UNIFYING FRAMEWORK FOR THE PREDICTION OF PROTEIN FOLDING PATHWAYS AND TERTIARY STRUCTURE FROM PRIMARY SEQUENCE

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY
AASHISH NATH ADHIKARI

CHICAGO, ILLINOIS
DECEMBER 2012
Copyright © 2012 by Aashish Nath Adhikari

All Rights Reserved
To my parents,

Mahendra Nath Adhikari and Padma Adhikari
# TABLE OF CONTENTS

LIST OF FIGURES ................................................................. vii

LIST OF TABLES ........................................................................ ix

ACKNOWLEDGMENTS ............................................................... x

ABSTRACT .............................................................................. xiv

1 BACKGROUND AND BASICS .................................................. 1
   1.1 What makes the protein folding problem challenging? ................. 5
   1.2 Some essential concepts .......................................................... 9
       1.2.1 Reduced representations, Ramachandran Maps and Nearest neighbor effects ................................................................. 9
       1.2.2 Statistical Potentials ............................................................ 12
       1.2.3 Sequential Stabilization ....................................................... 14
   1.3 Putting it all together: Ingredients of a protein folding algorithm ....... 15
   1.4 Some useful formalisms regarding protein structure ....................... 18
   1.5 Organization of Thesis ............................................................ 19

2 DE NOVO PREDICTION OF PROTEIN FOLDING PATHWAYS AND TERTIARY STRUCTURE USING THE PRINCIPLE OF SEQUENTIAL STABILIZATION .............. 21
   2.1 Introduction ........................................................................... 22
   2.2 Approach .............................................................................. 24
   2.3 Methods ............................................................................... 32
       2.3.1 Monte Carlo Simulated Annealing (MCSA) Minimization ........ 32
       2.3.2 Sampling using Ramachandran Maps ..................................... 33
       2.3.3 Move Sets ........................................................................ 34
       2.3.4 Secondary structure fixing .................................................... 35
       2.3.5 Tertiary structure fixing ....................................................... 36
       2.3.6 Energy Functions ................................................................ 37
       2.3.7 Biasing Energies used for sequential stabilization .................... 41
       2.3.8 Software .......................................................................... 41
   2.4 Results .................................................................................. 42
       2.4.1 Folding of Lambda repressor (λ6-85) ..................................... 42
       2.4.2 Immunity proteins Im7, Im9, Slm9 ........................................ 47
       2.4.3 Prediction of early events, foldons and non-native contacts ........ 51
   2.5 Comparison to all atom molecular dynamics simulations ................ 61
   2.6 Discussion ............................................................................ 68
   2.7 Conclusion ............................................................................ 71


3 INVESTIGATIONS OF NON-NATIVE INTERACTIONS IN FOLDING OF PROTEIN L AND PROTEIN G
3.1 Protein L
  3.1.1 Introduction
  3.1.2 Overview of experimental results
  3.1.3 Hairpin simulations and TS modeling of protein L
  3.1.4 Discussion
  3.1.5 Comparison to $\phi$ analysis
  3.1.6 Conclusion
3.2 Protein G
  3.2.1 Hairpin simulations: Preliminary Results
3.3 Comparative analysis of protein L, protein G and DesG: TerItFix predictions
  3.3.1 TerItFix Observations Summary
  3.3.2 TerItFix recommended experiments
  3.3.3 Explanation of non native N terminus hairpin in protein G

4 MODELING LARGE REGIONS IN PROTEIN STRUCTURES
4.1 Introduction
4.2 Approach
4.3 Methods
  4.3.1 Ramachandran Map (Pivot) Move Set and Sampling
  4.3.2 Energy functions
  4.3.3 Scoring
  4.3.4 MCSA Simulation Procedure
  4.3.5 Parallel Scripting
4.4 Results and Discussion
  4.4.1 Loops in crystal structures
  4.4.2 Ends in crystal structures
  4.4.3 CASP9 blind InsEnds predictions
  4.4.4 Blind Prediction of refinement targets in CASP9
  4.4.5 Molecular Replacement Results for CASP9 refinement targets
  4.4.6 Global InsEnds RMSD vs. local InsEnds RMSD
  4.4.7 Applications to protein folding simulations
  4.4.8 Simultaneous Folding of multiple InsEnds
  4.4.9 Protein Structure prediction pipeline
4.5 Conclusions

5 SUPPORT VECTOR MACHINES FOR ENERGY FUNCTION OPTIMIZATION
5.1 Support Vector Machines: Theory
5.2 Energy function optimization as a SVR problem
5.3 Results
  5.3.1 Decoy generation
  5.3.2 Individual Proteins
  5.3.3 Protein Classes
5.4 Some Comments
5.5 Conclusions ................................................................. 162

6 FUTURE DIRECTIONS AND CONCLUSION .......................... 163
6.1 Improvements to TerItFix for prediction of larger proteins ............ 163
   6.1.1 Multiple levels of local structure biases .............................. 164
   6.1.2 Clustering Before Averaging ......................................... 166
6.2 Extending TerItFix to obtain folding kinetics ............................ 169
   6.2.1 Constant Temperature Metropolis Monte Carlo ...................... 169
   6.2.2 Markov State Models ............................................... 173
6.3 Protein folding: Looking forward ........................................ 174
6.4 Conclusions .............................................................. 179

REFERENCES .................................................................. 181
## LIST OF FIGURES

1.1 Dichotomy in protein folding and protein structure prediction methods ............................................. 4
1.2 A Ramachandran map depicts the backbone dihedral angles $\phi, \psi$ of a protein's amino acids ................................................................. 11
1.3 Essential components of a protein folding algorithm ................................................................. 17
1.4 Contact map representation of a protein ............................................................................................... 18
1.5 Ramachandran basin occupancies from the protein amino acid sequence ........................................ 19
2.1 A folded protein has to optimize three competing properties together ........................................... 23
2.2 Flowchart outlining TerItFix algorithm ................................................................................................. 26
2.3 TerItFix protocol applied to Ubiquitin .................................................................................................. 27
2.4 Restriction of the backbone sampling distribution. ............................................................................... 28
2.5 Two kinds of backbone torsional angle moves are used in TerItFix .................................................. 29
2.6 The energy functions used in TerItFix. ................................................................................................. 31
2.7 Inclusion of a C$\beta$ multibody burial statistical potential improves structure prediction over pairwise additive DOPE-PW statistical potential .............................................................. 40
2.8 TerItFix prediction of $\lambda_{6-85}$ ......................................................................................................... 44
2.9 Evolution of inter-helical contacts in $\lambda_{6-85}$ as a function of TerItFix rounds .............................. 45
2.10 TerItFix prediction of Im proteins. ........................................................................................................ 48
2.11 Basin propensities and average contact maps in the last round of TerFix folding of Im proteins. ............................................................................................................................................... 49
2.12 TerItFix prediction of folding intermediates in Im proteins. ................................................................ 50
2.13 TerItFix prediction of Ub, CI2 and $\alpha^3d$ ......................................................................................... 52
2.14 Folding pathway for Ub. ......................................................................................................................... 54
2.15 Folding pathway for CI2 ......................................................................................................................... 56
2.16 Non-native hairpin formation in protein L ............................................................................................ 57
2.17 Folding pathway for proteinL ............................................................................................................... 58
2.18 Folding pathway for $\alpha^3D$ ............................................................................................................... 59
2.19 Folding pathway for protein A ............................................................................................................. 60
2.20 Lowest RMSD structures from TerItFix predictions of 12 fast folding proteins compared to DESRES all-atom molecular dynamics simulations ................................................................. 62
2.21 Comparison of DESRES pathways to TerItFix predicted pathways (Trp-cage, BBA, villin). ................................................................................................. 64
2.21 Comparison of DESRES pathways to TerItFix predicted pathways (WW, NTL9, BBL, protein B). ................................................................................................. 65
2.21 Comparison of DESRES pathways to TerItFix predicted pathways (Homeodomain, protein G, $\alpha^3d$, $\lambda$ repressor). ................................................................................................. 66
2.22 Transition state ensemble for Protein L and Protein A selected by clustering structures with $0.6 < RCO/RCO-Native < 0.8$ ......................................................................................................................... 72
3.1 Transition state of protein L according to $\phi$ analysis vs $\psi$ analysis .................................................. 79
3.2 Rate of $\beta$ hairpin formation in protein L. The carboxy-terminal hairpin forms in fewer simulation rounds than the amino-terminal hairpin ................................................................................................. 85
3.3 Simulations of the protein L $\beta$ hairpins using TerItFix and amm-atom molecular dynamics. ............................................................................................................................................... 86
3.4 Native (Tube) vs predicted (Trace) overlay of the C terminal hairpins for protein L. ................................................................. 87
3.5 TSE determined by $\psi$ analysis is more ordered than that identified by $\phi$ analysis. 91
3.6 Simulations of protein G hairpins using TerItFix and all-atom molecular dynamics 97
3.7 Comparative TerItFix predictions of protein L, protein G and DesG (redesigned protein G) ................................................... 99
3.8 Both native and non native turns are possible for N terminus hairpin of protein G 103

4.1 The InsEnds modeling problem .................................................. 109
4.2 The InsEnds local protein structure prediction algorithm ..................... 112
4.3 The ligation terms close the loop. .................................................. 113
4.4 Inclusion of SASA to pairwise DOPE-PW statistical potential improves decoy selection ...................................................... 120
4.5 Selection of the top 5 loop predictions for 1xnb. ................................. 121
4.6 CASP9 InsEnds blind predictions ............................................... 128
4.7 InsEnds predictions for CASP9 refinement targets ............................. 132
4.8 Global versus local RMSD. ......................................................... 137
4.9 InsEnds algorithm applied to protein folding pathways ........................ 140

