures that can inform us about dynamics by providing snapshots of a structure in its various conformations. These conformations are likely to be a meaningful sample of the inherent flexibility of the biomolecule. In order to better understand these functional motions, Principal Component Analysis (PCA) can be employed to convert the ensemble into a set of vectors that capture its spatial variance. These vectors can then be compared to computed models of RNA motion.
A computational method for representing the intrinsic flexibility of proteins that has achieved significant validation and widespread use with proteins is Normal Mode Analysis (NMA) using Elastic Network Models (ENMs) to obtain the motions of a structure at atomic, coarse-grained, or mixed resolution, based on the simplest force field (20, 34). ENMs, in particular the Anisotropic Network Model (ANM), have proven useful for obtaining the dominant functional motions of biomolecules through many studies (5-12). It has been shown that the motions computed using ENMs correspond well with the principal components of MD trajectories (13), can aid in molecular structure determination (16), refining of structural models (17), and in flexible docking(18, 19). In Yang, et. al. (8), we found that the spatial variance seen in the superposition of many structures of the HIV-1 protease have a close relationship with the normal modes computed with ANM, and Bakan and Bahar (14) investigated kinases with a similar approach. These studies have shown that the accessible conformations of a protein can be sampled by the dominant motions of one representative structure. The environments captured within this ensemble of structures include the apo, drug or RNA bound, and in complexes with other proteins (see Supplemental Table 1). Previous study has shown that these ENM models are capable of reproducing key aspects of nucleotide dynamics, but that they may not be as accurate for loosely packed structures as they are for densely packed ones, such as globular proteins (35). Here we investigate the ability of normal modes from ANMs to capture a sample set of structures for tRNA. If the same strong correspondence holds for this RNA structure as for protein, then a higher level of confidence will be gained for the application of ENMs to other folded nucleotide structures as well as co- RNA-protein structures.

How structure alignment is carried out can affect the results. Structure alignments have been carried out with many different published algorithms, varying from minimizing the Root Mean Square Deviation (RMSD) of all points in the system, to aligning secondary structures, and iterating to refine fits by down-weighting aligned pairs that are far apart spatially. Multiple structure alignment methods have also been devised (36, 37), but are not tested here. Despite the differences in these algorithms, analysis of the effect of alignment choice on structural studies has not often been performed or even considered. In this work we compare the effect of alignment choice on the overlap between normal modes from ANM models and PCs from the superposition of multiple structures. Additionally, we investigate the extent of loss of dynamics information with various levels of coarse-graining and the effects of including the crystal environment.
2.3. Methods

2.3.1. Dataset (Ensemble of Structures)

In this work we start from the tRNA structure described by PDB code 1TRA. Its sequence was submitted to BLASTn (http://www.ncbi.nlm.nih.gov) against the entire PDB, yielding 91 hits. Some of these were distant homologues or had substantial insertions. After removing these, there remain 75 tRNA structures, which are listed in Supplemental Table 2. Similar to the protein case(8), we capture conformations from wild-type, antibiotic bound, ribosome or elongation factor bound, cognate synthetase bound, etc. The Multiple Sequence Alignment of the 75 structures is shown in Supplemental Table 1 where the position of chemically modified bases is also marked. We not only consider principal components and ANM models built from phosphate atom coordinates, but also from the ribose sugar and backbone atoms (see Supplemental Figure 2). Only backbone atoms are considered for the more detailed analysis to avoid the complications that would arise from sequence variation and modified bases. Seven structures were not of sufficiently high resolution to resolve these backbone atoms and are thus excluded from the more detailed analysis (listed in Supplemental Table 3). Some of these remaining sequences have fewer than 76 nucleotides. To permit retention of these structures in our dataset we consider only the 73 nucleotides that best fit the 1TRA structure by allowing up to 2 nucleotides on the N-terminus and one nucleotide on the C-terminus to be missing.

2.3.2. Elastic Network Model

This section provides a full treatment of the common ENM methods utilized in Chapters 2 through 8. Specifically, it is a combination of Methods sections from Chapters 2, 6, 7, and 8.

Coarse-grained protein structures are often represented by $C^\alpha$ atom coordinates with harmonic springs to connect spatially close residues, since it has been shown that the dynamics of such coarse-grained structures closely resembles that of the atomic structures (9). This is an elastic network representation of the protein structure. The Gaussian Network Model (GNM), the earliest and one of the simplest of the elastic network models, is used to compute the relative magnitudes of motion. It was originally proposed by Bahar, Atilgan, Haliloglu and Erman for coarse-grained models in 1997 (38, 39), who applied the assumption postulated by Tirion (34) for atoms that the interactions of both bonded and non-bonded contacts in proteins can be represented by a single universal spring. The Anisotropic Network Model (ANM) proposed in(40), can be used to compute the directions of motions of all points in such a structure. To generate an ANM model we first construct a Laplacian (or Kirchhoff) matrix using
Equation 2.1 where \( r_c \) is a cutoff radius (typically 10-13Å), \( d_{ij} \) is the distance between atoms \( i \) and \( j \), and \( \gamma \) is the spring constant, taken to be identical for all interactions. The potential energy of such a system is given by Equation 2.2. We then compute a matrix of second derivatives of the potential energy - details are given in Reference (40) - the eigenvectors \( (Q_k) \) of which are called normal mode shapes and the eigenvalues \( (\lambda_k) \) are the corresponding square frequencies \( (\omega_i^2) \). Note that for clarity, \( i \) and \( j \) index atom positions, while \( k \) indexes the modes. Only a few modes are usually important since the contributions of modes to the total motion decrease rapidly with increasing mode index. For a given normal mode, \( Q_k \), fluctuations of the structure are computed with Equation 2.3. Low frequency normal modes represent the collective motions of the system and have been shown to be biologically relevant.

\[
\Gamma = \begin{cases} 
-\gamma & d_{ij} \leq r_c \\
0 & d_{ij} > r_c \\
-\sum_{k=1,j\neq k}^N \Gamma_{jk} & i = j
\end{cases} 
\]

\[
V = \frac{\gamma}{2} \Delta R^T \Gamma \Delta R 
\]

\[
\Delta R_i = Q_k A_i \cos(\omega_i t)
\]

Extensive applications of NMA to biological and chemical systems have been discussed in Cui and Bahar (41), Jernigan and Kloczkowski (11), and Sen et al. (9). For reviews of ENM methods and their use in structural biology, see references (5) and (7). Successes with these methods make it clear that functionally important motions of biomolecules are usually governed by packing density and their slowest motions. These and many other studies have enabled computations of the important motions on time scales beyond the usual reach of atomic Molecular Dynamics (MD). ENMs can be generated for small and medium sized proteins in seconds or minutes, a huge gain in comparison with the extremely long computational times required for corresponding MD studies. Recent work by Bakan and Bahar suggests that ANM may even sample conformation space more thoroughly than classic MD (14). It has been demonstrated that extremely large molecular assemblages can be even further coarse-grained without loss of the major important motions (9). More detailed analyses are available by use of elastic models that employ mixed-resolution models, where most of the structure is coarse-grained but with the regions of special interest remaining in atomic detail (see Figure 1.3).
The dominance of the low frequency normal modes is universal, and usually there are only a few of these characteristic motions that are truly important. These motions are also the most entropically favored due to the inherent geometry of the conformation. Higher frequency modes represent more local motions that is less likely to be representative of functional motions of the biological system, but may be useful in identifying nodes that are important for energy transfer through the structure (42).

To obtain the mean square fluctuations (MSFs), which are related to the B-factors (also known as temperature or Debye-Waller factors) from X-ray crystallography, the stiffness (Hessian in ANM) matrix must be inverted. A pseudo-inverse is computed using equation 2.4 since the stiffness matrix is singular (has zero-value eigenvalues corresponding to rigid body motions).

\[
\Gamma^{-1} = \sum_{k}^{1} \frac{1}{\lambda_k} (Q_k Q_k^T)
\]  

(2.4)

The summation of normal modes, \(Q_k\), and square frequencies, \(\lambda_k\) (the eigenvector and eigenvalues, respectively), is over all relevant modes (\([7 \ldots 3N]\) in ANM and \([2 \ldots N]\) in GNM). Modes 1-6 for ANM are the rigid body motions of the structure in three dimensions and mode 1 is the rigid body motion in the one dimensional GNM. The MSF of each atom in the system is then computed using Equation 2.5. When \(i = j\), \(\langle \Delta R_i \cdot \Delta R_j \rangle = \langle \Delta R_i^2 \rangle = \frac{3k_B T}{\gamma} [\Gamma^{-1}]_{ii}\). Changes in internal distance can also be calculated from \(\Gamma^{-1}\) using equation 2.6.

\[
\langle \Delta R_i \cdot \Delta R_j \rangle = \frac{3k_B T}{\gamma} [\Gamma^{-1}]_{ij}
\]

(2.5)

\[
\langle (\Delta R_i - \Delta R_j)^2 \rangle = \langle \Delta R_i^2 \rangle + \langle \Delta R_j^2 \rangle - 2 \langle \Delta R_i \cdot \Delta R_j \rangle
\]

(2.6)

ENMs can also be constructed by considering all pairs of points to be in contact and their interaction strength (the spring constant) to be a function of their separation. This alternative ENM is implemented as an inverse power dependence as in equation 2.7, but distance-dependent step-functions (43) and exponentials (44) have also been used.

\[
\Gamma = \begin{cases} 
-d_{ij}^{\chi} & i = j \\
-\sum_{k=1, k \neq j}^{N} \Gamma_{ij} & i \neq j 
\end{cases}
\]

(2.7)
Many other model types have been developed including methods where bond or dihedral angles are explicitly taken into account (45-47) and are implemented in MAVEN by incorporation of the freely available Spring Tensor Model developed by Lin and Song (45). We also implemented the nearest neighbor method which utilizes a coarse-grained model (usually one point per residue), but assigns contacts between residues based on an atomic model of the system. To facilitate further detailed analysis, one may employ mixed resolution modeling (48, 49). In this scheme most of the structure is coarse-grained, but any region of interest is included in greater detail (see Figure 2.1D). One may then choose different cutoff values for the two regions and use the geometric mean of the cutoff values as the cutoff for connections between two regions of different resolution (48).

In the present tRNA study, we also generate atomic and coarse-grained models with various levels of detail in order to learn about how best to represent RNA and also to learn about the amount of information lost upon coarse-graining. The crystal environment is also considered which Riccardi et. al. has shown improves protein models (43). We also test an alternative to the cutoff based ANM by defining spring interactions using distance dependent springs, as in equation 2.7 and explained further in Yang, et. al. (50).

2.3.3. Principal Component Analysis (PCA)

PCA is often used for dimensionality reduction on complex data where the idea is to rank order the contributions of granular variables. The first few PCs capture a significant part of the variance, with the first giving the largest contribution, the second capturing the largest part of the remaining variance, and so forth. If we begin with N-dimensional data vectors (in this case we have M atoms and therefore the modes are \( N = 3*M \) dimensional to capture the 3D coordinate of each), we will have N PCs. After extracting the rigid body motions, there are N-6 remaining internal modes of motion. This rank ordering of the PCs allows us to ignore most of them while retaining most of the important information. All structures are superimposed onto 1TRA using one of four different alignment algorithms. We then construct a matrix where each row holds the information for a single tRNA structure. Columns are then variables – one for each spatial coordinate. PCA is performed on this matrix using MATLAB (2010a, The MathWorks, Natick, MA).
Figure 2.1: PCs of the tRNA ensemble constructed from all common atoms and highlighting the physical limits of deformation which are not the same in each direction.

A) Scatter plot of the structures in Principal Component (PC) space. Each tRNA structure is plotted as a point in PC1-PC2, PC1-PC3, and PC2-PC3 space after being aligned using TM-align. Structure 1T3R is marked by a red star and is the central structure to which the ensemble was aligned. The structure closest to the origin in PC space is 2B9MW, which is marked with a blue star. Due to the proximity of 2B9MW to the origin, it is used for illustrating the PCs. B-G) The first three PCs are shown as deformations to 2B9MW in two orthogonal views using PyMOL(51). We show the unaltered structure in yellow (ribose and bases) and orange (backbone trace) with the PC deformations of phosphate atoms only that are of significant amplitude as blue arrows, and the resulting backbone trace in red. PCs are shown as B) +150 PC1, C) -150 PC1, D) +100 PC2, E) -100 PC2 F) +50 PC3 G) -50 PC3. Magnitudes used for each mode are taken from the maximum deviation seen in A. From these images we begin to understand the physical limits of each PC. For instance, the negative directions of PC1 and PC2 correspond to compression of the minor grove of the anticodon arm and simultaneous stretching of the anticodon stem-loop.
Neither of these actions can be supported to this degree by the structure. See text for more detail. PC3 consists of a smaller motion (and accounts for a smaller contribution to the total variance) that again describes communication between the acceptor arm and anticodon stem loop.

2.3.4. Structure Alignment

We utilize three established alignment methods, propose a new ENM-assisted algorithm, and consider their effect upon the relationship between PCs and normal modes. A minimum total root mean square deviation (RMSD) shape matching alignment is constructed and referred to here as “minR.” The second alignment procedure begins with a sequence alignment and then uses a weighted RMSD superpositioning of the well-aligned atoms (seqW). Finally, we use TM-align (52) which aligns secondary structure elements and employs a heuristic to refine the initial fit. After considering the performance of these three alignment types, a normal mode based alignment protocol is considered where atoms with lowest mean square fluctuation in the global modes are aligned by minimizing their total RMSD.

2.3.5. Structure Representation

ANM performance is tested across multiple structure representations at atomic and coarse-grained resolutions. These range from the simplest model which takes into account only the backbone phosphate atom positions, two points per base (P and C2’), three points (P, C2’ and O4’), all heavy (non-hydrogen) atoms, all atoms including hydrogen atoms (with and without covalent bonds), including the X-ray resolved waters, and with symmetry related intermolecular atoms within 5, 7, or 12Å filled in with solvent generated using Solvate by Grubmüller and Groll (53). See Supplemental Figure 2.2 for further explanation on these representations.

2.3.6. Comparing PCs and Modes

Tama and Sanejouand (54) defined a normalized overlap for comparing the \( i^{th} \) PC \( (P_i) \) to the \( j^{th} \) mode \( (M_j) \) and we use this here to compare the PCs from the experimental structures with the normal modes computed for a representative structure:

\[
O_{ij} = \frac{|P_i \cdot M_j|}{\|P_i\|\|M_j\|}
\]  

(2.8)

The cumulative overlap (CO) between the first \( k \) normal modes and a given PC measures how well the first \( k \) modes together can capture the variance represented within a single PC (equation 2.9). The overlap between the space spanned by first \( I \) PCs and the first \( J \) low frequency ENM modes was defined by the root mean square inner product (RMSIP) in Reference (55) and is shown here in equation 2.10.
2.4. Results

In this study we compare a collected ensemble of 75 tRNA structures superimposed with four different alignment methods with the modes of motion calculated from elastic network models. In Table 2.1 we report the percent variance captured by the first three PCs of the 75 structures across all four alignment types using phosphate atom positions only (further details are given in Supplemental Figure 1). More detailed structure models are also considered by including all backbone and ribose sugar atoms (visualized in Supplemental Figure 2) in the 68 structures having these atom positions reported (sufficient resolution), computing a second ensemble and comparing to atomic ANM. We find that 59 to 70% of ensemble variance is captured by only three PCs.

For the most detailed models we take all backbone atoms and align the different structures with TM-align. Then the covariance matrix is decomposed into Principal Components (PCs), and ANM models constructed from all heavy atoms in the representative structure 1TRA. Backbone atoms are considered for the PCs to avoid complications that would arise from sequence variation and any chemical modification of the bases. For comparing between these PCs and the ANMs, we select the subset of atoms form the normal mode corresponding to the atoms selected for PC generation and renormalize the mode to have unit length. Figure 2.1 displays the structural ensemble to show how the individual structures are distributed in PC space. We see that the structure to which others are aligned (1TRA) is not at the origin in PC space. Instead, 2B9MW is much closer. These two structures are only 1.6Å RMSD apart from each other, so the mode-shapes computed from each would be similar. For this reason, the deformations visualized in Figures 2.1 and 2.2 are considered using initial coordinates from 2B9MW. It is apparent that an asymmetry exists; the structures exhibit greater deformations in one direction along the PCs than the other. Most notably, PC1 shows the most extreme deviations occurring in the positive
Figure 2.2: Normal mode space and dominant motions from atomic ANM visualized similarly to Figure 2.1.
We visualize the heavy (non-hydrogen) atom ANM mode space similarly to Figure 2.1. A) All 68 structures are again shown in a plot of the 2D projections onto mode spaces. To determine the mode coordinates, we compute the minimum difference between 2B9MW and a given structure using mode 1, mode 2, or mode 3. Minimum difference is determined by the minimum RMSD conformation to the target structure after deforming 2B9MW by the given mode. The amplitude used is then the coordinate of each structure. We find that the first mode clusters all structures based on the relative orientation of the anticodon stem loop and the acceptor arm. Modes 1 (B & C), 2 (D & E) and 3 (F & G) are shown for their positive (B, D, F) and negative (C, E, G) deformations using the maximum amplitude seen in section A. These include compression of the stem loop minor groove, in or out-of plane motion of the acceptor arm, and minor rearrangements of the shoulder domain.

direction small deviations in the negative direction. This is due to the physical limits to compression of the minor groove in the anticodon stem loop in the negative direction of PC1, whereas the positive PC1 direction exhibits simultaneous unwinding and bending having less physical constraints.
ANM using atomic details accurately captures the ensemble. Model performance for the heavy atom ANM model and atom-derived PCs was evaluated for cutoff values up to 25Å, but little change was seen in either Cumulative Overlap (CO - equation 2.9) or Root Mean Square Inner Product (RMSIP - equation 2.10) above 8Å. CO quantifies how well an individual PC is captured by a set of modes, while RMSIP considers a subspace defined by a set of PCs and the extent that it is covered by a set of modes. We compare 3, 6, 10, or 20 modes to each of the first 3 PCs (CO) or the combination of up to 3 PCs (RMSIP) and find that considering 20 modes captures 75-80% of the dominant PCs. See Supplemental Figures 1 and 3 for overlaps of individual modes and individual PCs used for select cutoffs. Using fewer modes accounts for less of the ensemble variation, however, by using only 3 normal modes 55% of the space spanned by the first 3 PCs is covered. Figure 2.2 shows a visualization of the mode-space using a representation similar to Figure 2.1. We immediately see that the mode 1 projection is very different from the PC1 projection. Mode 1 partitions the structures into those that have the acceptor arm in an “up” or “down” position coupled to a corresponding extension or compression of the anticodon stem loop. Mode 2 involves the isolated movement of the anticodon stem loop, while mode 3 is dominated by tangential movement of the acceptor arm relative to mode 1. Interestingly, in order to achieve the amount of flexing of the anticodon stem loop that observed in the PCs, the ANM modes require the counterweight of a significantly exaggerated motion of the acceptor arm (mode 1). While the acceptor arm may appear to have exaggerated motions here, similar to the “tip effect” noted by the Ma group(56), nonetheless its large motion does appear to meet the required deformation it undergoes when binding inside the ribosome.

One of the more remarkable findings from studies on proteins is the excellent performance of coarse-grained models for representing the motions. Some loss in the fidelity of computed motions occurs, but nonetheless the essential aspects of these motions are observed in both the atomic and coarse-grained models. We have already demonstrated that the ANM motions of one representative structure capture the motions observed across a large set of experimental structures. Next, we will investigate the performance of coarse-grained representation of the tRNAs. Since RNA is a very different material from protein, we test multiple structural representations of tRNA to determine the level of detail that is most informative for ANM motion computations. We consider the effect on elastic network model performance, as the resolution with which we treat the tRNA and its environment is changed. In order to compare the same experimental data to various models, we construct the PCs by using the phosphate atom coordinates. The models used vary in resolution from one point per base, all heavy
atoms, including hydrogen atoms, or inclusion of increasing amounts of the crystallographic environment (see Methods and Supplemental Figure 2) and the results are summarized in Figure 2.3 for CO using either 6 or 20 normal modes. While coarser-grained models capture a large proportion of the PCs with a small number of modes, the more detailed heavy atom model exhibits some redundant coverage. That is, multiple mode motions in the atomic model capture an individual PC, resulting in redundant coverage of those deformations. This is particularly interesting because the normal modes are mutually orthogonal and form a basis set. However, when projected onto a three dimensional structure, there are portions of some modes with redundant structural motions. If we could identify modes having significant motions in common, we might be able to further simplify the mode-motion space. This is one advantage of the coarse-grained modes because that approach partially accomplishes this reduction directly by ignoring some of the local fluctuations. The single point per nucleotide does not fully capture the backbone motions; this differs from proteins where one point per amino acid can capture the backbone dynamics. In nucleotide systems, a minimum of two torsion angles are required for the backbone as well as a point corresponding to the base position (57). This may explain the redundancy we see when comparing to coarse-grained PCs. More detailed models that take into account the crystal environment or closely packed water molecules, do not outperform the phosphate only ANM model (data not shown). Since inclusion of crystal environment has shown improved results for proteins, this may point to a characteristically different behavior for RNA. The extent of differences between the best performing high and low resolution (more or fewer atoms) models appears to be greater for this tRNA structure than is generally observed for proteins when only Cα atoms are considered (9).

Distance dependent springs have consistently shown modest improvements for protein dynamics and are considered here. Results are summarized in Figure 2.3 using CO, Supplemental Figure 3 which shows individual mode-PC overlaps and the effect of the power dependence, and Supplemental Figure 4 which shows the effect of alignment algorithm choice on CO. Overall, this all-pairs model seems to perform well across parameter space. The highest degree of redundant motion is seen for a linear power dependence and atomic resolution. Interestingly, the cutoff methods limit to this case and do not
PCs are generated for various elastic network models with different sets of atoms, and then the phosphate atom positions only are used for comparisons of the first two principal components. Performance of each model is judged by the Cumulative Overlap (Equation 2.9) between 6 or 20 normal modes and the first and second PC. In some cases the lines do not extend to low cutoff distances because the models break up into separate pieces and have too many zero eigenvalues. Four representations of the 1T3R structure are considered; phosphate only, two or three points per nucleotide, and all heavy (non-hydrogen) atoms. A) Behavior of cutoff-distance based ANMs. While the coarser models capture a large percent of motion in the PCs, the heavy atom model can exhibit redundant coverage (CO > 1) even when only 6 modes are considered. More detailed models that take into account the crystal environment or closely packed water molecules do not outperform the phosphate only ANM model (not shown). Interestingly, there is not a loss of performance at high cutoffs, but generally a convergent limit is reached. Supplemental Figure 4 shows results for cutoffs up to 50Å. B) Behavior of distance-dependent spring elastic network models used in ANMs. Springs are assigned a stiffness of $d_{ij}^x$ where $x$ is the power dependence and $d_{ij}$ the distance between atoms $i$ and $j$. The heavy atom model exhibits the best coverage of PC1-PC2 space for the 20 mode comparison. Overall, the coarser models and cutoff based models appear to be better at representing the ensemble motions within the first 6 modes. However, the 20 modes are almost consistently better represented by the most detailed atomic model. See supplemental Figure 3 for the correlations between the PCs and individual modes for the power dependent spring models.
exhibit as strong of a performance decrease at long cutoffs as is usually observed for proteins. However, the modes with highest overlap are shifted to slightly higher frequencies than in the cutoff based models where the very lowest frequencies have the highest individual overlap with the PCs (Supplemental Figure 3). These results may be indicative of the increased importance of long range electrostatic interactions across RNA structures that play a more dominant role in fold determination and stabilization than in globular proteins.

Other representations and ANM methods were tested, but none outperformed the data presented in Figure 2.3. The Nearest Neighbor method can improve ENM performance in proteins by computing a coarse-grained ANM model, assigning contacts based on a more detailed representation. Such a model does not appear to outperform the simpler representations (not shown). Inclusion of hydrogen atoms or covalent bonds increases the redundancy within PC2, but does not improve the performance within the first 6 modes (not shown).

In Supplemental Figure 4 we show the effect of the alignment algorithm choice on the CO. While there is some loss of performance with the longest cutoff distances, the loss is not as large as generally is observed for proteins. Phosphate atom models with a cutoff value for interactions of up to 50Å were tested. Robustness even at high cutoff distances merges with the high performance at a power dependence of zero (maximum distance in 1TRA is about 75Å). Interestingly, atomic resolution ANM models also perform well at high cutoff values, but much of the performance improvement is in the form of increases in the redundant coverage of the PCs likely due to local motion that is not possible for the coarse-grained system. This pattern may point to the increased contribution from long range electrostatics in RNA structures in comparison with proteins. Supplemental Figure 1 displays comparisons of the principal components for ensembles aligned with the four different methods to the normal modes derived from phosphate atom positions and an ANM cutoff of 18Å. PCs using all four alignment procedures are similarly concise, but in the PC space they appear to be quite different. This underscores the importance of the alignment procedure for interpreting what PCs or computed motions may signify and the relative similarities of each structure within the ensemble – critical for interpretation. The distribution of structures exhibits a directional asymmetry seen in all cases, showing that the structure is easier to deform in one direction along a given PC than the other. Translating this asymmetry of motion (knowing the physical limits) to the modes would be a valuable improvement to future ENM studies.
It is important to make a direct comparison between the normal modes and the PCs. In structural studies the lowest frequency modes are often considered first by visualizing the normal mode distortions (animating the structure using the mode shape). It is observed in general that only the first few modes are the most important. We consider the effect of the ANM cutoff (22Å used here) and alignment algorithm choice (indicated by color) on the overlaps of the very lowest frequency modes with the PCs generated from the ensemble of structures. We also consider here two levels of detail for representing the structure – one point per base (A and B) or three (C and D), labeled 1P and 3P, respectively. Cumulative Overlap is calculated using equation 2.9 and either the 3 (A and C) or 6 (B and D) lowest frequency modes, labeled 3M and 6M, respectively. All three aspects (cutoff, structure representation, and alignment algorithm) are important for interpreting the computed motions. The ENM-assisted alignment yields the most consistent results, with the minimum total RMSD (minR) structure alignments achieving the maximum performance only for the coarsest of systems.

The strong relationship between the first PCs and normal modes points to the possibility of using the normal modes to align the structures. To test this hypothesis we first calculate the mean squared fluctuation of each atom in 1TRA from the two lowest frequency modes. We next select the 20th percentile of mean-square fluctuation (Supplemental Figure 5) and align the ensemble using least squares fit on only these atoms. The relationship between CO using the lowest frequency modes and all four alignment types is compared in Figure 2.4 where we find ENMa to be the most consistent. Thus, aligning using the atoms that move least in the global mode of motion results in increasing the overlap of the space spanned by the first few normal modes and PCs at the cost of some conciseness of the overlaps. This is a somewhat opposite approach from that of the recently published ALADYN method of Potestio and colleagues (58) which optimally matches the largest motions from two protein structures. Refinement of this method could lead to further improvements and insight into the structure-function relationship. Numerous studies have shown that the lowest frequency modes are biologically relevant.
ENM-assisted structure alignment further emphasizes that the global modes are indeed biologically meaningful.

To further investigate the amount of information lost upon coarse-graining of the system, we directly compare ANM modes from heavy atom and phosphate only models. Supplemental Figure 6 depicts the correlations between computed temperature factors (the relative amplitude of motion for each atom) for atomic and phosphate-only models and metrics that describe the agreement of their anisotropy. Anisotropy refers to how the atoms fluctuate to different degrees in each direction and is described by a 3x3 tensor. These tensors are compared using the modified real-space correlation coefficient and Kullback-Leibler distance, both of which are described by Zheng (59). We find that the models exhibit the most agreement with each other when the atomic cutoff is 7Å or larger and the coarse-grained is 20Å or greater. These parameters agree well with the parameters that best reproduce the structural ensemble (Figure 2.3). These cutoff values also correspond to the best performance as judged by correlation with crystallographic B-factors (Figure 2.5). Interestingly, the inclusion of hydrogen bonds does not significantly affect CO, but does have some modest impact in improving temperature factor correlation (data not shown). Thus, the cutoff based ANM models that best reproduce experimentally determined flexibility are also the most self-consistent in their computed motions.

The presence of bound protein can have a pronounced effect on tRNA structure. We find that structures far from the origin in PC space are almost always bound to proteins, but that numerous bound structures are close to the origin as well. Protein partners found far from the PC1-PC2 origin include cognate synthetases (3A2K, 1QF6, 1EIY), ribosomes (3FIH, 3E1A), and EF-Tu (1TTT). Protein partners found near the PC1-PC2 origin include ribosomes (2B64, 2HGR, 2HGI, 3DEG) and CCA-adding enzyme (1SZ1). If we measure the extent of interactions by the number of amino acids within a close radius of the tRNA, we find that the extent of protein binding is also not clustered in PC space, nor do the most highly bound structures (more than 10 Cα atoms within 7Å) represent outliers. Thus, neither the presence nor the extent of bound protein is observed to be a dominant factor in the sampled deviations.
Figure 2.5: Correlation between crystallographic B-Factors and computed fluctuations for A) cutoff based ANM and B) ANM with distance dependent springs where all atoms are considered to be interacting through connected springs with a weight of $d_{ij}^{-x}$ where $x$ is the power dependence.

A power dependence of zero results in a fully connected network where all springs are uniform in strength. Interestingly, coarse-grained cutoff-based models all appear to reproduce B-Factors equally well. Results differ somewhat from those in Figure 2.3 where cumulative overlaps were considered only for PC1 and PC2. First, the heavy atom model does not perform as well. Second, the power dependence of 6 best reproduces the B-factors, although in Figure 2.3 we see a smaller power dependence of 1 better matching the structural ensemble variance corresponding to PC1 and 2 for PC2.

2.5. Discussion

In this study we consider the ability of ENMs to capture the spatial variance seen in a superimposition of 75 experimentally determined (using X-ray crystallography) tRNA structures. ENMs have become an increasingly popular method for determining the dominant motion of macromolecules from the modest in size to increasingly large structures such as the ribosome. They have been confirmed previously to reproduce the Principal Components made from the covariance matrix of 164 superimposed HIV protease structures (8). We perform a similar analysis here for the tRNA to confirm that the low frequency ANM modes sample experimentally determined conformations of tRNA. The variation seen in the structural ensemble is visualized in Figure 2.1 and Supplemental Figure 1. The majority of this variation is from coupled motion between the acceptor and anticodon arms. After
comparing many ANM models we find that the regions of the tRNA that are constantly predicted incorrectly are the D and T arms. This region has relatively low experimental temperature factors and cutoff based ANM models tend to predict too much relative motion. This may be due to suppressed motion in the crystal due to packing, but may also be caused by the stabilization of non Watson-Crick (Hoogstein) interactions that are not explicitly present in our coarse-grained models. Due to the high performance of all atom representations at a low cutoff, a hydrogen bond ANM may exhibit further gains.

In many ENM studies, researchers begin by animating the biomolecule using the normal modes, visualizing the effects on the structure of excitation of one or more modes. It is most common to begin with the lowest frequency mode and visually compare the first few – perhaps up to a dozen. Thus, it is not as practically useful to show that 20 modes capture experimental ensembles when many researchers will only consider the first few. In Figure 2.4 we report the alignment-dependent performance using only 3 or 6 modes and either one or three points per base. We find that the minimum total RMSD alignment can give the most easy to interpret mode space (the very lowest frequency modes are most meaningful for capturing PCs), but that the ENM-assisted alignment is the most consistent. A notable consideration for interpreting the results presented here is that the actual functional meaning of each mode is not changed upon alignment algorithm choice, but their apparent meaning is. That is, the modes are strongly related to the structural ensemble, but one must properly align the ensemble in order for the metric to reflect this agreement.

Combining all available information, we find that the best model is an atomic level representation using a cutoff of 7Å. This is because it is a self consistent model, achieving similar results as coarse grained models. Also, it is capable of addressing more detailed questions concerning the tRNA motions compared to coarser models. We visualize the mode-space of this model in Figure 2.2, finding that the computed motions are very similar to the PC subspaces. The first mode of motion is noticeably dominated by a large motion of the acceptor arm which is coupled to motion of the anticodon stem loop. This motion groups the structures into two groups – those with the acceptor arm flexed down or upwards, relative to 2B9MW. The next two modes of motion refine the fit between structures by further modifying the anticodon stem loop minor groove compression along with out-of-plane motion of the acceptor arm.
Table 2.1: Percent of tRNA ensemble variance captured by the first 3 PCs using each alignment method. The percent variance captured by each of the first three PCs across four alignment methods and either a coarse-grained (phosphate atoms) or all-atom backbone. The backbone selection is defined as the ribose sugar and backbone phosphate and oxygen atoms, where the bases themselves have been excluded because of their variable atom types (see Supplemental Figure 2). The four alignment methods are abbreviated as: seqW, for sequence weighted – a sequence alignment is performed and atoms aligned by sequence are given more weight in the structural alignment; minR, for minimum total RMSD; ENMa, mean-square fluctuation in the first two modes is computed and the 14 points with the least motion are superimposed; TM denotes TM-align. From this data, no alignment method is clearly better than another, but the seqW method is the least consistent in its structural alignments. The conciseness of the modes differs and TM-align performs similarly to ENMa. For the coarse-grained ensemble, TM marginally outperforms ENMa and minR, but for an ensemble at atomic detail ENMa outperforms the others.  

This alignment protocol may appear to be the best by conciseness of the PCs, but a large portion of the variance is captured by this PC because of a poor structural alignment. Some of the structures exhibit the closest sequence match at the anticodon stem loop, leading to the larger part of the structure being under-weighted in the structure alignment. The first PC distinguishes these two groups – the bulk of the structure predominantly aligned, or the anticodon stem loop alone.  

For consistency, we align backbone atoms that meet the MSF criteria (see Methods).

Figures 2.3 and 2.5 point to the possibility of using multiple ANM models with different power dependences to predict the various motions captured by each PC. To test this possibility, we select a reduced set of modes from multiple ANMs by using their dissimilarity. Simply put, the chosen modes must predominantly move a different set of atoms, or move the same set in a different direction. The specific criteria used are 1) half of the atoms in the 50th percentile of amplitude must be different between the modes, or 2) atoms common to the 50th percentile of amplitude between the two modes have an average absolute dot product between their directions of less than 0.6. The second criteria is calculated by taking the absolute value of the dot product between the directions of motion (unit length) for a pair of modes and averaging across the atoms common to the 50th percentile of amplitude in both modes. Mutual dissimilarity is enforced by calculating, for each new mode under consideration, its similarity to all previously retained modes. If it is sufficiently similar to any of them, it is not retained. The results show improvement in the efficiency of the selected 6 modes to capture PC variance (see Table 2.2). From a pool comprised of the first 3 normal modes in each power dependent ANM, the modes retained as the 6 best are given as ordered pairs in the form (Power Dependence, Mode Index):
(1,1), (6,3), (7,1), (7,3), (8,1), (8,3). ANMs of various power dependences are capable of capturing a significant portion of the ensemble variance, but a combination of modes from multiple models is most efficient.

<table>
<thead>
<tr>
<th>Number of PCs</th>
<th>Power Dependence</th>
<th>3</th>
<th>4</th>
<th>8</th>
<th>Best6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.38</td>
<td>0.55</td>
<td>0.43</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.43</td>
<td>0.62</td>
<td>0.48</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.46</td>
<td>0.64</td>
<td>0.50</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.50</td>
<td>0.67</td>
<td>0.51</td>
<td>0.92</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2: Percent of variance apparent in the PCs captured by the first 6 normal modes shows that combining modes from ANMs with different power dependences is most efficient.

For springs with stiffness $d_{ij}^x$ where $d_{ij}$ is the distance between atom $i$ and atom $j$ and $x$ is the power dependence, we calculate the % Var captured by each model across the first 4 PCs (see Table 2.1 for the percent of ensemble variance captured by each PC). We show the three power dependent ANMs that individually can best capture the first 4 PCs using 6 modes. The 6 best modes (Best6) are selected from multiple ANMs based on mutual dissimilarity and are shown to provide a more efficient set (see text for details). For example, we capture 92% of the total ensemble variance within the first 4 PCs using the 6 best (mutually dissimilar) modes.

Considering the amount of redundant PC overlap when atomic models are employed, it may be possible to further reduce the mode space. The modes are always orthonormal vectors in the space that they are computed in. That is, for a system with $N$ atoms, a normal mode is a unit vector in a $3N$ space. When that vector is projected onto a three dimensional space (onto the structure) there are often large components of the motion that are similar. For instance, we have often observed the behavior where one mode describes anti-correlated motion of two domains and another correlated. Could such behavior be an indication of independent motion since linear combinations of the two motions could move one domain and not the other? Could the mode space be further simplified by filtering by common motions after the mode is projected onto the structure? We seek to explore these and other questions in future studies.