5.1 Support vector machines maximize the margin to separate the data ........ 147
5.2 Support vector machines map the input data into higher dimensional space to classify linearly inseparable data ............................... 150
5.3 Support vector regression framework for energy function optimization. 153
5.4 Correlation of true RMSD to the SVM predicted RMSD and DOPE-PW energy in $\alpha$ proteins ................................................. 157
5.5 Correlation of true RMSD to the SVM predicted RMSD and DOPE-PW energy in $\beta$ proteins ................................................. 158
5.6 Correlation of true RMSD to the SVM predicted RMSD and DOPE-PW energy in $\alpha/\beta$ proteins ........................................... 159
5.7 Correlation of true RMSD to the SVM predicted RMSD in different protein classes ............................................................. 161

6.1 Tertiary fixing can improve structure prediction even when all secondary structure is fixed ......................................................... 166
6.2 Clustering the low energy structures before obtaining average contact maps improves bias signal .............................................. 168
6.3 Constant temperature biased Metropolis Monte Carlo simulation to connect foldons in ubiquitin ............................................ 172
6.4 Obtaining the optimal temperature for Metropolis dynamics .................. 173
6.5 Most proteins fall somewhere in between on a spectrum that ranges from topology dominated to sequence dominated folding behavior 176
6.6 Changes in the primary sequence can affect the folding landscape in different ways 178
LIST OF TABLES

2.1 Parameters used in the TerItFix simulations .......................... 37

4.1 Prediction of 8-12 residue loops in crystal structures - Local RMSD (Align Loop, RMSD of Loop) .............................................. 122

4.2 Prediction of 8-12 residue loops in crystal structures - Global RMSD (Align rest, RMSD of loop) .............................................. 123

4.3 Prediction of 20 end residues in crystal structures - Local RMSD (Align ends, RMSD of ends) ......................................................... 125

4.4 Prediction of 20 end residues in crystal structures - Global RMSD (Align rest, RMSD of ends) ......................................................... 126

4.5 Blind InsEnds prediction of refinement targets in CASP9 .............. 130

4.6 Prediction of InsEnds in CASP8 structures (post-diction) - Local InsEnds RMSD (align InsEnds, RMSD of InsEnds) .......................... 135

4.7 Prediction of InsEnds in CASP8 structures (post-diction) - Global InsEnds RMSD (align rest, RMSD of InsEnds) .......................... 136

4.8 Prediction of InsEnds in CASP8 structures (post-diction) - Global RMSD (align all, RMSD of all) .............................................. 136

5.1 Pearson Correlation Coefficients of true RMSD values to DOPE-PW energy function as well as SVM predicted RMSD values ..................... 156
ACKNOWLEDGMENTS

When I started graduate school, I knew very little about proteins or the protein folding problem, besides the fact that it was a very challenging scientific problem. As it would to a naively ambitious first-year graduate student, a chance to tackle a problem that had confounded scientists for 50 years seemed like the perfect project to undertake. It didn't take a very long time for me to be thoroughly humbled by the fact that a lot of very smart people before me had developed, tested and analyzed a lot of the ideas that I was just beginning to develop and formulate. It was during numerous moments like those throughout my graduate school career, where it seemed like I had exhausted all possible ways of looking at a problem, that Tobin's flurry of insights and Karl's infinite wisdom almost always came to my rescue. Perhaps, more importantly, after most of our conversations, I would usually emerge with a rejuvenated sense of purpose and drive to keep pushing through and moving forward, in part thanks to Tobin's contagious passion and enthusiasm as well as Karl's encouraging words and sense of humor. The two of them have been my academic and scientific role models and I shall be forever grateful to them for their guidance and support throughout graduate school.

I would also like to thank Prof. Aaron Dinner for being part of my thesis committee and providing helpful suggestions regarding the thesis. I thank Prof. Greg Engel and Prof. David Mazziotti, whose classes in the first year provided me with some foundational background for topics in graduate level Chemistry. I want to thank Mike Wilde and Glen Hocky for introducing me to supercomputing and Mike for helping me implement my programs in
various supercomputing architectures. I thank Jinbo Xu and Jian Peng for providing the homology modeling data for InsEnds improvement.

Over the years in the Freed/Sosnick group, I’ve had the opportunity of knowing and working with some of the brightest people I know. I would like to thank the current members of the group (Dr. Chloe Antoniou, Dr. Michael Baxa, Alex French, Dr. Esmael Haddadian, James Hinshaw, John Jumper, Dr. Jack Skinner, James Townsend, Dr. Wookyung Yu, Josiah Zayner) as well as former members (Dr. Andres Colubri, Dr. Joe Debartolo, Dr. Haipeng Gong, Glen Hocky, Dr. Jouko Virtanen, Tae Yeon Yoo), who have taught and helped me in various ways. I have benefited immensely from my interactions with the experimental side of Tobin’s lab, who helped me align and orient my computational work towards directions that are experimentally testable, which were crucial for comparing my predictions with nature. I’ve learnt a lot about programming as well as problem solving from my interactions with Dr. Marc Parisien and I thank him for that.

Given the single-minded focus on research usually typical of graduate school, time has a way of blurring in terms of not distinguishing itself a week from the week before, a month from the month before and even a year from the year before. Reflecting back, the only way I seem to be able parse through the years and add graduations is through the various interactions I’ve had with the people I’ve met here. My life at the University of Chicago has been enriched by the friendships I’ve developed here. I want to thank Jay Foley, Kelly Fransted, Dugan Hayes, Gaby Avila-Bront, Alex Dickson, Toan Huynh, Mark Maienschein-Cline, Bo Qi, Pavlo Zolotavin, Srikant Veera Raghavan, Raman Shah and others for making
my life enjoyable and supporting me in various ways. As I type this sitting at a table in
the Cobb coffee shop, I remember those countless Cobb coffee runs with Jay and Kelly over
the years. Several fragmented memories ranging from potluck dinners to all-night problem
set sessions and pub wings to playing squash and happy hours to office hours, etc. all now
serve as markers through time. I want to thank Evan Miller for the insightful conversations
on how to positively change the world over countless whiskey/sodas at Jimmy’s. Thanks to
Aashish Jha for continuous encouragement and suggestions for extending my work into new
frontiers. Thanks to Arjun Sharma for years of friendship and shared adventures. Thanks
to Dan Herbst, Dave McCowan and Jouko Virtanen for being awesome officemates.

Music has been an important part of my life in Chicago. Jay, Dugan and Kelly have played
an integral part in that regard by engaging me in conversations about music, sharing new
music as well as accompanying me to several live music events in Chicago. I also immensely
enjoyed playing music with Jay, Dugan and Alex - it certainly was one of the highlights of
my time in Chicago and I wish we could have played more sessions together.

I also want to acknowledge people who have positively influenced my life before my time
in Chicago. I thank my mentors at Williams College, especially Prof. Dieter Bingemann and
Prof. Jay Thoman, who sparked my interest in Chemistry and more generally in scientific
research. My education at Williams, thanks to Profs. Willam Wooters, Duane Bailey, Lois
Banta, Frank Morgan, Jim Teresco and several others, helped me develop some foundational
skills regarding learning and thinking about ideas and problems, which I find indispensable
to this day. Also at Williams, I developed some lifelong friendships with some incredible
people, including Mike, Olivia, Kristof, Daumantas, Mascha, Upton, Edwin, Arjun and countless others which I continue to cherish.

Special thanks goes out to my fellow Electrickers, Suyog Shrestha and Asim Shrestha for years of continued friendship, support and inspiration. Our high school science project to demonstrate the photoelectric effect using some hacked together materials in Nepal might have not been successful, but it certainly successfully sparked the spirit of exploration and experimentation that continues to drive us today. I want to thank Amrit Tuladhar, Bigyan Bista and Abishkar Shrestha as well for their friendship and engaging conversations over the years.

Finally, I want to acknowledge the crucial and continuous role the support from my family has played towards the journey in which I find myself. My younger brother, Abhishesh, has been a constant positive presence throughout my life. To this day, he is the first person I instinctively turn to as a sounding board for matters ranging from the most menial to the most meaningful and I’m very thankful for that. And both of us owe most of our accomplishments to our parents, who always saw the value of a good education and toiled away towards providing us exactly that. I have learnt valuable lessons from them about handling hardships with resilience and grace. I shall be eternally grateful to them for their continued love, support, inspiration and encouragement throughout my life.
ABSTRACT

While we understand much more about today about how proteins fold than we did 50 years ago when the problem was first introduced, we still lack a reliable and general predictive framework to simultaneously predict both a protein’s tertiary structure and its folding pathway starting from its amino acid sequence alone. Most methods that successfully predict what a protein folds to (structure prediction) provide no information about how it gets there. Alternatively, most methods that are designed to predict how a protein gets there (protein folding mechanism) do so by using prior knowledge of the final structure. This dichotomy, however, doesn’t exist in nature because the tertiary structure of the protein is a consequence of the process of protein folding and the two are therefore coupled.

Motivated by this, I have developed a method, termed TerItFix, that unifies the prediction of a protein’s folding pathways and secondary/tertiary structure from only its primary sequence by applying basic principles of protein folding. The underlying algorithm is inspired by real proteins which fold via sequential stabilization of secondary/tertiary structure elements (foldons). Using a combination of single pivot and local crankshaft backbone torsional moves on a Cβ-level polypeptide chain, I perform multiple rounds of Monte Carlo simulations that involve progressive learning and building of tertiary motifs through contact and hydrogen bond matrix constraints.

TerItFix can predict the tertiary structures of several proteins using only their amino acid sequence to within 5 Å RMSD to their native structures. More notably, I demonstrate that the evolution of contact matrices in the TerItFix simulations capture the iterative fixing of
stable tertiary structure motifs, which in multiple cases are consistent with the major events along experimentally observed folding pathways (Ubiquitin, CI2, λ repressor). Since no prior knowledge about the either the secondary or tertiary native structure is assumed, TerItFix is able to characterize important kinetic traps as well as obtain the structures of intermediates in multiple cases, again supported by experiments (Im7, Im9, SI9). At the same time, I use TerItFix to fold 12 fast-folding proteins that were recently studied using all-atom molecular dynamics simulations and show that TerItFix is able to obtain comparable accuracy in terms of information about their tertiary structures and folding mechanisms, while requiring orders of magnitude less computational resources.