2.6. Conclusions

We find that the low frequency ANM modes reproduce experimentally determined structural variation for the tRNA. This is important in that it further justifies using these models for nucleotide and mixed nucleotide/protein systems. Previously, this type of confirmatory analysis had only been performed for protein. Now we can be more confident in the use of ENMs for nucleotide systems.
We also emphasize the importance of proper alignment protocols for biomolecules. Simpler alignment methods result in PCs that are similarly concise compared to more sophisticated methods. One should not consider only the conciseness of the PCs as this may in fact point to a poor structure alignment (Table 2.1). Alignment is important in interpretation of the results as it determines the relative similarity of each structure and their location in PC space. We also show that performing a simple least squares fit of the least mobile atoms in the dominant normal modes produces alignments that are comparable to other methods. However, at atomic detail, ENM-assisted alignment of the ensemble shows improved performance. This methodology may be useful for further exploring the conservation of functional motions in biomolecules. Contrary to what has been observed in proteins, inclusion of the crystal environment in the ENM did not improve performance. Similar to proteins, if significant coarse-graining is performed, the models still reproduce experimentally determined motion from the ensemble variance and crystal B-factors. However, one point per base coarse-graining results in more loss of dynamics information for the tRNA as is usually observed for proteins.

Considering all pairs as interacting, but with spring strength weighted by the inverse second power of the distance between atoms shows consistent gains for protein structures, but less clear dependence was observed here. A decrease in performance at long cutoffs is usually observed for protein structures, but this decrease was less for tRNA. This may point to the greater relative importance of long range interactions in defining and stabilizing RNA structures. However, from comparing the coarse and fine-grained models to each other directly, we find that their agreement is high. Utilizing mutually dissimilar modes from multiple ANM models is shown to result in the most efficient way to capture the structural ensemble’s variance. With the use of only 6 modes, we capture 92% of the variance within the first 4 PCs. Interestingly, the modes being retained as the 6 best are the ANM models having an inverse first power and an inverse 6th (or higher) distance dependence. Intuitively, this points to the importance of Coulombic interactions in RNA, which would be expected to show a sixth power dependence.
CHAPTER 3. Heavy metal ion transport complex CusCBA assembly, structure, and mechanism

3.1. CusA Structure Summary


Gram-negative bacteria, such as Escherichia coli, frequently use tripartite efflux complexes in the resistance-nodulation-cell division (RND) family to expel various toxic compounds from the cell. The efflux system CusCBA is responsible for extruding biocidal Cu(I) and Ag(I) ions. No previous structural information was available for the heavy-metal efflux (HME) subfamily of the RND efflux pumps. Here we describe the crystal structures of the inner-membrane transporter CusA in the absence and presence of bound Cu(I) or Ag(I). These CusA structures provide new structural information about the HME subfamily of RND efflux pumps. The structures suggest that the metal-binding sites, formed by a three-methionine cluster, are located within the cleft region of the periplasmic domain. This cleft is closed in the apo-CusA form but open in the CusA-Cu(I) and CusA-Ag(I) structures, which directly suggests a plausible pathway for ion export. Binding of Cu(I) and Ag(I) triggers significant conformational changes in both the periplasmic and transmembrane domains. The crystal structure indicates that CusA has, in addition to the three-methionine metal-binding site, four methionine pairs—three located in the transmembrane region and one in the periplasmic domain. Genetic analysis and transport assays suggest that CusA is capable of actively picking up metal ions from the cytosol, using these methionine pairs or clusters to bind and export metal ions. These structures suggest a stepwise shuttle mechanism for transport between these sites.

3.2. CusBA Structure Summary

Gram-negative bacteria, such as Escherichia coli, expel toxic chemicals through tripartite efflux pumps that span both the inner and outer membrane. The three parts are an inner membrane, substrate-binding transporter; a membrane fusion protein; and an outer-membrane-anchored channel. The fusion protein connects the transporter to the channel within the periplasmic space. A crystallographic model of this tripartite efflux complex has been unavailable because co-crystallization of the various components of the system has proven to be extremely difficult. We previously described the crystal structures of both the inner membrane transporter CusA and the membrane fusion protein CusB of the CusCBA efflux system of E. coli. Here we report the co-crystal structure of the CusBA efflux complex, showing that the transporter (or pump) CusA, which is present as a trimer, interacts with six CusB protomers and that the periplasmic domain of CusA is involved in these interactions. The six CusB molecules seem to form a continuous channel. The affinity of the CusA and CusB interaction was found to be in the micromolar range. Finally, we have predicted a three-dimensional structure for the trimeric CusC outer membrane channel and developed a model of the tripartite efflux assemblage. This CusC<sub>3</sub>–CusB<sub>6</sub>–CusA<sub>3</sub> model shows a 750-kilodalton efflux complex that spans the entire bacterial cell envelope and exports Cu<sup>2+</sup> and Ag<sup>+</sup> ions.

**Figure 3.1:** The swinging motion of PC2 coupled to TM8 is the dominant structural difference between the apo and copper bound forms and is captured by ANM. Computed with the same elastic network model used for figure 3.1, this motion is very similar to the difference between the apo and bound crystal forms. An arrow indicates the motion of swinging from the red conformation to the green. For clarity PC2 and TM8 are shown in thicker lines than the remainder of the monomer.
3.3. CusBA mechanistic insights derived from structural studies

The observed conformational changes between the Cu(1) bound and the apo forms, displayed in Figure 3.1, are shown to be intrinsically favored by the new structure through the use of normal mode analysis on the new structure. An elastic network model was investigated that is comprised of the Cα atom coordinates and a portion of coarse-grained lipid bilayer between the trans-membrane domains. The opening and closing of the space between PC1 and PC2 coupled with a shift of TM8 is reproduced by this simple model. The model does not show simultaneous opening and closing of all three periplasmic heavy metal entrance sites, but instead has a more complex opening/closing motion in which each pair of two adjacent is alternately open and closed (Figure 3.2). All three combinations of these coupled motions are obtained by using two of the most important normal modes.

![Diagram showing neighbor dependent opening and closing motions of the CusA structure using ENMs shows an asymmetric mechanism.](image)

We show the effects of the strong internal motions on the CusA trimer. In the center is a diagram of the periplasmic domain looking downward, towards the inner membrane. We build an elastic network model using the formalism of Atilgan (40) to predict the natural motions of CusA. On each of three sides we show the effect of two of the structure’s natural motions. These motions describe coupled opening and closing of adjacent periplasmic metal entry sites. Black wedges are shown in open and closed form, indicating the approximate angle between the edges of PC1 and PC2 in each motion. From these the alternating open/close motion is evident. Note that DC and DN also have motions in these modes, which is increased by combining modes 2 and 6. This makes the PC1/PC2 gap appear occluded, but this is only because we are viewing the molecule from the top down. The DC/DN trimeric ring moves almost rigidly, parallel to the membrane plane.
3.4. Modeling of the interaction of CusC with the CusA+CusB co-crystal and the injection mechanism

We begin by rigidly docking the structure of the CusC homolog, OPRM (PDB code 1WP1), to the structure obtained for the CusAB complex (figure 3.3). We find that this structure can rest on top of the CusB hexamer. This conformation may be an intermediate structure in pump formation. We then tried to open the ring formed by the six CusB monomers to allow CusC to bind much closer to CusA. Below, we show the basis of the calculations for opening the CusB to accommodate the injection of CusC, where two domains on each CusB monomer are pulled toward the corresponding sphere center using driving harmonic forces in NAMD and the CHARMM27+CMAP force field.

Figure 3.3: Docking of CusC to CusBA to assemble the full CusCBA system with CusC either on top of the CusB ring or injected into it.

(a) The α-helical end of CusC interacts with the α-helices (Domain 4) of CusB in the CusBA complex. (b) The funnel top of CusA, α-helical end of CusC and Domain 2 of CusB contact one another when the CusC channel slides further down the hexameric CusB channel. The surface rendering of the CusC3- CusB6-CusA3 complex is colored as follows: green, CusC trimer; blue, CusB hexamer; red, CusA trimer.

The motivation for CusC binding closer to CusA is motivated by the proposed sulfur transfer mechanism where silver ions are shuttled through the pump by passing them from one sulfur pair to the next. A homology model of CusC was generated by homology modeling to OPRM (44% identity) using I-TASSER. The top 5 models generated all have highly significant scores implying that the method is confident that the model is close to the structure that this sequence would adopt. In this model we
notice that there are no methionine residues in the lowest part nearest the inner membrane) of CusC (figure 3.4). This motivates the opening of the CusB ring to allow CusC to bind closer to CusA.

Rigid docking using Autodock and 3D-dock did not succeed in placing CusC at the desired location on the top of CusB. So, a steered molecular dynamics simulation was used to pull CusC towards CusB. CusB was held rigidly in this simulation. The resulting pose is shown in the left panel of figure 3.3. Next, we wish to determine a candidate structure where CusC is “injected” into the CusB annulus. Molecular Dynamics was employed with custom forces added to open the CusB ring (figure 3.5) Followed by steered MD, directing CusC toward CusA. To facilitate computations, only the middle section of the complex was modeled (figure 3.6). To allow the rest of the structure to be modeled back into the system the bottom 13Å and top 11Å of the protein was held rigid, while simulated annealing was run for a short time (marked with large brackets in figure 3.6). This refined CusB from its final SMD position.

Full relaxation of the complex to achieve an accurate description of the CusABC binding region has not been performed. This preliminary model may inform about the interaction, but is not likely to be accurate in detail to the lowest energy interaction between these proteins.
3.5. Extracting the ion-exchange mechanism

Dynamics information from simulations can often inform us directly about mechanisms. One of the most important lessons learned about protein structures is that their low frequency motions relate closely to their functional dynamics. For CusA, we would like to understand the details of how it moves to effectively transport heavy metal ions. One of the most informative approaches for tackling this problem has proven to be the use of elastic network models, categorizing structural motions by the normal modes to rank order them by frequency enables identifying the functionally important motions. The Jernigan group has pioneered the use of principal component analysis of multiple structures of the same or closely related proteins (8) and elastic network methods to extract molecular mechanisms, in

Figure 3.5: The top portion of the CusBA crystal structure is shown, displaying the directions with which the CusB annulus is computationally opened.

The CusB annulus is comprised of 6 CusB monomers, which we group into three clusters. Each cluster is pulled using customized MD toward the point in space with corresponding color.

Figure 3.6: Depiction of the MD environment used to refine the CusC-CusBA post-injection complex.

After the SMD simulation, the pose needs to be refined in order to make any detailed analysis of the full tripartite pump. Atoms within the bracketed regions are held internally rigid (i.e. CusC can move freely, but atoms toward the top will move rigidly) to allow the full-length molecules to be added back in at a later time.
applications from enzyme reactions (60, 61) to the intricate mechanistic steps of the ribosome (49, 62-66). These approaches have proven themselves and now many others have successfully been applying them to a broad range of problems (67, 68). The particular version of elastic network models that we employ here is the Anisotropic Network Model (ANM, (40, 50), which is used to extract the most probable directions of motions from a static structure. Interestingly, most functional biological structures have only a few types of motions that relate closely to their functional mechanisms. We employ ANM to analyze the CusA heavy metal ion efflux system, specifically focusing on how mode motions affect the five clusters of methionine residues that have been proposed to act to shuttle the ions through the structure. The intention is to delineate the details of the transfer of metal ions between the various binding sites.

**Figure 3.7**: Comparison between standard ANM modes and mode-motions where virtual bond lengths and approximated Van Der Waals exclusion are taken into account. (Figure 3.7: continued) We utilize the HIV-1 protease structure 1T3R, coloring the B) original structure tan (inhibitor green), with the A) positive and C) negative mode motions. The red backbone trace is the resulting conformation from standard ANM usage, while the blue trace corresponds to the corrective simulation described in the text.

First, we compare the two available structures to see what motions are indicated. A comparison against these changes (shown in Figure 3.8) ensures that the simulations are able to capture these motions, since it has been widely demonstrated that motions from multiple experimentally determined structures usually provide a sample of the most important motions. Second, we compute ANM modes to animate the static structures. We will compare these with those in Figure 3.8 to see if the motions of the two structures are similar or different. We anticipate that both structures will exhibit similar motions,
but this must be confirmed. For some conformational transitions, the motions computed from each end-point structure move in a direction toward one another.

Figure 3.8: Transition between the Cu$^{+1}$ bound and apo states, emphasizing the dynamic motion between PC2 and the ion-exchange.

The full CusA structure is shown (left) emphasizing one monomer (darker). A closeup view shows the changes near the binding site. The region PC2 in the bound state (pink) rotates to the apo (red) position nearer PC1 (green). The C$^\alpha$ atom of P708 (shown as sticks) in PC2 moves by nearly 20Å. Arrows show the largest motions. Methionine residue C$^\alpha$s are blue spheres and a Cu$^+$ is shown as an orange sphere.

ANMs are computed for coarse-grained structures using their C$^\alpha$ atoms connected by uniform harmonic springs if their centers are within 13Å. The best way to represent the membrane is a somewhat open problem. Complications include how to best account for the ease with which hydrophobic-hydrophobic interactions can exchange in the center of the membrane (oil drops - lubricated motions), while including greater constraints on the motions of the phospholipid head groups. If one ignores completely the presence of lipid around the CusA trans-membrane (TM) helices, the intracellular and TM regions exhibit impractically large fluctuations. Thus, it is essential to include the membrane in a well-validated way to ensure computationally accessing the meaningful dynamics of CusA. We add a cylinder of coarse-grained lipid around the CusA TM domain. Restraining springs are differentially stronger between the lipid head groups. The POPC lipid molecules used contain 134 atoms. We retain 11 atoms (names N, P1, O21, C32, C24, C28, C212, C216, C36, C310, and C314) from each molecule to make a coarse grained representation which we combine with the C$^\alpha$ atoms of the protein after removing the apex of intracellular loops (to avoid “tip effect”). These coordinates are used to make our ANM model. ANM modes are used as biasing directions for simulations where VDW and virtual bond
lengths are enforced. This results in a trajectory for the structure, and we monitor the locations of the methionines and Cu⁺.

**Figure 3.9: Directionality Imparted to the Cu⁺ Ion by Normal Modes 1-3 Suggests a Specific Structural Mechanism.**
The lowest frequency modes deform the structure around the methionine triad where Cu⁺ is bound. **A)** The Cu⁺ bound to the three methionines (blue) is buried. **B)** We see how the structural deformation pushes Cu⁺ outward away from the methionines and imparts a velocity to it, favoring movement to the upper methionine pair.

ANM provides an analytic set of motions for a given structure, but does not typically take into account virtual bond lengths (VBL) or van der Waals exclusion constraints (VDW). In order to account for these two terms, we first calculate a set of ANM modes. Each mode is then used as the biasing direction for a simulation. At each step in the simulation, we set the velocity of each coarse-grained point to be the mode direction plus an adjustment to correct VBL and VDW. VBL is corrected by adding a corrective direction to the velocity if the VBL exceeds 3.8Å. The correction effect becomes stronger for greater violations. VDW exclusion is accounted for by forcing atoms to “roll” along their VDW radius (set at 1.5Å for all atoms), rather than allowing them to penetrate. By taking small steps, the magnitude of the correction terms is small. Figure 3.7 compares these corrective simulations to the standard modes. The result is a more realistic trajectory for the structure than typical mode-motions.
Each of the three lowest frequency modes shows a similar motion for the methionines in the trimeric cluster (shown in Figure 3.9, where the entire structure has been deformed to 5.3 Å); imparting an outward directional velocity to the Cu⁺; where it moves from a partially buried position to the exterior of the structure, toward the neighboring methionine pair. Figure 3.10 shows the same two states, but in a wider context. Figure 3.11 quantifies the change in distance within the methionine pairs under the normal modes. We anticipate that successive steps in the transport will be uncovered in a similar way.

Figure 3.10: The motion from Figure 3.9 is shown from a wider view, showing the details included in our simulation.

A) We show motion in the negative direction of mode 2 as well as the B) positive where in each case the structure is deformed to an RMSD of 5.3Å. CusA domains are colored as PC2 red, PC1 green, PN2 orange, DN+DC purple, transmembrane domains in gray, and coarse-grained lipid in yellow. Spheres mark the location of methionine pairs.
Figure 3.11: Distance changes between paired methionine residues upon normal mode deformation relates to the ion-shuttling mechanism.

The structure is deformed to 7Å RMSD in either the positive or negative normal mode direction (sign indicated on abscissa). We find that the normal mode motions lead to increases or decreases in the distance between methionine residues and those changes are different between the three chains of the CusA trimer. Each methionine pair (MP) is designated by a different color while the methionine triad (MT) distances are distinguished by line type. A) For mode 1 within chain A we find that the periplasmic pair move together in the negative mode direction and further apart in the positive direction, whereas bottom (closest to intracellular side) move oppositely. There is also directionality to the MT where two atoms in the triad together move away from the third. B) For mode 1 within chain C, we find the opposite case. Thus, there is an anti-correlation between the methionine motions across the CusA trimer. C) For mode 3 and chain A we find a behavior similar to mode 1, but with some different characteristics. Now the MP within the TM domain, on the periplasmic side, shows coordination with the MP on the intracellular side. D) CusA is shown with chain A Cα atoms as small spheres and the Cα atom from methionine groups as colored spheres. The surface is shown for the other two monomers.
CHAPTER 4. Structural control over telomerase mechanism


4.1. Abstract

Telomerases constitute a group of specialized ribonucleoprotein enzymes that remediate chromosomal shrinkage resulting from the ‘end-replication’ problem. Defects in telomere length regulation are associated with several genetic diseases, and telomerase also plays important roles in aging and cancer in humans. Despite significant progress in understanding the structure and physiological roles of telomerase, the complete structure of the human telomerase enzyme bound to telomeric DNA remains elusive, and the detailed molecular mechanism of telomere elongation is unknown. By application of computational methods for distant homology detection, comparative modeling, and molecular docking, guided by experimental data from the literature, we have generated a three-dimensional structural model of a partial telomerase elongation complex composed of three essential protein domains bound to a single-stranded telomeric DNA sequence in the form of a heteroduplex with the template region of the human RNA subunit, TER. This model provides a structural mechanism for the observed processivity of telomerase and offers new insights into telomere elongation. In particular, we conclude that the RNA:DNA heteroduplex is constrained by the telomerase TEN domain through repeated extension cycles and that the TEN domain controls the process by moving the template ahead one base at a time by translation and rotation of the double helix. The RNA region directly following the template can bind complementarily to the newly synthesized telomeric DNA, while the template itself is reused in the telomerase active site during the next reaction cycle. This first structural model of the human telomerase enzyme provides many details of the molecular mechanism of telomerase and immediately provides an important target for rational drug design.

4.2. Introduction

Telomerases are essential for maintaining chromosome length and integrity (69-71). They complement the cellular DNA-dependent DNA polymerase replication machinery that is not capable of
fully replicating chromosomal ends, leading to telomeric DNA sequence erosion. Loss of telomerase activity is tolerated to some extent in yeast, worm, plant and mouse (72), but after a few generations, telomeres typically become too short to perform their essential functions. Excessive telomere shortening can result in chromosome degradation, illegitimate recombination and end-to-end fusion, the compromise of cell cycle regulation and, ultimately cell death (73, 74). Most eukaryotic organisms utilize telomerases for the successive synthesis and maintenance of telomeric DNA repeats at chromosome ends, replenishing the capability for further cell proliferation in stem cell lineages (75). Interestingly, the activity of telomerases in fully differentiated somatic cells is strongly down-regulated over time, in concert with ageing, and a direct correlation between telomere shortening and ageing has been demonstrated (76). In human cells, increased telomerase activity can increase renewal capacity in certain tissues, which has been interpreted as delaying the ageing process. However, pathogenic over-expression of telomerases is also a hallmark of many human cancers (77), and several studies have shown that telomerase malfunction can lead to diseases in humans (72, 78), including dyskeratosis congenita (79). Thus, telomerase appears to be a key player critical in maintaining the balance between normal cellular differentiation (and ageing) and the aberrant proliferation manifested in carcinogenic transformation (and immortality).

Influences of changes in telomerase activity have been observed in many biological processes not directly related to telomere maintenance, (80, 81). For example, Gonzalez-Suarez et al. found that induced somatic expression of telomerase led to increased cellular proliferation and growth, and consequently, enhanced wound healing in mice (82). Studies on promyelocytic leukemia cells revealed that telomerase expression may also inhibit apoptosis (83). When over-expressed, telomerase is directed to the mitochondria and appears to help protect cells from H$_2$O$_2$-mediated damage (84). Recently, Blackburn and colleagues have shown that changes in telomerase activity are associated with human stress-related syndromes, including major depression (85, 86).

Telomerases function as specialized reverse transcriptases (87), RNA-dependent DNA polymerases capable of synthesizing multiple copies of the telomeric DNA repeat sequence by using an intrinsic RNA template to direct telomeric DNA synthesis (69, 81). Telomere repeats are often shown as 5’-(TTAGGG)$_n$-3’ which is the DNA repeat unit. In this work we choose to focus on the RNA template, displaying its coding region as 5’-UAACCC-3’. For clarity, the alignment region 3’-AUC-5’ of hTR pairs to the DNA primer sequence 5’-TAG-3’ in order to synthesize the next DNA addition. The newly synthesized telomeric DNA repeats are added to the overhanging single-stranded 3’ end of the DNA at the
chromosome termini. In other respects, the telomerase reverse transcriptase mechanism appears to be similar to that of well-studied retroviral reverse transcriptases (69, 80, 88).

The human telomerase enzyme contains a template-encoding RNA molecule, TER (Telomerase RNA or hTR) and a primary protein component, TERT (Telomerase Reverse Transcriptase) with several functional domains: TEN (Telomerase Essential N-terminal domain), TRBD (Telomerase RNA Binding Domain), RT (Reverse Transcriptase domain) and the C-Terminal Extension (CTE) (69, 80, 81). The RT domain can be further divided into two distinguishable subdomains: the “fingers” involved in nucleotide binding and processivity and the “palm” providing the polymerase catalytic residues and DNA primer grip. The C-terminal extension is responsible for interaction with DNA and has been proposed to correspond to the RT “thumb” domain (88). Because this CTE appears to be structurally equivalent to the C-terminal α-helical thumb of retroviral RT (89), in describing the human TERT structure here, we refer to RT as fingers, palm and thumb for clarity and simplicity. CTE will be referred to as the RT thumb domain below. Both RT and TER are essential for telomerase activity, together forming the active site that catalyzes deoxynucleotide addition. The TRBD domain links the RT and TER components, and the TEN domain is proposed to facilitate the repetitive repeat addition mode of telomerases, which is one of the distinguishing features of telomerases, relative to classical reverse transcriptases (69, 80). Efficient repeat addition processivity is governed by multiple mechanisms involving both TER (90) and protein subdomains of TERT (91), as well as additional telomerase-associated processivity factors, notably TPP1-POT1 in humans, which enhances telomerase processivity by slowing primer dislocation and facilitating translocation (92). The TEN domain contains an “anchor” site (93) which is thought to help stabilize the bound single-stranded telomeric DNA substrate within the complex, while the intrinsic RNA template is realigned for the next, iterative reverse transcription cycle. Therefore, the complex is capable of processively synthesizing a long array of single-stranded telomeric DNA repeats by repeatedly copying the 6-nt long RNA template region within the TER component (69, 80). In addition to processivity factors, several species-specific accessory proteins are critical for telomerase assembly, subcellular localization and function in vivo (80, 94). In human cell lines, for example, the catalytically active form of telomerase includes dyskerin (95), which together with NHP2 and NOP10, are required for stability and accumulation of the RNA component of human telomerase in vivo (96).

Recently, X-ray structures of the full length T. castaneum telomerase (containing RT and TRBD domains) alone (3du5 and 3du6, (89)) and in complex with an RNA:DNA hairpin (3kyl, (97)) have been published. In addition, the crystal structures for separate TRBD (2r4g, (98)) and TEN (2b2a, (99)) domains
from *T. thermophila* are now available. Despite much experimental effort, the detailed molecular mechanism of human telomerase enzymatic activity and the structural details of the interactions between the TEN domain and the other components of the complex (RT, TRBD, TER and the telomeric DNA) are still not fully known. The absence of a high-resolution experimentally determined structure for the assembled telomerase core catalytic complex is a serious impediment to designing experiments that could further elucidate the molecular mechanism of telomerase action. Moreover, species-specific features of telomerase structure and function make obtaining a complete structure of the human telomerase enzyme particularly important. Thus we employ here a theoretical modeling approach to generate the entire 3D structure of the human TERT, TEN, and TRBD bound to a DNA substrate and its RNA template.

Automatic homology modeling (performed by our collaborators in Dr. Kolinski’s lab) using available web-based servers was not feasible because the amino acid sequence identities between the human telomerase domains and the corresponding RT and TRBD structures in the Protein Data Bank (PDB) are only 24% and 22%, respectively. This level of sequence similarity is near the limit for the reliable use of standard homology detection methods based on PSI-BLAST or RPS-BLAST. Furthermore, in our hands, such standard sequence comparison methods were unable to detect significant sequence similarity between the N-terminal domain of the human telomerase protein and the TEN domain of *Tetrahymena*, or any other known protein structure. To obviate this problem, we used an advanced meta-profile comparison method, Meta-BASIC (100), to map the human telomerase protein sequence onto sequences of the determined structures from *Tribolium* and *Tetrahymena*. The mappings obtained were confirmed by using a variety of fold recognition methods. Together with detailed manual inspection, these approaches allow us to generate highly accurate sequence-to-structure alignments between the human telomerase sequence and relevant structural templates. We then built three-dimensional models separately for TEN and the other components of the human telomerase complex including a hybrid RNA:DNA double-helix formed between the RNA template and the single-stranded telomeric DNA substrate, and assembled them by using protein-protein docking, guided by relevant experimental data. Based on the resulting structural model of the telomerase enzyme, we propose a mechanism for human telomerase action in which interactions between TEN and the RT:TRBD subcomplex play a critical role, and where the elongating telomeric DNA is stabilized by the TEN domain. We hypothesize that the helical structure of the heteroduplex formed between the RNA template and the telomeric DNA substrate is actively maintained during the individual repeated telomerase reactions producing a single
copy of the template. Following this, the RNA template must be repositioned relative to the active site, but we observe that its translocation proceeds along the extending helix due to constraints imposed by the TEN ‘anchoring’ domain.

4.3. Results and discussion

Figure 4.1: Partial model of human telomerase ribonucleoprotein complex generated computationally in this study.

The model includes the TERT protein component composed of the catalytic reverse transcriptase domain (RT), the RNA binding domain (TRBD) and the N-terminal ‘anchor’ domain (TEN), bound to a heteroduplex formed by oligomers corresponding to the template-encoding RNA (TER) and the single-stranded human telomeric DNA substrate. The template region (5’-UAACCC-3’) of the RNA is shown in yellow and the complementary region of DNA in magenta, while the partial template repeat (5’-UAAC-3’) in the RNA is shown in lighter yellow, and its complement in the DNA, in pink. The N- and C-terminal α-helices of TEN interact with the major groove of the heteroduplex. Residues with experimentally determined influence on telomerase function or assembly are labeled and shown in red. The orange dotted line shows the ~50Å distance between TEN (Q169) and the RT active site (D869).

The present work was motivated by the lack of a complete structural model that could explain the detailed functional molecular mechanism of human telomerase. By using distant homology detection, comparative structural modeling and computational docking, we developed a model of the human telomerase complex (Figure 4.1). Then by using elastic network models, we investigated the intrinsic motions of the modeled structure. Our goals were twofold: 1) to understand how the individual
telomerase protein domains and the intrinsic TER component interact in the assembled active human telomerase RNP enzyme, and 2) to generate a model illustrating how the telomerase RNP enzyme binds to and extends single-stranded telomeric DNA by reverse transcription of telomeric repeat sequences.

Several studies indicate that the TEN domain functions as an “anchor” to bind and stabilize the telomeric DNA substrate, contributing to the processivity of the repetitive reverse transcriptase activity (81, 91, 93). However, no structure-based explanation for how TEN contributes to the overall function of the telomerase enzyme or its processivity has been proposed previously. Here we suggest that TEN plays a critical role in controlling the processive step in which telomerase advances by one base on the strand being copied. Similarly, many experiments have determined the effects of specific amino acid substitutions or deletions on telomerase enzymatic activity and have provided us with useful information about key residues (see Supplementary Table 1). It is important to reconcile these against a structural model and a structural mechanism. To date, however, it has not been possible to understand the effects of these changes due to the lack of a complete structural model for the telomerase RNP. The present structure will facilitate such investigations.

To derive a 3D model for human telomerase enzyme we used the available structures for telomerase components from T. castaneum and T. thermophila, including the recently released structure for RT and TRBD domains solved with RNA template and telomeric DNA (97). The modeled RNA:DNA heteroduplex was extended as observed in the closely related HIV-1 reverse transcriptase structure (101), to construct interactions with the TEN domain (see Methods). The human TEN domain, which was modeled separately, seems to be designed to accommodate a nucleic acid double helix, as does the corresponding TEN structure from T. thermophila (99). It has a well-defined cleft to interact with the phosphate backbone and two helical segments that fit well into the major groove of the double helix. We used molecular docking to assemble the entire complex by fitting the TEN model into a position that ensured appropriate interaction with both the telomeric DNA and the other telomerase protein domains. We also considered available data suggesting plausible RNA binding sites, conservation of surface residues and certain amino acid mutations that have been shown to impact the enzymatic activity (see references in Supplementary Table 1). The final model shows that the central ring-shaped part of the human telomerase structure, formed by TRBD and RT domains, accommodates the RNA:DNA heteroduplex and provides catalytic residues (Figure S1). The ring is coupled with the TEN domain, which interacts with TRBD, the RT thumb and the RNA:DNA major groove, and helps stabilize the substrate within the active site during subsequent reaction cycles.
**Reverse Transcriptase and RNA Binding Domains.** The RT and TRBD domains form the central, ring-shaped core of the telomerase RNP complex (Figure S1). They provide the catalytic activity of the enzyme by bringing together the necessary active site residues. Previous analyses of sequence conservation within these domains highlighted several motifs, shared by the majority of reverse transcriptases, including telomerases (69, 102). As shown in Figure S2 starting from the N-terminus, these motifs are: CP, T, 1, 2, 3, A, IFD, B’, C, D and E. Motifs CP and T belong to the α-helical TRBD domain (Figure S1), which mediates interactions between the template-carrying RNA component, and the telomerase. The CP and T motifs have been shown to be directly involved in RNA binding (80, 103), and are required for the proper assembly of the complete telomerase ribonucleoprotein complex (81, 104). Motifs A and C in the RT palm contain the conserved active site signature sequence, KXD(X)\_nDD, in which three invariant aspartic acid residues (D712, D868 and D869) coordinate two Mg\(^{2+}\) ions, while the lysine (K710) provides the base for the deoxynucleotide condensation reaction (81, 89, 102) (see Figure S3 A).

The pocket surrounding the catalytic amino acids is lined with several residues that help position deoxynucleotide substrates (A, T and G) with respect to the complementary RNA template-encoded ribonucleotides in the active site, as in the *T. castaneum* structure (89, 97). Three conserved uncharged residues: Y717, Q833 and V867 (from motifs A, B’ and C, respectively), form a hydrophobic pocket adjacent to the catalytic aspartates and take part in nucleotide binding (97). Residue V867 has been shown to alter human telomerase substrate specificity (105), and residue Q833 corresponds to Q151 in HIV-1 reverse transcriptase, where mutations cause hypersensitivity to substrate analogs (106). This pocket appears to hold the incoming deoxynucleotide in close proximity to the active site, for coordination with one of the Mg\(^{2+}\) ions. In addition, Motif 2 residues K626 and R631 (from RT fingers), together with K902 from motif D (RT palm), may interact with both the sugar ring and phosphate groups, and provide stacking interactions with bases, of the incoming deoxynucleotide. These interactions likely stabilize the telomeric DNA substrate during catalysis (89). The relatively conserved residues C931 and G932 from the RT palm define a ‘primer grip’ (motif E) (97), which is essential for proper maintenance of telomeric DNA within telomerase active site (Figure S3 B). Additionally, R972 and K973 from the RT thumb, both located on an α-helix that packs into the minor groove of the RNA:DNA heteroduplex, interact with DNA backbone (Figure S3 C). These residues, present also in telomerases that lack the TEN domain, may contribute significantly to repetitive addition processivity.
The interactions between telomerase and telomeric DNA are mediated by a variety of critical residues grouped into motifs characteristic for this family of polymerases (69, 80, 81). The spatial arrangement of these motifs resembles the shape of double-stranded nucleic acid helix (89). The recently described ‘motif 3’ of the reverse transcriptase domain provides several residues that may interact directly with telomeric DNA (107). Notably, mutations of motif 3 residues (V658A and K659A from RT fingers, and R669A from RT palm) cause telomerase hyperactivity (107), apparently resulting from weaker interaction with the double-stranded heteroduplex, facilitating telomeric DNA release after reaction. In our model, the positively charged K659 and R669 side chains are directed toward both DNA and RNA backbones, compatible with their essential contribution to nucleic acid binding (Figure S3 A).

Increased repeat addition rate reduces processivity, however, possibly because the telomeric DNA cannot be stabilized sufficiently while the template-carrying RNA is realigned and prepared for the next reaction cycle (107). In support of this, Xie et al. (107) were able to obtain hyperactive and hyperprocessive human telomerase mutants by combining the V658A mutation with the deletion of residues 643-649 in the RT fingers and the hTR-U57C substitution in the RNA. The loop containing amino acid residues 643-649 precedes the motif 3 α-helix and is an intriguing structural feature: it may freely interact both with the telomeric DNA and RNA and likely stabilizes the position of the DNA substrate in the telomerase complex (Figure S3 D). Deletion of the 643-649 loop weakens this interaction and likely allows for more rapid dissociation of the heteroduplex, making the template binding site available for the next substrate deoxynucleotide and the next round of synthesis. Additionally, the hTR-U57C substitution results in extension of the RNA:DNA heteroduplex by an additional base, which could potentially form a classical Watson-Crick pair, possibly further stabilizing the telomerase-telomeric DNA association.

In contrast to T. castaneum telomerase (3kyl), the human protein contains two additional α-helices (residues 415-456) within the TRBD domain (similar to T. thermophila 2r4g, residues 333-371) which, according to our model, together with the final α-helix in TRBD, form a three-helix bundle that packs tightly against the RT fingers (Figure S4). This structural feature additionally stabilizes the interaction of RT and TRBD.

**Essential N-terminal Domain.** The TEN domain, composed of a central β-sheet flanked by α-helices on both sides, is the most divergent domain within the telomerases (69, 80, 81). Nevertheless, it appears to be essential for proper telomere maintenance, because it ‘anchors’ telomeric DNA (93). A recent study used a combination of comparative modeling and machine learning to identify several residues in
TEN that are likely to play a role in nucleic acid binding (108). The sequence diversity among TEN domains of different species may be related to differences in telomeric DNA repeat sequences. TEN recognizes telomeric DNA in a sequence-specific manner, and several experiments have revealed differences in DNA binding affinity to different telomeric repeat sequences (109, 110). The TEN domain is separated from TRBD by a linker region, predicted to be largely unstructured, ranging in length from 20 to more than 500 amino acids, depending on the species (69). The TEN domain is believed to contribute to the processivity of the enzyme, because several studies have identified mutations or deletions of TEN residues that lead to a reduced ability of the telomerase to synthesize more than one telomeric repeat (91, 111). The crystal structure of the TEN domain from T. thermophila (2b2a, 99) revealed certain features that adapt it for interaction with telomeric DNA. In particular, a deep cleft in the TEN domain surface is closely complementary to the shape of a double helix. Mutation of Q169, which is located in the central part of the cleft, compromises human telomerase processivity by hindering proper protein-DNA interaction (110, 112). In our modeled human TEN, Q169 forms a hydrogen bond with the backbone carbonyl group of P174 and the backbone amine group of L175, stabilizing the intervening loop, which establishes hydrogen bonds with another element of the TEN structure (Figure S5). These interactions thus bridge adjacent structural elements and stabilize the overall shape of this region. The cleft is flanked by α-helical extensions on both sides and engages the RNA:DNA double helix, fitting into its major groove (Figure 4.1). Jurczyluk, et. al. (113) recently performed mutational analyses on TEN. Two mutations of particular interest involve residues 8-13 and 170-175 of TEN. The former exhibits wild type processivity, but a decreased $K_m$ and the latter has significantly decreased processivity and an increased $K_m$. In our model, residues 8-13 make close contact with the heteroduplex, supporting this observation of decreased affinity. Residues 170-175 do have heteroduplex interactions in the model, but are also partly buried in the TEN domain.

**Interaction with the RNA Template.** According to our model the template-carrying RNA component (TER) interacts with the TRBD, RT and TEN domains of telomerase. Human TER is a structurally complex RNA molecule of 451 nucleotides, containing several conserved sequence and structural motifs (114). A characteristic pseudoknot domain, located in close proximity to the template repeat sequence and to an RNA loop domain designated CR4-CR5, is essential for telomerase catalytic activity (115). A 3′-terminal 'H/ACA box' in TER contributes to the assembly and maturation of the ribonucleoprotein complex (116). The RNA pseudoknot region was shown to interact with a C-terminal region (residues 150-159) of TEN, providing further insight into the localization of the template region within the human telomerase
complex (117). Our modeled structure of TRBD exposes a wide cleft opening towards the C-terminus of TEN, which might bind the pseudoknot domain. Furthermore, the surface of the cleft presents several lysines (K492, K493, K511) that could interact with RNA. The TER CR4-CR5 domain has been shown to bind to the CP and T motifs of the TRBD domain (69), which are located at a considerable distance (~23Å) from the putative pseudoknot-binding site in our model. Recent work by Egan and Collins provides insight into hTERT-hTR interactions and will be useful for future studies that attempt to model the full length hTR (118).

**RT:TRBD-TEN Interaction.** The shape of the TEN domain restricts its possible orientation with respect to the other domains and the RNA:DNA heteroduplex. Mutual positioning of the telomeric DNA substrate and RT, to ensure proper interactions within the reverse transcriptase active site, determines the distance between the surface of TRBD and the major groove of the bound double helix. Therefore, the possible TEN orientations are dramatically limited. Together, these constraints aided in the assembly of our model of the human telomerase RNP complex. The surfaces of RT, TRBD and TEN expose poorly conserved residues, hindering the modeling of interactions between TEN and the other two domains. However, detailed analysis of surface residues in the model reveals increased conservation of uncharged amino acids at the domain interfaces (e.g., G100, F101 in TEN and G967, V1025 in the RT thumb; T117, S118 in TEN and S504, L505, A542 in TRBD, see Figure S6). Despite poor surface conservation in TEN, the proposed interface represents an optimal structural fit between TEN and RT:TRBD and the RNA:DNA heteroduplex. Additional support for our proposed assembly is provided by Sealey et al. (110), who reported that T116A, T117A and S118A mutations in TEN compromise repetitive addition processivity but do not alter DNA binding affinity and thus probably affect interaction with the remaining protein domains of the telomerase. In the model, these three residues are located at the binding surface between TEN and the TRBD and RT thumb domains. This interface is far from the nucleotide binding region, with the closest nucleotide atom ~16 Å away. It is, however, involved in the slower motions, presumed to be functionally relevant (see below) and this may explain the role of these residues in the processivity.

**Mechanism of telomerase action.** To investigate the processive mechanism of telomerase, we used elastic network models to generate the mechanistic step shown in close-up in Figure 4.2. We utilized our Anisotropic Network Model (ANM) (40), with one node for each Ca′, P and O4′ atom, and with identical springs placed between pairs of these atoms within 13 Å. To generate the processive conformations shown, we used the global mode of motion to move the original structure. This reveals in detail how the
telomerase structure effects the motion of the template. Views of the whole structure can be seen in further detail in Figure S7 and in videos S1 and S2.

![Figure 4.2: Structural model for the processive motion of human telomerase.](image)

The procession of the DNA:RNA heteroduplex is a critical aspect of telomerase function. Rotation and translation of the heteroduplex is evident in the global mode of the elastic network. The effects of following the global mode in the A) negative (-1) and B) positive (+1) directions are shown. Termini closest to the viewer are highlighted: the 3’ end of RNA in red and the 5’ end of DNA in cyan. See Figure S7 for corresponding views of the entire structure, and movies m1 and m2 for two dynamic views of this motion. Panel C) offers a side view and D) a face view of the heteroduplex for three states of the negative global mode (-1), the original state (0), and the positive mode deformation (+1).

Our 3D model assumes that the intrinsic template-carrying RNA molecule forms a heteroduplex with the single-stranded telomeric DNA, while the architecture of both the RT active site and the TEN ‘anchor’ domain are adapted for double-stranded nucleic acid binding. The template-encoding region of TER contains one complete 6-nt repeat complementary to the human telomeric (TTAGGG)_n repeat sequence which can initiate RNA:DNA heteroduplex formation. Multiple sequence alignment of telomerase RNAs from several phyla (extracted from the Telomerase Database, [http://telomerase.asu.edu](http://telomerase.asu.edu) (119)) reveals that the telomere repeat template sequence is partially repeated, extending the potential length of the helical heteroduplex region, as was previously proposed.
In the human enzyme, such an extended RNA:DNA heteroduplex would contain at least 10 base pairs, exactly filling the space between the RT active site and the experimentally supported nucleic acid binding region of TEN (Figure 4.1). Thus, our model provides a strong structural basis for the mechanism shown in Figure 4.2, in which, after completion of a single telomere repeat synthesis cycle, the template RNA is moved ahead at the polymerase site to the next base. The TER template region must then shift relative to the substrate DNA, while maintaining the RNA:DNA heteroduplex (Figure S8). The second, partial repeat (5′-UAAC-3′) adjacent to the TER template region complements the newly synthesized telomeric DNA repeat, while the TER template region itself (5′-UAACCC-3′) forms a 5′ overhang ready for the next reaction cycle. Our model strongly supports an essential role of TEN in stabilizing the assembling heteroduplex in an orientation that promotes proper interaction with the RT active site. Further, our dynamics simulations suggest the important role of the structure in controlling the shift along the template to the next base to be copied. The current model does not provide insight into the mechanism by which the nucleic acid translocates after the complete synthesis of each complete template sequence; however, there is the possibility that the protein can extend together with the RNA through the repeated synthesis steps for one template cycle before recoiling to activate and reposition the template for the next cycle.

Notably, the template-encoding region of TER is followed by a sequence rich in uracil (U) residues that are capable of forming wobble base pairs with guanine (G). Therefore, the proposed RNA:DNA heteroduplex formed during reverse transcription may actually be longer, to help stabilize the helical structure during the RNA template realignment step between reaction cycles. Such an extended helix could contribute significantly to the extension mechanism.

Several telomerases (e.g., those from *C. elegans* and *T. castaneum*) lack a distinguishable TEN domain, which is essential for activity and processivity of the human and *Tetrahymena* enzymes (69, 80, 81). Despite this, the *T. castaneum* enzyme appears to be active *in vitro* (97) and genetic evidence suggests that the *C. elegans* telomerase is capable of synthesizing multiple telomeric repeats *in vivo* (121). Meier and colleagues found that TRBD domain in *C. elegans* is preceded by a domain that might be distantly homologous to TEN (121); however, our methods failed to detect significant similarity of this region to any known protein domains. The RT thumb and IFD motif (RT fingers) have been shown to play roles in repetitive addition processivity (88, 93) and, in the absence of TEN, could provide substrate stabilization during subsequent reaction cycles. Species-specific accessory factors that influence processivity could also compensate for the lack of a TEN domain (80, 81). Interestingly, we found that
telomeres lacking the TEN domain, e.g., 3kyl, also lack the $\alpha$-helical insertion within the TRBD (Figure S4), which would allow for a more elastic RT-TRBD interface (discussed above). Telomeres possessing TEN domains possibly do not require such elasticity because TEN aids the nucleic acid binding and chaperones the processivity.

Supplementary Data are available online at:
http://www.pnas.org/content/suppl/2011/05/23/1015399108.DCSupplemental

4.4. Conclusions

The availability of a structural model of the assembled human telomerase complex presented here provides information necessary for investigating its mechanism further, as well as for locating its interactions within the complex cellular signaling networks in which it is known to participate (76, 85, 86). A complete structural model of telomerase may also accelerate the development of new anti-cancer therapies that aim to abolish telomerase activity in proliferating tumor cells, or to augment enzymatic activity in cases of telomerase insufficiency diseases.

4.5. Methods

We have utilized the human telomerase protein sequence (GenBank accession # NM_198253.2). In order to study sequence conservation within the telomerase protein family, PSI-BLAST (122) searches were carried out against the NCBI non-redundant protein sequence database using this human telomerase sequence as query (3 iterations, E-value threshold 0.005). Multiple sequence alignments of collected sequences were prepared with PCMA, a progressive alignment method (123). Additionally, to obtain secondary structure conservation patterns useful for fold assignment validation and for manual curation of the sequence-to-structure alignments, for every aligned sequence, the secondary structure was predicted with PSI-PRED (124).

Templates for comparative modeling of human telomerase domains were identified using the Gene Relational Database (GRDB) system, which stores pre-calculated Meta-BASIC mappings (100) between Pfam families, conserved domains and PDB structures. Meta-BASIC is a distant homology detection method that exploits comparisons of sequence profiles enriched with predicted secondary structures (meta-profiles) (100). Meta-BASIC predictions were further validated with the consensus fold recognition server, 3D-Jury (125), followed by manual inspection of the hits obtained. All searches were carried out by using the complete telomerase sequence and the sequences of its three separate protein
domains: TEN, TRBD and RT. Sequence-to-structure alignments between human telomerase (and its closest homologs) and selected proteins of known structure were prepared using the consensus alignment approach and 3D assessment (126).

Three-dimensional models of human telomerase protein domains were generated with Modeller (127) based on manually curated, high confidence sequence-to-structure alignments. These models were built separately for: 1) the TEN domain, using *T. thermophila* TEN (2b2a, (99)) as template; 2) the RT:TRBD subcomplex, using the corresponding structure of *T. castaneum* RT:TRBD (3kyl, (97)) and the superimposed TRBD domain from *T. thermophila* (2r4g, (98)) as templates, since the *T. castaneum* TRBD domain lacks two critical $\alpha$-helices that are present both in human and *T. thermophila* TRBD. The resulting 3D models of TEN and the RT:TRBD subcomplex were then assembled manually after careful consideration of the CABS (128) results for protein domain docking (described below) and published experimental data regarding specific residues proposed to mediate inter-domain interactions (Supplementary Table 1). Additionally, conservation of surface residues was investigated with ConSurf (129, 130), in order to detect patterns of highly conserved amino acids that might suggest plausible interactions and the location of interfaces between domains (Figure S6).

Automatic CABS assembly of the modeled human telomerase domains was performed using a three-stage docking procedure. First, rigid docking via an exhaustive global search in a six-dimensional space of ‘ligand’ rotations and translations against the frozen structure of the ‘receptor’ was carried out using FTDOCK (131). Rotational space was scanned in 12-degree increments. For translations, sampling was performed on a cubic lattice with 0.875 Å spacings. For each rotation, the three top-scoring translations were saved for subsequent analysis. The resulting 10,000 FTDOCK top-scoring structures were re-scored with the CABS force field and grouped using hierarchical clustering. From each cluster, a representative with the lowest energy was selected. The number of models was thereby reduced from 10,000 to 30.

To account for protein flexibility upon complex formation, each resultant structural model was subjected to a short stochastic dynamics simulation with the CABS algorithm. From each simulation 1000 models located in the vicinity of the initial structure were collected. Hierarchical clustering was again used to select the most populated states. Structures from the centers of the clusters were extracted to represent variants of the final telomerase tertiary structure. These representatives were used as starting models for further manual adjustments based on the consistency of the model with available experimental data (Supplementary Table 1) and according to surface conservation in individual domains.
Positions of the intrinsic RNA template and single-stranded telomeric DNA substrate in the human telomerase model were copied from the *T. castaneum* telomerase structure (3kyl) after superposition of their RT and TRBD domains. The sequence of the RNA:DNA hairpin containing RNA template was modified: 5’-CUGACCUGA-3’ and the complementary telomeric DNA: 5’-TCAGGTCAG-3’ were replaced with 5’-UAACCCUAA-3’ and 5’-TTAGGGTTA-3’ sequences suitable for human telomerase. The hairpin loop (5’-CTTCGG-3’) was removed and, to allow for interaction with the TEN domain, the double helix was extended by 7 base pairs (resulting in DNA: 5’-GTTAGGGTTAGGGTTA-3’, RNA: 5’-UAACCCUAACUGAGAA-3’). The telomeric DNA is less complementary to the RNA at the 5’ end, however a nonclassical (e.g. wobble) pairing might still be sufficient for double helix formation. Finally, minor adjustments of the TEN domain were made to improve its fit between the nucleic acids and the protein domains in the complex.

To relieve steric clashes and improve internal packing, the 3D partial model of the human telomerase complex, comprising all three protein domains and the RNA:DNA heteroduplex, was energy-minimized with Tripos SYBYL using an AMBER force field (132), followed by a short molecular dynamics run (simulation time of 0.2 ns, a NTV ensemble, 100fs coupling, Boltzmann initial velocities, Amber7 FF99 force field, Gasteiger-Hückel charges and dielectric constant = 1).

The final model was used to investigate the motions of the structure using simple coarse-grained elastic network models in conjunction with normal mode analysis. This approach has been widely used to investigate important functional motions of biomolecular structures (67). It is based on a highly cohesive model of structure and investigates particularly the larger motions that are available within the constraints of the geometry of the structure. Residues close to each another in the structure are connected with identical springs and the vibrational motions of the set of springs are analyzed with a normal mode decomposition (40). The approach has proven useful for extracting the functional motions of large domains in many structures (67). In the case of the human telomerase complex, our modeled structure, together with these dynamics simulations, allowed us to investigate the functional role of the TEN domain in telomere elongation.
CHAPTER 5. Ribosome mechanism explored using coarse-grained models

5.1. Background

Many important experimental insights have been gained concerning the ribosome, its conformations during the stages of translation, as well as the association of the ribosomal proteins and their role in assembly, however, many of the details remain elusive. The Wilkerson group has determined many states of ribosome assembly at low resolution, extending knowledge of the steps, rates, and order of ribosome assembly (133, 134). This data complements the many studies that have determined dominant motions of the assembled ribosome through experimental structure determination. Motions are inferred by comparing ribosomes that have been stalled at various points in translation. Multiple labs are investigating the precise nature of these transitions with the hope of further understanding the mechanism with the aim of understanding diseases or developing drugs to specifically target motions of bacterial ribosomes. One of the major computational efforts is from the Sanbonmatsu lab (135). This group focuses on atomic Molecular Dynamics simulations. Being one of the largest molecular structures that science has determined, these simulations are computationally intensive and costly, and these scientists have access to the large computers at the Los Alamos National Lab. We aim to accomplish similar goals of filling in the details of translation, but using our simpler models.

In a recent paper by Voss, et. al. (136), the experimentally determined ribosome structures were computationally analyzed and features of the nascent chain (NC) exit tunnel elucidated. The geometry and surface properties of this tunnel have a large impact on ribosome fidelity as well as native function. However, not only the properties, but also the motions of the NC tunnel during the major steps in translation are important. This is one example of the details we are investigating.

5.2. Extracting the Ribosomal Mechanism from Simulations

Because the fluctuation simulations are independent of time, we investigate which motions occur in a coordinated way within a single mode, and thus necessarily occur at the same time.

We consider the ribosome model consisting of PDB files 1GIY and 1JGO. This model is of moderate resolution (5.5Å) and as such only contains phosphate atom coordinates for much of the rRNA. We construct an ANM model using RNA P atoms, protein Cα atoms, and cutoff values of 12Å, 20Å, and 17Å
for Cα-Cα, P-P, and Cα-P interactions respectively, without the proteins L9, L7, L12, and L11 that otherwise would dominate the motions.

Fifty low frequency normal modes are calculated using BLZPACK and analyzed further in an attempt to better understand mechanistic features of the ribosome. We want to learn about the coordination of the motions of various functional parts like the mRNA entrance and exit sites, tRNAs, nascent chain (NC) peptide tunnel, individual proteins along the lining of the mRNA and peptide tunnels, etc., and how these move in coordinated ways. These structural features are highlighted in figure 5.1. This information will lead to a deeper understanding of the mechanism of the ribosome.

Figure 5.1: The ribosome showing the mRNA and the nascent peptide channel (NC) tunnels.
The 30S subunit (1JGO) is shown in blues, 50S (1GIY) in oranges, the tRNAs green, the region that binds to the Shine-Dalgarno sequence cyan, and the mRNA red. Spheres represent phosphate and alpha carbons present in the mRNA, mRNA tunnel, and NC tunnel. The other proteins that we specifically consider (those associated with the NC and mRNA tunnels) are shown with transparent surfaces. The left panel shows the whole ribosome while the closer views of the mRNA tunnel are shown on the upper right and the peptide tunnel on the lower right.
Figure 5.2: Viewing the 70S ribosome with mRNA and nascent chain tunnel volumes.

The same ribosome structure as in Figure 5.1 is visualized to highlight the tunnel volumes. To generate tunnel volumes, cubic lattice points at 2Å separations were used. Any close or overlapping lattice points that were within 4Å of ribosome atoms were removed. The remaining lattice points in the tunnels provide a useful representation of the tunnels (colored red) with the surfaces of these volumes shown. This is an appropriate approach because of the low resolution of the structure (only phosphate atoms for rRNA and Cα atoms for r-proteins).

A) The whole ribosome and B) a close-up view of the nascent chain tunnel and C) mRNA tunnel in the same orientation. D) The entire structure from A) has been rotated 90°. We show the tunnel structures from B) and C) in rotated form E) nascent chain tunnel, and F) mRNA tunnel. The orientation in F) is slightly altered from D) to better show the tunnel’s variable width.

A low resolution description of the two major tunnels is shown in Figure 5.2. We investigate the mRNA entrance proteins S3, S4, and S5, the mRNA exit proteins S7, S11, and S18 as well as the region which binds the Shine-Dalgarno (SD) sequence, L3, L4, and L22 proteins which are on the surface of the 50S and each have a long appendage that reaches into the NC exit channel, and peptide channel exit proteins L23, L24, and L29. The NC tunnel is quite linear, although there is substantial variation in its
thickness (red volume in Figure 5.2). Because the PDB structures used are of relatively poor resolution, only phosphate (P) atoms for the RNA and C\(^\alpha\) atoms for the proteins are resolved in the structure we have used. To determine the volume of each tunnel we fill the space with lattice points spaced at 2Å apart. Any lattice point that is within 4Å of ribosomal atoms is discarded. The remaining points describe the tunnel. We define the nascent chain (NC) and mRNA tunnels as all atoms of the ribosome within 7Å of these remaining tunnel lattice points. It is interesting to follow how the tunnel walls move during translation and to learn whether these motions relate to the polypeptide’s motion exiting the ribosome, or its folding while in the tunnel. The wall selections include atoms from multiple r-proteins (dark green in Figure 5.2).

Since all of these ribosomal components have different sizes we cannot simply consider standard vector correlations to determine whether subunits are moving similarly. One way to consider the directional relationships of the motion in a given normal mode of two subunits is to consider how each atom (or coarse-grained point) moves in the mode. For each atom, we compute the dot product between its direction of motion and the direction of motion of each atom in another component. The result is a matrix of pair-wise dot products between the atoms of two components, an example of which is displayed in Figure 5.3.

![Figure 5.3: Two examples of matrices of vector overlaps using the most collective mode of motion; A) L3 and L22 and B) S5 and the mRNA tunnel. L3 and L22 move mostly together in a highly coordinated way. About 40\% of the S5 subunit moves together with the tunnel, but about 60\% moves in the opposite direction.](image)

In Figure 5.3 we show two examples of heatmaps for the correlations between components of the ribosome based on these matrices of vector overlaps generated using the most collective mode; one
case is simple with nearly completely correlated motions and the second is more complex. All values fall on the interval [-1,1] and each point indicates a vector overlap of two atoms in their displacement vectors for a given normal mode. We show examples of the complete vector overlaps of L3 and L22 proteins as well as the more complicated relationship between S5 and the mRNA tunnel. We can see that about 40% of the S5 subunit moves parallel to the tunnel, but about 60% moves in the opposite direction.

Figure 5.4: Visualizations of the first mode of motion – a ratchet motion with opposite rotation of the 50S and 30S subunits.

A) This matrix displays the extent of pair-wise vector overlaps that are above a given threshold (0.5 was used). Entries in the upper triangle, colored from blue to red, indicate more positive directional correlations the lower triangle shows negative correlations. Thus, red in the upper triangle corresponds to dominance of positive correlation in the directions of motions of the two parts and red in the lower triangle to a high degree of opposing vector overlaps. NC-T and mRNA-T are the peptide and mRNA tunnel-wall atoms. Self interactions (diagonal) are colored black. B) Coloring is the same as Figure 5.1. For each atom in the model we plot a directionless vector connecting its position in the positive and negative extremes of its motion. This mode corresponds well with the established ratchet-like motion. We find that the motion of S4 is anti-correlated with the mRNA exit site proteins S7, S11, and S18. The mRNA exit site proteins exhibit strong positive correlations with one another, as do the NC associated proteins L4, L22, L23, and L24. Interestingly, the anti-Shine-Dalgarnosequence (the ribosomal RNA that binds to the Shine-Dalgarno sequence) exhibits strong correlation with the mRNA exit proteins and a strong negative correlation with the mRNA entrance protein S4. Proteins S7 and S3 display positive correlation of motion with the large subunit proteins, pointing to allosteric communication through the ribosome.

To develop a wider view of these correlation matrices, we consider the amount of area within each matrix that is positive or negative and whose magnitude is greater than a threshold (here 0.5 was used). In this simple way, we learn about the strongest correlations and anti-correlations. For the cases in
Figures 5.3 and 5.4, we consider the percent of the matrix with a magnitude greater than the threshold; the upper triangular entries are for positive vector overlaps and the lower for negative cases. In the following set of figures colors indicate the fractions of the heatmap matrix area, so 1 means a complete overlap.

From Figure 5.4 we see that the mRNA exit proteins S7, S11, and S18 move in tandem with each other and the SD sequence, but also move strongly in the opposite directions from the mRNA entrance protein S4. Likewise, L4, L22, L23, and L24 move together. The mRNA entrance protein S3 moves in the same directions as the NC exit and tunnel proteins (except L3). Tunnel atoms exhibit more complex motions in this mode of motion.

**Figure 5.5:** The 5\textsuperscript{th} mode of motion – a rocking hinge motion which alternately opens and closes the two ends of the mRNA exit tunnel.
See Figure 5.4 for colors and explanation of upper and lower halves of the matrix. The 30S proteins (except S3) move in a coordinated fashion with each other, while the 50S proteins behave similarly. The anti-correlation between the small and large subunit proteins is evident in the lower triangle. Moderate anti-correlations are seen between the mRNA, its tunnel lining, and the 50S proteins.

In Figure 5.5 we analyze another mode of motion; the first that exhibits non-ratchet behavior. This mode describes a coordinated hinge motion within the 30S and 50S that alternately opens and closes the two ends of the mRNA tunnel. Interestingly this motion when combined with the ratchet motion will require some stretching of the mRNA within the ribosome during the synthetic cycle. We can see from Figure 5.5 (left) that there are some similar cohesive motions (strong red blocks in their upper triangular entries) as in Figure 5.4, but with differences. The SD and mRNA sequence as well as the mRNA tunnel wall atoms are all well correlated in this mode with the entrance and exit proteins (with the exception of
S3 and to some extent S7). There is some degree of anti-parallel motion between the mRNA-associated proteins and the NC-associated proteins. The strongly correlated blocks are S4, S5, S11 and L3, L4, L22, L23, and L24. There is some anti-correlated motion between the mRNA-associated proteins and the NC associated proteins – showing that there is significant coordination between the motions at the two distant sites.

The NC and mRNA tunnel walls do not exhibit strong directional agreement in the first 3 modes. For example, near the peptide reaction center the lining has negative overlap with L4 and the three exit proteins (L23, L24, L29), while it has positive overlap elsewhere.

In mode 5 we begin to see stronger directional preferences within the tunnel walls as well as a diminishing of the well defined ratchet-like rotation between the 30S and 50S. Many modes show relatively low or seemingly random distributions of overlap area. From visual inspection of the modes this seems to be due to the predominant motion occurring elsewhere in the structure. One example is mode 6 where the dominant motion is a rotation of L1 coupled with bending of the 30S. In many of the modes the core of the 30S moves semi-rigidly, so this flexing accounts for the unusual behavior.

Some modes exhibit exceptionally strong overlaps, but do not correspond to the ratchet motion or any other semi-rigid motion of the 30S and 50S. Mode 11 is such a case. Interestingly, the groups we consider here (NC and mRNA entrance and exit) move rigidly. If we define the NC tunnel as “vertical” then there is a horizontal shearing motion that moves the NC tunnel mouth and surrounding proteins as a unit. Simultaneously, the 30S and portions of the 50S flex away from each other, opening the tRNA binding regions. This could be indicative of moving the NC forward to accommodate the ratchet motion, eject the E-tRNA, and move the A and P tRNAs forward.

It has been hypothesized that the Shine-Delgarno (SD) sequence may bind non-specifically to the mRNA and helps pull it through the structure. Modes 5-18 and selected higher frequency modes depict strong overlaps between the motion of the SD and mRNA sequences.

The preliminary investigations here have uncovered significant information about the correlations among the motions of the functional components and have demonstrated why it is important to consider the structure as a whole. Further, some modes of motion exhibit similar frequencies which may indicate a change in timescale of the described motion (Figure 5.6.). We will now begin the more detailed, including the more definitive investigations about the motions of the ribosome and how its motions control its function.
Figure 5.6: Distribution of eigenvalues. Jumps in eigenvalues (squares of frequencies) indicate a change in the time scale of these motions.
This information may indicate which events are likely to happen concurrently or on the same time scale.

5.3. The ribosome is like a clock-work

Here we further describe the pulling simulations that we have performed and briefly reported in section 5.2. Figure 5.7 illustrates the direction of the added force. The correlation between the force-induced motions and select normal modes is listed in Table 5.1. Figure 5.8 gives further details about the mRNA gating mechanism that we observe in simulations as well as the force pulling deformation. We compare the present study to past ENM studies of the ribosome carried out in the Jernigan lab in Figure 5.9 where it was determined that S3 was the most mobile of the mRNA tunnel gating proteins. S4 was shown to be more mobile than S5. This is in agreement with the mechanism described here where the mRNA gating is predominantly driven by S3 and S4 moving together (Figure 5.8). The degree of difference in mobility of these proteins is influenced by the presence of tRNAs or other proteins in the A, P, and E sites, which may help to direct and regulate the translation cycle.
Figure 5.7: External force imposed to pull the mRNA ahead. We pull the ribosome bound mRNA in the direction $F_{\text{app}}$, which induces two major types of motion. The terminal mRNA (red) residue is pulled in the direction $F_{\text{app}}$. The first is the 30S-50S ratchet (large arrows) and the second is the alternating opening and closing of the mRNA tunnel (smaller arrows) – when the mRNA entrance is open, the exit is closed and visa versa.

Table 5.1: Pulling on the ribosome bound mRNA induces functional motions that are similar to the dominant normal modes, especially the 30S-50S ratchet and the mRNA tunnel gating.

<table>
<thead>
<tr>
<th>Mode</th>
<th>Correlation with Deformation</th>
<th>Mode Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.65</td>
<td>30S-50S ratchet</td>
</tr>
<tr>
<td>2</td>
<td>0.45</td>
<td>alternating A/E open/close</td>
</tr>
<tr>
<td>3</td>
<td>0.12</td>
<td>alternating A/E open/close</td>
</tr>
<tr>
<td>6</td>
<td>0.21</td>
<td>mRNA tunnel gating</td>
</tr>
<tr>
<td>13</td>
<td>0.19</td>
<td>mRNA tunnel gating</td>
</tr>
<tr>
<td>14</td>
<td>0.25</td>
<td>E site opening</td>
</tr>
<tr>
<td>16</td>
<td>0.17</td>
<td>interface C/E</td>
</tr>
<tr>
<td></td>
<td></td>
<td>interface C/E and peptide</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tunnel compression</td>
</tr>
</tbody>
</table>
Figure 5.8: Two major types of motion within the 30S shown by our models are displayed – the 30S head-swivel as the mRNA tunnel gates.

A) The 30S head-swivel was first analyzed by (137) by comparing experimentally determined structures and aligning the bulk of the 30S – the body (b), spur (sp), shoulder (sh), and platform (pt) – and seeing a difference in the position of the head (h) and beak (bk).

B) The mRNA gating is displayed, focusing on the mode-induced effects at the entrance and exit. A role for the flanking ribosomal proteins is proposed where they act as gate keepers, partially controlling this process.

Figure 5.9. Previous work in the Jernigan lab (Figure 7 in Kurkuoglu et. al. (65) reproduced with permission) using ANM showed that S3 is the most mobile of the mRNA entrance proteins, followed by S4, which is in agreement with our observed dynamics (Figure 5.8).
5.4. Assembly Order of Ribosomal Proteins Relates to Their Computed Protein Dynamics

(Manuscript in preparation for submission)

“Assembly Order of Ribosomal Proteins Relates to Their Computed Protein Dynamics” by Brittany Burton, Michael T. Zimmermann, Robert L. Jernigan and Yongmei Wang

5.4.1. Abstract

Assembly of the ribosome from its protein and RNA constituents has been studied extensively over the past 50 years, and experimental evidence suggests that prokaryotic ribosomal proteins undergo conformational changes during assembly. However, to date, no studies have attempted to elucidate these conformational changes. The present work utilizes computational methods to analyze protein dynamics and to investigate the linkage between dynamics and binding of these proteins played during the assembly of the ribosome. We studied the dynamics of three primary proteins from *E. coli* and *T. thermophilus* 30S subunits (S15, S17, and S20) with atomic molecular dynamic simulations and other analysis methods developed in our group. We find that r-proteins contain a higher than average percentage of positive residues (Lys + Arg is 18.7% for *E. coli* and 21.2% for *T. thermophilus*). Also, positive residues constitute a large proportion of RNA contacting residues (39% for *E. coli* and 46% for *T. thermophilus*). This indicates the importance of charge-charge interactions in the assembly of the ribosome. We show that solvent-exposed proteins (S15 and S17) tend to adopt more stable solution conformations than an RNA-embedded protein (S20). We also found protein residues that contact the 16S rRNA are generally more mobile in comparison with the other residues. This is because there is a larger proportion of contacting residues located in flexible loop regions. By the use of elastic network models, which are computationally more efficient, we show that this trend holds for most of the 30S r-proteins.

5.4.2. Introduction

Ribosomes are the macromolecular machines that synthesize proteins in all living organisms. They are composed of ribosomal RNA (rRNA) and ribosomal proteins (r-proteins) that self-assemble into functional units. The efficient and accurate self-assembly of the active ribosome in vivo is essential for cell growth because new ribosomes and proteins must be produced in order for cells to grow. It is estimated that approximately 60% of all cellular transcriptional activities have been attributed to the synthesis of rRNA in a rapidly growing cell (138) and 40% of the total energy of an *E. Coli* cell is directed
towards the synthesis of proteins (139). It is therefore not surprising that ribosome biogenesis in cells is intricately regulated. Elucidating this complex regulation network has become the focus of a rapidly developing field.

The assembly of the ribosome requires the orchestration of highly coordinated events that involve both rRNA folding and r-protein binding. While many cofactors have been identified that participate in assembly in vivo, active functional units can be assembled in vitro in the absence of these cofactors (140). The small 30S subunit of the bacterial ribosome (see Figure 5.10), which is composed of 16S rRNA and 21 r-proteins, has been more extensively studied than other structural assemblages and is a good system to analyze in order to determine what is important for the ribonucleic particle (RNP) assembly. In particular, the 30S subunit was the first to be reconstituted from purified components by the Nomura group in the late 1960’s (141). The reconstituted 30S active particles showed nearly the same activities in all performed biochemical assays. This ability to reconstitute active particles in vitro allows for in-depth exploration of the roles of the individual components in ribosome assembly and their functions by the combinatorial addition and omission of individual components (140, 142). These experiments revealed that the 30S subunit assembles in a sequential and ordered process (140). The Nomura group also provided a detailed assembly map describing the sequential and interdependent binding of all r-proteins (143). The map also classified the proteins as primary, secondary, and tertiary binders, depending on their ability to bind to 16S rRNA. The primary proteins bind to bare rRNA, secondary proteins can bind to 16S rRNA after at least one primary protein has already bound, and tertiary proteins require at least one primary and one secondary protein (144).

The Nomura assembly map reflects the equilibrium thermodynamics of r-protein binding with 16S rRNA to intermediates. Using chemical probing methods, these binding kinetics were more recently studied by Powers et al (145). Based on their experimental results, the r-proteins were divided into early, mid, mid-late and late binders. The kinetics data were partially in agreement with thermodynamic data in that the tertiary binding proteins were consistently found to be late binders. The availability of atomic structures of the 30S subunit (146) provided tremendous new opportunities to understand the assembly mechanism (147). Most of the knowledge gained in earlier experimental studies was found to be consistent with the determined structures.
Figure 5.10: The 30S \textit{T. thermophilus} subunit (1J5E), interface side colored by domain.

The 16S rRNA and r-proteins of interest are highlighted: 5’ Domain yellow, Central Domain grey, 3’ Major Domain orange, and 3’ Minor Domain purple; S15 blue, S17 dark green, and S20 dark red. The \textit{E. coli} structure (2AVY) is nearly identical, but slight structural differences for the proteins of interest are discussed in the text and visualized in Figure 5.11. The remaining r-proteins have been removed for better visualization of the 16S domains.

In the meantime, significant progress was made with experimental methods to probe the ribosome assembly mechanism. Time-resolved X-ray-dependent hydroxyl radical footprinting (148) provides resolution on the order of milliseconds, much shorter than other chemical probing methods (145). Directed hydroxyl radical probing (149) allows for the detection of specific interaction sites between proteins and RNA. The Williamson group used PC/QMS (pulse-chase followed with quantitative mass spectrometry) to measure the kinetics of individual protein binding during the assembly of the full ribosomal complex (150). New experimental data suggest that ribosome assembly proceeds via multiple parallel pathways (150, 151) rather than a single pathway involving the formation of a single rate-determining “reaction intermediate” RNP (152). Current understanding of the ribosome assembly process suggests it is similar to protein folding in that it can proceed via multiple pathways across a rugged energy landscape.
Despite significant progress in recent years, the understanding of ribosome assembly remains limited. One major obstacle in this field is elucidating the mechanisms of coordinated RNA folding, protein binding, and the associated conformational changes of RNA and r-proteins (153). Although earlier studies suggested (154) that r-proteins adopt the same structures in solution as in the assembled ribosome, more recent studies suggest (153) that there are conformational changes in the r-proteins and rRNA upon forming the complexes. Predicting RNA structure is also one of the most challenging topics in structural biology because a single stranded RNA can adopt a variety of secondary and tertiary structures. The 16S rRNA molecule in a ribosome is divided into four domains: the 5’ domain, the central domain, the 3’ major domain and the 3’ minor, each with a well-defined structure (see Figure 5.10). Magnesium ions are thought to stabilize the secondary structure of RNA and many r-proteins are thought to stabilize the tertiary structures. Many of the r-proteins interact with and bind to only one domain, but a few associate with more than one, such as S20 which interacts with both the 5’ and the 3’ minor domains. Stagg et al. (155) analyzed the atomic contacts of r-proteins with RNA in the 30S subunit structure and reported the interesting observation that most of the late binding r-proteins were found to bind at the 3’ end of 16S RNA. This observation was consistent with the earlier understanding that 16S RNA folds with 5’ to 3’ polarity (144, 156). The Stagg group further used coarse-grained representations of RNP structures to examine the potential fluctuations of binding sites when proteins were removed or added. Their study shows that the binding sites of primary proteins are formed first and, once associated, these proteins help organize the late binding sites. Trylska et al. (157) calculated the binding energy of individual r-proteins with the 16S RNA by solving the Poisson-Boltzmann equation, which accounts for electrostatic interactions. Though the calculated binding energies varied, some late binders were found to have less favorable binding free energies while the early binders were found to be more favorable, an observation consistent with known experimental results. Other studies used various coarse-grained representations to explore the global motions of the ribosome (62) and the assembly of the 30S (158). Despite the coarse representations of ribosomal structure, some of the known dependencies of r-protein and rRNA binding were captured in these computational studies.

Ribosome assembly remains an active research field. A better understanding of its assembly mechanisms will provide valuable biochemical insight into cellular regulation and will allow for the optimal development of ribosome-targeted drugs. While experimental studies continue to make great progress, computational studies reported so far are still limited. Most of the earlier reported computational studies have used coarse-grained representations of the ribosome. To truly understand
the specific binding of r-proteins with 16S RNA, atomistic details need to be considered. Because assembly involves both RNA folding and protein binding, the examination of individual components before and after binding in atomistic detail is necessary. Here we specifically investigate the potential correlation between r-protein dynamics properties and their binding properties. The aim is to answer the following specific questions: what are the key residues that bind to the 16S rRNA? Are these key residues more flexible than the others? Do free r-proteins adopt the same conformations as those found in the assembled 30S subunit? To explore the answers to these questions, we rely on the use of atomistic molecular dynamic simulations of r-proteins as well as another method for investigating protein dynamics using elastic network models.