I also present TerItFix simulations of protein L and protein G, two proteins with identical topologies. TerItFix predictions of both these proteins suggest an important role of non-native interactions, in contrast to previous native-centric simulations and mutational experiments. TerItFix’s prediction of a non-native C terminus hairpin in the transition state of protein L is consistent with the recent experimental ψ analysis studies. Alternatively, TerItFix’s observation of a possible non-native N terminus hairpin geometry in protein G is an active prediction that needs to be experimentally tested.

I also apply similar principles used in TerItFix for structure prediction of large local regions in proteins, including loop regions in crystal structures as well as insertion regions in homology models. By refolding the local regions onto the rest of the structure using backbone torsional moves and Cβ-level statistical potentials, I show that we can predict the structures of much larger loop regions than existing loop modeling methods can handle. As a further
validation, our local structure prediction was ranked on of the best in the CASP9 tournament for improving local regions in homology models for crystal structure determination.

I also briefly discuss how the use of machine learning techniques can help extract better information from the data generated using our simulations. Particularly, I show that use of support vector machines can optimize our statistical potentials to help select better protein models. Finally, I present some ideas for extending the current work to tackle larger and more complex protein topologies as well as to obtain more detailed information regarding folding kinetics.
CHAPTER 1

BACKGROUND AND BASICS

“You see, proteins, as I probably needn’t tell you, are immensely complicated groupings of amino acids and certain other specialized compounds, arranged in intricate three-dimensional patterns that are as unstable as sunbeams on a cloudy day. It is this instability that is life, since it is forever changing its position in an effort to maintain its identity—in the manner of a long rod balanced on an acrobat’s nose.” - Isaac Asimov, Pebble in the Sky

The genetic information of life is transferred in a unidirectional process from DNA to RNA to proteins, as indicated by the central dogma of molecular biology [1]. Built from a linear chain comprised of twenty common naturally occurring amino acids, proteins are biomolecules whose role in life is ubiquitous, demonstrated through a wide range of functions that include translation, regulation, signaling, transport, catalysis and support. These functions however often require very well-defined three dimensional structures. Perhaps even more interestingly, for any given protein, the chemical information defined by its sequence of amino acids can uniquely determine the functionally relevant structure for that particular protein.

Given only a linear sequence of amino acids, how a protein attains a well-defined three dimensional or tertiary structure has been one of the fundamental puzzles in biology. While being able to predict the tertiary structure of a protein from its sequence (structure prediction) can provide valuable insights about its function, the understanding of how it folds to that particular structure (folding mechanism) can provide clues to protein misfolding and
aggregation, which are often implicated in various amylogenic diseases. Obviously, the two problems of structure prediction and folding mechanism are intertwined in nature. Given the biological importance coupled with the complexity of the problems, over the last two decades, numerous computational and theoretical approaches have been developed to tackle these issues, although generally separately (Figure 1.1).

Homology-based methods, which harness the power of evolutionary relationships present in the ever-growing databases of protein sequences and their corresponding structures have been the most successful in the prediction of tertiary structure from primary sequence[2]. In the absence of existing fragments or already-solved structures in the databases, however, most of these methods are handicapped, as best evidenced by the lack of methods able to tackle the free modeling targets in CASP9[3], a recent iteration of the biannual community wide assessment of structure prediction methods. Homology-free structure prediction methods, which are aimed at prediction of the tertiary structure without the use of homology information in existing sequence and structural databases, have not been able to perform significantly better on most of the challenging free modeling targets either [4]. While most structure prediction methods are not expected to predict folding mechanisms, the fact that most of these fail in the absence of homology perhaps highlights a deeper problem, namely their inability to capture the underlying physiochemical principles that govern protein folding.

Alternatively, there have been continual efforts to develop theoretical models that aim to extract general principles regarding how proteins fold, the earliest of which are inspired
by simulations of polymers on lattices. Guided by the notion that proteins fold on funneled energy landscapes with minimal energetic frustrations [5] and the observation that the native topology correlates strongly with protein folding rates [6], the majority of these so-called Gō-like methods [7] make the fundamental assumption that protein folding is driven by native interactions [8, 9, 10]. Hence, these native-centric simulations not only assume the knowledge of the final tertiary structure but proceed through a strong energetic bias to that native state. Whereas the landscape might describe the gross features of folding, its predictive power is limited by the fact that the ruggedness in landscape often shapes a particular protein’s folding mechanism and is required to capture the kinetic traps that might be specific to that protein. Moreover, in recent years, the importance of non-native interactions have been realized [11] and slight changes in the amino acid sequence have been known to drastically alter the folding properties in various cases [12, 13]. Although variants of Gō models exist that supplement it with sequence dependence as well as detailed representations [14, 15], they still are not fully able to overcome the inherent native bias in these models.

A third approach to the problem involves atomistic molecular dynamics simulations (MD), which in recent years has been able to use the increase in computing power to fold multiple proteins to their native states [16]. Although most of the proteins that MD is able to fold have simple topologies and are naturally fast-folding, the results are impressive nonetheless considering that these simulations make no prior assumptions. However, as was recently pointed out, challenges still exist for MD simulations to properly tackle the protein folding problem [17]. In particular, issues with accuracy of the force fields still remain,
Figure 1.1: Dichotomy in protein folding and protein structure prediction methods. Computational methods that specialize in obtaining folding mechanism generally assume knowledge of the native structure. Alternatively, methods that specialize in prediction of the native state from sequence in general do so without any explicit consideration of the folding mechanism.
while slow-folding and larger proteins are still out of reach. Regardless, extracting the general principles of protein folding from these simulations still requires a level of abstraction that involves going beyond the mechanistic details of the observed folding process of specific proteins in the MD trajectories, which may not be straightforward.

Given what we understand about the basic physical principles underlying protein dynamics from theory and experiments developed in the last five decades, one could reasonably expect that it should be possible to devise a computational framework that can predict both the folding pathways and tertiary structure starting from the primary sequence. My work presented in thesis describes such an approach, motivated by the dichotomy in existing protein folding and structure prediction methods, as outlined above and in Figure 1.1.

1.1 What makes the protein folding problem challenging?

The protein folding problem has gained notoriety as one of the most challenging problems in structural biology and over the years attracted researchers from areas as diverse as physics, chemistry, biology and computer science. This often results in the challenges in protein folding research being framed from different perspectives. While this thesis primarily describes methods to investigate the protein folding mechanism computationally by encoding various aspects of protein physics and chemistry, it will be useful to briefly consider the various ways in which the protein folding problem is considered challenging. Particularly, an understanding of which challenges are due to the limitations of current experimental, theoretical and computational techniques and which ones are fundamentally tied to the process of folding.
itself can help devise better methods for tackling the problem.

- **Conformational search.** One of the fundamental reasons why protein folding is hard is due to the astronomical number of possible conformations available to a protein in the absence of well-defined pathways, articulated by Levinthal’s paradox\[18\]. It would be impossible for a protein, in vivo or in vitro, to find its native state in microseconds to seconds simply through an unbiased or random search of the conformational landscape, given its vastness. Hence, it is reasonable to assume that without a good search strategy to explore the conformational landscape, attempts to find the native state starting from the denatured state will likely be impossible.

- **Delicate balance of forces.** In their natural state, proteins are embedded in water. With this in mind, it is relevant to note that the twenty amino acids, while all chemically different, can be classified broadly into polar and non-polar types. Additionally, each amino acid contains a N-H and a C=O group capable of hydrogen bonding to water. A folded protein is generally only marginally stable (\(\Delta G \sim 5-10\) kcal/mol) compared to the denatured state. A consequence of this is that a very delicate balance of forces is maintained between the native (folded) and denatured (unfolded) states of a protein, which provides another reason why protein folding is fundamentally challenging. From a perspective of energetic benefits, in the denatured state, water-protein hydrogen bonding is prevalent whereas in the folded state, protein-protein hydrogen bonding to form stable secondary structural elements is prevalent. From an entropic perspective, the protein chain loses configurational entropy when it goes from a de-
natured to a folded state, since the denatured state is free to move in the aqueous environment. Another relevant force often discussed in context of folding is the hydrophobic effect, which describes the exclusion of water molecules by aggregation of non-polar entities. In a folded protein, this often results in a hydrophobic core characterized by the burial of the non-polar amino acids. Quantifying the contributions of these various competing forces relevant in protein folding can be very difficult. This is especially relevant for computational folding simulations where one needs to design force fields and potential functions to accurately model the various interactions. A properly calibrated force field that can accurately characterize both the denatured and native state of a protein is still an active area of research.

• Technical issues. Besides the fundamental challenges discussed above, there are issues which complicate experimental or computational investigations of the folding process. For example, a lot of small, globular proteins fold cooperatively with two-state kinetics, which means the kinetics is usually determined by a single barrier in the free energy surface. As a consequence, during the process of folding, the protein must first climb uphill in free energy to the transition state before it can get to the native state. While stable intermediates can be populated on the native side of the barrier and studied using experimental techniques like native state hydrogen exchange, the early folding steps leading to the transition state are very hard to characterize experimentally and novel experimental techniques are warranted. Analogously, an example of a challenge computational methods face would be balancing the tradeoff between the
level of representation or sampling with accuracy of the information obtained. While all-atom molecular dynamics simulations represent all the atoms in the system including water and are generally what other simulations are benchmarked against, they can be computationally slow and can’t yet be used to study slower folding proteins. On the other hand, by formulating reduced representations of the system, approximate potentials as well as faster sampling techniques, useful (albeit approximate) information can be obtained in systems inaccessible to molecular dynamics.