5.4.3. Methodology

Analysis of Contacts in the Assembled 30S subunits

In the current study, we analyze the crystal structures of the 30S subunits from the *Escherichia coli* (PDB ID 2AVY (146)) and *Thermus thermophilus* ribosomes (PDB ID 1J5E (160)). Structural and sequence alignments of r-proteins found in the two species were done with Molecular Operating Environment (MOE) software (161). Contacts between r-proteins and 16S rRNA were analyzed using our own computer program. A contact point was defined as any atom of a protein residue found within a 3.5 Å cut-off distance from any 16S nucleotide atom. That amino acid was labeled as a “contact” residue. The total number of “contacts” between one r-protein and the 16S rRNA may exceed the total number of contacting residues identified in the protein because an amino acid may be within cutoff distance of more than one nucleotide, thus counting as more than one contact. The identity and position of these contact residues found in the assembled 30S subunit were recorded and used for further analysis.

Molecular Dynamics Simulations

Molecular dynamics (MD) simulations were run using the AMBER 10 software package (162) and the parmbsc0 force field (163), an optimization of the Amber99 force field for nucleic acids and proteins. The starting conformations of r-proteins for the MD were obtained from the crystal structures of the 30S subunits (*T. thermophilus* 1J5E and *E. coli* 2AVY). Counterions were added to neutralize the charge of the protein, and an additional 10 potassium and 10 chloride ions were added to create a low salt concentration. The protein systems were then solvated using a rectangular box of TIP3P water. The systems were subjected to two minimization cycles: 1000 steps with the protein fixed and 5000 steps unrestrained. Afterward, a 100 ps warm-up MD simulation was run at constant volume by increasing
temperature from 0 to 300K, with the protein fixed using a restraint constant of 10.0 kcal·mol$^{-1}$·Å$^{-2}$. The MD simulation then switched to the NPT ensemble (p=1.0 bar), using the Langevin thermostat with a collision frequency of 1.0 ps$^{-1}$, to equilibrate the ions and water density for 2 ns. The restraint force on the protein was then removed and the production run began with the NPT ensemble (p=1.0 bar) using a time step of 2 fs. All simulations used the SHAKE algorithm to constrain covalently bonded hydrogen atoms and the Particle Mesh Ewald (PME) method to calculate long-range electrostatic interactions, with a cutoff distance of 10.0 Å. Histidines are represented as HIE (neutral charge: hydrogenated N$^\circ$, aromatic N$^6$). Duplicate MD simulations were performed to verify that the reported dynamic behaviors of each protein are representative in the final MD runs. MD production runs were performed for at least 200 ns.

Using Ptraj to monitor the overall structural changes in reference to the starting structure, the root-mean-square deviation (RMSD) for each protein was calculated as a function of production run time. If the plot of the RMSD versus time forms a plateau, the protein likely adopts a solution-stable conformation; however, a widely fluctuating RMSD plot indicates a flexible protein in solution. To quantify the mobility of each residue, root-mean-square fluctuations (RMSF) were calculated using the average protein conformation as the reference state. The RMSF values presented in this paper are calculated from 10 ns to the end of each simulation (approximately 200 ns) to allow adequate time for the protein to fully adopt its stable solvated conformation, if one was at all achieved. This ensures that the RMSF plot differentiates flexible residues from stationary residues during the time that the protein samples its solution-stable conformations. In both RMSD and RMSF calculations, all atoms were included.

The RMSF is related to the experimental B-factors reported by crystallographers, through a simple relationship ($B$-factor = $(8/3)\pi^2(RMSF)^2$), which could be compared with the experimental measured B-factors reported in the PDB files of the 30S subunits. However, the experimental B-factors for each r-protein found in the 30S subunits were nearly featureless for individual proteins, probably because the reported B factors reflect the mobility of the atoms within the whole assembled subunit and are not representative of the individual r-proteins. Hence, we did not compare the B-factors calculated from MD simulations with the experimental B-factors.

Snapshots of each protein at various stages throughout the simulations were visualized using Visual Molecular Dynamics (164) (VMD) to identify the flexible and stable regions of the protein. All images
were made with VMD, which is developed with NIH support by the Theoretical and Computational Biophysics group at the Beckman Institute, University of Illinois at Urbana-Champaign.

**Elastic Network Modeling**

Because the Molecular Dynamics simulations require significant resources, we have also chosen to model the dynamics of the complete set of 30S ribosomal proteins with the more computationally efficient elastic network models (34), using the Anisotropic Network Model in particular (38). ANM models permit us to investigate the dynamics of all of the 30S proteins more quickly but with less accuracy on the observed dynamics than MD. ANM models are constructed using the crystallographic $C^\alpha$ coordinates of each protein and a cutoff of 13Å. Due to its coarse-grain design, the ANM is subject to the “tip effect” (6) in which highly extended points ($C^\alpha$) experience exaggerated motions, which would place exaggerated weight on the most mobile residues. To compensate for this effect, we calculate the RMSF of each residue position in each structure and remove extreme outliers from subsequent analyses. The “tip effect” residues removed in this study are Arg88 and Gly89 from *T. thermophilus* S15, and Gly8, Val9, Val10, and Val11 from *T. thermophilus* S17. We also use RMSF to make comparisons between 16S rRNA contacting residues and non-contacting or highly conserved residues. The same definition for contacting residues and conserved residues are used in the ANM calculations.

**Statistical Analysis of Contact Residue Mobility**

To statistically determine linkages between highly mobile and contacting residues or conserved residues from both ANM calculation and MD simulation, we calculate an enrichment factor for each protein defined as the ratio of the average RMSF for contacting over non-contacting residues. An enrichment factor greater than 1 implies that the contacting residues are more mobile than the non-contacting residues. However, an enrichment factor less than 1 implies the reverse. The statistical significance ($p$-value) of the enrichment factor is calculated based on the permutation test explained as follows. For a protein of $N$ residues, $C$ of which are contacting, we have an observation of the enrichment of RMSF at the contacting residues relative to the non-contacting residues. Let this ratio be $O$. We then randomly select $C$ residues from the protein and calculate the analogous ratio between this random set and its compliment. Performing the random selection 10,000 times, we construct a distribution of enrichment values within random sets of $C$ residues. The significance ($p$-value) of our initial observation, $O$, is then the proportion of random samples that have an enrichment greater than $O$. A small $p$-value (e.g., $p<0.01$) implies that a random set of $C$ residues is unlikely to have an
enrichment factor equal or greater than the observed ratio O. This not only means that the contacting residues are more mobile than the non-contacting residues, but that there are very few subsets of size C exhibiting the same magnitude of mobility.

5.4.4. Results and Discussion

Ribosomal proteins are enriched with positively charged amino acids

Most ribosomal proteins are positively charged. We performed a simple calculation of the net charge of ribosomal proteins based on the sequences reported for the 2AVY and 1J5E structures, counting Asp and Leu as -1, Lys and Arg as +1, with all other residues treated as neutral. Of course some of these residues might have some charge because of shifted pK\textsubscript{a} values due to their location in the tertiary structure, but we will ignore these minor effects at present. Table 5.2 presents the net charge of r-proteins for the two species. The two r-proteins that are not positively charged could be explained by their special positions in the assembly map: S2 is the last protein to assemble (143) and S6 is known to form a dimer with S18 (165), which is positively charged, before associating with rRNA. The remainder of the r-proteins are all positively charged. We also note that the charge on r-proteins from T. thermophilus is on average higher than that for the E. coli proteins, which may relate the general observation that ribosomal subunits for thermophiles such as T. thermophilus are more stable than those of mesophiles such as E. coli. (166) Moreover, ribosomal proteins are enriched with positively charged amino acids. The typical percent of amino acids for Lys, Arg, Leu and Asp are 5% each for cytosolic proteins (167). However, in the case of r-proteins, the total percentage of Lys and Arg is approximately 20% (18.7% for E. coli and 21.2% for T. thermophilus), while the sum of Leu and Asp percentages remained near 10%.
Table 5.2: Net charge of r-proteins from sequences

<table>
<thead>
<tr>
<th>r-Protein</th>
<th>E.Coli (2AVY)</th>
<th>T. Thermophilus (1J5E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2</td>
<td>-1</td>
<td>-7</td>
</tr>
<tr>
<td>S3</td>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td>S4</td>
<td>17</td>
<td>23</td>
</tr>
<tr>
<td>S5</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>S6</td>
<td>-12</td>
<td>0</td>
</tr>
<tr>
<td>S7</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td>S8</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>S9</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>S10</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>S11</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>S12</td>
<td>21</td>
<td>27</td>
</tr>
<tr>
<td>S13</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>S14</td>
<td>15</td>
<td>16</td>
</tr>
<tr>
<td>S15</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>S16</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>S17</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>S18</td>
<td>12</td>
<td>19</td>
</tr>
<tr>
<td>S19</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>S20</td>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td>S21</td>
<td>14</td>
<td>12</td>
</tr>
</tbody>
</table>

Charge is calculated based on sequence as the simple sum of the number of Lys and Arg minus the number of Asp and Leu residues. As RNA is negatively charged, most r-proteins are positively charged.

We have further examined the contacts made between r-proteins and the RNA based on the atomic structures of the 30S subunit from the two species. Here, a contact is defined as having any atoms of a protein residue within 3.5Å of any rRNA nucleotide atoms. Table 5.3 presents the number of contacts made by each r-protein, along with the number of contacts with positively charged residues. It is clear that a high percentage of contacts between r-proteins and rRNA are made by positively charged residues. The total average percentages of contacts made by positively charged residues are 39% for *E. coli* and 46% for *T. thermophilus*, and both are significantly higher than the total percentage of the positively charged amino acids in r-proteins for the two species.
Table 5.3: Breakdown of total contacts in terms of contacts made by positively charged residues.

<table>
<thead>
<tr>
<th>R-proteins</th>
<th>E.Coli (3.5Å cut off)</th>
<th>T. Thermophilus (3.5Å cut off)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2</td>
<td>19</td>
<td>7</td>
</tr>
<tr>
<td>S3</td>
<td>40</td>
<td>7</td>
</tr>
<tr>
<td>S4</td>
<td>64</td>
<td>23</td>
</tr>
<tr>
<td>S5</td>
<td>46</td>
<td>13</td>
</tr>
<tr>
<td>S6</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>S7</td>
<td>29</td>
<td>15</td>
</tr>
<tr>
<td>S8</td>
<td>37</td>
<td>10</td>
</tr>
<tr>
<td>S9</td>
<td>81</td>
<td>44</td>
</tr>
<tr>
<td>S10</td>
<td>42</td>
<td>15</td>
</tr>
<tr>
<td>S11</td>
<td>52</td>
<td>19</td>
</tr>
<tr>
<td>S12</td>
<td>75</td>
<td>28</td>
</tr>
<tr>
<td>S13</td>
<td>48</td>
<td>22</td>
</tr>
<tr>
<td>S14</td>
<td>54</td>
<td>23</td>
</tr>
<tr>
<td>S15</td>
<td>42</td>
<td>8</td>
</tr>
<tr>
<td>S16</td>
<td>42</td>
<td>20</td>
</tr>
<tr>
<td>S17</td>
<td>32</td>
<td>14</td>
</tr>
<tr>
<td>S18</td>
<td>30</td>
<td>16</td>
</tr>
<tr>
<td>S19</td>
<td>37</td>
<td>17</td>
</tr>
<tr>
<td>S20</td>
<td>52</td>
<td>24</td>
</tr>
<tr>
<td>S21</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>836</td>
<td>330</td>
</tr>
</tbody>
</table>

Note: The primary r-proteins are highlighted in gray. The total number of protein contacts for S15, S17, and S20 above differs from the total number of contact residues presented in Tables 5.4-5.6 because some protein residues are in contact with more than one nucleotide, which are presented here as multiple contacts.

Structures and contact residues are more conserved than sequences

Figure 5.11 shows structural alignments for the three proteins from the two species. The sequence similarities between the two species are 60% for S15, ~40% for S17, and ~28% for S20, but the percentages of conserved amino acid side chain types (large, hydrophobic, aliphatic, acidic, etc.) are considerably higher: 75% for S15, ~58% for S17, and ~47% for S20. Thus, the structures for the three ribosomal proteins are well conserved, with RMSD values of 1.1 Å for S15, 1.4 Å for S17, and 2.1 Å for S20. In the cases of S17 and S20 from T. thermophilus there are extra C-terminal regions, as shown in Figures 5.11b and 5.11c.

Residues that contact rRNA exhibit higher than average sequence conservation. For S15, the percent of conserved contact residues is about 54% (52% for E. coli and 56% for T. thermophilus), which is just under the overall sequence conservation. For S20, the percentage of conserved contact residues is 38% for E. coli and 35% for T. thermophilus, both of which are considerably higher than the overall sequence conservation. For S17, the percentage of conserved E. coli contacting residues (52%) is higher than the overall sequence conservation, whereas that for T. thermophilus contacting residues (31%) is less. The
conserved contact residues percentages for S17 and S20 from *T. thermophilus* are lower than those for *E. coli* because *T. thermophilus* has extra C-terminal regions that make several additional non-conserved contacts. (Supplementary Tables S1-S3 presents the contact residues for S15, S17 and S20 for the two species, with conserved residue identities in red and conserved side chain types, largely Lys/Arg substitutions, colored green.)

**Figure 5.11: Crystal structure comparisons for A) S15, B) S17, and C) S20.**

Comparisons of proteins from two different species; *E. coli* proteins are shown in the lighter shade and *T. thermophilus* in the darker shade. Contact residues are shown as stick representations and some important parts of the proteins, discussed in the text, are labeled.

Further analysis of the identities of these contact residues reveals that, aside from the positively charged residues, also His, Thr, Ser, and Gln are common, all of which are polar and can form hydrogen bonds interaction with rRNA. For example, of the twenty-seven *E. coli* S15 contacts, five are basic (Lys48, Arg54, Arg64, Lys65, and Lys73), five are histidines (His38, His42, His46, His50, and His51), ten are polar (Ser2, Thr5, Thr8, Thr22, Ser24, Gln28, Gln35, Ser52, Ser61, and Gln62), and one is aromatic and polar (Tyr69). The remaining six contacts are acidic (Asp21 and Asp49) or nonpolar (Gly23, Leu31, Leu39, and Gly55). Therefore, most contacts between the r-proteins and the rRNA are either charged interactions, or hydrogen bonds, with few aromatic stacking or nonpolar interactions.
Figure 5.12: RMSD calculated from the X-ray crystal structure for A) S15, B) S17, and C) S20 across MD trajectories and for both species. 

_E. coli_ proteins are represented by lighter squares and _T. thermophilus_ by darker triangles. Panel (b) shows the RMSD for the whole S17 _T. thermophilus_ protein in olive green as well as that for just the part of the structure that is homologous structure (dark green) to _E. coli_ S17 (omitting the extra _T. thermophilus_ C-terminal part). (Refer to Figure 5.11b for visualization of the homologous region.) Notably this C-terminal part of S17 causes the _T. termophilus_ to greatly increase its overall mobility.

**Dynamics and conformational Changes of S15**

S15 is a primary binding protein which binds in the 3’ major domain of 16S RNA. In the assembled 3OS subunit, S15 is solvent-exposed and located on the back of the 3OS subunit body. The 16S RNA binding site of S15 is at the three-way junction of helices 20, 21, and 22 in the 16S central domain. The primary, secondary, and tertiary structures of S15 are highly conserved across species: four bundles of α-helices are connected by short loops (Figure 5.11a). All 16S rRNA contact residues are found on one side of S15, located on helices 1, 2 and 3 and the loops connecting the three helices, but helix 4 does not have any contacts with rRNA.
In previous structural studies, X-ray (168) and NMR (169) derived structures were reported and the only significantly different conformation reported was in the crystal structure Clemons et al. (168), where helix 1 was rotated 90° away from the remaining bundled helices. Additional studies have been published about the role of S15 in ribosome assembly and antibiotic responses with mutagenesis studies (170) and MD simulations, studying the effects of Mg²⁺ ions on the protein alone and with its rRNA binding site (171). It has been suggested that this protein acts as a bridge between the large and small subunits in the fully assembled ribosome (172).

Root-mean-square deviations (RMSD) were calculated from the molecular dynamics simulations of the S15 protein and are presented in Figure 5.12a. The S15 from the two species exhibit relatively low RMSD values during MD simulations, with values remaining below 5 Å. Figure 5.13 presents the root-mean-square fluctuation (RMSF) values calculated over the period of time from 10 ns until the end of the simulation. Contact residues are shown as solid symbols in the plot. High RMSF values were observed for the loop connecting helices 2 and 3, which have several conserved contact residues. The contact residues found on helices 2 and 3 have very low RMSF values, whereas helix 1 and the loop connecting helices 1 and 2 have a few contact residues with moderate RMSF values. Helix 4, which has no contacts with 16S RNA retains its helical structure during the MD simulation and has moderate RMSF values. Representative backbone structures for *E. coli* and *T. thermophilus* S15 are depicted in Figure 5.14. The proteins retain their secondary and tertiary structures during the MD simulations and only small conformational changes are observed for either S15 protein. This indicates that the S15 protein from both organisms is a relatively stable protein in solution and that the conformations observed during the simulations are similar to that of the attached protein in the assembled ribosome.

Table 5.4 compares average RMSF data for all protein residues, contacting residues, and conserved residues. Comparison of these residue subsets will inform us about whether contact residues or conserved residues are more or less mobile. The average RMSF value for all *E. coli* S15 residues is 2.11 Å and for all contact residues 2.24 Å. For *T. thermophilus* S15, all residues have an average RMSF of 1.84 Å and for all contacts the average RMSF is 2.37 Å. These differences are small, but statistical analysis shows that S15 contact residues are positively enriched with mobile residues, as indicated by enrichment factors greater than 1 for both species (Table 5.4; 1.08 for *E. coli*; 1.46 for *T. thermophilus*; see Methodology for explanation of enrichment factors). The reported p-values for these enrichment factors signify that the mobility enrichment of *T. thermophilus* contact residues is more significant than for *E. coli*. In this sense, p-values represent a comparison against the proportion of randomly selected
residues that would be likely to have enrichment factors higher than contact residues. Conserved residues, on the other hand, tend to exhibit lower RMSF values: *E. coli* conserved residues average RMSF is 1.84 Å and for *T. thermophilus* is 1.54 Å. The mobility enrichment factors show a negative enrichment for conserved residues (0.73 for *E. coli*; 0.68 for *T. thermophilus*), and p-values indicate ~99% of random selections would exhibit larger enrichment values. Thus, on average, one may conclude that S15 contact residues have higher RMSF whereas conserved residues tend to be less mobile.

Table 5.4: Average MD RMSF values (Å), (standard deviation) and enrichment factors

<table>
<thead>
<tr>
<th></th>
<th>All Residues</th>
<th>All Contacts</th>
<th>Conserved Residues</th>
<th>Contact Enrichment</th>
<th>P-value</th>
<th>Conserved Enrichment</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S15</td>
<td>2.11 (1.24)</td>
<td>2.24 (1.27)</td>
<td>1.84 (1.06)</td>
<td>1.077</td>
<td>0.217</td>
<td>0.735</td>
<td>0.988</td>
</tr>
<tr>
<td>S17</td>
<td>1.85 (0.94)</td>
<td>2.28 (0.92)</td>
<td>1.79 (0.89)</td>
<td>1.103</td>
<td>0.199</td>
<td>0.953</td>
<td>0.603</td>
</tr>
<tr>
<td>S20</td>
<td>8.82 (3.17)</td>
<td>9.14 (3.34)</td>
<td>8.65 (3.05)</td>
<td>1.055</td>
<td>0.215</td>
<td>0.974</td>
<td>0.560</td>
</tr>
<tr>
<td><em>T. thermophilus</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S15</td>
<td>1.84 (1.29)</td>
<td>2.37 (1.66)</td>
<td>1.54 (0.96)</td>
<td>1.461</td>
<td>0.008</td>
<td>0.680</td>
<td>0.995</td>
</tr>
<tr>
<td>S17</td>
<td>4.68 (2.93)</td>
<td>5.74 (3.40)</td>
<td>3.43 (1.29)</td>
<td>1.398</td>
<td>0.008</td>
<td>0.631</td>
<td>0.998</td>
</tr>
<tr>
<td>S20</td>
<td>6.96 (2.94)</td>
<td>7.62 (2.80)</td>
<td>7.37 (2.94)</td>
<td>1.152</td>
<td>0.057</td>
<td>1.083</td>
<td>0.196</td>
</tr>
</tbody>
</table>

Dynamics and conformational Changes of S17

In the 30S subunit, S17 is also solvent exposed and is located near S15 in the 5’ domain of the 16S rRNA. To date, no X-ray crystal structures have been determined for S17 alone, but a low resolution NMR solution structure has been presented for *Bacillus stearothermophilus* S17 (173). The S17 structure found in the *E. coli* 30S subunit is comprised of a small β-barrel and an extended β-hairpin loop (Figure 5.11b). The contact residues are located on one end of the β-barrel and in the extended β-hairpin loop. The S17 from *T. thermophilus* has an extra C-terminal α-helix which makes additional contacts with the 16S rRNA (Figure 5.11b). Thus, *E. coli* contact residues exhibit somewhat higher conservation than the overall sequence does, whereas *T. thermophilus* contact residues are slightly less conserved than the sequence of the full-length proteins. In the *E. coli* 30S subunit, the S17 β-hairpin loop is embedded in rRNA and contains five contacts, three of which are found contacting helix 11 of the central domain with two contacting the 5’ domain at helix 21. The axis of the β-barrel is oriented into the main part of the rRNA, and the end of the barrel nearest the RNA contains the remaining contact points, all of which contact the 5’ domain of 16S rRNA along helices 7, 9, and 11. Because these contacting residues associate with both the 5’ domain and the central domain, *E. coli* S17 is a plausible anchor between them. The *T. thermophilus* S17 also contacts these two 16S domains but includes an additional ten
protein contacting residues in its C-terminal α-helix and coiled tail. These residues have a larger extent of contact with helix 11 and strengthen the association with the central domain at helices 20 and 27. Research indicated that the 30S subunit assembly begins at the 16S rRNA 5’ end (145) and, S17 appears to organize the 5’ region (156), so it is clear that the cooperative conformational changes and rRNA binding of this protein are likely to play an important role in the early stages of ribosome formation.

Figure 5.13: RMSF values for A) S15, B) S17, and C) S20 calculated from MD trajectories. RMSF values for S15: results for E. coli are in red with the squares indicating contact residues and for T. thermophilus proteins colored blue with triangles for contacts. The proteins have been aligned to demonstrate the behaviors of the conserved structural elements. Aligned Residue Number, therefore, does not necessarily reflect the actual residue index in the protein sequence.

During the MD simulation of E. coli S17, the β-sheet structures remained stable: the average RMSD for this protein was relatively low (below 5 Å; lime green plot, Figure 5.12b). Conversely, a much higher
RMSD was observed for S17 from *T. thermophilus* (olive green plot, Figure 5.12b), although the protein did take on a relatively stable conformation after ~80 ns of simulation. Further investigation reveals that the extra α-helix in *T. thermophilus* S17 is responsible for the high RMSD values. The structurally homologous portions of the proteins have comparable RMSD values (*T. thermophilus* homolog: dark green plot, Figure 5.12b), both around 4 Å. The backbones of structurally homologous portions both retain their overall shape during the MD simulations.

S17 RMSF values (Figure 5.13) were calculated from the MD simulations starting from the 10 ns point until the end of the trajectory. While the *T. thermophilus* S17 generally exhibited larger deviations from its starting structure than did the *E. coli* S17, when sequentially aligned, the RMSF values for the structurally homologous portions of the proteins correlate well. For *E. coli* S17, the loops connecting the β-strands, the extended β-hairpin loop, and both termini exhibit comparably high RMSF values, whereas the β-strands participating in the β-barrel (valleys in Figure 5.13) have low RMSF values. The same pattern is true for the homologous portion of the *T. thermophilus* RMSF plot, and the extra C-terminal region exhibits very large RMSF values. The contact residues in the *E. coli* S17 are located in the highly mobile β-hairpin, the moderately mobile Loops 1 and 6, as well as the least mobile β-strands of β-barrel: β5, the last residue of β1, and the first of β2. In *T. thermophilus* S17, there are four regions of the protein with high RMSF (the N-terminus, the β-hairpin loop, Loop 4, and the C-terminus), all of which contain contact residues. In fact, every residue in Loop 4 is a contact residue, and residues close to each end of the loop also have high RMSF values. The three contact residues in the α-helix have high RMSF and the ten residues in the C-terminal coil have some of the highest RMSF, seven of which are contact residues. The low and moderate contact residues are found in the β-barrel: β1, Loop 1, β2, and β3.

Representative structures seen throughout the *E. coli* and *T. thermophilus* S17 simulations are shown in Figure 5.14. The RMSF data and these images indicate that the structurally homologous regions of the S17 protein behave similarly in solution and that the β structures of both homologs retain their overall shape throughout the simulations, whereas the flexible C-terminal α-helix in *T. thermophilus* loses its helical structure. These data imply that the β-barrel confers good stability in solution for the two species.

Further analyses of the relative mobility of contact residues and conserved residues shows similar trends as S15. The average RMSF (Table 5.4) for all residues in *E. coli* S17 is 1.85 Å and 2.28 Å for all contacting residues; for *T. thermophilus*, the average for all residues is 4.68 Å, and 5.164 Å for all contacting residues. The differences in these values, while small, indicate that contact residues are, on
average, more mobile than all residues for both S17 proteins. Enrichment factors for S17 show positive mobility enrichment for contact residues in both species (Table 5.4; 1.10 for \textit{E. coli}; 1.40 for \textit{T. thermophilus}), with p-values indicating that \textit{T. thermophilus} enrichment is more significant than \textit{E. coli}. All \textit{E. coli} conserved residues have an average RMSF of 1.79 Å and an enrichment factor of 0.95. The \textit{E. coli} S17 p-value for conserved residues signifies that approximately 60% of random selections of residues would show stronger mobility enrichment. The average RMSF for \textit{T. thermophilus} conserved residues is 3.43 Å, with an enrichment factor of 0.63. More than 99% of random residues selections are likely to be more positively enriched for mobility than these conserved residues. Therefore, like S15, the S17 contacts are statistically more mobile residues and conserved residues statistically less mobile.

\textit{Dynamics and conformational change of S20}

In the 30S subunit crystal structures from both species, protein S20 is found deeply embedded in the 16S rRNA. This protein contacts 16S RNA helices 6-9, 11, and 13 in the 5’ domain and is the only r-protein to contact helix 44 in the 3’ domain. The structure of S20 consists of a unique set of three bundled α-helices, with helix 1 twice as long as the others, the N-terminus most deeply inserted into the subunit, and only a small portion of the three-helix bundle exposed to solvent. While the \textit{E. coli} and \textit{T. thermophilus} S20 proteins have a generally conserved tertiary body (Figure 5.11c), the \textit{T. thermophilus} S20 crystal structure is missing its first seven residues and has an additional 15 residue C-terminal tail which the \textit{E. coli} protein does not have.

The simulation RMSD values for S20 from both species oscillate wildly (Figure 5.12c), indicating the proteins conformation vary broadly from their starting conformations (up to ~20 Å). Multiple length simulations (at least 200 ns) show that while S20 RMSD may remain within a range of 5-10 Å for a time, the protein does not adopt a solution-stable conformation. The S20 RMSF plots (Figure 5.13) have similar trends for both \textit{E. coli} and \textit{T. thermophilus} S20 proteins, and aside from the first portion of α1, the three α-helices are primarily located at valleys in the plots. The highly flexible region of α1 binds to rRNA helices 6, 7, and 13, whereas the nearby, more stable contact residues in α1 contact the tip of rRNA helix 44, a helix that has no contacts with any other small subunit proteins. The remaining contacts have relatively moderate or low RMSF values. As seen in the other proteins, the loop regions between the stable secondary structures are located at peaks in the RMSF plot, whereas the α-helical regions themselves correspond to the RMSF valleys. Visual inspection of the trajectories suggests that the major contributor to S20 flexibility is helix 1 (Figure 5.14), which extends deeply into the rRNA. The N-terminal
portion of helix 1 bends and swings wildly during the MD simulations. *E. coli* helix 1 bends near Arg24 and Thr30 and *T. thermophilus* near Lys29.

Previous studies (174) have shown that the free S20 protein in solution does not exhibit the high percentage of α-helical regions as seen in the crystallized structure. The conformational variation exhibited by S20 in the work here is consistent with this data, and this flexibility coupled with the deep insertion of the protein into the folds of RNA in the fully-assembled ribosome indicate that S20 is stabilized primarily by its large number of contacts with the RNA.

The average RMSF trends (Table 5.4) for S20 contacting and conserved residues are generally in agreement with the results presented for S15 and S17. For *E. coli*, the average RMSF for all residues is 8.82 Å and for all contacting residues is 9.14 Å. In *T. thermophilus*, the average value for all residues is 6.96 Å and 7.62 Å for all contacts. These data show that the mean RMSF for all contacts is greater than that for the whole structure, consistent with the results for S15 and S17. Both *E. coli* and *T. thermophilus* S20 proteins show positive enrichment of mobility in their contact residues (Table 5.4; 1.06 for *E. coli*; 1.15 for *T. thermophilus*). Like S15 and S17, the *T. thermophilus* S20 contact residues are more significantly enriched, indicating that there is a smaller chance that a random set of the same number of residues would have a greater enrichment factor than do the contact residues. For *E. coli*, the conserved residues have an average RMSF is 8.65 Å and for *T. thermophilus* is 7.37 Å. The *E. coli* conserved residue enrichment factor of 0.97 indicates that the conserved residues are somewhat negatively enriched with mobility whereas the *T. thermophilus* conserved enrichment factor of 1.08 corresponds to a small positive enrichment. This difference in conserved residue mobility between the two species could be due to their respective thermal stabilities, or it may be related to the major conformational changes associated with S20.
Figure 5.14: Structural variations during MD simulations for all six proteins. Backbone snapshots of *E. coli* S15, S17, and S20 are shown in A, C, and E, and *T. thermophilus* S15, S17, and S20 in B, D, and F, respectively. Backbone starting structures are shown in yellow.

*General trends based on Elastic Network Modeling*

To rapidly assess the potential connection between contacting residues and their mobilities, we use elastic network modeling which can also compute RMSF values using only a fraction of the computational resources required for the MD simulations. The dynamics calculated via the Anisotropic Network Model (40, 175) correlate reasonably well with those from the MD simulations. The elastic
network models have been applied previously to the ribosome by us (62, 63, 65, 66). For example, the correlation coefficient between RMSF values calculated for E. coli S15 is 0.57, for S17 is 0.63, and for S20 is 0.81. ANM and MD predict similar patterns of mobility and stability for the aligned sequences, with most of the discrepancy at the terminal residues and highly flexible regions (such as S20 α-helix 1 and S17 β-hairpin loop). In fact, if the first two and last two residues of E. coli S15 are excluded from the correlation, the correlation factor increases to 0.67. The MD simulations typically predict greater terminal residue mobility (except for the highly mobile S20 helix 1) and the ANM calculations consistently predict higher fluctuation values for extended residues in the middle of the protein.

ANM analysis confirms the observation from the MD simulations: rRNA contacting residues are enriched for high mobility. Mobile residues for proteins S15, S17, and S20 from both species are enriched 123-148% in contacting regions. P-values for the enrichment of each protein are calculated using a permutation test (see Methodology) with all six proteins showing significance. Overall, conserved residues tend to exhibit higher than average mobility, but the pattern is complex. Judging conservation using the ConSurf server (130), we find that some proteins (such as S20) exhibit increased enrichment for high mobility for the most highly conserved residues compared even to the moderately conserved residues. However, other proteins (such as S15 and S17) actually exhibit lower than average enrichment for the most conserved residues despite the moderately conserved residues (this set also includes the highest) being more mobile. See Figure 5.15 for visualization of the ConSurf conservation scores on the 30S structure. There appears to be a complex relationship regarding conservation. Some residues are conserved because they are important for the tertiary structure or fold of the protein, while others are conserved for rRNA binding. Thus, general rules that relate conservation and mobility of r-protein residues remain elusive. E. coli protein contact residues show significance at the 0.05 confidence level whereas those from T. thermophilus are enriched at the 0.01 level, demonstrating a slightly higher degree of control over the placement of mobile residues in the thermophile relative to the mesophile.
Figure 5.15: Sequence conservation shown by using ConSurf for the 30S r-proteins, displaying S15, S17, and S20 specifically with rRNA contacts.
Protein C\(^\alpha\) atoms are shown as spheres colored by their respective ConsSurf conservation score (inset). RNA is shown as a pale orange transparent molecular surface. The ConSurf server (130) generates multiple sequence alignments for each r-protein using between 150 and 300 sequences.
A) Full view of the 30S in a top-down view placing the 50S behind the 30S. B) Rotated to show the surface of the 30S that faces the 50S. C) Specific views of S15, S17, and S20 shown as cartoons. D) Rotated view. It is apparent for S15 and S20 that the rRNA interacting residues are conserved, but this effect is not so strong for S17.

ANM mobility enrichment was also calculated for all 19 r-proteins in the two 30S X-ray structures (Table 5.5). *T. thermophilus* proteins show a slightly increased enrichment relative to *E. coli*; with average enrichment factors of 1.51 and 1.46, respectively, with medians of 1.43 and 1.33. This analysis shows that most r-proteins are significantly enriched for mobile residues at the rRNA contact points at the 0.05 level. That is, contacting residues are not only enriched, but they make up a subset of residues that is maximally enriched, for the given structure. Of the 6 proteins categorized as being early by Stagg (155), two *E. coli* and five *T. thermophilus* are significant at the 0.05 level. Of the six primary proteins identified by Nomura (142), three *E. coli* and and five *T. thermophilus* are significant at the 0.05 level. Proteins involved later in assembly are not differentially significant between the two species. This
suggests a novel adaptation in thermophiles – the increased control over the placement of highly mobile residues.

<table>
<thead>
<tr>
<th>Table 5.5: ANM enrichment factors and significance for 30S proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>S02</td>
</tr>
<tr>
<td>S03</td>
</tr>
<tr>
<td>S04</td>
</tr>
<tr>
<td>S05</td>
</tr>
<tr>
<td>S06</td>
</tr>
<tr>
<td>S07</td>
</tr>
<tr>
<td>S08</td>
</tr>
<tr>
<td>S09</td>
</tr>
<tr>
<td>S10</td>
</tr>
<tr>
<td>S11</td>
</tr>
<tr>
<td>S12</td>
</tr>
<tr>
<td>S13</td>
</tr>
<tr>
<td>S14</td>
</tr>
<tr>
<td>S15</td>
</tr>
<tr>
<td>S16</td>
</tr>
<tr>
<td>S17</td>
</tr>
<tr>
<td>S18</td>
</tr>
<tr>
<td>S19</td>
</tr>
<tr>
<td>S20</td>
</tr>
</tbody>
</table>

Note: EF is the enrichment factor, defined as the ratio of root mean square fluctuations for contacting over non-contacting residues. The P-value computed with a permutation test is also reported. See text for details.

5.4.5. Conclusion

Several important conclusions can be reached based on the above reported results. First, the positively charged residues on r-proteins must play important roles in binding with 16S rRNA. A significantly higher percentage of contacts between r-proteins and rRNA are formed by these positively charged and hydrogen bonding residues. We also see that r-proteins from a thermophilic species (T. thermophilus) have more positively charged residues than a mesophilic species (E. coli), which correlates with the fact that thermophilic ribosomes must maintain stronger (or a larger number of) interactions in order to function at considerably higher temperatures. Second, as previously discussed (153), conformational changes of r-proteins could take place during 16S rRNA binding. Our study clearly shows that the α-helix 1 of S20 is unstable in solution by itself and exhibits large conformational changes. In
contrast, S15 and S17 adopt stable conformations in solution, which agrees with the earlier suggestion (154) that ribosomal proteins do not undergo structural changes during the assembly. We attribute the differences in these behaviors to the level of solvent exposure the protein experiences in the assembled subunit. In the ribosome, S15 and S17 are primarily solvent exposed so their solution structures more closely resemble their bound structures, whereas S20 is deeply embedded in the 16S RNA and its association with its RNA binding site stabilizes the flexible portion of α-helix 1. Third, analyses of residue mobilities reveal that RMSF values for contact residues are statistically higher than those for all residues. This means that contacting regions are more enriched with mobile residues than non-contacting regions, which supports previous observations (154) that the flexible regions of ribosomal proteins are usually the locations of RNA contacts. This trend holds for all 30S r-proteins, and the increased mobility of contact residues could ensure more efficient binding and even aid in the binding site preparation for later binding proteins by actively associating with their 16S binding partners and helping to fold and maintain the appropriate rRNA tertiary structure. The thermophile exhibits increased control over the placement of highly mobile residues within proteins that bind rRNA early in assembly. However, this does not mean that all contact residues are located in the flexible loop regions. It is important to point out that there are many contact residues found in α-helices and β-sheets that exhibit low to moderate RMSF values. Interestingly, for some proteins such as T. Thermophilus S15 and S17, the most conserved residues tend to have lower average RMSF than all residues despite the overall trend that conserved residues are more mobile. Conserved residues in S17 are located on β-sheets that help organize the structure, but they are not crucial for binding with RNA.
CHAPTER 6. Functional insights into Immunoglobin-G from structural dynamics


6.1. Abstract

Motions of the IgG structure are evaluated using normal mode analysis and a new time dependent form of the elastic network model, to detect hinges, the dominance of low frequency modes, and the most important internal motions. We also evaluate the protein crystal and its packing effects on the experimental temperature factors and disorder prediction. We find that the effects of the protein environment on the crystallographic temperature factors may be misleading for evaluating specific functional motions of IgG. The extent of motion of the antigen binding domains is computed to show their large spatial sampling. We conclude that the IgG structure is specifically designed to facilitate large excursions of the antigen binding domains. Normal modes are shown as capable of computationally evaluating the hinge motions and the spatial sampling of domains of the structure. The antigen binding loops and the major hinge appear to behave similarly to the rest of the structure when we consider the dominance of the low frequency modes and the extent of internal motion.