- **Methodological differences and biases.** A final barrier towards resolving some of the existing challenges in protein folding have to do with the contradictory conclusions obtained for various aspects of protein folding when different techniques are used, both experimentally and computationally. For example, small-angle X-ray scattering and single-molecule FRET spectroscopy produce highly divergent views of the low-denaturant unfolded state [19] with SAXS observing no initial collapse during folding and SM-FRET arguing for an initial collapse. Similarly, complementary experimental techniques to study the transition state of proteins, $\phi$ and $\psi$ analysis produce contradictory pictures for the transition states of various proteins [20, 21]. $\phi$ analysis, which investigates the structural content in transition state of proteins through effects of various mutations on the transition state, seems to underestimate the structural content in folding transition states of multiple proteins compared to $\psi$ analysis, which infers the structural content of the transition state using metal binding properties. Analogously, various theoretical methods and computer simulations can produce diverging
results regarding the folding of the same proteins. For example, different computational folding studies of a five helix bundle protein, λ repressor disagree on the folding mechanism [22]. In fact, molecular dynamics simulations of the protein performed by three different research groups [16, 23, 24] propose contradictory mechanisms. These examples provide a note of caution that while a wealth of data from wide range of experimental and computational protein folding methods exist, one should carefully examine and understand the biases in interpretation arising from peculiarities of the different methods.

1.2 Some essential concepts

In this section, I outline three concepts which will I be using in later chapters to tackle some of the challenges raised in the earlier section.

1.2.1 Reduced representations, Ramachandran Maps and Nearest neighbor effects

The level of detail at which a protein is represented in a simulation is crucial to the information that can be obtained from it. In an ideal scenario, one would represent the system in the highest level of detail by including all the atoms and their interactions. However, increasing the detail of representation typically increases the associated computational cost and complexity of the problem as well. Thus, the right choice regarding the level of representation in a simulation should to be determined by the questions one intends to address.
In the work described in this thesis, I will use a Cβ level representation of the protein, where all the backbone atoms are present and the sidechain of each amino acid is represented by a single Cβ atom. The chemical information lost by removing the sidechain atoms beyond the Cβ atom is compensated by treating the pairwise and multibody interactions of the Cβ atoms of the 20 amino acids separately, discussed in later chapters. Because the protein backbone is treated explicitly, the intramolecular hydrogen bonding that stabilizes the various secondary structural elements is possible. Similarly, the Cβ level energy function allows for implicit treatment of the solvent by ensuring proper burial and packing of the different kinds of amino acids.

To sample the conformational space, for each amino acid, I allow two degrees of freedom through their backbone dihedral angles $\phi, \psi$. Developed from the original insights by Ramachandran that any amino acid can adopt only certain regions of its $\phi, \psi$ space (thus called Ramachandran plots/maps) largely due to steric constraints in a protein chain, this sampling space greatly restricts the conformational space that can be sampled. Figure 1.2 shows the preferred regions of the Ramachandran space using a heatmap obtained by plotting the $\phi, \psi$ angles of all amino acids in known high resolution crystal structures from the Protein Data Bank (PDB) database.

One of the most important insights regarding Ramachandran maps that I will use in the subsequent chapters is the notion that an amino acid’s conformational preferences is influenced by the chemical identity and conformation of its immediate neighbors [25, 26]. These
Figure 1.2: A Ramachandran map depicts the backbone dihedral angles $\phi, \psi$ of a protein’s amino acids. In this case, the plot shows the distribution of dihedral angles when all amino acids from folded proteins in the Protein Data Bank structural database are taken together.

nearest-neighbor (NN) identity and conformation dependence is utilized in our simulations by pregenerating Ramachandran distributions for all 8000 possible amino acid triplets (central residue and its two neighbors) from a high resolution dataset of crystal structures. These Ramachandran distributions can identify the conformational preferences of each amino acid in a sequence and therefore can provide a natural way to sample the conformational space. I will utilize two kinds of moves in the space of these NN Ramachandran distributions: a global “pivot” move as well as a local “doublecrank” move (introduced in Chapter 2) to generate new proposals inside the simulations discussed in later chapters.
1.2.2 Statistical Potentials

The Protein Data Bank (PDB) contains experimentally determined native structures of more than 6000 proteins. Statistical potentials were developed with the aim of harnessing the wealth of information regarding preferred chemical interactions in the database of folded protein structures. Specifically, using the assumption that the protein’s native state usually corresponds to the lowest free energy state, statistical potentials are parameterized functions which reflect the statistics of certain features in the PDB database. The use of frequency of contacts in X-ray structures to extract effective interactions originated decades ago [27]. Miyazawa and Jernigan in 1985 used the quasi-chemical approximation that contact pair formation resembles a chemical reaction to obtain effective inter-residue contact energies [28]. Perhaps the most successful variation was due to Sippl [29], where he used probability distribution of pairwise distances in a database of crystal structures and using analogy from statistical physics of liquids, to obtain a “potential of mean force”. One crucial assumption Sippl makes is that the distribution of pairwise distances (or other structural features) in a sample of native structures follow a Boltzmann distribution. The validity of such an assumption is a topic of debate since the native structures in the PDB database are not part of a canonical ensemble, as required for the Boltzmann assumption. Furthermore, to normalize the probability distribution a proper reference state needs to be defined. While some fundamental theoretical issues remain regarding the soundness of assumptions in statistical potentials, over the last decade they have proven to be one of the most successful formalism for approximating free energies or scoring functions in a wide range of applications, ranging
from structure prediction, decoy selection, protein design, protein docking studies, etc. In the simplest form, the energy is obtained from the probability distribution of some feature observed in the native structures using:

$$E(feature) \approx -k_B T \times \ln \left( \frac{N(feature)}{N_{Ref}(feature)} \right)$$ (1.1)

where $N(feature)$ and $N_{Ref}(feature)$ are the observation counts of that feature in the native structure ensemble and some reference state, while $k_B$ and $T$ are the Boltzmann factor and the temperature.

In this work, I will use a pairwise orientation dependent statistical potential developed in our group called DOPE-PW. Originally parameterized by Shen and Sali [30] according to the distribution of pairwise distances for different atomtypes in the PDB database, the DOPE potential was reformulated into DOPE-C$\beta$ by Fitzgerald et. al. [31] to work on a backbone + C$\beta$ level representation of the protein. DOPE-PW was developed by adding orientation as well as secondary structure dependence to the DOPE-C$\beta$ potential [32]. The reference state for the potential is formulated by making the assumption that all atoms are uniformly and spherically distributed when no nonbonded interactions are present. In the case of DOPE-PW we use a sphere whose radius is proportional to the radius of gyration of the native protein, more specifically $\sqrt{5/3}R_g$ [31]. I also introduce a simple multibody statistical potential to quantify burial level of amino acids in Chapter 2.
1.2.3 Sequential Stabilization

While multiple views on how a protein resolves Levinthal’s paradox exist [33, 34, 35], one such view involves sequential stabilization of structural subunits along a folding pathway [36, 37, 38, 39]. Supported by observations native-state hydrogen exchange experiments [38, 39, 40] in multiple proteins, this view argues that most proteins fold along a major pathway by stepwise addition of cooperative folding units called “foldons”, guided by a process of sequential stabilization where prior native-like structure templates the formation of subsequent complementary structure [36]. There is growing evidence that proteins generally do fold along a few well-defined pathways [41, 42], including recently from all-atom folding simulations of 12 proteins [16]. Some evidence for native-like long range strand-strand interactions have been obtained using engineered bihistidine metal binding sites through $\psi$ analysis even in early, unstable intermediates [43, 44] (which are generally harder to characterize), leading to a conjecture that sequential stabilization might be operational throughout the folding pathway (both pre- and post- transition state). Since the principle of sequential stabilization provides a very simple and direct prescription for a computational strategy to overcome the conformational search problem, I use it as the guiding mechanism in the current work for obtaining folding pathways.
1.3 Putting it all together: Ingredients of a protein folding algorithm

Armed with the above concepts, I develop a new method, termed TerItFix (details in Chapter 2), that unifies the prediction of a protein’s folding pathways and secondary/tertiary structure from its primary sequence by applying some basic principles of protein folding. The underlying algorithm is inspired by real proteins which fold via sequential stabilization of secondary/tertiary structure elements (foldons). The input to the method is simply the amino acid sequence and the output is predicted tertiary structure as well as the steps that build up the structure.

It will be instructive to first list the three fundamental properties of proteins that I believe any proper folding algorithm should encode: sequence-dependent backbone torsional angle preferences that facilitates backbone motions, backbone hydrogen bonding that facilitates secondary structure formation, and chemical properties of different amino acids that facilitates the packing of tertiary structure.

These properties are well-represented in TerItFix in the following way: First, I use single nearest-neighbor (NN) dependent amino acid backbone $\phi/\psi$ distributions derived from a high resolution PDB database to sample the conformational space, which has been previously shown to mimic realistic backbone motions in various applications [32, 45, 46]. Second, I use a simple angle and distance dependent definition of a hydrogen bond, supplemented by a penalty for any buried NH or CO groups that aren’t hydrogen bonded (i.e. desolvation
term), following the realization that the likelihood of finding an unsatisfied hydrogen bond in a protein is insignificant [47]. Finally, to recover the amino acid properties lost by the removal of sidechains, I use two main energy functions that capture the properties of the various amino acids: a pairwise-additive distance/orientation dependent statistical potential, DOPE-PW, developed elsewhere [32] that aids the formation of a proper chain topology and a novel statistical multibody energy term that captures the burial properties of the various amino acids.

The above three features often interplay in various ways and so the challenge becomes combining these effectively into a folding algorithm. From a computational perspective, this translates to the challenge of devising a search strategy where one needs to combine three main aspects - sampling, scoring and search - to find the lowest minimum on a complex landscape (Figure 1.3). The sampling and scoring follow directly from the features just discussed. Also, I use a combination of a backbone $\phi/\psi$ motions to explore the conformational space: large-scale single amino acid pivot moves and local three amino acid “doublecrank” moves.

As mentioned in the earlier section, I use the principle of sequential stabilization to tackle the third aspect, conformational search. As a protein folds from its unfolded to native state, local and non-local structure builds incrementally, while continually changing the environmental context that guides the folding process. Mirroring this, in TerItFix, the backbone $\phi/\psi$ sampling distributions (which guide local structure) as well as energy functions (which guide tertiary structure) evolve continually as the structure builds along a folding
pathway. The approach is outlined in detail in the next chapter, where I start from the primary sequence and perform multiple rounds of Monte Carlo simulations to evolve and apply constraints for both local and non-local structure incrementally, and show that the evolution of these local and non-local constraints parallels the evolution of foldons along the folding pathway.

While there have been previous methods that use hierarchical approaches to build protein structure as well as methods that integrate secondary and tertiary structure prediction [48] [49] [50] [51] [52] including our own [32, 53], I believe the present work is novel in its capability to simultaneously predict both the protein structure and folding mechanism with impressive experimental agreement in multiple and diverse systems starting from a reasonable unfolded state, while using a side-chainless model that runs order of magnitudes faster than atomistic...
1.4 Some useful formalisms regarding protein structure

I briefly introduce two formalisms that will be used in later chapters to quantify protein structural information and to help guide the reader. The first is a two dimensional representation of a protein’s tertiary structure using contact maps that indicate which amino acids are in contact in the protein structure. A benefit of this representation is that one can easily pick out different secondary structure motifs as well as easily delineate long vs. short range contacts using this representation, as shown in Figure 1.4. Typically, two amino acids are considered to be in contact when their $C_\beta$ atoms are within 7 - 8 Å of each other.

![Figure 1.4: Contact map representation of a protein](image)

Another useful representation deals with visualization of the local torsional angle propensities of a protein backbone. Given the amino acid sequence, one can calculate the nearest-neighbor dependent backbone torsional angle distributions or Ramachandran distributions. By dividing the Ramachandran map into different regions (basins) based on secondary struc-
ture definitions, one can make a bar plot to represent the basin occupancies (probabilities) for the amino acids in a given sequence, as shown in Figure 1.5.

These two representations will help guide the reader track the evolution of tertiary and secondary structure observed during our protein folding simulations in the subsequent chapters.

Figure 1.5: Ramachandran basin occupancies from the protein amino acid sequence

1.5 Organization of Thesis

The thesis is organized in the following way:

In Chapter 2, I will describe in detail the development TerItFix method, which uses the principle of sequential stabilization to predict both folding pathways and tertiary structure from sequence. In Chapter 3, I will describe the folding studies of topologically identical proteins, protein L and protein G, where the role of non-native interactions appear to be more important than previously thought. I will explain the origin of signatures of non-native
behavior in both proteins. In Chapter 4, I will describe an adaptation of the folding principles for the problem of local protein structure modeling. This includes prediction of loops in protein structures as well as large insertion regions in homology models that correspond to gaps in sequence alignments. In Chapter 5, I will describe the use of support vector machines, a machine learning technique to optimize the various energetic contributions in our potential function for better selection of the generated protein decoys. Finally, in chapter 6, I will discuss some future directions as well as concluding remarks.
CHAPTER 2

DE NOVO PREDICTION OF PROTEIN FOLDING PATHWAYS AND TERTIARY STRUCTURE USING THE PRINCIPLE OF SEQUENTIAL STABILIZATION


Motivated by the relationship between the folding mechanism and the native structure, we develop a unified approach for predicting folding pathways and tertiary structure using only the primary sequence as input. Simulations begin from a realistic unfolded state devoid of secondary structure and use a chain representation lacking explicit side chains, rendering the simulations many orders of magnitude faster than molecular dynamics simulations. The multiple round nature of the algorithm mimics the authentic folding process and tests the effectiveness of sequential stabilization (SS) as a search strategy wherein 2 structural elements add onto existing structures in a process of progressive learning and stabilization of structure found in prior rounds of folding. Because no a priori knowledge is used, we can identify kinetically significant non-native interactions and intermediates, sometimes generated by only two mutations, while the evolution of contact matrices is often consistent with experiments. Moreover, structure prediction improves substantially by incorporating information from prior rounds. The success of our simple, homology-free approach affirms the
validity of our description of the primary determinants of folding pathways and structure, and the effectiveness of SS as a search strategy.

**ABBREVIATIONS:** aa, amino acid; $E_{contact}$, $E_{HBond}$, contact and hydrogen bond biasing terms; MCSA, Monte Carlo Simulated Annealing; Rama, Ramachandran; RCO, relative contact order (RCO); Ub, ubiquitin; transition state ensemble (TSE); sequential stabilization (SS)

### 2.1 Introduction

Despite numerous advances since the original sequence-to-structure folding paradigm was proposed over 50 years ago [54], we still lack a general framework that enables simultaneous prediction of the folding mechanism and structure using only the amino acid (aa) sequence (notwithstanding recent successes of all-atom simulations to fold small, fast-folding proteins [16]). An obvious obstacle is the astronomical number of conformations available to a polypeptide. Proteins overcome this obstacle by sampling a limited set of conformations, guided by the folding process itself. However, most successful structure prediction methods do not consider the folding mechanism when sampling conformations. Conversely, many methods for predicting folding mechanism rely on knowledge of the final structure (e.g., Gō models). Another obstacle emerges because many non-native and near-native conformations often differ by only a few RT, which is at or beyond the ability of current energy functions to reliably distinguish. A related difficulty arises because the native state is the global free energy minimum even if three competing properties local backbone torsional angle preferences,
hydrogen bonded secondary structure, and tertiary packing - are not individually optimized (Figure 2.1). For example, tertiary context can overcome local biases in determining the final secondary structure [53]. Hence, a successful framework should couple tertiary context to secondary structure formation, rather than relying on a strict hierarchical approach.

Figure 2.1: A folded protein has to optimize three competing properties together.

Sequential stabilization (SS) provides one mechanism for coupling secondary and tertiary structure formation during folding and guiding the search process [36][55]. Supported by native state hydrogen exchange experiments, $\psi$ analysis, and other observations [40][38], this view argues that proteins predominantly fold along one or a few low energy pathways determined by the stepwise addition of cooperative units of structure or foldons (e.g., a helix or a strand). Prior emergence of hydrogen bonded structure serves as a template for the formation of additional structure that may only exist as a minor population in isolation.
Here we describe an iterative framework, termed TerItFix, to test whether the combination of SS with basic principles of protein chemistry can be used to predict folding pathways and structure using only the sequence as input. The principle of SS is implemented by using the statistics of folding trajectories garnered from prior rounds of simulation to bias the subsequent sampling of backbone dihedral angles [32] and the energies of tertiary contacts and hydrogen bonds. The approach combines simple backbone torsional $\phi/\psi$ moves, a polypeptide chain with no side chains beyond C$\beta$ carbons, and multiple rounds of simulation with the progressive learning and building of tertiary motifs through constraints imposed by data from prior rounds. We predict the secondary and tertiary structures and pathways for 8 proteins using only around 1000 CPU hours per protein. The results are largely consistent with experimental data, even in the presence of kinetically significant non-native interactions.

2.2 Approach

Initially, 500 individual Monte Carlo Simulated Annealing (MCSA) folding simulations are performed using specialized $\phi/\psi$ backbone moves and energy functions appropriate for a reduced chain representation consisting of the backbone plus C$\beta$ atoms, as discussed below. The best final structures (lowest energy quartile) are then examined for recurring secondary structures, backbone hydrogen bonding, and tertiary contacts. After modifying the move set and energy functions to promote these recurring features, another round of 500 folding simulations is performed. The passing of information from one round to another is repeated until convergence (Figure 2.2). This iterative, multi-round learning and biasing procedure
equates to a search strategy involving sequential stabilization, as illustrated with the folding of ubiquitin (Figure 2.3).

The folding simulations employ move sets and energy functions that are designed to describe three competing protein properties: $\phi, \psi$ preferences, secondary structure, and tertiary packing. Angle preferences are incorporated by sampling conformational space using neighbor dependent $\phi, \psi$ distributions derived from the PDB (Figure 2.4). These angles are used for pivot moves, where only a single residue’s $\phi, \psi$ angles are changed, as well as for double crankshaft local moves, where two consecutive peptide groups are rotated [45][22] (Figure 2.5). In the initial round, angles are chosen from a coil library that accurately describes the structure of the chemically denatured state [56]. Later rounds use information about the secondary structures garnered from the prior round to restrict the sampling distribution (Methods).

Our energy function is composed of three statistical potentials and two biasing terms (Figure 2.6), which guide the formation of secondary structure and tertiary packing. The 1st potential [53] describes the residue-residue interactions according to the distance distributions in the PDB, contingent on secondary structure and the relative orientation of the two residues’ Cα-Cβ vectors (Figure 2.6). The 2nd potential describes each residue’s burial propensity, as calculated using the number of heavy atoms surrounding each Cβ atom in a 11 Å hemisphere defined by orientation of the Cα-Cβ vector (glycines are ignored). The 3rd term is associated with backbone desolvation and backbone hydrogen bonding. The desolvation term assigns a penalty for the loss of water-peptide hydrogen bonds when there
Figure 2.2: Flowchart outlining TerItFix algorithm. A. The algorithm iterates through multiple rounds of MCSA simulations, where information about secondary and tertiary structure from prior rounds is utilized. B. A single MCSA simulation involves a two stage energy minimization, first using pivot moves followed by a refinement stage using doublecrank moves.
Figure 2.3: **TerItFix protocol applied to Ubiquitin.** Secondary and tertiary structure co-evolve as the rounds proceed with local and nonlocal constraints derived from the prior round. The $\phi, \psi$ sampling distribution is initially obtained from a coil library. It contains angles from all regions of the Ramachandran map at a frequency given by the relative height of the blue, green, pink and red bars, color coded according the legend in the lower left. The sampling distribution becomes more restricted as secondary structure is progressively fixed after each round. E.g. L43 preferentially adopts $\beta$ conformations as the rounds progress and its distribution shifts to the $\beta$ basin (left column). The contact maps identify the order of tertiary structure formation along the pathway, which can be used to construct potential folding intermediates (bottom).
Figure 2.4: Restriction of the backbone sampling distribution. The distribution becomes more refined as the dependence on type and secondary structure of the aa and its neighbors are included.
A single $\phi, \psi$ pivot move permits large scale motions

A three residue doublecrank (DC) move for local refinement

Figure 2.5: Two kinds of backbone torsional angle moves are used in TerItFix. The pivot move involving two angles is effective in exploring large regions of conformational space while doublecrank changes four angles while minimally perturbing the rest of the protein structure.
is no compensating protein-protein hydrogen bond (Figure 2.6).