6.2. Introduction

Immunoglobulin Gamma (IgG) is one of the principal players in the adaptive immune system and is commonly referred to as an antibody. It is produced in a vast array of diverse antigen binding forms by B-cells through the process called V(D)J recombination. This produces a molecule consisting of two heavy and two light chains with highly variable complementary determining regions (CDRs), shown in the context of the full structure in Figure 6.1. The heavy chains of a given antibody are identical by sequence to each other and are comprised of four immunoglobin folds (a two layer sandwich of 7 anti-parallel beta strands), one of which is variable in sequence. Light chains are also identical in sequence and consist of two immunoglobin folds, one of which is variable. The C-terminus of each light chain usually has a cysteine to form a disulfide link to the C-terminal region of the second heavy chain domain. These four immunoglobin domains come together to form an antigen binding domain (one for each
pair). The remaining two heavy chain immunoglobin folds interact with each other to form a third domain that is often referred to as the constant region. IgG has been studied by proteolysis which cuts the structure into the three described fragments. For this reason the two types of domains which comprise the IgG are referred to as the antigen binding fragments, or F_{ab}, and the constant fragment, or F_{c}.

![Figure 6.1: The structure of IgG (1IGT) highlighting the domains and hinges.](image)

Heavy and light chains are distinguished from one another by colors. The F_{ab} domains have orange ellipses indicating the locations of the hypervariable loops (the CDRs) and green and yellow circles identifying the major and minor hinges. F_{ab} domains consist of one light chain and half of a heavy chain and are connected to the F_{c} and each other by the major hinge.

The region connecting the F_{ab} domains to the F_{c} is known to be a highly flexible hinge. This hinge region has 2-4 disulfide bonds bridging the heavy chains. Previously it has been excised from IgG and used in protein design as a molecular linker. This sequence has been extensively characterized and even synthesized (176, 177). While the primary characteristic of this region is its hinge flexibility, which was an impediment to resolving the structure early on, it has also been shown to have somewhat unique hydrophobic binding that allows it to be selectively bound to a stationary membrane so that either the F_{ab} or F_{c} fragments can be cut off with different proteases and recovered (178).

Here we compute a hinge map of IgG using Elastic Network Models (ENMs), show the extensive spatial freedom of the unrestrained F_{ab} domains that presumably facilitates binding, analyze the internal changes of the structure and how they affect the hypervariable Complementarity Determining Regions (CDRs), and apply a recently derived normal mode based kinematic simulation to generate motions of the structure. Motions of the structure are analyzed and a description of the high mobilities of the CDRs is provided.
6.3. Materials and methods

6.3.0. Elastic Network Model – See Chapter 2.3.2 for details.

6.3.1. Kinematics of Proteins

Our method for solving the kinematics of coarse-grained protein structures is based on the Lagrangian equation for the potential and kinetic energy of the system, as described by Chirikjian and coworkers (179-183). First, a rigid body translation and rotation of the structure is performed to place the origin of the coordinate system at the center of mass and so that the moment of inertia tensor is diagonal. The potential energy of the system of \( N \) points can then be written as in Equation 2.2. The displacement vector of the system, \( \Delta R(t) \), could be calculated as

\[
\Delta R(t) = \sum_{i=1}^{3N} \left[ \frac{1}{\sqrt{\omega_i}} \sin(t \sqrt{\omega_i}) Q_i Q_i^T \Delta R(0) + \frac{\cos(t \sqrt{\omega_i})}{\sqrt{\omega_i}} Q_i Q_i^T \Delta R(0) \right]
\] (6.1)

Where \( \omega_i \) and \( Q_i \) are the eigenvalues (square of angular frequencies) and eigenvectors (normal modes) of the system, respectively. This facilitates performing time-dependent kinematic simulations with the ENMs using any desired combination of normal modes fixed by the index \( i \) and by choosing appropriate phase angles to describe the displacements between the phases of the different normal modes.

6.3.3. Computing Changes in Internal Distances

We also consider the displacements of the positions of points in the structure with ANM. The mean-square change in internal distance (MSID) is computed as

\[
< (\Delta R_i - \Delta R_j)^2 > = < \Delta R_i ^2 > + < \Delta R_j ^2 > - 2 < \Delta R_i \cdot \Delta R_j >
\] (6.2)

These values are obtained directly from the inverse of the Hessian matrix, \( \Gamma^{-1} \), from which the normal modes are derived:

\[
< (\Delta R_i - \Delta R_j)^2 > = (3k_B T/a) \left[ \Gamma_{ii}^{-1} + \Gamma_{jj}^{-1} - 2 \Gamma_{ij}^{-1} \right]
\] (6.3)
where $k_b$ is the Boltzmann constant, $T$ absolute temperature, and $\gamma$ the spring constant. We can also consider the normalized change in internal distances. This metric can be used to compare the magnitude of internal distance changes.

$$< (\Delta R_i - \Delta R_j)^2 > = \frac{< (\Delta R_i - \Delta R_j)^2 >}{\sqrt{\Delta R_i^2 \Delta R_j^2}}$$ \hspace{1cm} (6.4)

### 6.3.4. Fractal and Spectral Dimension

As early as 1980 the fractal dimension of myoglobin was studied (184), and it was found to be about 1.65. Experimental analysis of the spectral dimension of lysozyme was recently performed (185). This study revealed not only that proteins may exhibit a mix of phonon (exhibiting discrete vibrational modes) and fractal character but also that the spectral dimension is relatively low and shows only moderate sensitivity to temperature. This finding provides an explanation for the efficient information transfer through protein structures. More recently, Granek and Klafter showed mathematically that certain fractal structures (and not uniform lattices) will experience the type of autocorrelation decay that is observed in protein experiments (186). The compactness of protein structures is represented by a fractal dimension between 2.3 and 2.7 (see Enright and Leitner (187)). Investigation of the spectral dimension of elastic networks and explaining its relation to real structures has also been carried out (188). Spectral and fractal dimensions were related to one another in recent papers by Reuveni and colleagues (189, 190). An equation was proposed that relates the two dimension metrics that fits well with the 5794 surveyed protein structures (191). Here, we utilize the methods described in Ref. (189) for calculating the spectral and fractal dimensions of IgG and relate these findings to the CDR motions.

The fractal dimension describes how the mass captured within concentric spheres scales with the radius of these spheres. It is calculated here by finding the ten points closest to the proteins center of mass. Concentric spheres with incremental radii of 1Å are constructed and the total mass captured within each is recorded. Linear regression is performed ten times, once for each of the points closest to the protein’s center. The average slope of the log-log plot of sphere radius versus mass captured is taken as the fractal dimension. The spectral dimension describes how the frequencies of vibration for the structure scale with the density of modes. That is, one performs a linear regression against the log-log plot of frequency versus the cumulative number of modes at each frequency. The spectral dimension is then the slope of this regression.
6.4. Results and Discussion

We seek quantification of the motions of IgG in its dominant normal modes, which correspond to the flexing about the major hinge, particularly to see how this affects the spatial freedom of the CDR, both overall and internally.

Figure 6.2 shows the impact of the six slowest normal modes on the motion of the IgG and points out the CDRs. We see that these six normal modes account for nearly all of the motion of the Immunoglobulin, above 85% of the total motion for all residues and greater than 90% for the majority. This means that residues of the immunoglobulin move in a highly coordinated motion and that the loops do not act as in polymers, to randomly sample their dihedral angles. A large body of evidence shows that the ENM generates low modes that correspond with known biochemical functions of proteins. This gives us confidence to conclude that correlations of motion within a low frequency mode are pertinent to the function of the IgG.

We have also computed correlations between mean-square fluctuations calculated using only the six lowest frequency modes and using all modes, for all residues, and the CDR (see Figure 6.2). Similarly, we compare correlations computed by using all normal modes in equation 6.1 with those obtained by using only the slowest modes (see details in the Methods section).

Figure 6.2 shows that the mean impact of the first 6 normal modes on the total motion is about 96% and that the lowest frequency modes do provide an excellent representation of the overall motions of the system. From a visual inspection we see the low frequency modes are associated with domain motions, a behavior that is usually seen in multidomain structures. For this reason normal modes have been used to identify hinges within structures (192). We perform similar computations to confirm the presence of the hinges within IgG (see Figure 6.3). In order to determine the extent of CDR sampling we generate conformers using the normal modes. The magnitude of deformation in each mode is set by choosing the largest deviation that does not substantially deform the sequential virtual bond lengths. The lowest frequency modes corresponding to the global motions are collective in nature and exhibit comparatively low virtual bond stretching. We find that overall the three domains are anti-correlated with one another. Figure 6.3 part C displays a representation of CDR sampling after following normal modes. It is apparent that the structure of IgG is designed to span the maximal space for the CDRs, presumably to aid binding. A similar approach for understanding the spatial freedom that the structure can sample would be to alter dihedral angles; Figure 6.4 illustrates this mapping approach, which is only
preliminary. Other considerations would be required for more realistic dihedral sampling (steric hindrances, backbone constraints, energetic and inertial effects, etc.).

Figure 6.2: The first six modes of motion capture most of the total motion in IgG and the CDR. Mean correlations of the motion derived from the first six normal modes with the total motion for (A) the entire structure, (B) the six CDR loops from chains A and B, (C) the six CDR loops from chains C and D. 1IGT has 1316 residues in total. The mean correlation over all residues is 0.96 showing that the slowest modes strongly dominate the intrinsic motions of the structure. Behaviors for each of the individual CDRs are shown.
Figure 6.3: Hinges and the structure behavior identified by ANM computations.

(A) Correlation matrices (dot products between normalized pairwise displacement vectors). All values fall in the range [-1, 1]. The average over the first 9 modes is displayed, with red corresponding to the motion of $C^\alpha$ pairs positively correlated, blue for negative correlations, and white uncorrelated. From the block structure of the diagram and the changes in sign we can identify the hinge regions within the structure. (B) Similar to (A), but for individual modes; 1 (top) and 9 (bottom). (C) We use the 12 slowest modes to compute conformers of IgG because the majority of the motions in these modes are localized at the three prominent hinge regions. The $F_c$ is aligned in all conformers. Structure coloring shows the $F_c$ in blue, the major hinge in red, CDRs in green, and the remainder of the $F_{ab}$ in orange (in two perpendicular views).

Normal mode calculations are often performed to elucidate which residues or atoms in a molecular structure are the most mobile. Mobile active site residues may play roles in binding or substrate selection, whereas rigid regions are more likely to play key stabilization roles in the structure as a whole,
as in a scaffold. An important exception to this occurs for the catalytic residues within an active site cleft that are relatively rigid.

Figure 6.4: The effect of sampling hinge dihedral angles on the position of the CDR.
In this figure the red amino acids psi-angle at the major hinge is varied in 15° increments and the resulting position of the CDR on the right side is accumulated. The collection of all of these CDR coordinates is shown as the blue volume similar to Figure 6.3(C). This visualization could be useful for IgG hinge analysis, but would require inclusion of the limitations imposed by torsion angle availability (Ramachandran space) and steric clashes.

Another quantity that is informative about internal conformational changes is the mean-square internal distance (MSID) changes, \( \langle (\Delta R, - \Delta R)^2 \rangle \), given by equation 6.2. MSID changes can be calculated directly from the Hessian matrix that is used to generate normal modes in ANM with equation 6.3. This quantity describes the changes within a structure; how the normal modes stretch, compress, or otherwise alter the pair-wise distances between points in the structure. If this change in internal distance is zero for a given \((i, j)\) pair, then the two points move together rigidly (the distance between them remaining unchanged). We have analyzed structures and seen that (data not shown) the areas of a protein with the smallest internal mean square distance changes are the cores of domains with these values increasing further away from stable cores. We have employed ANM models built with uniform springs with cutoffs ranging from 10-15 Å and with springs having inverse square dependences on distance. All of these yield similar results. Figure 6.5 shows this quantity averaged across all pairs of points within 7Å of one another. The CDR and major hinge are shown separately. We see that the CDR and hinge regions do not have significantly lower or higher average RMSID. We find that the \( F_{ab} \) domains experience more internal motion than the \( F_c \), but that the two \( F_{ab} \)s are not symmetric in their motions. This is likely due to the asymmetry in the initial structure. But, other feasible structures might be expected to actually behave in a symmetric way. \( F_{ab}^2 \) is closer to the \( F_c \) than \( F_{ab}^1 \) and has more
connections (higher stiffness) with it. Notably, we find that the internal distance changes at the hinge, as usual, are relatively small.

![Image](image.png)

**Figure 6.5: Mean-square internal distance changes within the entire immunoglobulin.**

(Inset) The CDRs from Fab1 are comprised of the variable loops in chains A and B, and the Fab2 CDRs are in chain C and D of 1IGT. We find that F\textsubscript{ab} domains experience more internal motion than the F\textsubscript{c}, but that these are not symmetric. The inset shows the same quantity but specifically for the two CDR regions and the major hinge.

For many proteins, it is common to compare motions from the computations with the crystallographic temperature factors, the B-factors. The B-factors describe the uncertainty assigned to a given atom, usually by assuming it originates from relevant internal fluctuations. Rigid body contributions are often removed by the crystallographer, but the successes of the TLS (193, 194) and vGNM (195) methods provide strong evidence that B-factors often contain significant rigid body contributions. In the case of the 1IGT crystal we find that the B-factors may not be representative of the solution dynamics since the CDR of each F\textsubscript{ab} is strongly bound to the F\textsubscript{c} of another IgG (see Figure 6.6). While the experimental B-factors do highlight the major hinge as the most flexible part of the structure it is important to note that this is not because it is allowed to flex in the protein crystal, since the molecules are highly restrained by intermolecular interactions.

Intrinsic disorder in proteins is a topic of growing popularity. Two disorder predictors were applied to the IgG structure, DisEMBL (196) and POODLE (197). Interestingly, both methods predict the most mobile part of the structure, the CDR, to be the least likely to be disordered. Both methods have components in their scoring scheme that are knowledge-based; learned from scanning the PDB. We believe the CDR is predicted to be so stable because of the abundance of IgG structures and because the CDR is rarely unbound. Because the CDR is almost never free to move, it is always ordered in the known
structures. DisEMBL predicts disorder while the POODLE prediction predicts 3 quantities; the secondary structure as coil, the residues un-resolved in a crystal, and the probability of residues having a high B-factor. Both methods employ a probability cutoff of 0.5; any residue above the threshold is deemed to be disordered. Interestingly, the CDR is predicted to be the least disordered part of the structure. It is possible, due to the ambiguity that remains in defining protein disorder and the complexity of crystal B-factors, that some knowledge based disorder predictions may not be predicting exactly what one expects.

Figure 6.6: Intermolecular crystal packing in 1IGT shows that substrate free structures are still bound CDR bound.  
(A) One F_{ab} domain in 1IGT is shown in blue with its CDR as an orange surface. A symmetry related IgG is interacting with this CDR in the protein crystal. It is shown as a red surface and green sticks for the bound N-Acetyl-D-Glucosamine. (B) The F_{c} domain of 1IGT is shown as a red cartoon with gray transparent surface. Two intermolecular interacting F_{ab}s are shown in blue whose CDRs are highlighted in orange. These two views highlight crystal packing via CDRs contacting F_{c}.

In Figure 6.7 we show the mean-square fluctuations of the IgG variable fragments computed with the ANM model. We find the parts of the structure that are most variable in sequence, the CDRs, are also the most mobile. It is interesting to note that ANM indicates 4 loops with high spatial mobility (and also the N-terminus), but there are only 3 CDR loops in the variable domain of each chain. The fourth loop is colored purple in the inset molecular images in Figure 6.7. Each F_{ab} domain then has two of these conserved mobile loops with one on either side of the CDR. It is interesting to note that the F_{ab}-like T-cell receptor (TCR) has the same spatial arrangement of loops, but the fourth loop found to be mobile here is also hypervariable in sequence (198). While this loop does not usually bind antigen, it is involved in nonspecific antigen binding of TCRs.
Figure 6.7: The computed positional mean-square fluctuations of F\textsubscript{ab} and CDR residues shows increased mobility of the sequence-variable loops.

CDR residues highlighted with a thicker line. Molecular structures are shown with a semi-transparent surface colored blue to red for low and high computed B-Factors, respectively. Each plot has 4 peaks. Three of these correspond to the CDR and are colored yellow while the fourth is marked with an asterisk and colored purple. The remainder of the structure is colored green. In A, B, and C the CDR faces to the left while in D it faces right. This has been done because in D the back side is less mobile and the yellow CDR loops are less distinguishable when looking through the blue surface. In the F\textsubscript{ab}-like T-cell receptor the fourth loop (*) is also variable in sequence.

To further investigate the motion patterns within the CDR and whether these may indicate that the IgG structure itself facilitates excursions of the hypervariable loops, we consider the difference in the mean square fluctuations, internal distance changes, as well as spectral dimensions and fractal dimensions of the full IgG structure and of the individual F\textsubscript{ab} domains. GNM usually predicts the mean square fluctuations of each point more accurately (as judged with crystallographic B-factors) than ANM. Motions of the CDRs from the GNM are shown in figure 6.8. Curves from the full IgG structure and for the F\textsubscript{ab}1 domain have been scaled overall to match the range of the crystal B-factors. Including the whole structure yields a correlation with experiment of 0.87 for the F\textsubscript{ab}1, whereas utilizing only the structure of the F\textsubscript{ab}1 for the computations yields a correlation of 0.5. We have previously noted that the B-factors in this structure have the F\textsubscript{ab} locked in a bound state. Interestingly, the computed CDR motion appears to be captured better by the full IgG model than by use of the Fab only. Internal distance changes, computed by using equation 6.2, indicate the extent of deformation between pairs of points in the structure. In figure 6.9, the effect of the full structure on the magnitude of change within and between hypervariable loops is summarized. The mean change within a (or between) loop(s) is plotted with error bars indicating one standard deviation. Before computing the extent of fluctuation, we
rescale each mode to agree with equipartition wherein each internal degree of freedom would be assigned $k_B T/2$ energy (199). The full IgG structure amplifies internal distance changes within and between hypervariable loops, relative to a single Fab. This is further indication that the whole antibody structure may facilitate CDR configurational sampling so that a proper binding pose is found.

![Figure 6.8: Correlation of experimental and computed temperature factors of the CDRs is stronger when the full IgG structure is considered than when parts of the structure are considered.](image)

Experimental temperature factors from the CDR are plotted in the heavy line. The first 25 residues plotted are from the heavy chain and the next 22 from the light. Heavy (left) and light (right) chains are delineated by black bars underneath the residue index and individual CDR hypervariable loops are shown by gray bars. CDR motion computed from the full IgG has a higher correlation (0.87) with experimental data than motion computed for the CDR using only the Fab domain (0.50) when the 50 lowest frequency normal modes are used. The correlation between the two theoretical curves is 0.60. (inset) The correlations between temperature factors from the experimental B factors and computed from GNM models are shown. The magnitude of the correlation is not significantly affected by the choice of cutoff value. “+s” indicates the inclusion of coarse grained points from sugar molecules that attached to the Fc. It is evident that the motions available to IgG in the crystal environment are not likely to be identical to those in solution.

Information transfer within molecular structures has been the focus of numerous studies including the consideration of protein structures being fractal in nature (see Methods). Following previously established algorithms, we compute the fractal and spectral dimension of IgG and single Fab domains (Figure 6.10). Briefly, the fractal dimension describes how the mass captured by concentric spheres scales with the radius of these spheres, and the spectral dimension describes how the frequencies of
vibration for the structure scale with the density of modes. We find that the full length IgG behaves nearly like a 2D object in terms of its ability to transfer information from one part of the structure to another. Information transfer is thus significantly faster than one would expect from a uniform crystal lattice. This spectral dimension is in the range expected for proteins (191). Interestingly, we find that the full structure has a lower spectral dimension than any single domain, again pointing to the possible utility of the whole structure for finding the right binding pose.

Figure 6.9: Inclusion of the full IgG structure diminishes magnitude of internal distance changes.
Normalized internal distance changes within the CDR calculated from ENM modes are scaled according to equipartition and constructed from the full IgG structure (blue) and from a single Fab domain (black). The IgG structure itself appears to facilitate significantly larger excursions of the CDR loops away from their native positions, but does so without any significant internal rearrangements. The non-normalized changes are significantly greater for IgG.
Figure 6.10: Spectral and Fractal dimension of IgG and its domains shows the full structure to have the most efficient communication.

The full length IgG has a lower spectral dimension no matter the cutoff employed. For the typical GNM of 7.3 the spectral dimension of IgG is 1.7. This is in good agreement with experimental measures on other proteins. The Fab domain alone has a larger spectral dimension of about 1.9, but all structures considered have a similar fractal dimension. “+s” indicates the inclusion of coarse-grained points from sugar molecules that are affixed to Fc.

6.5. Conclusion

Normal mode analysis using ANM is shown here to detect the hinge motions within the dominant low frequency motions, as well as the internal motions of the IgG structure. We have also evaluated the protein crystal and compared against the experimental temperature factors and disorder predictions. We find that the protein environment may be misleading in the crystal regarding the actual functional motions. Crystallographic temperature factors also reflect the crystal intermolecular interactions, which are extensive in this structure. Modeling approaches such as those applied here can provide a more comprehensive view of the biomolecule and its functional motions. The spectral dimension relates the density of vibrational modes to their frequency and can be used as a judge of the efficiency of energy transfer through a structure. Since this quantity is lower for IgG than for any individual domain, the hypothesis that hinge motions (the dominant computed motions) facilitate CDR motion is strongly supported. We conclude that the IgG structure is specifically designed to facilitate large excursions of the Fab domains, as shown with the present methods for computational evaluation of the extent of hinge motions and the spatial sampling by components of the structure. Normal modes derived from the simplest potential function afford a good approximation to the total hinge motion and predict the most sequence-variable regions also to be the most spatially mobile – facilitating the binding of the Fab's. Our
results may impact immunology by suggesting ways to include flexibility in the docking to predict the bound structures of IgGs.
CHAPTER 7. Combining 4-body energy with entropies calculated from elastic networks

7.1. Incorporating vibrational entropy into decoy discrimination


7.1.1 Abstract

We propose a novel method of calculation of free energy for coarse grained models of proteins by combining our newly developed multibody potentials with entropies computed from elastic network models of proteins. Multi-body potentials have been of much interest recently because they take into account three dimensional interactions related to residue packing and capture the cooperativity of these interactions in protein structures. Combining four-body non-sequential, four-body sequential and pairwise short range potentials with optimized weights for each term, our coarse-grained potential improved recognition of native structure among misfolded decoys, outperforming all other contact potentials for CASP8 decoy sets and performance comparable to the fully atomic empirical DFIRE potentials. By combing statistical contact potentials with entropies from elastic network models of the same structures we can compute free energy changes and improve coarse-grained modeling of protein structure and dynamics. The consideration of protein flexibility and dynamics should improve protein structure prediction and refinement of computational models. This work is the first to combine coarse-grained multibody potentials with an entropic model that takes into account contributions of the entire structure, investigating native-like decoy selection.

7.1.2 Introduction

One of the most striking properties of globular proteins is their high packing density, observed at both atomics and amino acid resolution. Because of hydrophobic clusters and the network of hydrogen bonds, proteins can be expected (and have been shown) to behave in highly collective ways. Yet, the main tools that are used to assess these structures are pair-wise interactions for energies and local
environments for entropies. There is a clear need for more rigorous methods for evaluating the free energies of protein structures. Coarse-grained models of proteins have found favor due to their ability to satisfactorily reproduce results obtained at atomic detail and also their computational speed. This article discusses an investigation of two highly cooperative representations using coarse-grained models: four-body contact potentials for energy and elastic network models for entropy. The latter have been shown through multiple studies that most important motions of globular proteins are the large collective (often domain) motions, indicating that entropy evaluations should be based on the entire structure. Here we propose the integration of these two methodologies to evaluate protein free energies that should offer improvements for the evaluation of protein structure predictions, comparison of structures, and the refinement of computational models of proteins. Success with this approach would improve upon the present parlous state of protein thermodynamic assessment.

Different types of computational protein studies have been performed using knowledge-based potential functions, including structure prediction (200-204), design (205-208), docking prediction (209-212), and folding (213-216). Atomic (217-219) and coarse-grained potential functions (220-223) have been developed utilizing diverse methodologies. Knowledge-based potentials not only can significantly reduce the computational cost of modeling but can also improve predictions by selecting good predictions from a set. Extraction of better, computationally less expensive coarse-grained potentials that are able to perform as well as atomic potentials is an important challenge in computational biology.

Both coarse-grained and atomic structures use many different types of potentials in assessing protein models and for native structure recognition. The Miyazawa-Jernigan potential (221) is one of the most widely used coarse-grained two-body potentials. However, as suggested by Betancourt and Thirumalai (224), pair-wise potentials are unlikely to be sufficient for threading applications. In principle, multi-body potentials should better account for the more complex three dimensional interactions in densely packed structures, and more importantly, capture the strong cooperativity operative within protein structures. Three-body potentials proposed and developed by Munson and Singh (225) as well as by Li and Liang (226) all showed improvements over two-body potentials. Four-body potentials by Krishnamoorthy and Tropsha (227) (first derived in the context of Delaunay tessellation) also performed better than two-body potentials.

Our group, through a simple geometric construction, recently developed four-body contact potentials (32) incorporating sequence information and details of interactions between backbones and
side chains. These potentials also enable us to distinguish between different levels of solvent accessibility for the residues.

Overall performance has been enhanced (228) by combining the four-body sequential (32) with the four-body non-sequential potentials (229) and with a short range potential (230). The results for the rankings of the best models are obtained by combining these three sets of terms, and globally optimizing the weights for each term based on performance.

7.1.3. Four-body contact potentials

Most two-body potentials neglect the sequence information of proteins while both types (sequential and non-sequential) of four-body contact potentials derived by our group (32) consider the interactions between the backbones and side chains and the long-range interactions between side chains. These four-body contact potentials give a more cooperative representation of protein interaction energies and can discriminate well between native structures and partially unfolded or deliberately misfolded structures.

7.1.4. Geometric construction of four-body contacts

Residues are all represented by the geometric centers of the side chain heavy atoms, except for Glycine, where the alpha carbon atom is used.

<table>
<thead>
<tr>
<th>4-Body Class</th>
<th>Residues</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>E, D</td>
<td>Acidic</td>
</tr>
<tr>
<td>B</td>
<td>R, K, H</td>
<td>Basic</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>Cysteine</td>
</tr>
<tr>
<td>H</td>
<td>W, Y, F, M, L, I, V</td>
<td>Hydrophobic</td>
</tr>
<tr>
<td>N</td>
<td>Q, N</td>
<td>Amide</td>
</tr>
<tr>
<td>O</td>
<td>S, T</td>
<td>Hydroxyl</td>
</tr>
<tr>
<td>P</td>
<td>P</td>
<td>Proline</td>
</tr>
<tr>
<td>S</td>
<td>A, G</td>
<td>Small</td>
</tr>
</tbody>
</table>

Table 7.1. The eight classes of residue types used in the 4-body potential, chosen to reduce the amino acid alphabet.
The reduced alphabet is used for the non-bonded amino acid to reduce the number of parameters that are required to be fit.

The geometric construction of four-body contacts is shown in Figure 7.1. Three residues form a sequence triplet of a four-body, whose residue types were reduced to eight classes (Table 7.1). The fourth point in the “4-body” is then one of the closest nonbonded alpha carbons to the centroid of the three sequential residues. The non-bonded point retains its specific amino acid type as one of the 20
amino acids. Thus a four body set for side chain-backbone interactions always has three sequential points and one non-sequential in the quartet of interacting residues. All possible four body sets are taken into consideration. This is repeated to derive the non-sequential 4-body interactions, where the 4 residues involved are not close in sequence.

The specific sequence order of the three residues within each backbone triplet is ignored. As a result there are only 120 different triplets instead of $8^3 = 512$. Data is collected by including the fourth point (one of the 20 types of residues), which is within the cutoff distance (8 Å) from the coordinate center (red point in Figure 7.1), and assigning it to one of the corresponding four tetrahedra defined by the vectors originating from the red point to three of the black points. Thus, four body sets comprised of the three sequential residues and a non-sequential nearby residue are obtained. This procedure is then repeated for all quartets defining close interacting residues and for the entire set of proteins.

**Figure 7.1. Identification of residue points for use in the four-body contacts**

The four black points represent the side chain geometric centers of sequential residues $i$, $i + 1$, $i + 2$, and $i + 3$. The red point is the geometric center of the four black points which is chosen as the center of the interacting group. The six planes, defined by each of the six combinations of pairs of black points and the central red point, fully subdivide the space surrounding the red point into four tetrahedra. The blue points represent other residues in close proximity to the red point, within the interaction range of 8.0 Å from it. An example of a set of four contacting residues used for potential extraction is shown by the four residues in boxes.

Each of the residues can be fully or partly exposed to solvent when it is on the surface of the protein or buried inside the protein core. These three situations were considered separately for each type of 4-body potential. Relative solvent accessible surface area (RSA) is used to group the triplet into three groups corresponding to buried (with all three residues in the triplet having RSA < 20%, denoted as Bu), exposed (with all three having RSA < 20%, denoted as E), and intermediate (some of the residues in the
triplet being exposed, and some being buried, denoted as \( I \). Better results were obtained in
discriminating native structures from a large number of decoys by incorporating surface exposure.

### 7.1.5. Four-Body Contact Potential Energy Function

Inverse Boltzmann principle is used to calculate a four-body contact potential energy. First, the
probabilities of \( P_{4|X} \), \( P_{3|X} \) and \( P_A \) are calculated using the equations shown below. Here \( P_{4|X} \) and \( P_{3|X} \) are
respectively the frequencies of quadruplets and triplets in each of the sets specified by \( X \) (= \( Bu \), \( E \), or \( I \))
and \( P_A \) is the frequency of amino acid type singlets in the protein datasets.

\[
P_{4|X} = \frac{\text{number of specific quadruplets given } Bu, E, \text{ or } I \text{ in the data set}}{\text{total number of all types of quadruplets given } Bu, E, \text{ or } I \text{ in the data set}}
\]

\[
P_{3|X} = \frac{\text{number of specific triplets given } Bu, E, \text{ or } I \text{ in the data set}}{\text{total number of all triplets given } Bu, E, \text{ or } I \text{ in the data set}} \quad (7.1)
\]

\[
P_A = \frac{\text{number of specific type of amino acid } A \text{ in the data set}}{\text{total number of all amino acids in the data set}}
\]

Then, the four-body contact potential energy is calculated using the inverse Boltzmann relationship:

\[
E_{4|X} = -RT \ln \left( \frac{P_{4|X}}{P_{3|X} P_A} \right) \quad (7.2)
\]

The total energy for a protein is obtained by summing the four-body contact potential energies over
all quadruplets \( n_q \).

\[
E_{\text{total}} = \sum_{n_q} E_{4|X} \quad (7.3)
\]

This is the equation used to estimate the free energy of native structures and their decoys.

The results are shown in reference (32) (see Figure 3 in reference (32)) where the relative values of
these four-body contact potentials are shown, as a heat map. For these sequential four-body potentials
we require the triplet of amino acids to be sequential, but for the non-sequential four-body potentials this requirement is no longer enforced.

Performance of different knowledge-based potential functions has been compared (219, 231, 232) on large data sets of protein models. They have carried out evaluations by finding the success in the ranking of the native structure as the conformation having the lowest energy, and also by obtaining the average Z-score between the energy of the native structure and the next most favorable structure. The larger the Z-score, the better is the performance.

7.1.6. Performances of different individual potential functions for model ranking

In evaluating the performance of two-body and four-body potential functions in identifying the native (or near native) protein structure CASP8 decoy sets were used. Altogether twenty-three different two-body (see Pokarowski et al. (233) for details) and four-body potentials (both sequential (32) and non-sequential (229)) were used (228).

All knowledge-based coarse-grained potentials based on coordinates of Cα (sometimes Cβ) atoms that are usually designed to capture the statistics of contacts, are tested. For template modeling targets, the BT potential derived by Betancourt and Thirumalai (224) performs best (in terms of correlation coefficients, average Z-score and average RMSD) individually in comparison with other two-body potentials and the two four-body potentials; the best RMSD values being in the range of 4 Å to 5 Å (228). Four-body potentials perform well in the identification of native structures. A few two-body potentials show similar performances with RMSD in the 4 Å range.

The performance for targets from template-free modeling is not as good as that for the homology based targets. However, potentials that perform better for template-free modeling targets also perform better for homology modeling targets but do not yield results that are as good as for the latter. This is due to the fact that the models submitted to CASP8 usually deviate significantly from the native protein structures for the template-free modeling cases, more than for the homology modeling ones, and are usually more poorly packed and/or poorly folded. Therefore empirical potentials which are derived based on real globular protein interactions do not perform well when applied to these cases. Rankings, RMSDs and correlation coefficients all show that both four-body sequential and four-body non-sequential potentials, on average, perform better than or as well as the two-body potentials (228).
7.1.7. Obtaining an optimized potential

The four-body sequential, four-body non-sequential and short-range potentials were combined linearly using a different weight for each potential according to the following formula:

\[ V = w_{4\text{-}body\text{-}seq} V_{4\text{-}body\text{-}seq} + w_{4\text{-}body\text{-}nonseq} V_{4\text{-}body\text{-}nonseq} + w_{SR} V_{SR} \]  

(7.4)

An optimization of the weight for each term was performed to find an optimized potential.

The optimization was carried out using the Particle Swarm Optimization (PSO) technique (234). The weight of the four-body sequential term was set to 1.0 \( (w_{4\text{-}body\text{-}seq} = 1) \) while the weights for the other two terms \( (w_{4\text{-}body\text{-}nonseq} \text{ and } w_{SR}) \) were varied by using the PSO.

For each combination of terms, the average RMSD for the best ranked model and the Z-scores for all CASP8 targets were calculated. CASP8 targets were divided into two subsets according to the method used to generate decoys. One set was comprised of models obtained using homology modeling (153 cases) while the other was obtained from template-free modeling approaches (12 cases).

For the homology modeling targets, the optimized weights obtained for the four-body non-sequential and short-range potentials were 0.28 and 0.22 respectively. For template-free modeling targets, the corresponding weights were different at 1.01 and 0.56.

The native structure rankings obtained for the optimized potential were compared with those obtained for other coarse-grained potentials and for the empirical atomistic potential DFIRE (219). The Decoys 'R' Us dataset (231) was used for the comparison with the atomistic potential. Both single and multiple decoy sets were used. The weights obtained for homology modeled targets were used in assessing the quality of our optimized potential.

7.1.8. Performance of the optimized potential

The resulting combined potential performs better than the two four-body potentials individually and better than all other coarse-grained potentials (with an average RMSD \( \sim 3.7 \) Å for the homology modeling targets using CASP8 decoys), and almost at the same level of performance as the empirical atomistic potentials (using Decoys 'R’Us database). For template-free modeling targets the Betancourt-Thirumalai (224) potentials perform almost as well as the optimized potentials but for homology modeling targets the improvement found for the RMSDs with the optimized potentials is significantly better.
For the misfolded, asilmarh and Pdberr&sgpa data sets from the Decoys 'R'Us database the optimized potentials identify all native structures from these datasets and thereby perform as well as the other empirical atomistic potentials (228) like RAPDF (231), atomic KBP (218) and DFIRE. The native structure ranks and the Z-scores are compared for the above atomistic potentials and for our optimized potentials using multiple decoy sets (228). Optimized potentials are able to predict all native structures in the lattice-ssfit decoy set, and they fail to identify only two native states in the 4-state reduced decoy set. The average Z-score for the optimized potentials for these decoys is 1.87. Multi-body potentials perform well, if the protein structures are large enough, sufficiently compact, and well-packed with many multi-body contacts.

The explosive growth in the number of protein structures (33), presents many new opportunities. The deeper comprehension of the functional role of a protein requires not only the structure but also information about its dynamics. Dynamics information can be extracted directly from the structures if sufficient numbers of structures of the same protein have been determined (8, 235, 236), but these are not so commonly available, and hence, computational approaches are usually used for this purpose. An important lesson from the use of coarse-grained models of proteins is that their motions are dependent on the entire structure, and consequently their entropies should also be dependent on the whole structure. This provides an important new way to extract entropies of protein structures.