Even for small proteins, the exploration of the folding landscape poses a formidable search problem, and thus requires additional constraints. The principle of SS provides a realistic method of guiding the search by continually refining the $\phi, \psi$ sampling distributions, which determine local structure, and biasing the energy for recurring motifs to guide secondary and tertiary structure formation. After each round, the lowest energy quartile is evaluated to identify secondary structure preferences and popular tertiary contacts and hydrogen bonding. These items are used to restrict the backbone sampling library (Figure 2.4) and to generate two energetic biasing terms, $E_{\text{contact}}$ and $E_{\text{HBond}}$, that are employed in the next folding round (Methods).

This iterative process incrementally fixes secondary structure and biases tertiary structure and hydrogen bonding as the rounds proceed, producing a series of sequential steps that may correspond to the authentic folding pathway (Figure 2.3). Individual biases may strengthen or weaken in subsequent rounds because of the emergence of competing contacts. In principle, if no major kinetic traps impede the pathway, the final sampling distribution and contact probabilities should converge to produce the native secondary and tertiary structures, respectively.
Figure 2.6: **Energy functions used in TerItFix.** Examples of the energy profiles for the three statistical potentials used in TerItFix. All simulations use the same DOPE-PW parameters except for Ub where a non-continuous parallel strand is set higher to reduce the non-native antiparallel β1-β3 pairing. A burial penalty profile similar to NH group is obtained for the C=O group as well.
2.3 Methods

2.3.1 Monte Carlo Simulated Annealing (MCSA) Minimization

Using the algorithm described by Aarts and Korst [57], the MCSA procedure for each round starts with an assignment of dihedral angles based on the Ramachandran Maps for each amino acid. Moves and energy functions are described below. The structure converges to a local energy minimum by gradually diminishing the annealing temperature, which controls the fraction of allowed moves that increase the energy of the system. A fixed number of moves are computed in order to achieve thermal equilibrium at each annealing temperature. The annealing temperature is decreased using the formula

\[
t = \frac{t'}{1 + \frac{t'\log\delta}{3\sigma_t'}}
\]  

(2.1)

where \(t'\) and \(t\) are the old and new annealing temperatures, respectively, \(\sigma_t\) is the standard deviation for the energy distribution at temperature \(t'\), and \(\delta\) is a tunable parameter for the cooling schedule. The temperature is decreased according to a Cauchy cooling schedule, until either convergence is reached or the total number of Monte Carlo steps reaches a specified value. The convergence criterion is based on the magnitude of the energy fluctuations for each annealing temperature, and annealing stops when the inequality,

\[
\sigma_t \leq \epsilon t
\]

(2.2)
is satisfied, where $\epsilon$ is the convergence tolerance, also a tunable parameter. The tunable parameters in the simulations are set as: $\delta=0.05$ and $\epsilon=1$. After every move, the energy of the new conformation is evaluated, and the change is accepted with probability,

$$P = \min \left[ 1, \exp \left( -\frac{\Delta E}{\ell} \right) \right]$$

(2.3)

where $\Delta E$ is the energy difference between the new and old conformations.

### 2.3.2 Sampling using Ramachandran Maps

In the backbone plus C$\beta$ representation used in TerItFix, only backbone $\phi, \psi$ angles are changed during the simulation with all other angles and bond lengths fixed at their ideal values. Backbone conformations are sampled using $\phi, \psi$ distributions, or Ramachandran maps, obtained from high resolution PDB structures (resolution < 2.5 Å homology below 90%). Individual Ramachandran maps are generated for each position conditional on both the amino acid and its nearest neighbors’ chemical identity and, and their secondary structure specification (Figure 2.4).

Because TerItFix proceeds by eliminating secondary structure options at a given position, six possible categories of secondary structure are considered for the construction the sampling distributions (H: helix, E: strand, C: coil, A: everything, O: not helix, and Q: not strand). This description requires a total of 1,728,000 possible Ramachandran maps for the 203 amino acid triplets. Each Ramachandran map is divided into 722 bins, with each of these $5^\circ \times 5^\circ$ bins assigned a probability determined by frequency of $\phi, \psi$ angles in a set of high resolution
structures appropriate for amino acid identities and their secondary structure. The secondary structure is ascertained from structures obtained in the prior round, as described in the next section. In the initial round, the secondary structure for each amino acid is assumed to be in a coil conformation and the Ramachandran maps are generated accordingly.

The sampling during the MCSA simulation is determined by the probabilities proscribed by a particular Ramachandran map. A particular bin in the Rama space is first selected according to the probability assigned to that bin (e.g., a bin that contains 1% of the angle counts for the distribution at that position has a 0.01 probability of being selected). A random angle within the $5^\circ \times 5^\circ$ bin is selected.

### 2.3.3 Move Sets

The two possible moves involve backbone $\phi, \psi$ angles rotations (Figure 2.5). The first is a pivot move, which involves changing the $\phi, \psi$ angle of a single residue, with the angles selected according to the bin probabilities in the associated Ramachandran map. This move twists the protein backbone drastically and is effective in exploring large regions of the energy landscape. The second move is a three-residue local crankshaft motion termed a double-crank or DC move. It allows for the selection of new angles described by the Ramachandran maps while minimally perturbing the rest of the protein structure. Each double-crank move involves the selection of a new $\phi_i, \psi_i$ angle for the central residue $i$, followed by compensating counter rotations in $\psi_{i-1}$ and $\phi_{i+1}$ of the flanking residues. All the MCSA simulations begin with using pivot moves, followed by a subsequent refinement step using double-crank moves.
2.3.4 Secondary structure fixing

The frequencies of helix, strand and coil structure, as determined by the DSSP (Dictionary of Protein Secondary Structure) [58], are used to update the consensus secondary structure assignments. At each position, one of the three types is eliminated as a sampling option when its frequency falls below a threshold, e.g., remove helix if frequency lies below 1%. The consensus secondary structure restricts the $\phi, \psi$ sampling library employed in the subsequent folding round (Figure 2.4). If the subsequent round of TerItFix eliminates a secondary structure option at a position, the Ramachandran map at that position is changed accordingly. The protocol for eliminating a secondary structure option at a position is determined using the secondary structure frequencies at a given round, $P_X$ ($X=E$, $H$ or $C$), calculated using DSSP for the lowest energy quartile according to the following main criteria. For $i$ consecutive positions (in order of precedence):

- For $i > 3$: [HEC] $\rightarrow$ [HC] if $PH > 0.50$ or ($PH > 0.25$ and $PE < 0.005$)

- For $i > 0$: [HEC] $\rightarrow$ [HC] if $PH > 0.1$ and $(i+1)$ is (Proline and [HC])

- For $i > 4$: [HC] $\rightarrow$ [H only] if $PH > 0.4$

- For $i > 0$: [HC] $\rightarrow$ [H only] if $PH > 0.9$ and $(i-1,i+1$ both [HC])

- For $i > 0$: [HC] $\rightarrow$ [H only] if $PH > 0.25$ and $(i$ is Proline) and $(i-1,i+1$ both [HC])

- For $i > 4$: [HEC] $\rightarrow$ [EC] if $PH < 0.02$

- For $i > 3$: [HEC] $\rightarrow$ [EC] if $PH < 0.2$ and $PE > 0.05$
• For $i > 0$: $[\text{HEC}] \rightarrow [\text{EC}]$ if $PH < 0.01$

• For $i > 0$: $[\text{HEC}] \rightarrow [\text{EC}]$ if $PH < 0.05$ and $PC(\text{Turn}) > 0.3$

• For $i > 0$: $[\text{EC}] \rightarrow [\text{E only}]$ if $PE > 0.10$

• For $i > 0$: $[\text{EC}] \rightarrow [\text{E only}]$ if $PE > 0.05$ and $(PE > 0.05$ for $i-1,i+1)$

• For $i > 0$: $[\text{HC}] \rightarrow [\text{C only}]$ if $PH < 0.30$ and $PC(\text{Turn}) > 0.25$ and $(i-1,i+1$ not $[\text{HEC}])$

• For $i > 0$: $[\text{HC}] \rightarrow [\text{C only}]$ if $PH < 0.05$ and $(i-1,i+1$ not $[\text{HEC}])$

• For $i > 0$: $[\text{EC}] \rightarrow [\text{C only}]$ if $PE = 0.0$ and $PC(\text{Turn}) > 0.25$ and $(i-1,i+1$ not $[\text{HEC}])$

2.3.5 Tertiary structure fixing

The frequency of contacts between residues $i$ and $j$, $E_{\text{contact}}$, $i<j$, in the lowest energy quartile serves as a bias for the next round, with a contact defined by $C\beta_i-C\beta_j$ separation below 7.5 Å (only for $|i-j| > 3$). A similar hydrogen bonding bias between residues $i$ and $j$ is given by

$$-E_{HBond_{ij}} = 25p_{ij} + 2(1-p_i),$$

where $p_{ij}$ is the probability that the NH of residue $i$ bonds to the CO of residue $j$, and $(1-p_i)$ is the probability that the NH of residue $i$ lacks a hydrogen bond. This functional form ensures a minimal contribution even when $p_i$ is low in the prior round. The total energy is given as the weighted sum of the biasing terms $E_{\text{contact}}$, $E_{HBond}$, plus the three statistical potentials. Each trajectory comprises two stages, with the first stage having higher contributions from the two biasing terms. This stage produces partially structured conformations that form the starting points for the second stage in which the
weights of the statistical potentials are increased. Each stage concludes with a refinement step where the minimization is rerun using the double crankshaft local move.

### 2.3.6 Energy Functions

The energy function contains three statistical potentials as well as two biasing terms:

\[
E_{MCSA} = w_1 E_{DOPE-PW} + w_2 E_{Burial} + w_3 E_{desolvation} + w_4 E_{HBond} + w_5 E_{Contact} \quad (2.4)
\]

The relative weights, \( w_i \), of the different energy terms for the two stages are listed in Table 2.1, which were derived semiempirically. The biasing energies are more heavily weighted in the 1\(^{st}\) stage.

<table>
<thead>
<tr>
<th>Table 2.1: Parameters used in the TerItFix simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCSA parameters</td>
</tr>
<tr>
<td>Total MC Steps</td>
</tr>
<tr>
<td>Temperature Update Interval (Steps)</td>
</tr>
<tr>
<td>Initial Temperature</td>
</tr>
<tr>
<td>Energy Function Weights</td>
</tr>
<tr>
<td>DOPE-PW, ( w_1 )</td>
</tr>
<tr>
<td>Burial, ( w_2 )</td>
</tr>
<tr>
<td>Desolvation ( w_3 )</td>
</tr>
<tr>
<td>HBond Bias (( E_{HBond} )), ( w_4 )</td>
</tr>
<tr>
<td>Contact Bias (( E_{Contact} )), ( w_5 )</td>
</tr>
<tr>
<td>MoveSet MoveSet</td>
</tr>
<tr>
<td>then doublecrank</td>
</tr>
<tr>
<td>blecrank</td>
</tr>
</tbody>
</table>

Residue-residue interaction Potential. DOPE-PW is a pairwise additive statistical
potential based on observed distance distributions in the PDB between all atoms in our C\(_\beta\) representation. It is designed to reflect the basic protein structural principles of chemical complementarity, packing, and secondary structure conformation (Figure 2.6). As detailed in our previous studies [32] [53], each interaction in DOPE-PW is classified according to atom type, residue type, secondary structure assignment, and side-chain orientation defined by the C\(_\alpha\)-C\(_\beta\) vector. The orientational dependence is characterized using the relation,

\[
\rho = \sqrt{(\rho_{1-2} - 90)^2 + (\rho_{2-1} - 90)^2}
\]  

where \(\rho_{1-2}\) is the angle between the C\(_\alpha\)-C\(_\beta\) vector of residue 1 and the C\(_\alpha\)-C\(_\alpha\) vector from residue 1 to residue 2. Similarly, \(\rho_{2-1}\) is the angle between the C\(_\alpha\)-C\(_\beta\) vector of residue 2 and the C\(_\alpha\)-C\(_\alpha\) vector from residue 2 to residue 1 (Figure 2.6). The \(\rho\) value is binned into three orientational regions: low (\(\rho < 40^\circ\)), medium (\(40^\circ < \rho < 70^\circ\)) and high (\(\rho > 70^\circ\)), which are accordingly parameterized from a high resolution database of PDB structures. Figure 2.6 illustrates a sample DOPE-PW energy profile for ILE-LEU C\(_\beta\) atoms for high and low values of \(\rho\).

**Burial energy**. \(E_{\text{burial}}\) energy is obtained from the burial properties for each of the 20 aa’s from a database of around 1000 single domain high resolution PDB structures with sizes ranging from 50-150 aa (Figure 2.6). The burial level for residue type \(i\) is determined from the number of heavy atoms, \(N_i\), around the C\(_\beta\) in a 11 Å radius hemisphere in the direction defined by the C\(_\alpha\)-C\(_\beta\) vector. The burial energy contribution of \(N_i\) is given by
where \( f(N_i) \) and \( f(N_{\text{any}}) \) are frequencies for residue type \( i \) and any residue type, respectively.

The original motivation behind the multibody statistical burial energy term was to enhance the cooperative behavior observed in experiments, resulting from the association of hydrophobic groups. Most potentials used in protein folding simulations mimic this effect by packing all the hydrophobic groups towards the center of the protein in a uniform manner. However, while it is true that most globular proteins have a single, central hydrophobic core, these are often built together from smaller clusters of hydrophobic groups [59]. Hence, a second motivation for our burial term was to localize the burial of various amino acids to promote formation of hydrophobic clusters rather than directly enforcing a central core by overdriving the collapse of the system. As an example, in Figure 2.7 we show how the addition of the burial term improves the folding of 1di2, a small protein which is slightly elongated with distributed hydrophobic clusters, compared to other globular proteins of similar size. While the best structure obtained using only the pairwise DOPE-PW potential is greater than 6 Å, adding the burial term to DOPE-PW increases the population of low RMSD structures with the best under 4 Å.

**Backbone desolvation** A similar term is calculated for hydrogen bonded and non-hydrogen bonded amide nitrogen and carbonyl oxygen atoms. The number of heavy atoms is calculated in 5 Å radius hemisphere in the direction defined by the N-H and C=O vectors,
Figure 2.7: Inclusion of a Cβ multibody burial statistical potential improves structure prediction over pairwise additive DOPE-PW statistical potential respectively. The N-H and C=O hydrogen bond energy terms are specified as whether or not they form protein-protein hydrogen bonds, and an energy penalty is assigned to NH and CO groups with unsatisfied hydrogen bonds (Figure 2.6). A hydrogen bond is defined for H-O distances of $< 3.5 \, \text{Å}$ and H-O=C angles of $> 120^\circ$. Sample profiles of the burial energies for LEU and GLU Cβ atoms are shown in Figure 2.6, along with the NH burial penalty energy profile.

The present approach differs from our earlier ItFix protocol in which the local sampling distribution is fixed [32] [53] but tertiary contacts and hydrogen bonds are not explicitly biased. We now include peptide group desolvation and burial potentials in lieu of a globularity term.
2.3.7 Biasing Energies used for sequential stabilization

$E_{Contact}$: After each round, the average contact matrix from the ensemble of the lowest 25% energy structures is used to generate the contact bias energy term for the next round,

$$E_{Contact} = \sum_{i=1}^{n} \sum_{j>i}^{n} -w_{ij} \quad (2.7)$$

where $w_{ij}$ is the probability of contact between residues i and j ($C\beta$ separation < 7.5 Å.

$E_{Hbond}$: The hydrogen bonding energy bias is,

$$E_{HBond} = \sum_{i=1}^{n} \sum_{i=1}^{n} [H_{max}p_{ij} + H_{min}(1 - p_{i})] \quad (2.8)$$

where $p_{ij}$ is the probability that the NH of residue i bonds to the CO of residue j, and $p_{i}$ is the probability that the NH hydrogen bonds to any other residue in the current round.

The weights $H_{max}$ and $H_{min}$ are set to -25.0 and -2.0. Since a hydrogen bond is generally favorable, the above functional form ensures a contribution to the hydrogen bonding energy even when the specific probability $p_{ij}$ is low.

2.3.8 Software

The simulation protocol is implemented in C, and input/output is managed with PDB tools from the BioPython package[60]. To facilitate job submission in high performance supercomputing architectures, the programs are wrapped using the parallel scripting language, Swift [61]. Simulations are performed using the PADS and Beagle clusters at the Computation
2.4 Results

We apply TerItFix to deduce the folding pathways of eight proteins and simultaneously to predict their tertiary structures. The different levels of information accessible are demonstrated by studying the fast folding five helix subdomain of lambda repressor (λ6−85) and comparing the predictions to experiments [62] [63] and molecular dynamics (MD) simulations [16] [24] [62]. Next, TerItFix is applied to describe the folding pathways of two homologous immunity proteins, Im7 and Im9, along with a double point mutant, SIIm9, to demonstrate that our method is sensitive enough to capture the kinetic consequences of slight aa variations and to predict the presence of kinetic intermediates. Finally, TerItFix is used to describe five other proteins, Ub, chymotrypsin inhibitor 2 (CI2), Protein L and two three helix bundle proteins, Protein A and the designed α3d.

2.4.1 Folding of Lambda repressor (λ6−85).

Starting with a φ,ψ distribution generated from the coil library, a folding pathway emerges after five rounds of folding with a 4.3 Å Cα RMSD (best) structure (Figure 2.8). The initial φ,ψ distribution provides little indication of the positions or propensities of the helices or their order of formation because most angles in the initial distribution are non-helical. A clear pathway emerges as the rounds proceed, with helices H3 and H4 appearing first and interacting. As the probabilities of these two helices increase in progressive TerItFix
rounds, H1 gradually appears and docks against the H3-H4 motif by round R3. While the number of helices remains largely unchanged after R3, the helices lengthen, and their contact probabilities continue to increase for the next few rounds, as evident by the evolution of the $\phi, \psi$ distributions, average contact maps, and the centroid of the largest cluster formed from the structures generated in each round (Figure 2.8A). Although H2 and H5 appear in some trajectories, the population of structures containing these helices remains insufficient to justify restricting the sampling distributions in those positions to the helical basin.

These simulations highlight the interplay between secondary and tertiary structure formation. The helical probability for the residues of helix H2 in R1 exceeds that in subsequent rounds. This loss of native-like structure suggests that H1 and H2 initially interact, but tertiary contacts between H1 with H3 and H4 dominate in later rounds. At the same time, the average contacts between H4 and H5 continue to rise until round R4, even though residues in H5 never become highly helical (Figure 2.9). Besides predicting the pathway, incorporating the strategy of SS into TerItFix improves the predicted structures (Figure 2.8B).