7.1.9. Elastic network model (ENM) entropy

The details of ENM are left to Chapter 2.3.2. What follows here is the extension to entropy.

The stiffness matrix describes how resistant to deformation each point in the structure is, within the context of the whole structure and the cooperative interactions within it. In other words, given a deformation with a certain energy \(k_B T\), the model can be used to determine how far each point will be displaced and, for ANM, in what direction. The GNM stiffness (Kirchhoff) matrix \(\Gamma\) is given by equation 2.1. To obtain the mean square fluctuation (MSF) of each point in the structure, the stiffness matrix must be inverted using equation 2.4. However, if \(\Gamma\) is expressed as \(\Gamma = \Gamma_1 + \Gamma_2 = \Gamma_1 - (-\Gamma_2)\) where \(\Gamma_1\) represents the diagonal elements of \(\Gamma\) and \(\Gamma_2\) the off-diagonal ones, then \(\Gamma^{-1}\) can be approximated as a Neumann series (237, 238) as
\[ \Gamma^{-1} = (I + \Gamma_1^{-1}\Gamma_2)\Gamma_1^{-1} \]
\[ = \left[ \sum_{i=0}^{\infty} (\Gamma_1^{-1}\Gamma_2)^i \right] \Gamma_1^{-1} \]
\[ = [(\Gamma_1^{-1}\Gamma_2)^0 + (\Gamma_1^{-1}\Gamma_2)^1 + (\Gamma_1^{-1}\Gamma_2)^2 + \ldots] \Gamma_1^{-1} \]
\[ = [I + \Gamma_1^{-1}\Gamma_2 + \ldots] \Gamma_1^{-1} \]
\[ = \Gamma_1^{-1} - \Gamma_1^{-1}\Gamma_2\Gamma_1^{-1} + (\Gamma_1^{-1}\Gamma_2)^2\Gamma_1^{-1} - (\Gamma_1^{-1}\Gamma_2)^3\Gamma_1^{-1} + \ldots \]

where \( I \) is the identity matrix. A first order approximation is to replace \( \Gamma^{-1} \) with \( \Gamma_1^{-1} \) under the assumption that the \( \Gamma_1^{-1}\Gamma_2 \) terms is small. This corresponds to an assumption that the entropy of each point (atom) is independent but provides a simple way to relate it to the freedom of each point. From its definition (equation 2.1) it is evident that \( \Gamma_1 \) contains the degree (number of edges) of each atom along its main diagonal and zeros elsewhere. The degree is often referred to as the atom’s coordination number, \( z_i \), as it is a count of the closely packed atoms.

The equation for the change in entropy for point \( i \) is computed with the GNM, and the details can be found in reference (9). The basic assumption behind this is that the fluctuations of point \( i \) about its mean position obey the Gaussian distribution

\[ W(\Delta R_i) = \exp\left\{ \frac{-3(\Delta R_i)^2}{2<(\Delta R_i)^2>} \right\} \]

(7.6)

The corresponding change in entropy is:

\[ \Delta S_i = k_B \ln W(\Delta R_i) = \frac{-\gamma(\Delta R_i)^2}{2T[I^{-1}]_{ii}} \]

(7.7)

where \( \Delta R_i \) is the deformation vector representing the changes in positions, \( T \) the temperature, \( k_B \) the Boltzmann constant, and \( \Delta S_i \) the change in entropy for the \( i^{th} \) point. Upon deformation of the structure, the change in entropy of point \( i \) then originates directly from the inverse connectivity of that point (to a first approximation). The deformation is given by the normal mode shape as the deformation vector in equation 2.3, where the amplitude factor, \( A_i \), can be based on a fixed energy for each mode,
the RMSD from the native positions, or by the inverse of the corresponding eigenvalue (the mode’s square frequency).

7.1.10. Energy and entropy changes for different structure pairs of the same protein

For the assumption of first approximation by Neumann series given above, there is perfect energy/entropy compensation as seen in the following equation for the change in entropy \( \Delta S \), and change in energy \( \Delta V \). Beginning with equation 12 we can substitute in the first order approximation \( \Gamma^{-1} = \Gamma_{1}^{-1} \) (equation 7.5).

\[
\Delta S_{i} = \frac{-\gamma(\Delta R)^{2}}{2T \Gamma_{ii}} = \frac{-\gamma(\Delta R)^{2}z_{1}}{2T} = \frac{-\gamma(\Delta R)^{2}z_{1}}{2T} = \frac{-\gamma}{2T} (\Delta R)^{2} \Gamma_{ii} = \frac{-\Delta V_{i}}{T} \quad (7.8)
\]

where the coordination number of point \( i \) is \( z_{1} \). Also, note that the spring constant \( \gamma \) used is arbitrary and usually set to 1, and the change in potential energy is that of a Hooke’s law spring \( \Delta V = \frac{\gamma}{2} \Delta R^{2} \). Since the change in free energy is

\[
\Delta G = \Delta V - T \Delta S \quad (7.9)
\]

the contributions from energy and entropy terms to the free energy change are exactly equal. In Dubois, et. al. (237) we find that the Neumann series approximation will hold if the largest magnitude eigenvalue of \( \Gamma_{1}^{-1} \Gamma_{2} \) is less than 1. For the 10 structures initially considered (see Table 7.2) we find that the largest magnitude eigenvalue is exactly equal to 1 and that the expansion may not converge. For these reasons we will use the pseudo-inverse definition given by equation 2.4.

Our initial exploration of the energy and entropy relationship from the ENMs utilizes five structure pairs from the database molmovedb (239). This is a database of known conformations and computed interpolations between structures that employs a short range energy minimization for each intermediate structure to attain feasible conformations. The protein names, PDB IDs, and the number of residues are given in Table 7.2. The changes in potential energy for each point \( \Delta V_{i} \) are calculated by assuming that each point moves independently in a given normal mode. That is, we initially calculate the amplitude that will deform the structure by \( k_{B}T \) energy, i.e., thermal energy. A temperature of 300 K was used for all calculations. From this deformation, we then calculate the potential energy change for each
atom assuming all others remain fixed. We calculate the deformations based on the entire structure experiencing the thermal background energy. The change in entropy for each point in the ENM was calculated using equation 7.7 with the pseudo-inverse defined in equation 2.4.

<table>
<thead>
<tr>
<th>Structure Pair</th>
<th>PDB 1</th>
<th>PDB 2</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biotin Carboxylase</td>
<td>1DV1A</td>
<td>1DV2A</td>
<td>433</td>
</tr>
<tr>
<td>ATP Sulfurlase</td>
<td>1I2DA</td>
<td>1M8PA</td>
<td>573</td>
</tr>
<tr>
<td>Elongation Factor 2</td>
<td>1N0VC</td>
<td>1N0UA</td>
<td>842</td>
</tr>
<tr>
<td>Adenylate Kinase</td>
<td>1AKEA</td>
<td>4AKEA</td>
<td>214</td>
</tr>
<tr>
<td>Acetyl-CoA-synthase*</td>
<td>ACS1</td>
<td>ACS11</td>
<td>728</td>
</tr>
</tbody>
</table>

**Table 7.2. Five structure pairs from the molmovedb database (239) used in this study.**

Four letter PDB IDs are given as well as the chain ID from the structure.

*For this structural transition, no PDB files were listed. Instead, we chose the first and last frame (1 and 11) from the morph. \( N \) is the number of matching residues between the two structures.

<table>
<thead>
<tr>
<th>PDB</th>
<th>( \sum_{i=1}^{3} -T \Delta S_{i}^{f} )</th>
<th>( \sum_{i=1}^{3} \Delta V_{i}^{f} )</th>
<th>%S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1DV1A</td>
<td>0.202</td>
<td>0.350</td>
<td>41</td>
</tr>
<tr>
<td>1DV2A</td>
<td>0.203</td>
<td>0.250</td>
<td>46</td>
</tr>
<tr>
<td>1I2DA</td>
<td>0.233</td>
<td>0.780</td>
<td>24</td>
</tr>
<tr>
<td>1M8PA</td>
<td>0.248</td>
<td>1.060</td>
<td>23</td>
</tr>
<tr>
<td>1N0UA</td>
<td>0.243</td>
<td>0.840</td>
<td>23</td>
</tr>
<tr>
<td>1N0VC</td>
<td>0.243</td>
<td>0.840</td>
<td>30</td>
</tr>
<tr>
<td>1AKEA</td>
<td>0.155</td>
<td>0.290</td>
<td>35</td>
</tr>
<tr>
<td>4AKEA</td>
<td>0.187</td>
<td>0.550</td>
<td>26</td>
</tr>
<tr>
<td>ACS1</td>
<td>0.226</td>
<td>0.390</td>
<td>36</td>
</tr>
<tr>
<td>ACS11</td>
<td>0.210</td>
<td>0.270</td>
<td>45</td>
</tr>
</tbody>
</table>

**Table 7.3. Total change in entropy and energy upon deformation by the first three normal modes for the five structure pairs.**

The effect of deforming the structure by the first three modes is determined for each residue and summed. The rightmost column is the average percent of the total change in the entropic component of the free energy. All modes are assigned \( k_{B} T \) total energy for their deformation.

We calculate the effect of deforming the structures based on the three lowest frequency normal modes, and the results are summarized in Table 7.3. For each normal mode used, we deformed the structure so that the total change in potential energy is \( k_{B} T \). The mean change in potential energy is estimated here with the GNM, which does not have directions of motion, so this may not be so precise and may overestimate the energetic contribution to \( \Delta G^{*} \) (which is defined in equation 7.9). Despite this, the entropic changes still amount to 20-50% of the free energy changes.
Figure 7.2 compares the change in energy (A) and entropy (B) upon deformation by the first three normal modes for the ATP sulfurylase structure 1I2DA (structure shown in panel C). We show the entropic change on the structure corresponding to mode 1. Coloring is spectrally, from red, to yellow, to green, and finally to blue. The red side of the spectrum corresponds to zero and blue to the largest change in entropy. The structure 1M8PA is shown in gray. This structure pair is the most similar of the five. Parts D, E, and F of Figure 7.2 are similar to A-C, but for the elongation factor 2 structures and panel F shows the structure pair 1N0VC. This structure pair has the largest total RMSD of the five pairs. In both examples, we find that the part of the structure that requires the largest fluctuations to attain the other conformation also has the largest change in entropy in the first mode of motion.

In Table 7.4 we show the differences in energy and entropy between the structure pairs for the slowest global mode of motion. Differences in entropy and free energy changes between the structure pairs do not appear to relate closely to either the sizes or the RMSD changes (shown computed in two
different ways). This is possibly because in this exploration we are considering only the single difference between two structures whereas each of the individual structures should be more properly represented as an ensemble of conformers. The background energy here is simply taken as $k_b T$ at $T=300$, or about 0.6 kcal/mol. In calculating the amplitude of deformation we have scaled the total deformation to be $k_b T$ in one normal mode. Normal modes represent the natural vibrational frequencies of the structure.

The background energy here is simply taken as $k_b T$ at $T=300$, or about 0.6 kcal/mol. In calculating the amplitude of deformation we have scaled the total deformation to be $k_b T$ in one normal mode. Normal modes represent the natural vibrational frequencies of the structure.

<table>
<thead>
<tr>
<th>Structure Pair</th>
<th>$\Delta V$</th>
<th>$-TAS$</th>
<th>$\Delta G$</th>
<th>$\text{RMSD}_{CE}$</th>
<th>$\text{RMSD}_{total}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biotin Carboxylase</td>
<td>0.087</td>
<td>0.003</td>
<td>0.165</td>
<td>4.17</td>
<td>11.48</td>
</tr>
<tr>
<td>ATP Sulfurylase</td>
<td>-0.253</td>
<td>0.011</td>
<td>0.271</td>
<td>2.63</td>
<td>6.41</td>
</tr>
<tr>
<td>Elongation Factor 2</td>
<td>0.226</td>
<td>-0.034</td>
<td>0.267</td>
<td>2.49</td>
<td>23.61</td>
</tr>
<tr>
<td>Adenylate Kinase</td>
<td>-0.188</td>
<td>0.034</td>
<td>0.256</td>
<td>3.57</td>
<td>9.35</td>
</tr>
<tr>
<td>Acetyl-CoA-Synthase</td>
<td>0.093</td>
<td>-0.034</td>
<td>0.339</td>
<td>5.14</td>
<td>8.06</td>
</tr>
</tbody>
</table>

Table 7.4. The mean difference in computed energy ($V$) by ENM, entropy ($S$), and free energy ($G$) for the five structure pairs.

Absolute differences are also listed since positive and negative differences can mask the presence of large total difference. Entropy is calculated using equation 7.7 with the pseudo-inverse defined in equation 2.4. Equation 7.9 gives the change in free energy. $\text{RMSD}_{CE}$ is the root mean square deviation returned by the CE algorithm for superimposing the two structures. It is thus the RMSD of the aligned section. $\text{RMSD}_{total}$ is the RMSD of all matching alpha carbons. In this case, matching is for the residues chosen for comparison. For example, the two biotin carboxylase structures have 433 and 450 residues. We choose for comparison the 433 residues that most closely match after alignment with CE.

In Tables 7.3 and 7.4 we report the average change in entropy across all residues in the structure, whereas in Table 7.5 we calculate the total entropic contribution to free energy. Combined with the average change in contact potential energy, average changes in residue entropy provides insight into the relative contribution of energy and entropy for each deformation of each structure. Within the ENM framework, we can see from Table 7.3 that the entropic contribution to free energy is significant. Thus, the development of better methods to take into account entropic as well as energetic contributions is needed. In Table 7.5 we present a first look at combining contact potentials for energetic evaluation with entropy calculations from ENMs. The entropy values compared here are difference in the total free energy change of entropic origin if the structure was influenced by all three of the lowest frequency normal modes. That is, we first compute

$$\Delta S_T = \sum_{k=1}^{3} \sum_{i=1}^{N} \Delta S_{kl}$$  (7.10)
for each of the structure pairs, where \( k \) indexes the mode and \( i \) the residue. The difference in \( \Delta S_T \) between the structure pairs is reported and indicates the difference in total entropy change upon excitation of the first three modes.

<table>
<thead>
<tr>
<th>Structure and PDB IDs</th>
<th>RMSD_{CE}</th>
<th>RMSD_{Total}</th>
<th>( \Delta E )</th>
<th>-T( \Delta S )</th>
<th>( \Delta G )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biotin Carboxylase (1DV1A-1DV2A)</td>
<td>4.2</td>
<td>11.5</td>
<td>-14.6</td>
<td>-2.8</td>
<td>-17.4</td>
</tr>
<tr>
<td>ATP sulfurylase (1I2DA-1M8PA)</td>
<td>2.6</td>
<td>6.4</td>
<td>-12.4</td>
<td>-6.3</td>
<td>-18.7</td>
</tr>
<tr>
<td>Elongation Factor 2 (1N0UA-1N0VC)</td>
<td>2.5</td>
<td>23.6</td>
<td>-15.1</td>
<td>26.9</td>
<td>11.8</td>
</tr>
<tr>
<td>Adenylate Kinase (1AKEA-4AKEA)</td>
<td>3.6</td>
<td>9.4</td>
<td>5.5</td>
<td>-7.2</td>
<td>-1.7</td>
</tr>
<tr>
<td>Acetyl-CoA-Synthase (ACS1-ACS11)</td>
<td>5.1</td>
<td>8.1</td>
<td>-8.2</td>
<td>24.6</td>
<td>16.4</td>
</tr>
</tbody>
</table>

Table 7.5. Free energy changes by combining the optimized four body energies with the entropies estimated from the elastic network model.

Inspection of the results in Table 7.5 show some interesting features. In the case of biotin carboxylase the structures are with and without ATP bound. The contact potentials show a lower energy for the ATP bound form and also that the ATP bound form has a negative entropic contribution. Likewise for ATP sulfurylase the T state has both lower energy and lower entropy than the R state. For elongation factor 2, the inhibitor bound form is favored energetically but not entropically, presumably because of tight binding that reduces conformational freedom. The two adenylate kinase structures are open and closed. The energy favors flap closing but entropy disfavors closing. However, the net free energy difference between open and closed forms is quite small, perhaps pointing to the relative ease of this structural transition. For acetyl co-A synthase the transition is favorable energetically but is overall unfavorable because of a large unfavorable entropy change.

The ENMs are entropic models that depend on the shape and packing density of the modeled structures. They can be used to extract entropic changes for the various deformations of the structure. The energies used for the exploration in Table 7.4 are not fully reliable. The entropies computed here can, however, be combined more reliably with energies from pair-wise, 4-body (32), or distance dependent potentials (240) to provide a better understanding of protein structural energetics and their dynamics. This has been done in Table 7.5. Also, it would be possible to generate ENM models using all
atoms, this approach could allow comparison and evaluation with commonly used atomic force fields such as AMBER or CHARMM to increase the accuracy, specificity, and breadth of the free energy analyses.

The inclusion of $\Delta S$ in the calculation of $\Delta G$ is used to improve two aspects of structure prediction. First, the more standard metric of structure prediction is compared in Figure 7.3 using the CASP9 dataset, where the target sequence is modeled as a monomer and we use $\Delta S$ alone as a classifier for decoys being native-like. Testing numerous aspects of the $\Delta S$ profiles, we find that a local character index (LC) similar in concept to the metric used by Brooks, et. al. (241) and Lu and Ma (242) provided the best classification based on the number of times we find the decoy with lowest RMSD to the target in the top 10 ranked decoys. The LC index used here is computed as $LC = \Sigma (\Delta S^{1/p})$, where $p$ is the LC parameter. In the following, $\Delta G$ is always normalized by the number of residues in the decoy.

![Figure 7.3: Performance of classifying decoy structures by Local Character of $\Delta S$.](image)

The local character (LC) index is conceptually similar to the metric used in (241, 242), see text and Supplemental Figure 1 for further detail. The percent of targets for which the lowest RMSD decoy appears in the top 10, 50, or 100 is shown. For a parameter value of 4, the LC index has nearly converged to its limit of classification power. From this data, it is evident that $\Delta S$ alone is capable of ranking decoys in a meaningful way.
Figure 7.4: Result of classifying CASP9 monomer decoys using vibrational entropy, energy, and a free energy.

A) The 45 targets that have the lowest \( \Delta G \) calculated using only \( \Delta S \) are selected. We plot the RMSD to the native structure of the decoy with lowest \( \Delta G \) as calculated using only \( \Delta S \) (•), only \( \Delta E \) (◊), and a combination (-) using \( \delta = 1 \) in equation 7.11. For these targets, \( \Delta S \) outperforms \( \Delta E \) yielding a mean decoy RMSD100 of 5.2Å and 7.9Å, respectively. The combined method does not significantly outperform the 4-body potential for these targets. 

B) A Similar plot, but now the 26 targets with the best classification using only \( \Delta E \) are shown. For these structures, the 4-body potential significantly outperforms the entropy and combined classification, with mean RMSD100 of the lowest RMSD decoy of 2.1Å, 14.5Å and 4.0Å, respectively. Other interesting patterns emerge for different subsets; see text for details.

In Figure 7.4 a comparison between \( \Delta E \) (from 4-body potentials), \( \Delta S \), and \( \Delta G \) for classification of decoys is made. For clarity, we denote changes in potential energy from the ENM as \( \Delta V \) and energetic changes calculated from the 4-body potentials as \( \Delta E \). An important initial point is the lack of convergence in decoy prediction as judged by the range of RMSD100 values in Supplemental Figure 2. The RMSD100 is a normalized RMSD proposed in Ref. (243) and is interpreted as the RMSD between two 100 residue proteins of equivalent similarity to the structure pair in question. Some targets have no decoys within 5 Å RMSD100. Utilizing the energy or entropy alone, we achieve an average classification accuracy of 10.5 Å and 17.1 Å, respectively, across all targets. By combining the two metrics into one classifier, the average RMSD100 remains at 10.5 Å. This is due to increased performance for some targets, but decreased performance for others (Figure 7.4), showing that the entropic contribution to \( \Delta G \) upon mode motions is meaningful for structure prediction, but our present treatment is likely still too simple. In this case, we observe in Figure 7.4 that \( \Delta S \) performs poorly for some targets, but significantly outperforms \( \Delta E \) for others. These tests were performed on 110 out of the 129 published CASP9 targets. We report RMSD100 values, but the CASP typically uses a Z-score describing the significance of the structure alignment. The Z-score is more lenient for unpredicted parts of a structure. For instance, most
of the RMSD100 values that are over 50 Å are due to the decoy containing a long (mostly linear) terminus. These extended termini are residues that were not significantly modeled, but were retained in the deposited prediction so that the sequence would match that of the target. Their inclusion in RMSD-based comparisons may be misleading. See Supplemental Figure 4 for a listing of the RMSD for the best decoy for each target. The dataset was limited by the resolution of the resolved experimental structure, presence of oligomeric state prediction (see below), and size of the target sequence.

Across most subsets considered, there were no size or “presence of good decoy” effects seen. The latter refers to the case where a target has no decoy within 5 Å RMSD. An interesting exception is when the targets are limited to those where the ΔG of entropic origin is below a threshold – then there is a strong correlation between the presence of native-like decoys and the computed ΔG. For instance, if the threshold is set to -1, the correlation of the RMSD of the most native-like decoy with either ΔE or ΔS is 0.69 and 0.53, respectively. Such a metric might be useful to determine whether a set of decoys contains any native-like predictions and if so, how native-like they are likely to be.

We also perform a test using a newly available metric – oligomeric state prediction. In the most recent CASP competition, contributors had the option of predicting not only the structure of target sequences in monomer form, but also of the higher order oligomers. In our dataset there are a total of 65 monomers, 33 dimers, 4 trimers, and 8 tetramers, as determined by the experimentally determined structure corresponding to the CASP9 target sequence. Correspondence between the target and decoy structures was obtained using the Smith-Waterman (244) local alignment algorithm and a low gap opening penalty to allow for unresolved loops in the crystal structure or parts of the sequence not modeled by the decoy. Following sequence alignment, a structure alignment between reasonable pairs was performed and normalized using the RMSD100 metric of Carugo and colleagues (243). For all combinations of ΔS and ΔE, the monomer predictions are correctly identified because these structures were never predicted to occupy higher oligomeric states. Using only energetic contributions to ΔG, we obtain an average classification index of 2.0, 5.3, and 1.0 for dimers, trimers, and tetramers, respectively. If the entropic contribution is used alone, the average classification index moves to 3.0, 4.5, and 7.1 – worse for even numbered oligomers. If the two contributions are combined in equal proportion, an improved classification index of 1.9, 5.3, and 1.0 is obtained. Scale values (δ) from 0 to 100 in increments of 0.1 were tested which relate the contribution to ΔG of ΔE and ΔS by:

$$\Delta G = \Delta E - \delta \left| \frac{\Delta E}{\Delta S} \right| kT \Delta S$$  \hspace{1cm} (7.11)
The prediction values listed previously were for $\delta=1$ and are quite similar for $\delta \leq 1.8$. However, for $\delta > 1.8$, there is improved accuracy for trimer predictions (to 4.5Å), and decreased performance for even numbered oligomeric states. Interestingly, these minor gains in monomer and oligomeric state prediction point to the possibility of improving evaluations of protein thermodynamics by utilizing information from the entire structure.

There are a number of aspects of $\Delta G$ that are not yet accounted for. For instance, in calculating $\Delta E$, the theory applied here assumes that the contribution is a change in energy from a disordered state to the folded state. However, the entropic contribution to free energy is computed as the change in entropy upon excitation of the normal modes. A more accurate model for the inclusion of the entropic penalty for folding (entropy of the denatured state) might yield a $\Delta S$ that is more comparable to the $\Delta E$ computed by the 4-body potentials and improve oligomeric state prediction and decoy selection in the future. A further consideration is the change in molecular volume. For heat pumps, engines, and many other objects the change in internal energy is often calculated in a way that takes into account the work done by changes in pressure and volume within the system. This aspect of entropy is usually ignored for protein systems, but must be a contributor as the compactness (volume) of a protein changes upon folding. Also, a protein structure exists as an ensemble of structure sampling from the feasible motions. As the structure samples conformations it also may change the molecular volume, which could be another contributing factor.

The proposed integration of statistical contact potentials with elastic network models of proteins will potentially improve coarse-grained modeling of protein structure and dynamics. The consideration of protein flexibility and its fluctuation dynamics should improve protein structure prediction, and should lead to a better refinement of computational models of proteins, demonstrated here to improve the selection of native-like decoys.

Supplementary Data available online at:

http://www.springerlink.com/content/503728q621w25r14/
7.2. Incorporating mean square fluctuation as entropy

In the previous section we worked entirely with one type of entropy – vibrational entropy. While there are theoretical and practical reasons for considering this type of entropy, it is not the only contributor. Entropy can be thought of as the redistribution and spreading out of energy into a set of microstates (granular conformations). This makes a straight-forward relation (at least conceptually) between entropy and the amount of volume swept out by equilibrium motions – less motion leads to fewer states. Using mean square fluctuations calculated using anisotropic network models we find improvement over the 4-body energy alone, again using the CASP9 decoy sets (Table 7.6). These preliminary results are encouraging and point to the large gains that are possible to achieve from adding entropic considerations to structural studies.

<table>
<thead>
<tr>
<th>-TΔS</th>
<th>ΔE</th>
<th>ΔE - 0.5TΔS</th>
<th>ΔE - TΔS</th>
<th>ΔE - 1.5TΔS</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSD to Target of Lowest ΔG Decoy</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>23.4</td>
<td>13.4</td>
<td>13.4</td>
<td>12.2</td>
</tr>
<tr>
<td>Median</td>
<td>20.3</td>
<td>11.4</td>
<td>11.4</td>
<td>9.7</td>
</tr>
<tr>
<td>Mode</td>
<td>3.5</td>
<td>1.7</td>
<td>1.7</td>
<td>0.8</td>
</tr>
<tr>
<td>#Decoys with Lower RMSD than Lowest ΔG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>393</td>
<td>249</td>
<td>249</td>
<td>189</td>
</tr>
<tr>
<td>Median</td>
<td>371</td>
<td>235</td>
<td>235</td>
<td>151</td>
</tr>
<tr>
<td>Mode</td>
<td>500</td>
<td>179</td>
<td>179</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 7.6: Decoy discrimination is significantly improved compared to 4-body energy alone by incorporating fluctuation-based entropy.

To generate the fluctuation-based entropy, we compute mean squared fluctuation (MSF) using the Gaussian Network Model. This differs from our previous computations that relied on the theory of vibrational entropy. MSF is then used as an estimate of the entropy change for each residue. The relative contribution of energy and entropy is unknown. We show here the result of mean centering the MSF distribution on the ΔE value, 50% less or 50% greater contribution to ΔG, compared to ΔE. Optimal discrimination is found when ΔS and ΔE have equal contribution to ΔG.
7.3. Discriminating docking poses with computed free-energies

We apply a similar methodology for selecting computational docking poses that better reflect those observed in experimental structures. Candidate poses are generated using ClusPro (245) and Zdock (246). ClusPro data consists of the models returned, and is therefore already a distillation of the raw poses. Thousands of initial poses are generated by ClusPro which are then clustered and ranked by energy functions. The centroid of each cluster is then returned to the user (10 to 30 models for each set). We aim to improve the ranking of these refined poses compared to the ClusPro ranking. Our initial results use two datasets and were generated using an electrostatically favored docking mode. These ClusPro docking jobs were submitted to the ClusPro server by Usha Muppirala of Iowa State University who then distributed the results. The energy of each docked pose is evaluated with the 4-body potential of Feng, et. al., while the entropy is evaluated as above using our MSF approach. We use energy, entropy, and a combination of both to judge each of the binding poses and find that the combined method significantly outperforms either of the marginal metrics. See Figures 7.6 and 7.7 for more details. In Figure 7.6 we have used ClusPro results where no constraints regarding which residues are
involved in the binding interface were given. For Figure 7.7, constraints from the X-ray crystal contacts (list of residue pairs that are close in space) were provided to ClusPro.

A more thorough test of the discriminating power of our method is to apply it to unprocessed poses. For this test we generate 2000 poses using Zdock and attempt to rank them using our combined method. These tests are at a preliminary stage. PDB structure 1BKD is our first test. It contains a heterodimer (part of a hetero-octamer, but we consider only one pose of the two possible) that is split into a “receptor” and “ligand.” We arbitrarily call the larger chain the receptor. Using our combined method, we deem the 4th lowest L-RMSD to be the lowest energy. L-RMSD is defined as the RMSD of the ligand to the X-ray structure after the receptor has been aligned. This is a significant improvement over the 4-body potential’s performance which picks the 249th lowest L-RMSD structure to be the lowest energy pose.

Figure 7.6: Selecting docked forms using computed free energies with mean square fluctuation based entropy improves selection of docking poses in unconstrained docking, showing advantages compared to energy alone. ClusPro returns 10-30 docking poses (each is a cluster centroid from a total pool of 1000 poses) which are ranked by energy and our computational measure of entropy. The ability for entropy to improve selection over energy alone is evident.
Figure 7.7: Using mean square fluctuation based entropy improves selection of docking poses in constrained docking.

Pairs of residues across the X-ray crystal structure interface that are close in space are provided to ClusPro. ClusPro returns 10-30 docking poses (each is a cluster centroid from a total pool of 1000 poses) that agree with the provided constraints. They are ranked by energy and our computational measure of entropy. In the case of constrained docking we show even better performance than unconstrained. This further supports the initial observation that our combined method performs better when the dataset includes more accurate models.
CHAPTER 8. MAVENs: Motion analysis and visualization of elastic networks and structural ensembles


8.1. Abstract

The ability to generate, visualize, and analyze motions of biomolecules has made a significant impact upon modern biology. Molecular Dynamics has gained substantial use, but remains computationally demanding and difficult to setup for many biologists. Elastic network models (ENMs) are an alternative and have been shown to generate the dominant equilibrium motions of biomolecules quickly and efficiently. These dominant motions have been shown to be functionally relevant and also to indicate the likely direction of conformational changes. Most structures have a small number of dominant motions. Comparing computed motions to the structure’s conformational ensemble derived from a collection of static structures or frames from an MD trajectory is an important way to understand functional motions as well as evaluate the models. Modes of motion computed from ENMs can be visualized to gain functional and mechanistic understanding and to compute useful quantities such as average positional fluctuations, internal distance changes, collectiveness of motions, and directional correlations within the structure.

Our new software, MAVEN, aims to bring ENMs and their analysis to a broader audience by integrating methods for their generation and analysis into a user friendly environment that automates many of the steps. Models can be constructed from raw PDB files or density maps, using all available atomic coordinates or by employing various coarse-graining procedures. Visualization can be performed either with our software or exported to molecular viewers. Mixed resolution models allow one to study atomic effects on the system while retaining much of the computational speed of the coarse-grained ENMs. Analysis options are available to further aid the user in understanding the computed motions and their importance for its function.

MAVEN has been developed to simplify ENM generation, allow for diverse models to be used, and facilitate useful analyses, all on the same platform. This represents an integrated approach that incorporates all four levels of the modeling process - generation, evaluation, analysis, visualization - and
also brings to bear multiple ENM types. The intension is to provide a versatile modular suite of programs to a broader audience. MAVEN is available for download at http://maven.sourceforge.net.

8.2. Introduction

One of the first dynamic computations of a protein was published by Levitt and Warshel in 1975 (247), a folding a coarse-grained peptide chain. The first publication widely recognized as a Molecular Dynamics (MD) simulation came two years later (248) and presented a simulation of the 56 residue bovine pancreatic trypsin inhibitor. Today’s most advanced simulations, such as the fully atomic model simulation of the ribosome (135), (2.6 million atoms for a period of $10^6$ CPU hours) represent significant improvements in computational technology and advance our understanding of the behavior of molecular systems.

There is growing evidence supporting the collectiveness of the motions of biomolecules (23-27). Atomic MD takes into account the detailed modeling of the intrinsic randomness of the short time-scale motions. While true to the underlying atomic theories, this level of randomness may be distracting when analyzing motions of biomolecules on the longer time scale. To overcome such randomness, methods have been developed to determine the dominant motions within the trajectory - Principal Component Analysis (PCA) based on average covariances across the trajectory or Essential Dynamics (13). The results of PCA have been shown to agree well with ANM modes (8, 14). The rigor of molecular dynamics simulations makes them ideal for investigating specific events, but less applicable for extracting mechanisms, and overly detailed for general purposes.

Normal Mode Analysis (NMA) using Elastic Network Models (ENMs) shifts the focus from simulating the motion of all atoms based on a detailed empirical force field to the harmonic motions of a set of springs and masses representing the starting structure. ENM is also an analytic solution yielding a basis set of orthogonal independent motions, rather than a simulation over time. The differences between various ENM types specify how the masses (computed from atomic coordinates or density contours) and springs (their harmonic interactions) are assigned or modeled; the exact chemistry of the object is not usually considered. Despite their lack of detailed chemistry, ENMs have proven themselves. Yang, et al. (8) as well as Bakan and Bahar (14) find that the motions computed using ENMs correspond well to the principal components of MD trajectories as well as to the spatial variance observed when multiple crystallographically determined structures of the same protein are superimposed or taken from an NMR ensemble. Sen et al. (9) emphasized the cooperativity of protein motions and how atomic detail is not
necessary, while Yang, et. al. (10) shows the ability of elastic models to handle proteins across a broad range of sizes. From Jernigan, et. al. (11) we learn that functionally important motions of biomolecules are often governed by packing density: the basis for ENMs. Lu and Ma show that shape also plays an important role (249). Early studies by Gō and Scheraga computationally defined the difference between local and collective motions (250), while Gō, Noguti, and Nishikawa demonstrated how the low frequency harmonic vibrations in proteins can be computed (251). From these and other studies, it is becoming increasingly evident that molecular structures exhibit collective motions and that these motions can be sampled well by using ENMs. Methods that include distance dependent springs (43, 50, 252) or torsional angles (45-47) show improvements over the uniform springs and are also implemented in our software.

ENMs are capable of computing the important motions of biomolecules on time scales beyond the usual reach of atomic MD, and do so efficiently. Generation for small and medium sized proteins takes only seconds or minutes on a standard desktop computer, while typical MD studies require days or weeks of computer time on high performance systems or clusters. The largest molecular assemblages may require more time, but they can be further coarse-grained without loss of the major motions (20), once again making the computation of the dominant motions tractable. It is also possible to more efficiently solve numerically for only a subset of the normal modes. The low frequency modes contribute much more to the total motion than do the high frequency modes. Thus, many of the high frequency modes can be ignored without loss of important information. Detailed analyses can be achieved with elastic models by employing mixed-resolution models where most of the structure is coarse-grained but regions of special interest remain at atomic detail. Even with the proven efficiencies GPU computing brings to MD, the simplicity and effectiveness of NMA using ENMs will ensure their continued use. We aim here to make them more accessible.

8.3. Implementation

Here we present a platform for Motion Analysis and Visualization of Elastic Networks and Structural Ensembles (MAVEN), licensed under the Lesser General Public License so that MAVEN is freely available. Figure 8.1 provides a brief overview of the MAVEN interface and some of its features. It is built using MATLAB ® (2010a, The MathWorks, Natick, MA) and compiled into a standalone application, which can be run by the MATLAB Component Runtime (MCR). The MCR is freely provided from MathWorks and is distributed with MAVEN. Thus, MATLAB is not needed to run our application. The source code contains
numerous self-documented MATLAB functions, which can be used without modification to extend other user’s programs. Certain functions have been written in Perl or C++ to improve performance. MAVEN, its source code, MCR, User’s Guide, and video tutorials are available on our website.

Figure 8.1: MAVEN interface overview and examples of model generation features.
A) Screenshot of MAVEN having completed an ANM model of the HIV-1 protease 1T3R using 198 alpha carbons and $d_{ij}^2$ weighted springs. Motilities of each residue are shown: B-factors from the PDB file (blue) and mean squared fluctuations computed from the ENM (red). The correlation coefficient for these two curves is 0.70. B) Selecting by atom type within the Prepare Files module is shown with all atoms from 1T3R displayed. Methods to generate other coarse grained systems are provided and explained in the User’s Manual including the average sugar and base position in nucleotides or C) residue or side chain centroid positions. D) Mixed resolution systems are also supported. We show the protease inhibitor in red and surrounding atoms as blue sticks with the rest of the structure coarse-grained to Cα atoms shown as green spheres. E) MAVEN has the unique ability to convert electron density maps into coarse-grained points. EMDB structure 1800 is shown as a density contour followed by an approximate surface filled with spherically coarse-grained points. All of these examples are given in greater detail in the User’s Manual.
8.4. Elastic Network Model

Details of the elastic network model are given in Chapter 2.3.2. MAVEN implements all model types discussed and aids the user in model generation and analysis. For further details, see the User’s Guide available in supplementary material online at:

http://www.biomedcentral.com/1471-2105/12/264

8.5. Comparing ENMs to Ensembles of Structures

Principal Component Analysis of a set of structures can be used to describe the functional ensemble of the biomolecule. This is performed in MAVEN by computing the covariance matrix for atomic positions within the ensemble and applying singular value decomposition to it. The result is a set of vectors that capture the ensemble variance as efficiently as possible; the first PC explains the most variance possible by a linear approximation, the second as much of the remaining variance as possible, and so on (Figure 8.2E). These vectors can be compared to the normal modes from an ANM model using equations 2.8, 2.9, and 2.10 which describe the Overlap between the \( i^{th} \) PC \( (P_i) \) and the \( j^{th} \) mode \( (M_j) \), Cumulative Overlap between the first \( k \) normal modes and the \( i^{th} \) PC (Figure 8.2F), and overlap between the space spanned by first \( I \) PCs and the first \( J \) low frequency modes (their Root Mean Square Inner Product), respectively, and are further described by Tama and Sanejouand (54) and Leo-Macias, et. al. (55).

8.6. Model Generation

Perhaps the most important step in ENM modeling is choosing which points will represent the system. Originally, ENMs were proposed for use with all atom coordinates (34), but when it was learned that nearly the same motions are obtained with coarse-grained structures (9, 20), this became the more common approach. Within MAVEN, one may retain all atoms, choose standard representations like \( C^a \) atoms, pick specific atom types, or generate centroid points from residues, side-chains, or bases (for examples see Figure 8.1). A set of points can be further coarse-grained using spherical coarse-graining. This task is accomplished by selecting an initial point (or set of points) that will be retained. All points within a given radius of the retained point will be removed from the model. The closest point that was not removed is then added to the retained set. This process continues until no more points can be removed. The result is a spatially more uniform distribution of points than one would likely have after selecting points linearly along the sequence. Finally, MAVEN allows the user to employ low resolution
density maps (Figure 8.1E); a data source that has rarely been used with these methods. Since packing density and shape are the properties most critical to ENMs, model points picked along the desired density contour should provide a reasonable coarse-graining. Doruker and Jernigan have shown that similar motions are extracted from proteins and from the protein's molecular volume filled with lattice points (253), further showing the potential usefulness of density maps in ENM modeling. This method represents an extension to understand the dynamics of very large structures where there are no atomic coordinates.

**Figure 8.2: Analysis features of MAVEN.**  
A) Within MAVEN, we plot the anisotropic displacement tensors from PDB file 1T3R in red and computed from $d_{ij}^2$ weighted ANM in blue. B) How strongly the directions of motion of eight parts of the protease structure are correlated with one another in the first mode of motion is summarized. C) We display three frames in a top-down view of the animation of mode 1 that was exported to PyMOL: the negative mode direction (left), the initial structure (center), and the positive mode deformation (right). Coloring and tube thickness is by B-factor. D) Visualization of the NMR ensemble described in PDB file 2KTD in PyMOL. Coloring is blue to red from the N to C-terminus. E) Principal Component Analysis of the ensemble is performed and the variance in each PC and the cumulative variance plotted. F) Computation of the cumulative overlap between a set of low frequency modes generated from the first member of the ensemble and the first 3 PCs, using equation 2.9. MAVEN always displays a heatmap, but only displays text for significant relationships (CO > 0.5).
8.7. Discussion

Because of the usefulness of the ENM method, web and standalone applications have been constructed (46, 254-258). Web servers have the advantage of near universal accessibility, but often lack flexibility and extensibility. Existing standalone applications tend to only implement one ENM type and often force the user to use one representation, for instance alpha carbons only, thereby preventing the use of nucleotides, sugars, or small molecules. In developing MAVEN we seek to incorporate many ENM methods including support for dihedral angle and mixed resolution modeling, as well as to facilitate model generation and shape based coarse-graining (see Methods).

The first major feature of this platform is the ability to construct many types of ENMs whereas other servers and applications available are restricted to one or two types. These include the standard cutoff based models, distance dependent springs, nearest neighbor, Spring Tensor (45), and mixed resolution. The nearest neighbor method generates a coarse-grained model, but uses an atomic model for determining connectivity. The Spring Tensor model expands the energy function of ENMs to account for bond and torsion angle changes. Mixed coarse-graining represents a compromise: part of the system remains in high detail with the remainder coarse-grained. With mixed resolution, one is able to analyze molecular effects on motions such as chemical modifications, mutations, drug binding, proline isomerization, or post-translational modifications, while retaining nearly the computational efficiency of coarse-grained NMA. A second feature of this application is the ability to handle large systems through sparse matrix methods and the ability to calculate only a set of the lowest frequency modes. Since the contribution of each mode to the total motion decreases quickly, calculating only the lowest frequency modes captures the majority of dynamics while requiring considerably less computer resources. A further benefit of MAVEN is that it is setup to accept protein, RNA, DNA, and small molecule coordinates. From an unprocessed PDB file, one can generate a standard alpha carbon model, our atom selector can be used to save a subset of atom types for use in any ENM, points can be picked from electron density contours, united atoms representing the centroid of a set of atoms can be generated, or one may compose or edit an initial model using other software (such as a molecular viewer) and use MAVEN for ENM generation and analysis. See Figure 8.1 for examples of these model types.

Multiple analysis features are presently included. Selected methods are shown in Figure 8.2. These include the ability to analyze Principal Components constructed from multiple static structures, an NMR ensemble, or frames from an MD trajectory and compare them to the normal modes. Multiple studies including (8) and (14) have shown that the variance seen in ensembles of structures derived from MD
trajectories, NMR, or X-ray crystallography can be reproduced with ENMs. This represents an important method for ENM model validation and further exploration of functional motions. MAVEN also has the ability to compare two ENMs of different types or having different parameter choices, and to analyze the effect of the mode-motions on subsets of the structure, comparing within or across subsets. For a full list of our analysis features and examples of their use, please consult our user’s guide (Additional File 1). Future additions are likely to include automated methods for batch model generation and comparison, spectral analysis, or Markov Propagation Model simulations which probe paths of information transfer within the structure and were recently cast into the ENM framework (42).

Analysis and visualization of the resulting data can be performed within MAVEN. Alternatively, PyMOL (51), VMD (164), and other molecular viewers specialized for visualization of molecular systems and can be used. For this reason, animations of the modes are saved in PDB file format (each frame is a separate MODEL) so that any molecular viewer can be used to visualize them. The MAVEN interface is configured so that generation of animation files, loading them into a molecular viewer (PyMOL has been our preferred viewer), and setting up an appropriate initial view is performed.

8.8. Conclusions

MAVEN implements multiple types of ENMs for atomic, coarse-grained, and mixed coarse-grained representations and assists the user in generating these, permitting selection by atom or residue type, spherical coarse-graining, and united atom modeling (combining multiple atoms into one placed at their mean position). By implementing these and other methods for ENM model generation, MAVEN allows for diverse and detailed hypothesis testing. One may use sparse methods for fast mode generation, making large systems more tractable. Analysis of internal motions, directional correlations within the structure, comparing the mode shapes to the variance within a structural ensemble, and comparing anisotropies of motion are presently included. MAVEN, source code, and all optional components are freely available to assist the scientific community with dynamic studies of biomolecules.

Supplementary Data available online at:

http://www.biomedcentral.com/1471-2105/12/264
CHAPTER 9. Exploring protein fold space

9.1. Discrete versus continuous


The prevailing view concerning protein structure space is that it is comprised of a finite number of discrete folds or states. These states have been classified by a number of structural ontologies including the popular CATH (259), SCOP (260), and Pfam (261) that group all known protein structures into fold types based upon combinations of sequence homology, structural topology, and function. This recent work by Skolnick et al. (262) challenges this view by constructing a graph based on pair-wise structural relatedness judged by TM-score of 5906 proteins and showing that the average shortest path in this network is seven. This value was arrived at after analyzing the size of the largest strongly connected component in the network as well as how many nodes are within \( k \) edges of one another. We believe that this may not necessarily be informative about protein structure space, but instead is likely a general network property since the same result can be obtained in a simpler way using the approximation of Watts and Strogatz (263) for random graphs implying that the average path length is \( \ln(N)/\ln(n) \) with \( N \) being the total number of nodes and \( n \) the average node degree. We find that the average path length estimated by this method varies considerably with the TM-score threshold for retaining edges, but lies between 1.4 and 7 for most graphs. In reference (262) the outdegree (the number of directed edges originating at a given node), is used, but the total degree (indegree + outdegree) could have been used and would result in decreasing the estimated average shortest path by a factor of \( \ln 2 \). With increasing TM-score the edge set gets sparser, approaching a cardinality of zero. This is shown in Figure 9.1 where the average path length asymptotically approaches infinity as the threshold passes 0.8 (outdegree) or 0.84 (total degree).

It is not necessarily true that a short average path means protein structure space is continuous. If extent of clustering is high then we would have situations where dense clusters exist that are loosely connected. If this is the case then either the dataset is not a complete representation of fold space or the space is not continuous. We are presently investigating the cluster structure in the TM-score based interaction network.
Figure 9.1: The relationship between average shortest path computed using the Watts-Strogatz approximation and the TM-score threshold for retaining edges in the graph. Inset shows that most paths belong to [1.4, 7] interval, but once the number of edges approaches the number of nodes, the average shortest path increases sharply. Strict TM-score thresholds result in disconnected graphs.

9.2. Modularity of genes and its effect on protein structure

(In preparation for submission): “Characteristics of Protein Fold Space Exhibits Close Dependence on Domain Usage” by Michael T. Zimmermann, Fadi Towfic, Andrzej Kloczkowski, and Robert L. Jernigan

With the growth of the PDB and the simultaneous slowing discovery of new protein folds, we may be able to answer the question of how discrete protein fold space is. Recent studies by Skolnick et al. (262) have concluded that it is continuous. Here we extend our initial observation (264) that this conclusion depends upon the resolution with which structures are considered, leaving open the determination of what resolution is most useful. Here we utilize graph theoretical approaches to investigate the connectedness of the protein structure universe, showing that the modularity of protein domain architecture is of fundamental importance for future improvements in structure matching, impacting our understanding of protein domain evolution and modification. We show that state-of-the-art structure superimposition algorithms are unable to distinguish between conformational and topological variation. This work is not only important for an understanding of the discreteness of protein fold space, but also informs about the more critical question of what precisely should be spatially aligned in structure superimpositions. The metric-dependence is also investigated and leads to the conclusion that fold usage in homology reduced datasets is very similar to usage across all known structures and should not be ignored in large scale studies of protein structure similarity.
9.3. Introduction

The three dimensional structures of proteins are often grouped into hierarchical classifications in order to facilitate our understanding of their relationships with one another. Thus can one envision a "fold space" for protein structures where a fold is defined as a specific spatial arrangement of secondary structures. The folds of single domain proteins have been classified by a number of structural ontologies, including CATH (259) and SCOP (260) that group most known protein structures on the bases of combinations of sequence homology, structural topology, and function. Pfam (261) is another heavily used resource, which instead focuses more upon functional classification, rather than structural (though the two are often be related). Presently, structural classifications such as CATH and SCOP still rely heavily upon expert manual curation. The prevailing view concerning protein fold space is that it is comprised of a finite number of discrete folds as described by these structural ontologies. Recent updates have yielded increased coverage of the diverse types of folds that proteins can assume (259), with a notable saturation being reached. The results of such efforts, largely driven by structural genomics initiatives (265), may imply that we are reaching full enumeration of single domain folds (266). One of the interesting and important implications that has been drawn from these works is that fold space is discrete and not continuous.

Recent work by Skolnick et al. (262) challenged this viewpoint concerning the discreteness of protein structure space based on using a graph theory approach for analyzing the topological relatedness of protein structures. By considering a large representative set of structures, and the pair-wise structural relatedness as judged by the TM-scores (52) of 5906 protein chains having low sequence similarity (from the PDB, (159)), they showed that the average shortest path in this network is seven. In the graph, nodes represent structures and edges are placed between nodes if the pair-wise TM-score is greater than 0.4. We believe that this may not necessarily be fully informative about protein structure space (264), but rather is likely to be a general network property since the same result can be obtained in a simpler way by using the approximation of Watts and Strogatz (263) for random small world graphs. Multiple questions still arise because of the metric-dependence of this conclusion, if state-of-the-art structure matching algorithms can distinguish topological diversity from conformational, and the overall role of domain architecture. In this work we seek a more detailed understanding of the properties of fold space graphs and their implications for our perceptions of protein structure relatedness. Our main contributions are as follows (1) The graphs generated based on various TM score cutoffs show a high degree of modularity; however (2) we show that the TM algorithm is not so well suited for distinguishing
topology from conformation based on our comparative analysis (with TM align) of reverse transcriptase (RT) structures gathered from Pfam to manually curated categories in CATH. Thus (3) we explore structure space using a domain-based comparison utilizing CATH and SCOP categories. Our comparison shows that there exists one dominant, modular cluster with some discontinuities in structure space outside of the larger cluster. Thus, we conclude that the continuity (or discreteness) of protein structure fold space depends strongly on the resolution used for distinguishing folds.

Modularity, graph partitioning efficiency, and community detection are three terms that refer to roughly the same concept. For community structure to be prevalent there must be groups of structures that are closely related, but few structures that are simultaneously similar to members of a different group. High modularity combined with a relatively large number of clusters would point to a discrete fold space. Low modularity or high modularity with very few clusters would point toward a continuum. Various metrics to evaluate the community structure in graphs have been developed including the modularity score of Newman and Girvan (267) that we apply here. The logic behind community structure and graph clustering to explain a small average shortest path is the following: Consider a cluster A that is well connected. That is, for every node \( n_i \) in A, any other node \( n_j \) in A is reachable, on average, via a greater number of shorter paths compared to another node, \( n_x \) that is a member of a different cluster B. This means that any neighbor of any node in A is reachable from any node in the cluster via a short path. Strong community architecture does exist in fold space graphs and further analysis is performed by employing the Markov Cluster (MCL) Algorithm (268, 269). If a large number of well connected clusters exist in the graph and relatively few edges connect them, then either the dataset is not a complete representation of fold space or the space is not continuous.

Many methods to determine the relatedness of proteins and protein structures have developed. These are dominated by sequence algorithms because the data is abundant and algorithms are computationally efficient and fairly intuitive. One such scheme is VAST, Vector Alignment Search Tool (270), which incorporates statistical significance thresholds and estimation of random interactions. The widespread use of PSI-BLAST (122) and similar string algorithms in structure classifications like CATH and SCOP are further examples. Matches based on sequence homology represent a conservative subset of similar proteins due to the fact that the inverse folding problem, determining how many sequences can assume a given 3D shape (fold), is unsolved in general. Many cases exist where sequences with little to no homology assume nearly identical folds; i.e. ubiquitin (1UBI) and SUMO (1WM2) have 15% sequence identity, but fold to practically similar structures differing only by 1.5Å Cα RMSD. Many structure
alignment procedures exist that are widely used in structural biology. In this work, we will primarily use TM-align (52), which has been shown to give excellent alignments and is used for template detection, for example by I-TASSER (271), currently ranked among the best performing 3D structure prediction servers.

What exactly should be compared? Proteins with different numbers of amino acids are, mathematically, objects with different conformational dimensions; therefore we commonly reduce the problem to simply finding the best superimposition of two structures. Interestingly, there may be patterns in other mathematical spaces that simplify the analysis of structures, such as the relation between spectral dimension (related to allostERIC efficiency) and fractal dimension (related to packing density) in protein structures (190, 191). However, the details of the structure can influence the energy transfer (allostery) pathways within the structure (272) or their interactions (273). A much coarser view might consider proteins as approximate globules – amorphous 3D blobs whose surfaces are semi-molten (2) and have mostly polar character but with some non-polar groups exposed to water. We have just described two very different views of protein structures where the details can strongly bias an analysis toward a specific conclusion. In the first case, it is intuitive that relatively few structures will be similar whereas in the second case, many proteins could resemble each other’s shape. Thus, the resolution or the extent of coarse-graining with which we treat structures will affect our conclusion about the discreteness of fold space. In this work, we investigate structure superimposition using TM-align and domain similarity. Our application of the TM-align procedure is similar to that used by Skolnick et al. (262), except that we consider various thresholds instead of a single, fixed TM score cutoff. We also collect a representative from each known fold type and apply the same graph analysis to this smaller dataset. These representatives have been deemed by expert manual curation to comprise distinct fold types and thus represent a biased scenario for concluding that fold space is discrete. The results of this analysis are utilized to interpret data obtained by using the complete protein dataset. For domain similarity, we analyze fold space independent of any structure-based comparison by connecting nodes if two proteins share a common CATH or SCOP annotation. Annotations are taken from CATH at the Topology level and from SCOP at the Fold level. Such an analysis provides an impartial baseline for how any structure similarity metric that seeks to approximate CATH or SCOP-level fold similarity will perform.

Defining what is compared is important: complete PDB files, PDB chains (individual polypeptides), or single domains. Much effort has been expended to develop methods for computational domain prediction. Early contributions such as FSSP using Dali (274, 275) have been very influential, while newer
algorithms such as DomNet (276) show increased refinement and better agreement with manual curation. However, in this study we will focus on the manual curation levels of CATH and SCOP. If whole PDB files or chains were used, there would be cases where the peptide chain could be assigned to two or more domains. Such structures can act as cluster-linkers in the fold space graph since one domain may have a significant score with structures in one cluster, while the other domain will have strong structural relation to a different cluster. Alternatively, a single domain could require the interaction of more than one chain. Such proteins complicate the relationship between sequence-homology reduced datasets and fold usage. Considering the size of a protein may also be important since a small protein is more likely to possess a topology that is a subset of a larger protein.

9.4. Results

For any approach that relies on graph theory, understanding the structure of the graphs is essential. Figure 9.2 and Supplemental Figure 1 show us that, for any TM-score threshold, there exist a relatively small number of nodes possessing a high degree of connectivity and that the graphs exhibit a large amount of community structure. See Methods for a description of $F_{area}$ and $F_{mass}$ and their meaning in graph clustering. We have investigated these hub nodes and draw two conclusions. Some of them are hubs because they are among the smallest proteins in the dataset with approximately 50 residues. It is more likely for a small protein to be topologically similar to a subset of a larger protein than for two proteins of equal size to match. Others have high connectivity because they have multiple domains. Each domain can individually manifest a significant alignment with other structures, which inflates the connectivity relative to those of single domain chains.

In our previous work (264), the relationship between average shortest path computed using the Watts-Strogatz approximation (263) and the TM-score threshold for retaining edges in the graph was investigated. We found that the average shortest path is less than seven for cutoffs below 0.75. Stricter cutoffs result in large areas of the graph becoming disconnected. With increasing TM-score, the edge set gets sparser, approaching a cardinality of zero. This is shown in Supplementary Figure 2 where the number of nodes with no edges increases as the TM-score threshold increases.

Since TM-scores are numerical, defined on the interval from zero to one, and are not symmetric, pairwise scores can easily be interpreted as a directed graph where we use TM-scores as edge weights. In the MCL algorithm an edge weight is the probability of a random walk traversing a given edge. We construct unweighted graphs by assigning all edges a weight of one and undirected graphs by linking
nodes (with or without edge weights) based on the larger of their two TM-scores. From Supplemental Tables 9.1 and 9.2, it is evident that the edge treatment has a minimal impact upon MCL clustering.

![Figure 9.2: Metrics for Analysis of MCL clustering on TM-score graphs.](image)

Uniformly weighted (UW) representations are used where a weight of one is assigned if an edge exists in the graph. Area Fraction is defined by Equation 9.2 and relates to the cluster size. Mass Fraction is the fraction of total edge weight that is captured within clusters and is formally defined in Equation 9.3. Including the edge weights does not impact these metrics (see Supplemental Table 1 and 2).

Since the sequence-structure relationship is not fully understood, sequence-homology reduced datasets are not necessarily the same as topology reduced datasets. The effect on graph behavior of a topologically reduced dataset is of interest for comparison to the homology reduced dataset. Parameter choices that yield expected results in the topologically reduced dataset will help us to better interpret the meaning of clusters for the homology reduced set. For this reason, we also compare distinct topologies to each other by gathering 1233 CATH version 3.3 topology representatives; a collection of manually curated topology representatives that span all of the PDB, performing the same procedure. Interestingly, this dataset of distinct topology representatives exhibits a high modularity, indicative of community structure (Supplemental Table 2). We calculate a modularity score defined in (267) by comparing the number of edges within clusters to the number of edges that link clusters to one another. At low TM thresholds (0.4), the graph exhibits high connectivity (57550 edges) and the majority of the nodes are included in the largest cluster. The extent of community structure is less than for the PDB300 dataset (as judged by $F_{\text{Mass}}$ and $F_{\text{Area}}$ – see Methods), but remains high. The MCL inflation parameter
determines the granularity of the clustering with a low inflation yielding few large clusters and high inflation producing many small clusters. Even for the high inflation value of 5, the largest cluster still contains 855 structures, whereas a low value of 1.2 retains 1217. At a TM score cutoff of 0.6 we find that MCL consistently distinguishes many of the topologies from each other (only 334 edges between the 1233 nodes). Thus, these graphs may either be modular because they are significantly related (suggesting a continuous structure space) or because they are mutually distantly related (pointing towards a discrete fold space).

Conformational variability is also an important consideration for comparing topologies. Are our structure comparison metrics able to distinguish between conformational variations and larger topological differences? To address this question, we compare 283 reverse transcriptase (RT) structures gathered from Pfam (261) family PF00078 to one another and to the CATH topology representatives to investigate the ability of structure comparison metrics to distinguish between conformations (within the RT family) and topologies (between RT and fold representatives). The polymerase part of the RT structure is often described as analogous to a human hand where the active site is in the center of the palm and the fingers and thumb “grip” the substrate. The Pfam family set used corresponds to two fingers and the palm, thus containing sequence (average sequence identity 67%) and conformational variants. We find that all members of the reverse transcriptase family have TM-scores above 0.4, but there is significant diversity of scores within the family (Figure 9.3). Roughly half of the pairwise comparisons are between 0.4 and 0.7 corresponding to different finger conformations (generalized from visualizing 100 randomly chosen pairs from this group). Higher scoring pairs are characterized by the structures having the same general finger conformation. The subgroup at about 0.82 has a greater representation of a more extended finger conformation (again from visualization of 100 randomly chosen pairs; data not shown), with a representative pair is shown in Figure 9.3. Further, each reverse transcriptase domain has a TM-score between itself and a topology representative of at least 0.53, but none is higher than 0.76. Therefore, all RT structures have a significant structure alignment to a topology representative. One might expect that because all RT structures share a common fold, one topology would be the best match to most of the RT structures. However, matching each of the 283 RTs to its highest scoring topology yields 277 different topologies. Thus, large TM-scores, while relatively few, are not occurring because of any single (or even a small set) of RT-like topological representatives. Further, TM (and likely any rigid superimposition algorithm) is, in general, unable to distinguish between conformational and topological variation. Methods like Fr-TM-align (277), FATCAT (278), or IDSS (279)
that are capable of accounting for flexibility of the biomolecule may perform better in this specific test, but fast and accurate methods for incorporating flexibility in structure matching are still being improved. Current structure comparison algorithms have difficulty in distinguishing between conformational and topological differences.

Figure 9.3: Structural comparison of the reverse transcriptase fold and conformational variation – a test of state-of-the-art structure superposition.

A) Heat map of TM-scores between 283 reverse transcriptase (RT) structures and 1233 topology representatives from CATH v3.3. Rows are arranged in the same order as the columns of sections 1 and 2. Sections 1 and 2 correspond to two orientations of the RT “fingers,” while section 3 is the TM score between the topology representatives and the RT structures. B) Histogram of the TM-scores within the set of 283 reverse transcriptase structures. The set can easily be split into structures that are related to each other at a TM-score of greater or less than 0.7. No pairwise scores fall below 0.4. Low scores correspond to the two finger domains being in different positions, while the higher scores correspond to the two fingers in nearly the same orientation. The high scoring population can be roughly considered to consist of two groups; one where the fingers are more closed, and one where they are both extended. C) We show a representative of the lower TM-score population; 1RW3 aligned to 1JLA with a TM-score of 0.56. This view highlights the different finger positions that are characteristic of the lower scoring group. D) Histogram of maximum TM-score between each reverse transcriptase domain and the topology representatives from CATH (max for each row of Section 3). Each reverse transcriptase domain has a TM-score between itself and a topology representative of at least 0.53, but none are higher than 0.76. There are 277 topologies matched to the 283 reverse transcriptase structures. Thus, large TM-scores, while relatively sparse, are not because of any single (or even a small set) of reverse transcriptase like topologic representatives.

Metrics similar to Silhouettes (280) have also been generated (not shown). These are basically average path lengths from a node to any other node within a given cluster compared to the average path length from a node to every node that is not in that cluster. Evidence of the high number of
connections within each cluster is seen since the average out-of-cluster path is only slightly longer than the average within-cluster path.

A critical point of the above analysis hinges on the efficacy of the TM score algorithm in quantifying the fold space of proteins. Thus, it is reasonable to ask: are these investigations of protein fold space dependent upon the metric used? We have already shown that the state-of-the-art structure comparison method has difficulty in distinguishing topological and conformational differences, but can we explore fold space independently of structure superimposition? One way is to make a graph where each protein chain is represented by a node and nodes are connected by edges if the two proteins share a common fold. Common folds are determined by a shared CATH topology or SCOP fold using CATH version 3.4 and SCOP version 1.75. In the PDB300 dataset 90% of the protein chains are annotated with at least one of these ontologies, while all of the PISCES proteins are annotated (see Methods for dataset details). Unannotated nodes are neglected in the following analysis. Using the same graph analysis procedure, we find that this domain based graph also has a very high degree of connectivity and modularity. See Table 9.1 for details. We again find that there exists one dominant cluster. It has been shown that MCL usually generates a dominant cluster and for some applications modifications that generate a more uniform granularity are preferred (281). However, using these approaches is equivalent to assuming fold space to be discrete. Another explanation for the dominant cluster is the imbalance in topology usage. Table 9.2 summarizes the usage of the ten most used topologies across all of CATH, the PDB300, and the PISCES dataset. Seven of the ten most frequently used annotations across all of CATH are also in the top ten most used topologies in the datasets used here. Furthermore, if we sort the topology classes by their use across all of CATH, and compare with the topology use in each of our datasets, PDB300 and PISCES have a correlation coefficient with CATH of 0.94 and 0.93, respectively. Thus, the relative distribution of domain types is similar in these datasets compared to that in the whole PDB. We conclude that the reason for the observed shortest paths in TM-score based graphs is the modularity of proteins and the bias in topology usage. Protein structures exhibit variations upon themes – stable domains develop and are embellished upon evolutionarily for further modification of function.

Viksna and Gilbert (282) proposed a new method of assessment of domain evolution by measuring the rate of certain kinds of structural changes that can lead to novel fold development. Birzele et. al. (283) find interesting evidence that alternative splicing plays a role in protein structure evolution by developing transitional structures between fold types. Fong and coworkers (284) emphasize the modularity and importance of domain fusion events for the evolution of protein domains. Meier et al.
suggested a link between conformational flexibility and domain evolution where the native state ensemble can partially occupy at least two intermediate fold types and the relative population of each may be influenced by single amino acid mutations. The results presented here combined with these studies point to the importance of considering protein folds more rigorously in structure matching. It is not only important for our understanding of the discreteness of protein fold space, but informs the more critical question of what precisely should be spatially aligned in structure superimposition.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Inflation</th>
<th>$\text{Mod}_5$</th>
<th>$\text{Mod}_{\text{all}}$</th>
<th>$\text{Eff}$</th>
<th>$\text{F}_{\text{Mass}}$</th>
<th>$\text{F}_{\text{Area}}$</th>
<th>#C</th>
<th>Max</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDB300</td>
<td>1.2</td>
<td>.98</td>
<td>.99</td>
<td>.21</td>
<td>.98</td>
<td>.58</td>
<td>137</td>
<td>3617</td>
<td>34.8</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>.89</td>
<td>.92</td>
<td>.33</td>
<td>.91</td>
<td>.25</td>
<td>183</td>
<td>2155</td>
<td>26.1</td>
</tr>
<tr>
<td></td>
<td>4.0</td>
<td>.75</td>
<td>.82</td>
<td>.49</td>
<td>.81</td>
<td>.10</td>
<td>325</td>
<td>1429</td>
<td>14.7</td>
</tr>
<tr>
<td></td>
<td>6.0</td>
<td>.51</td>
<td>.59</td>
<td>.49</td>
<td>.71</td>
<td>.07</td>
<td>450</td>
<td>1179</td>
<td>10.6</td>
</tr>
<tr>
<td>PISCES</td>
<td>1.2</td>
<td>.96</td>
<td>.97</td>
<td>.34</td>
<td>.97</td>
<td>.27</td>
<td>138</td>
<td>864</td>
<td>15.7</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>.89</td>
<td>.94</td>
<td>.59</td>
<td>.93</td>
<td>.09</td>
<td>174</td>
<td>465</td>
<td>12.4</td>
</tr>
<tr>
<td></td>
<td>4.0</td>
<td>.78</td>
<td>.86</td>
<td>.66</td>
<td>.88</td>
<td>.06</td>
<td>236</td>
<td>357</td>
<td>9.2</td>
</tr>
<tr>
<td></td>
<td>6.0</td>
<td>.61</td>
<td>.68</td>
<td>.65</td>
<td>.82</td>
<td>.04</td>
<td>289</td>
<td>303</td>
<td>7.5</td>
</tr>
</tbody>
</table>

Table 9.1: Clustering metrics for graphs generated from common CATH or SCOP annotations across both datasets

Column titles are: Inflation for the MCL parameter that determines granularity of the clustering, $\text{Mod}_5$ - modularity using the 5 largest clusters, $\text{Mod}_{\text{all}}$ - modularity using all clusters, $\text{Eff}$ - efficiency of the clustering, $\text{F}_{\text{Mass}}$ and $\text{F}_{\text{Area}}$ are given in Equations 9.2 and 9.3, #C - number of clusters, Max - number of nodes in the largest cluster, Avg - average number of nodes across all clusters.

<table>
<thead>
<tr>
<th>CATH</th>
<th>PDB300</th>
<th>PISCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Usage</td>
<td>Architecture</td>
</tr>
<tr>
<td>3.40.50</td>
<td>19229</td>
<td>3-Layer(aba) Sandwich</td>
</tr>
<tr>
<td>2.60.40</td>
<td>13806</td>
<td>Sandwich</td>
</tr>
<tr>
<td>3.20.20</td>
<td>6106</td>
<td>Alpha-Beta Barrel</td>
</tr>
<tr>
<td>3.30.70</td>
<td>4236</td>
<td>2-layer Sandwich</td>
</tr>
<tr>
<td>2.40.10</td>
<td>3954</td>
<td>Beta Barrel</td>
</tr>
<tr>
<td>2.60.120</td>
<td>3433</td>
<td>Sandwich</td>
</tr>
<tr>
<td>2.40.50</td>
<td>2244</td>
<td>Beta Barrel</td>
</tr>
<tr>
<td>3.30.200</td>
<td>2012</td>
<td>2-layer Sandwich</td>
</tr>
<tr>
<td>1.10.510</td>
<td>1992</td>
<td>Orthogonal Bundle</td>
</tr>
<tr>
<td>1.10.490</td>
<td>1983</td>
<td>Orthogonal Bundle</td>
</tr>
</tbody>
</table>

Table 9.2: Top 10 domains by CATH Topology Usage

Usage is the number of protein chains that are annotated with the given CATH topology ID. C_r is the rank of this topology ID in CATH.
9.5. Discussion

We have shown that graphs generated either from TM scores or domain annotation show a high degree of community structure (modularity). TM-scores alone are not able to fully distinguish manually curated topology representatives from each other at the same threshold levels that have been used to analyze fold space across the PDB. This is partly due to the effect of conformation on TM-score. It is shown here that conformational variability within a set of reverse transcriptase structures can lead to very different conclusions about which CATH topology is the closest representative. It is important to realize that since we do not fully understand the relationship between protein sequence and structure, homology reduced datasets may not correspond strictly to topology reduced. This has been shown by analyzing graphs generated by connecting nodes if they share any common CATH topology or SCOP fold and showing that they have similar modularity and graph structure compared to graphs based on structure superimposition (see figure 9.4). It is possible that improving coarse-grained representations like TOPS strings (286, 287) will be useful in the future for handling the multi-resolution complexities of structure comparison.

![Figure 9.4: Frequency of CATH topology usage in three datasets.](image)

Using CATH version 3.4, we consider the 1282 topology IDs, counting the number of times a structure in each dataset is annotated with each topology ID. For both of the datasets used in this study, we find that many topologies are not used at all (abscissa value of 0), and that relatively few topologies have a high rate of use. In the CATH database, most topologies have a high rate of use. Interestingly, the PISCES dataset is more topologically diverse than the PDB300 dataset.

Classifying protein tertiary structures into a discrete set of domains is useful because it aids the conceptual understanding of protein structures, helps by reducing the possible outcomes for sequence based folding procedures, adds to our understanding of the structure-function relationship, as well as
many other possible applications. Whether protein fold space is continuous or discrete depends upon the resolution with which it is viewed. We believe that a more fundamental understanding may come from the usage of topologies.

9.6. Methods

9.6.1. Datasets

We use the PDB300 dataset from (262), which consists of 5906 protein chains of lengths between 40 and 300 residues sharing less than 35% sequence identity. Here, the word “chain” refers to a single polypeptide. It is worth noting that numerous individual protein chains in this dataset contain more than one domain (topology or fold) as defined by CATH or SCOP.

Domain centric datasets are constructed in two ways. The first is to cut the CATH hierarchy at the topology level resulting in 1233 or 1282 representative domains for version 3.3 and 3.4 respectively. The second utilizes the PISCES server (288) to gather a representative set of chains that have better than 2.5Å resolution, less than 20% mutual sequence identity, and a crystallographic R-factor of less than 0.25. This dataset contains 4750 PDB chains. We then use CATH to identify individual domains at the Topology level within this set.

The final dataset used is Pfam family PF00078, corresponding to reverse transcriptase (RT). There are 283 members with full 3D coordinates. These structures were downloaded and the subset of points agreeing with the Pfam family definition was retained.

9.6.2. Protein Structure Evaluation Metrics

In this paper, the TM-score defined in (52), is used to analyze the structural similarity of protein structures. This metric is not symmetric; TM(A,B) does not necessarily equal TM(B,A) particularly when proteins A and B are of different lengths. The TM-score is defined as:

$$ TM-score = \max \left[ \frac{1}{L_{\text{Target}}} \sum_{i=1}^{\ell_{\text{align}}} \frac{1}{1 + \left( \frac{d_i}{d_0 \left( L_{\text{Target}} \right)} \right)^2} \right] $$

(9.1)

Where $L_{\text{align}}$ is the length of the alignment, $L_{\text{Target}}$ is the sequence length of the target structure, $d_i$ is the Euclidean distance between aligned points, and $d_0$ is a normalization factor based on $L_{\text{Target}}$. 
9.6.3. Analysis of Structural Classification

A number of structural classification schemes exist including the CATH database (259), SCOP (260), and PFAM (261). Both CATH and SCOP are hierarchical in nature and utilize a combination of homology, topology, and biochemical function to organize protein structures. The first level of CATH and SCOP classification consists of 4 classes, separating structures into predominantly α-helix or β-sheet content, presence of both, or lack of secondary structure elements. The second level of SCOP, as well as the second and third levels of CATH, are based on overall secondary structure orientation. These levels are manually curated and place proteins into general categories like beta-barrel and two-layer sandwich. The third level of SCOP takes into account the topology and function of a given protein to decide how related they are evolutionarily. All subsequent levels in both classifications are decided by sequence identity or, in some cases, other sequence based scoring schemes. Pfam families are generated by manual functional curation, multiple sequence alignments, and Hidden Markov Models and come in two varieties: Pfam-A for only manually curated entries and Pfam-B where automated methods are also used to extend the sequence space covered by classification. All three of these databases rely on sequence homology and biochemical functions to group proteins into fold types rather than directly comparing quantitatively the topology of the biomolecules.

9.6.4. Graph Construction

We define a graph based on TM-score as \( G_t = \{E, V\} \) where \( e \in E \) is an edge in \( G_t \) if it connects two vertices \( a \in V \) and \( b \in V \) and \( TM(a,b) > t \). Each PDB chain in the dataset is represented by a single node. The TM-score threshold \( t \) is initially set at 0.4 as in (262), but values up to 0.9 are also considered in the analyses of graph structure. Graphs are either undirected or directed. To make the directed graphs we consider \( t = \max(t_1, t_2) \) where \( TM(a,b) = t_1 \) and \( TM(b,a) = t_2 \).

9.6.5. Cluster Generation and Comparison

To investigate the community structure of graphs we first employ the Markov Cluster Algorithm (MCL) (268, 269). In this procedure, graphs are clustered based on random walks that simulate flow along the graph edges. Nodes that are well connected will exhibit more flow between them than nodes with few connections. As the algorithm progresses, nodes that share high amounts of flow (many common walks) are grouped together into clusters.
MCL has evaluation protocols to explore the relatedness of clustering with different parameters. In MCL each edge has a weight. Here we use uniformly weighted (UW) graphs or we use the TM-score as the edge weight. Defining cluster size as the number of nodes within a cluster, MCL computes the Area Fraction ($F_{\text{Area}}$) defined by Equation 9.2. This metric gives an indication of the size of clusters, as many small clusters or isolated nodes result in a low $F_{\text{Area}}$. The Mass Fraction ($F_{\text{Mass}}$) is the sum of all edge weights within clusters and is shown in Equation 9.3 where $w_i$ is the edge weight of edge $i$ such that edge $i$ is in Cluster $c$.

\[
F_{\text{Area}} = \frac{\sum \text{clusterSize} e^2}{N(N-1)} \tag{9.2}
\]

\[
F_{\text{Mass}} = \sum_{c=1}^{[E]} \sum_{i=1}^{[E]} w_i \quad \text{s.t. } w_i \in C_c \tag{9.3}
\]

A value of $F_{\text{Area}}$ near zero implies that the graph has been clustered into many small clusters, while a value of one implies that all nodes are included in a single cluster. A value of $F_{\text{Mass}}$ near one indicates that clusters are tightly connected with relatively few edges connecting them. How the algorithm treats the length of a walk (number of edges traversed) is important for the process and is controlled by a parameter called Inflation. Penalizing longer walks produces a large number of small clusters. Allowing longer walks generates fewer, but larger, clusters. It is informative to compare results across multiple inflation values to better understand the organization of the graph.
CHAPTER 10. Conclusions

This dissertation explores the functional dynamics of important biomolecules while extending the utility of Normal Mode Analysis using Elastic Network Models through the application of novel analysis methods. We have described the software tool MAVEN which integrates and automates many of the steps in model building and analysis. Licensed under the Lesser General Public License, MAVEN and its source code is available free of charge to the academic community. As it has been written as a set of modular scripts, we hope it will be useful as a standalone tool as well as for extending the functionality of software developed by other research groups.

We have discerned the structural characteristics of antibodies, ribosomes, telomerase, and metal efflux systems. Normal Mode dynamics can be used to probe allosteric communication pathways and analyze the efficiencies of energy transfer through the structure, improving our understanding of the structure-function relationship. Further, in the case of antibody structure, we have related an important functional sequence characteristic, the hypervariable complementarity determining regions, to structural dynamics. The structure has been refined to facilitate large excursions of the sequence variable regions which likely aids in antigen recognition. Ribosome simulations have revealed coupled dynamics throughout the structure including the peptide exit channel twisting concurrently with the 30S-50S ratchet. Future work will investigate more of these details. The role that the ribosome-associated proteins play is also a topic of debate. Some have defined functions, but others do not. We have investigated all 30S ribosomal proteins between mesophilic and thermophilic organisms, determining a role for the placement of highly mobile residues in adaptation in high temperature organisms. Our study of the telomerase mechanism helps explain much of the data present in today’s literature and acts as a platform from which future studies can launched for determining the functional roles of the other constituents of the mammalian telomere maintenance system. Finally, increasing the detail with which we understand cellular efflux systems has the potential to impact the development of new therapies for treating drug resistant bacteria. These studies represent gains in our understanding of the structures that carry out necessary cellular functions. Mechanistic understanding of biomolecular motion aids in the understanding of physiology, disease states, and our ability to engineer new structures exhibiting novel functions.

The paradigm in most structure modeling procedures is to generate one model that exhibits best overall performance. This has been the usual treatment in ENMs studies as well. It is commonly observed that some proteins require different parameterizations to achieve good performance to fit
experimental data, but the most common practice is to use parameters that perform best, on average. By utilizing the various modes from multiple ANMs we capture collective as well as local motions which accurately, describing an experimental tRNA structure ensemble. To our knowledge, this is a novel technique for evaluating structure dynamics using ENMs. Utilizing multiple experimental structures to describe the native ensemble has the potential to improve the accuracy of future computational studies. The sets of derived structures could also be used to improve the other methods presented in this dissertation; namely, entropy estimation, identification of communication pathways through the structure, sampling of conformational changes, and more.

The ability to distinguish native-like structures from a set of computational predictions is important not only in structure prediction, but also for molecular docking and the prediction of conformational changes. We propose a new algorithm for evaluating the entropy of biomolecules, showing that it leads to increased discrimination between native-like and non-native-like models in CASP structure predictions. We are also able to better distinguish native-like protein-protein docking poses taken from either ClusPro results or Zdock poses. Our findings indicate that the shape of a protein or complex contains enough information to distinguish it from poorer quality predictions. Graph theoretical approaches have also been employed to investigate the connectedness of the protein structure universe, showing that the modularity of protein domain architecture is of fundamental importance for future improvements in structure matching, impacting our understanding of protein domain evolution and modification. We also investigate the limitations of current structure alignment algorithms and the effect they have on our conclusions about the nature of protein fold space.
APPENDIX A. Computational free energy evaluation to explore energy minima

Many problems in computational biophysics can be thought of in terms of an energy landscape where valleys represent favorable states with energy barriers. Traversing this landscape is then analogous to following conformational transitions. One can even think of the cross-sectional area at a given energy height as a measure of entropy or conformational diversity (depicted in figure A.1.). We will use this model here to discuss two different major aims:

- **Aim 1: Conformational Transitions**
  - Discover Multiple Minima from One
  - Constrained Energy Barriers
  - Determine Number of Intermediate Types

- **Aim 2: Structure Refinement**
  - Coarse Grained Potentials
  - No Target

![Figure A.1: Energy funnel analogy depicting the two aims – searching for alternate conformations of a structure and refinement of a modeled structure. Image is modified from Ken Dill’s energy landscape figure (289).]
Figure A.2: Graphical depiction of the two aims: structure refinement and conformational change.

Aim 1 involves the use of constrained ANM simulations and energetic evaluation using 4-body potentials to traverse between minima on the energy landscape. The goal is to enable the discovery of other minimal structures (functional conformations) from one starting minima. Energy barriers (red) will need to be overcome. Aim 2 uses the same methodology, but the goal will be to refine a structure down one dominant energy minima, not to traverse large energy barriers.

A1. Sampling multiple minima from a starting structure

Aim 1 involves the use of constrained ANM simulations and energetic evaluations using 4-body potentials to traverse between minima on the energy landscape. The goal is to become capable of discovering other minimal structures (functional conformations) from one starting minima. We will begin with a set of known experimental structures, such as adenylate kinase, where multiple functional conformations are known. By iteratively applying constrained ANM simulations and evaluating the energy of each conformer with 4-body potentials, we aim to pick out known intermediate forms. Constraints added to the ANM simulations involve adding a virtual bond length (VBL) and van Der Waals (VDW) criteria. The VBL criteria will prevent the structure from breaking apart (known as the “tip effect”) as sometimes occurs in extended structures, and is often seen in iterative ANM evaluation. Energy barriers will need to be overcome.

ENMs provide a local solution of dynamics, though the directions of motion used often point in the direction of meaningful conformational changes. For this reason we propose the use of iterative ENM where the modes are used to deform the structure a small amount, the modes are re-evaluated, and the process is repeated. The questions of which modes and how far to proceed by iterating is still open. Initially we will adopt a sampling procedure where modes are chosen based on a probability that reflects the contribution to the total motion of that mode. The amplitude will also be a random variable.
Our initial results are favorable and are shown in Table A.1. We begin with the open form of adenylate kinase and expand the initial structure into an ensemble. The ensemble is expanded until 1000 conformers are in the ensemble; afterwards the ensemble size remains constant. Throughout the ensemble of trajectories, we check the RMSD between each member of a set of known intermediates. There are members of the ensemble which are closer to the intermediates than the initial structure. As with other sampling procedures, the largest challenge is to be able to discriminate which members of the ensemble are “better” members. We believe our combined entropy and 4-body energy method shows promise.

<table>
<thead>
<tr>
<th>PDB ID</th>
<th>Initial RMSD</th>
<th>Iteration 13</th>
<th>Iteration 26</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inter 2rh5a</td>
<td>7.2 7.2</td>
<td>-1.1 1.5</td>
<td>-1.7 2.6</td>
</tr>
<tr>
<td>Inter 2rh5b</td>
<td>6.9 6.9</td>
<td>-1.0 1.4</td>
<td>-1.5 2.6</td>
</tr>
<tr>
<td>Inter 2ak3a</td>
<td>7.9 7.9</td>
<td>-0.2 0.9</td>
<td>-0.1 1.3</td>
</tr>
<tr>
<td>Open 1akea</td>
<td>0.0 0.0</td>
<td>0.6 3.4</td>
<td>0.9 5.1</td>
</tr>
</tbody>
</table>

Table A.1: RMSD change from the initial adenylate kinase structure to members of the structural ensemble generated by iterative ANM.

A2. Structure refinement

Aim 2 will use the same methodology, but the goal will be to refine a structure down to one dominant energy minima, and not to traverse large energy barriers. This has been done in the case where the target structure is used, but to improve a structure without knowing the target is difficult and remains an area of ongoing research. The good performance of 4-body potentials combined with entropic calculations using mean square fluctuations might provide the necessary improvement in discriminatory power required to achieve this goal.
Figure A.3: Preliminary simulations to refine a predicted structure using ANM and 4-body potentials, without using the target structure. Predicted structures are taken from a recent paper by Stumpff-Kane (290) where CHARMM normal modes were used to refine modeled structures toward a target. Our goal is to improve the structure without using the target. Our proof-of-concept test took one of their models and was able to improve it by following ANM modes and selecting new conformations with lower 4-body energy, though the amount of improvement was not large.
APPENDIX B. Developmental differences in IL1B expression


B.1. Abstract

IL1β is an important protein in vertebrates. It is a member of the cytokine protein family and is involved in generating an inflammatory response to infections. Researchers have found that there are two porcine IL1β proteins expressed – one in embryos and the other in macrophage and endometrial tissues. These two proteins have about 86% sequence identity. In this paper, we attempt to describe how these two proteins might differ structurally and functionally. We find that 1) A predicted binding site appears to have different side chain arrangements that might lead to different binding efficiencies for the same protein or even to different partners. 2) The Caspase 1 cleavage site in the precursor proteins differs in a way that has previously been experimentally determined to reduce the cleavage activity by one order of magnitude for the embryonic IL1β, conferring a significant advantage to the protein (embryonic IL1β).

B.2. Introduction

Living organisms possess immune systems ranging from simple selective membranes to biological entities that can detect and protect the organism against disease-causing pathogens or rogue tumor cells. Leukocytes, also known as white blood cells, are one important part of the adaptive immune system. Macrophages are one type of leukocyte that engulf foreign bodies (phagocytosis) and isolates them within cellular membranes. Many cells in the immune system, including other types of leukocytes, secrete substances called cytokines. These cytokines, including interleukins (ILs), signal neighboring cells and elicit a response to the invading pathogen. Interleukins form a primary communication channel for the immune system. The interleukins are synthesized by macrophages and endothelial cells such as the inner lining of the mammalian uterus (endometrium). IL1β is a member of the IL-1 cytokine protein family produced by activated macrophages as a proprotein. The IL1β proprotein is proteolytically
cleaved by Caspase 1 to activate it. In humans, this protein is related to many diseases such as major depressive disorder (291), osteoporosis in post-menopausal women (292), lung cancer in a Japanese population (292), increased bleeding after cardiac surgery (293), chronic and aggressive periodontitis (294), chronic inflammatory conditions of the brain (295), etc. It also takes part in a variety of cellular activities such as cell proliferation, differentiation, and apoptosis. It increases cell-surface adhesion by helping to synthesize cell-surface adhesion molecules and helps recruit leukocytes to the infected site. In the bloodstream it can cause fever and helps synthesize proteins that can activate transcription factors to stimulate gene expression. The porcine genome has just been sequenced. We followed up on some prior QPCR (quantitative real time polymerase chain reaction) and sequencing work that indicated there are two different IL1β RNAs expressed. Two genomic copies have been identified and appear to be tandem duplicates. The RNA sequence data indicates that one is expressed in macrophages and endometrium tissues and the other in the developing embryo at implantation. The two protein sequences are 86% identical. However, in the embryonic case, there is a proline inserted just 2 amino acid away from the predicted Caspase 1 cleavage site that activates the protein. Here we seek answers to the following questions: (1) How do these two proteins different structurally from one another? (2) Do these two proteins have any apparent differences in functions?

The tyrosine kinase family to which IL1β belongs, has two subclasses: receptor and non-receptor. Receptor class proteins play pivotal roles in diverse cellular activities including growth, differentiation, metabolism, adhesion, motility, and death. So it is possible that the activity of the kinases in embryos and macrophages might be different or at different levels. In this paper we present our findings that support our hypothesis that the small differences between the sequences of these two proteins, mIL1β (expressed in macrophages) and aYL1β (expressed in embryos), there may be some significant differences in the activity or level of specific functional role of these two proteins.

B.3. Methods

If the structure of a protein is not available in the PDB, we use comparative modeling approaches to predict the structure (128, 296). For structure prediction of the proteins, we rely upon Zhang’s ITASSER server (296-298) (http://zhanglab.ccmb.med.umich.edu/ITASSER/) to predict the structures of mIL1β and aYL1β. This algorithm gave the best protein models at the recent Critical Assessment of Structure Predictions (CASP 7 and CASP 8), a community-wide, world-wide experiment designed to obtain an objective assessment of the state-of-the-art methods in structure prediction (299-301). The I-TASSER
algorithm consists of three consecutive steps: threading, fragment assembly, and iteration. During threading, I-TASSER generates template alignments by a simple sequence Profile-Profile Alignment approach constrained with the secondary structure matches. Fragment assembly is performed on the basis of threaded alignments and the target sequences are divided into aligned and unaligned regions. The fragments in the aligned regions are used directly from the template structures and the unaligned regions are modeled in ab-initio simulations. Clusters of decoys are generated with the use of a knowledge-based force field. The cluster centroids are generated by averaging the coordinates of all clustered decoys and ranked based on structure density. In the iteration phase, the steric clashes of the cluster centroids are removed and the topology is refined. The conformations with the lowest energy are selected. The I-TASSER server returns the best five models with a C-score attached for each model and the top ten templates used. The C-score is a confidence score that is used to estimate the quality of the predicted model by I-TASSER. The C-score is based on the significance of the threading template alignments and the convergence parameters of the structure assembly simulations. When selecting one of these models, we select the model that comes from the largest cluster and has the best C-score. Reasonable structures are usually found to have a C-score in the range [-5,2], with larger C-scores indicating a better model (298).

B.3. Results

Structure Similarity

We predict the structures of the cores of the two porcine proteins, mIL1β and ayIL1β, and find that structurally they are very similar with an all atom RMSD of 0.48 A as shown in Figure B.1. We also predict the precursor protein structures. The C-scores for the structures of the precursors of mIL1β and ayIL1β are -3.07 and -3.06, respectively. These low C-scores indicate that structure prediction for the leading sequences of the precursor proteins was mostly ab initio structure prediction. We separately consider the mature part from each of the precursor structures and superimpose them. They have an all-atom RMSD of 1.61 A, which is significantly larger than the 0.48 A found when the cores were

Function Similarity

We also predict the binding sites and functions of these two proteins. First we used sequence based methods for function prediction of the two proteins with the ExPaSy server (http://ca.expasy.org/prosite). This server gives the motif, interleukin-1, for each of the sequences. For the mIL1β protein sequence, amino acids 112-132 form the motif and for ayIL1β the motif region is
In addition to this prediction, we have the same functional site predicted using the I-TASSER server. Both proteins are predicted to have two functions with EC-score > 1.1. A functional prediction with EC-score > 1.1 is considered to be a prediction with high confidence. One of these predicted functions is the enzymatic activity in the following phosphorylation reaction:

\[
\text{ATP} + [\text{protein}]\text{-L-tyrosine} \leftrightarrow \text{ADP} + [\text{protein}]\text{-L-tyrosine-phosphate}
\]

The other function is the limited hydrolysis of proteins of the neuroexocytosis apparatus but no action is detected on small molecule substrates.

**Determining differences in enzymatic cut sites**

Expasy PeptideCutter web server (302) is used to predict cut-sites in the proteins for use in distinguishing either their physiologic roles, or differential tests for their presence. Cysteine proteases have a common catalytic mechanism that involves a nucleophilic cysteine thiol in a catalytic triad. Both proteins under consideration possess cysteine protease cut sites. We investigate whether or not the structure may occlude the cut site in one tissue, but not the other. Figures B.3 and B.4 display homology models highlighting the cut site residues, and showing no significant occlusion. Many other enzyme cut sites are predicted by PeptideCutter, some of which are affected by the sequence differences. Further analysis may lead to the development of differential enzymatic tests to distinguish between the two proteins.

**Figure B.1: Comparison of the predicted structures of the two porcine IL1β structures.**

(a) Predicted structure of mIL1B with Cscore 1.76 (b) Predicted structure of ayIL1β with C-score 1.65 (c) Superimposition of the predicted structures in (a) and (b); all atom RMSD = 0.48 Å
B.4. Discussion and conclusions

We have used I-TASSER to predict the structures of the two porcine IL1β proteins. As these two proteins have high sequence similarity (90%) with human IL1β proteins, for verification purposes, we predicted the structure for human IL1β using the ITASSER server and the RMSD for this predicted structure against the experimental structure of human IL1β is 0.332 Å which is quite low. This says that the structure predictions for the two porcine IL1βs are likely to be reliable. It is always useful to utilize predicted structures for building new hypotheses. In this case, we are able to postulate two possible hypotheses.

**Hypothesis 1**

In order to achieve their biological function, IL1β proteins need to be cleaved (303). We also found that for these proteins an enzyme that cuts IL1β sequences is ICE (IL1β converting enzyme) (303). The two porcine proteins, mIL1β and ayIL1β, have cleavage sites differing slightly in sequence - FVCD*ANVQ and FLCD*ATPV (* indicates cleavage site). It was experimentally found that the right side (ANVQ or ATPV) of a cleavage site does not play an important role, and therefore only left side is the important one (304). We can see that for these two proteins, the only such difference in the cleavage site is between leucine(L) and valine(V). In mIL1β sequence, this position is occupied by valine (V) and in ayIL1β it is leucine (L). The substrate activity research shows that V in this position is highly promoted, and the activity of mutants with any other amino acid in this position is significantly lower (~one order of magnitude) (305). It can be explained by a bigger size of the side chain of L in comparison with V, which can lead to worse packing of the binding site of ICE. More efficient cleavage of the precursor for the macrophage, mIL1β, would mean a greater abundance of the mature form.

**Hypothesis 2**

Embryos and macrophages are two different stages of the same organism. In each case, the set of performed biochemical pathways or their productivity could be different. So in such cases, there is some possibility to control the efficiency of the pathways. One possibility is to modify the activity of the enzymes by changing their binding sites which might inhibit such an enzyme’s activity or change its biological activity. Therefore, because the mIL1β and ayIL1β proteins differ in the 2nd predicted binding site and one of them is predicted to have a modified amino acid (133rd Proline in ayIL1β) in the same binding site; this can affect the selective activity between the two, which may be the reason why they are expressed differentially in these two cases.
Figure B.2: We superimpose onto 1ITBA the 5 best ITASSER homology models of each of the two pig IL1-β pre-cleavage protein sequences using the CE algorithm. The core of each model is colored red and the cleaved N-terminal portion blue. CASP1 cleavage sites are identified with the pre-side ASP is colored green and post-side ALA yellow. To emphasize their solvent exposure, the molecular surface of the cleavage site residues is shown.

Figure B.3: Structure analysis to determine if a differential proline residue induces changes that could prevent cleavage by CASP1. Above (left) we show the 5 M sequence models and (right) the 5 AY sequence models with P116 () as sticks. There does not appear to be a large effect on the CASP1 cleavage site by the introduction of this proline.
APPENDIX C. Structural regulation of kinase activity

Many biological pathways involve kinase signaling. Kinases are proteins that transfer phosphate groups from a donor (often ATP) to a substrate. The result of the transfer is often a signal that leads to changes in the cells’ metabolism. We focus on the Src and Tec tyrosine families which have a head-SH3-SH2-KINASE-tail domain structure (see Figure C.1) and distinct mechanisms of activation (306). Src kinase domain is intrinsically active. The tail domain contains Y527, which when phosphorylated, allows intramolecular binding of the SH2 domain, deactivating the kinase. Trp in the linker between SH2 and Kinase is required to mediate this allosteric transition. Two examples of Tec that will be analyzed here are Itk and Btk. Unlike Src, the Tec kinase domain is intrinsically inactive. They still require Trp in the linker between SH2 and Kinase and SH2 docking to the kinase domain to stabilize the active conformation. There appear to be key differences between Itk, Btk, and Tec: Itk extensively self-associates, which seems to turn off activity, whereas Tec and Btk do not self-associate in that manner and require different mechanisms of inactivation. The focus of this project is differentiating the activation and inactivation rules for Itk, Btk, and Tec.

Figure C.1: Domain structure of Itk kinase reproduced from (307).

We aim to explain the relationship between conformational variability and activity level for protein kinases through collaboration with the experimental lab of Dr. Amy Andreotti (Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University) and her graduate student Scott Boyken; it is significant for understanding cell signaling at the molecular level and two of the proteins of interest in the study are potential drug targets for autoimmune diseases.

One important aspect of this study is more accurately characterizing the active and inactive states. First, we define the major subdivision of the catalytic domain into four minimal parts based on internal conformational change accompanying activation. The solvent-exposed activation loop region contains a tyrosine residue, which can be (de) phosphorylated to modulate activity. A buried activation loop region, the α-C helix, and the N-terminal hinge are also important for determining the activity of the kinase. See Figure C.2 for details of the structure. Two recently determined Itk X-ray crystal structures have been
released that do not fully conform to canonical active or inactive forms. We aim to explain these structural differences, their origin, and their effect on kinase activity. Other structural characteristics are important for kinase stability and activation including a salt bridge between the body and the α-C helix, a set of residues which has become known as the “hydrophobic spine” consisting of a chain of hydrophobic stacking interactions (see Figure C.3), the N-terminal linker, and the “gatekeeper” residue. It has been shown that certain residues in the N-terminal linker are necessary for modulating the activity of the kinase by facilitating or enabling self interactions with the SH2 domain. The gatekeeper is a phenylalanine residue that, if mutated, can compensate for linker absence. It controls the shape of the substrate entry point. Therefore mutations to small amino acids can lead to intrinsic activation, while other gatekeeper mutations are intrinsically inactive, regardless of the phosphorylation state of the activation loop. Using computational approaches, we seek a more fundamental understanding of the relationship between structure and activity in these kinase families.

![Figure C.2: Itk structure with regulatory and functional regions highlighted – reproduced from (308).](image)

Subdivision of catalytic domain into four minimal moving parts based on internal conformational change accompanying activation. Solvent-exposed activation loop region (residues 412–424) in red, buried activation loop region (residues 403–411) in orange, αC helix (residues 299–320) in green, and the N-terminal hinge (residues 253–262) in blue.

Large-scale Replica Exchange Molecular Dynamics (REMD) simulations are proposed to analyze the stability of six kinase structures using XSEDE (the new generation of TeraGrid) resources (grant for time has been obtained). REMD is a specialized MD protocol that is used to increase the efficiency of conformational sampling. We will run simulations with and without the SH2 linker to probe the role it
plays in stabilizing the overall structure as well as propagating the activation signal. The effect of phosphorylation of the active loop tyrosine will also be tracked in independent REMD trajectories. Further, the long range electrostatic network (309) will be used to track the activation/inactivation state of each member throughout the simulation. Some initial results using in-house test simulations are displayed in Figure C.3. We find that the structural ensemble samples many states, including denatured states at moderate to high temperatures, and that these states have a relation to the hydrophobic spine and the orientation of the $\alpha$-C helix.

Figure C.3: Snapshots from a preliminary REMD trajectory on the Itk free-no-linker structure showing diversity in hydrophobic spine alignment and loop melting. The structure used is based on 3MJ2, but with the inhibitor removed. This structure lacks the linker region, which is believed to be important for stabilizing the hydrophobic spine (orange spheres) and therefore overall stability and activity. Gatekeeper F435 colored blue, salt bridge forming residues in red. A) Initial conformation shows partial spine alignment and lack of salt bridges. B) In this member of the REMD ensemble the $\alpha$C helix has moved out of position, destabilizing the hydrophobic spine. C) Here the hydrophobic spine is more tightly packed than the initial structure, but this has come at the cost of partial unfolding of the N-terminus.
APPENDIX D. RNA nanostructure assembly and verification

A collaboration with Dr. Bruce Shapiro (CCR Nanobiology Program, National Cancer Institute, Frederick, MD) and based on his paper with Grabow, et. al. (310)

Nanodesign of biomolecules is becoming increasingly popular and prevalent due to the many possible applications that it affords. These include the targeted delivery of drugs, DNA, or RNA (see Figure D.1), as well as the generation of structures (either protein or nucleic acid) with novel folds or functions that can be designed for specific purposes. One of the groups working in nanodesign is the Shapiro lab at the NIH. We have been collaborating with them to develop computational models that explain the assembly of RNA structures including nanorings. The Shapiro lab has developed a database of RNA parts – small structural units, the topologies of which are determined by the sequence. These parts can be assembled into many types of structures.

The first nanostructure that we considered was called the technosquare. From Atomic Force Microscopy images it was known that this four sided RNA frame properly assembled in vitro, but the computational models did not predict closure of the ring (see Figure D.2). Our dynamics models (ANM) were employed to model the loop closure (Figure D.2) as Molecular Dynamics simulations did not accomplish closure in a reasonable time. All-atom ANM is being employed to analyze the differences between rings with different numbers of monomers, as well as different monomer junction types (Figure D.3). We have also analyzed the effect of minor changes to the structure. Figure D.4 shows the dot product between modes of motion from two structural models – one where atoms from all bases are included (3 points per base), and another model where extended single stranded RNA tails were removed (the tails were present in the experimental structure). Little change in the dominant motions was seen upon tail removal. This differs from what is normally seen in globular proteins where extended termini lead to “tip effect” or the domination of the lowest modes by large amplitude non-physical motion of these termini.
Figure D.1: Schematic of possible uses for RNA nanorings in delivering therapeutic small interfering (si) RNA - reproduced from (310).

Figure D.2: The technosquare test case ANM showing ring closure using the dominant normal mode.

The initial structure is shown in (B) with the structure following the negative direction in (A) and the positive in (C). From (A), it is evident that the dominant normal mode will favor the ring closure event. Once the kissing-loop bases are brought into close proximity they will interact favorably and close the loop.
Figure D.3: Example of all-atom models of pentameric nanorings where ANM again shows ring closure, but is insensitive to alterations in the monomer junction interactions. Three all-atom models are tested. The first is a complete ring structure where no gaps have been added. Either one or three nucleotide gaps are added to the sequence by shortening the C-terminus resulting in a one or three nucleotide single stranded region in the middle of the monomer.

Figure D.4: Unlike globular proteins, individual modes of motion for the technosquare are not significantly affected by removal of extended termini. The heatmap displayed is the mode-mode dot product. For the ANM with tails, we compute the modes using all bases, but the component of each mode that corresponds to motion in the tails are removed and the mode re-normalized to unit length. The dot product is then computed.
APPENDIX E. Animating cellular and sub-cellular images and tomographic maps

The first cell model we study is displayed in Figure E.1 and is derived from a set of 13 cellular images taken as a vertical stack through a membrane leaf (data obtained from Federica Brandizzi, Michigan State University). A 3D model is derived from the known dimensions of the images and density information from the individual pixels. Any 5x5 set of pixels whose average density is above a threshold is assigned a point in our model. For image generation we assign each point an atomic radius of 2Å. The data is a rather stark cuboid from a larger structure (see Figures E.2 and E.3). We introduce a coarse graining threshold, T, which is used to determine which pixels are retained upon coarse graining. XY planes are denser than the Z-slices. To compensate for this we collapse 5x5 matrices of pixels to one pixel and retain them as a model node if the number of non-zero parent pixels is greater than T. Functional dynamics are observed in Figure E.4 where the leaflet displays pore opening and closing. The next stage in this project will model entire subcellular compartments such as the Golgi apparatus, described in Figure E.5.

<table>
<thead>
<tr>
<th>GNM(10)</th>
<th>GNM(12)</th>
<th>ANM(12)</th>
<th>ANM(15)</th>
<th>wpANM(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNM(10)</td>
<td>0</td>
<td>0.9610</td>
<td>0.6075</td>
<td>0.6086</td>
</tr>
<tr>
<td>GNM(12)</td>
<td>0</td>
<td>0</td>
<td>0.6240</td>
<td>0.6264</td>
</tr>
<tr>
<td>ANM(12)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.9997</td>
</tr>
<tr>
<td>ANM(15)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table E.1: Correlation of mean square fluctuation for 5 types of ENMs applied to the membrane leaflet using a coarse-graining threshold of 9 units.

Figure E.1: We display three views of the 3D model generated from cellular images. This membrane leaf is relatively flat with a slight concavity. This is the model used in the time-dependent ENM simulations and corresponds to a pixel density threshold of 9.
Figure E.2: The coarse graining threshold, T, is varied and the resulting 3D models are displayed to show the relative density of each region.

Figure E.3: Different modes of motion of the image. We plot the structure colored by $\log_{10}(\text{Temperature Factor})$ where blue are small temperature factors and red large. The GNM and ANM models are consistent within the same model type, but show differences between model types. The dominant modes for ANM models are bending and twisting of the structure orthogonal to the above view. The appearance of a linear hinge through the middle of the structure is due to the distribution of points; the structure is thicker along this region. The minor hinge implied by the white band on the upper left of the ANM models is evident in the slow modes due to the low connectivity and extreme motion seen in the handle.

Figure E.4: ANM motion of the structure shows the opening and closing of pores along the membrane leaflet.
Figure E.5: The next phase of this project will focus on sub-cellular structures, animating Golgi apparatus and endoplasmic reticulum tomographic models. These models have been made available by Alberto Luini (311) (the above image is reproduced from (311)) and Byung-Ho Kang (312).
APPENDIX F. Characterization of YAP and YBHB by Mass Spectrometry

Manuscript to be submitted.

Excerpts specifically relevant to M. Zimmermann’s contributions are presented here for brevity.


F.1. Abstract

The YbhB protein is conserved from bacteria to mammals and is believed to regulate cellular proliferation via mitogen-activated protein kinase, but remains poorly characterized. The YbhB protein of E. coli and its Arabidopsis thaliana homolog were analyzed by ion mobility mass spectrometry and matrix assisted laser desorption ionization mass spectrometry to determine the extent of multimer formation. Collision cross sections were measured for positive proteins ions in various charge states using native electrospray ionization (ESI) conditions in ammonium acetate buffer and compared to calculations using crystal structure data. In native ESI mass spectral experiments, the E. coli YbhB protein formed dimers readily, whereas the Arabidopsis thaliana homolog preferred monomers, yielding cross sections for the 12+ charge state of 2480 Å² and 2560 Å², respectively, and confirmed by gel-shift. The Arabidopsis YbhB homolog was also analyzed under acid denatured conditions; dimers were more abundant and the cross sections were greater than those obtained under native ESI conditions.

F.2. Manuscript Excerpts

The present work studies two related proteins, the YbhB protein of Escherichia coli and its homolog from the plant Arabidopsis thaliana, the protein encoded by the Arabidopsis gene locus, At5g01300. The exact biochemical and physiological functions of these proteins are unknown. Amino acid sequence alignment of these two proteins (NCBI database BLASTp) reveals that they belong to a highly conserved family of bacterial/archial phosphatidylethanolamine-binding proteins. The E. coli YbhB gene is located within the biotin biosynthesis operon. They are also related to their eukaryotic homolog Raf Kinase Inhibitor Protein (RKIP). The substrate of RKIP, Raf-1, in turn regulates mitogen-activated protein (MAP) kinase via MEK (MacDonald, 1993 in MCB and Gardner, 1994 in MBC). The MAP kinase cascade has been
shown to regulate many aspects of cellular function including growth, differentiation, and compartmentalization (Davis, 1993 in JBC). Thus, improved characterization of YbhB and its homologues has the possibility of producing new molecular targets for the treatment of MAP kinase associated diseases.

Here we report and discuss results from our home-built ion mobility TOF MS and a MALDI TOF MS to characterize these two homologous proteins. This is the first publication detailing any characteristics of the plant protein. The focus of the discussion covers a) m/z values of protein ions, b) extent of multimer formation under different solution conditions, c) the collision cross sections for monomer and dimer ions in various charge states, and d) comparison of gas-phase cross-sections to analogous data derived from published crystal structure of the E. coli YbhB protein dimer (313). These results are confirmed by gel-shift assays and computational predictions of collision cross-section.

F.2.1. Collision cross section measurements

<table>
<thead>
<tr>
<th>Mobcal Method</th>
<th>Calculated Cross Section (Å²)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Monomer</td>
</tr>
<tr>
<td>PA</td>
<td>1375</td>
</tr>
<tr>
<td>EHSS</td>
<td>1732</td>
</tr>
<tr>
<td>TM</td>
<td>1771</td>
</tr>
</tbody>
</table>

Table F.1: Calculated cross sections for YbhB protein from E. coli using crystal structure data. The values include the sizes of the hydrogen atoms.

Collision cross sections (Ω) were determined experimentally using the measured ion m/z values, IM drift times (t_d) and instrument parameters. The following equation was used to calculate collision cross section:

$$
\Omega = \frac{3}{16} \sqrt{\frac{2\pi}{\mu k_B T n L^2}} \frac{Q t_d U}{L}
$$

(F.1)

The mobility equation variables are listed in Table F.1. The number gas density, n, was calculated using the pressure and temperature of the gas in the drift tube. The mobility drift time, t_d, for the protein ions was between 1800 and 4500 µs. Initially, the measured t_d values included the time the ion spent between the exit of the mobility cell and the source region of the TOF analyzer. This “extra” time was calculated using the m/z value and the applied voltages and was subtracted from the total t_d value.
before the latter was used in Equation F.1. The correction to the total flight time ranged from 100 to 200 µs.

F.2.2. Collision cross section calculations

The Mobcal program (314, 315) was used to calculate cross sections for the YbhB protein from *E. coli* using crystal structure data (PDB 1FJJ) (313). Hydrogen atoms were included using CHARMM27+CMAP force field with psfgen from VMD version 1.9.

Values from Mobcal are shown in Table F.1 for the *E. coli* YbhB protein, for which the crystal structure is known (PDB 1FJJ). The calculations were done in several ways. The exact hard spheres scattering (EHSS) method generates 1730 Å² for monomer and 2930 Å² for dimer. These values are similar to those from the trajectory method (TM) and are plotted in a subsequent figure for comparison with experimental measurements. The projection approximation (PA) method generates values significantly lower than the other two, more rigorous methods, as observed by others. Values are given in Table F.1 and compared to the experimental values in Figure F.1.

![Figure F.1: Measured cross sections for YbhB protein monomer and dimer from *E. Coli*.](image)

The charge state axes are superimposed so that cross sections from overlapping mass spectral peaks are plotted at the same horizontal position. Mobcal values are for TM calculations including H atoms using the crystal structure 1FJJ.
F.3. Conclusion

The \textit{E. coli} YbhB protein formed mostly dimer by native ESI, which is consistent with the crystallographic results of Serre et al. (313). For \textit{E. coli} YbhB in low charge states, e.g., D10+, the cross section (2080Å$^2$) determined by IM is near the value of about 2400 Å$^2$ estimated from the x-ray crystal structure (313). Under native nESI conditions, the \textit{Arabidopsis} homolog of the YbhB protein is mostly a monomer. This suggests that solution phase protein-protein interactions in the \textit{E. coli} YbhB dimer are stronger than those in \textit{Arabidopsis} homolog of the YbhB protein dimer. Under acid denatured conditions such as MALDI both \textit{Arabidopsis} homolog of the YbhB protein and \textit{E. coli} YbhB form dimers and trimers. The cross sections of the denatured \textit{Arabidopsis} homolog of the YbhB protein are larger than the ones generated from native ESI conditions. The molar mass of the \textit{E. coli} YbhB is about 20200 Da and 40400 Da for the monomer and dimer, respectively. The \textit{Arabidopsis} homolog of the YbhB protein has molar mass about 20900 Da. The measured cross sections for the lowest charge states of the \textit{Arabidopsis} homolog of YbhB provide quick measurements of the overall size of the molecule that could be useful to improve structure prediction in molecular dynamics calculations or the determination of molecular assemblies. The stoichiometric preferences of these two homologues may point to subtle differences in the up-stream regulation of the MAP kinase cascade between \textit{E. Coli} and \textit{A. Thaliana}, which may impact the understanding of mammalian disease states affected by this cascade.