Although the TerItFix algorithm produces a folding pathway, identifying the TSE is not straightforward due to the multi-round nature of the method. Using an observation deduced from our $\psi$ analysis studies of four proteins that obey the correlation between $\log(k_{fold})$ and relative contact order (RCO) [64] [42] [21], we identify a TSE by the condition $RCO_{TSE}/RCO_{Native} \sim 0.7$. The number of long-range contacts in $\lambda_{6-85}$ sharply increases in the low energy structures of R4, and some achieve the 70% RCO level (Fig 2.8C). Examination of such structures indicates that H1, H3, H4, and potentially H5 are
Figure 2.8: **TerItFix prediction of $\lambda_{6-85}$.** A. The initial $\phi, \psi$ sampling distribution has low helical propensity. A pathway emerges with the docking of helices H3 and H4, followed by H1. By Round 6, H2 and H5 emerge albeit weakly. Contact maps and illustrative structures are shown in the center and right columns, respectively. B. The RMSD distribution of the endpoints of the 500 trajectories improves with each round. C. The $\langle \text{RCO} \rangle$ reaches 70% of the native level in R4 and a representative, TS analog containing H1+H3+H4+H5 is identified.
Figure 2.9: Evolution of inter-helical contacts in λ repressor as a function of TerItFix rounds
present in this predicted TSE. Using mutational $\phi$ analysis, Oas and coworkers suggest that H1 and H4 are formed in the TSE, while the presence of H2 and H3 is unclear ($\phi = 0.2$ and 0.3) and the status of H5 is ambiguous [62]. Our kinetic amide isotope effect data indicate that the TSE contains 80% of the native helical hydrogen bonds, consistent with the TSE burying 70% of the total denaturant sensitive surface area [65] [66]. This high level of hydrogen bond formation suggests that at least one more helix is present in the TSE, potentially H3 because $\phi(H3) = 0.3 > \phi(H2) = 0.2$. Hence, predictions from TerItFix appear to be largely consistent with experiment. MD simulations by Shaw and coworkers find that the TSE contains H1-H4 and that these four helices are at least partially formed in the denatured state [16]. The presence of H1 and H4 in the TSE is consistent across the TerItFix trajectories, MD simulations, and experiments. But the MD simulations appear to disagree with TerItFix and experiments (at 310K) in predicting H2 to be significantly populated in the TS. This difference might be a consequence of the high helical propensity of H1-H4 in the unfolded state in the MD simulations (at 350K). Experimentally, the helical content in an unfolded analog is 16% at 310K [67], which is much lower than the 42% helical content in the MD simulations but accords with the low 15% helical angle content in TerItFix’s initial $\phi, \psi$ coil sampling library. A Markov state analysis of 3265 relatively short ($\mu$ sec) MD simulations identifies a TS structure (Pfold =0.53) having only 1-2 turns of helices H1-H4 and two adjacent $\beta$ strands [23]. Another set of MD simulations using a new tempering method finds that H1-H3 are formed prior to H4 and H5 [24]. Further experiments should be performed to permit more accurate assessments of the disparate results obtained by TerItFix.
and the three different MD simulations.

2.4.2 Immunity proteins Im7, Im9, SIm9

The homologous immunity proteins Im7 and Im9 highlight a case where TerItFix is advantageous over native-biased methods. Im7 and Im9 display different folding kinetics despite being nearly identical four-helix bundles with 60\% sequence identity. Im7 folds in a three-state manner with an intermediate containing helices H1, H2 and H4, while Im9 folds in a two-state manner [68] [69]. Im7’s three helix intermediate is misfolded in the sense that the three helices must at least partially separate in order to accommodate H3. The importance of sequence is further highlighted by the fact that only two conservative mutations in Im9 (SIm9) induce a three-state mechanism akin to Im7’s [12]. That such slight variations of the aa sequence can alter the folding behavior reflects the challenge of reproducing these results. After only three rounds of simulations, all four helices form and interact in Im9, whereas H3 fails to form in Im7 (Figure 2.10A, 2.11,2.12). The sampling distribution for the residues in H3 of Im7 never evolves beyond the coil specification, and the protein becomes trapped in an intermediate structure containing H1, H2 and H4 (Figure 2.12). Thus, TerItFix correctly captures the energetic frustration of the folding landscape of Im7 that is absent in Im9.

Next, TerItFix is applied to the folding of SIm9 which has the conservative V37L and V71I substitutions in H2 and H4, respectively, and folds with the accumulation of the same three helix intermediates as Im7. Remarkably, the TerItFix results for SIm9 are very similar to those for Im7, successfully predicting the same three helix intermediate as observed
Figure 2.10: **TerItFix prediction of Im proteins.** A. Contact maps. By the last round, R3, TerItFix predicts that the H3 is formed in Im9 (green circle) whereas H3 is absent in Im7 and SIm9, as observed experimentally. B. In SIm9, the interaction energy between V37L strengthens the H1-H2 contacts but destabilizes the H2-H3 contacts between V37L & I53.
Figure 2.11: Basin propensities and average contact maps in the last round of TerFix folding of Im proteins.
Figure 2.12: TerItFix prediction of folding intermediates in Im proteins. Evolution of inter-helical contacts in λ repressor as a function of TerItFix rounds A. Evolution of contact maps in Im7, Im9 and SIm9. H1-H2 contacts (red ovals) in Im7 and SIm9 are more populated than in Im9. B. H1-H2 contact frequency in rounds R1 & R2, calculated by summing the contact probabilities between the H1-H2 in the low energy population in each round. C. The interaction energy between V71I and the rest of the sequence strengthens the H1-H4 contacts in SIm9.
experimentally (Figure 2.10A) and demonstrating a high level of sensitivity of our method to changes in sequence and the energy landscape. The origin of the sensitivity to two conservative mutations is deduced from the differences in the H1-H2 contacts for Im9 and SIm9 (Figure 2.12B). SIm9’s two mutations promote docking of these helices in a geometry that precludes the addition of H3. Specifically, the two mutations alter the pairwise DOPE-PW energies between the helices (Fig 2.10B). In SIm9, the interactions between V37L, which lies in H2, and residues in H1 are stronger, while the interactions between V37L and I53 in H3 are weaker. These two differences provide an explanation for the disparate folding mechanism induced by only two aa mutations.

2.4.3 Prediction of early events, foldons and non-native contacts.

TerItFix simulations begin from a conformation devoid of regular structure. Hence, the method can provide insights by identifying motifs that form at the earliest stages of folding for five proteins: Ub, CI2, Protein L, α3D [70], and Protein A.

Ub is a 76 residue α/β protein with a relatively complex topology and a folding pathway that has been extensively characterized by ψ analysis and native state hydrogen exchange [55] [64] [38]. The TSE contains four adjoining strands, β1-4, and part of the major α helix. Folding from the TSE to the native state occurs in a stepwise manner with the addition of the small 3_{10} helix followed by the β strand. However, the early events leading to the TSE are difficult to identify due to the their intrinsic instability and the ensuing two-state kinetic folding behavior.
Figure 2.13: TerItFix prediction of Ub, CI2 and 3d. The average contact maps in R5 (upper) are compared to the native contact maps (lower). For Ub, the contacts are circled according to experimental $\psi$ values (green: $\psi = 0$, absent in TS; red: $\psi = 1$, present in the TS). For CI2, the non-native contacts are noted with the red arrow.
The first motif to form in the TerItFix simulations is the $\beta_1$-$\beta_2$ hairpin (Figure 2.13, 2.14), followed by the addition of the $\alpha$ helix and the interaction between the two terminal strands, $\beta_1$-$\beta_3$. The early interaction between the termini is significant because long-range contacts generally form with greater difficulty, especially when 30+ intervening residues are still unstructured. Although $\beta_1$-$\beta_3$ form a parallel arrangement in the native structure, we observe some non-native anti-parallel arrangements. The subsequent steps include the formation of contacts between $\beta_3$ and $\beta_4$ and the strengthening of contacts between the helix and $\beta_4$ in later rounds. By R6, both the secondary structure distribution and average contact maps plateau. While the two remaining foldons, the $3_{10}$ helix and $\beta_5$, maintain low populations in the contact maps, enough steps along the folding pathway are resolved to obtain the correct fold and a best C$\alpha$ RMSD structure of 4.6 Å. The TerItFix results are consistent with experiments; in particular, the foldons known to be in the TSE are predicted to form prior to the two foldons known to fold after the transition state.

CI2 contains both parallel and antiparallel $\beta$ strands onto which a single helix and an active site loop G3 are packed. CI2 folds in a two state manner, with a TSE characterized using $\phi$ analysis [71] [72]. The helix has the highest $\phi$ values, followed by strands $\beta_3$ and $\beta_4$. In the first TerItFix round, interactions appear throughout the protein (Figure 2.13). By R2, the helix begins to emerge (Figure 2.15), followed by the $\beta_3$-G3-$\beta_4$ motif. The $\beta_3$-$\beta_4$ interactions intensify as the carboxy terminus of the helix docks to $\beta_3$, forming a hydrophobic cluster and stabilizing interactions between the helix and $\beta_3$-$\beta_4$. Both experiment and previous simulations [73] suggest that this motif is present in the TS. We concur that the
Figure 2.14: Folding pathway for Ub. Secondary structure and tertiary contacts becomes increasingly determined through the six rounds of folding.
helix forms first, followed by $\beta3-\beta4$. The experimental studies with double mutant cycles also implicate interactions between A17, L50, and I59 in the TSE. Although A17-I59 interactions are absent in our simulations, we observe interactions between residues around A17 and L50. By R4 of TerItFix, non-native contacts between the active site loop G3 and $\beta3-\beta4$ emerge because G3 forms a $\beta$-hairpin structure with either $\beta3$ or $\beta4$. This non-native signature is consistent with the previous simulations [73] and is rationalized by the native G3 loop having an extended geometry that can readily hydrogen bond with either the $\beta3$ or $\beta4$ strands.

TerItFix simulations for Protein L converge within three rounds, with the formation of hairpin 2 followed by the formation of hairpin 1, producing a best RMSD structure of 3.2 Å (Figures 2.16 and 2.17). The TSE of Protein L has recently determined using $\psi$ analysis to be extensive, consisting of all four strands [21]. This result updates earlier $\phi$ analysis studies which indicate that the TSE is small and polarized [74]. Our simulations predict a non-native registry for hairpin 2, consistent with the experimental finding of non-native structure in the TSE for this region of the protein. The non-native structure arises because the native turn for hairpin 2 consists of three unfavorable consecutive positive dihedral angles, whereas TerItFix predicts a canonical Type I $\beta$ turn (Figure 2.16). This result agrees with all atom simulations [21] and can explain the origin of the non-native behavior observed experimentally. However, the non-native register is never fully resolved in the TerItFix simulations.

The two 3-helix bundles are relatively easy targets for TerItFix. According to TerItFix, the pathway for the 73 aa $\alpha3d$ begins with the docking of H2 and H3 (Figure 2.13, 2.18). As the contacts between these helices increase, the amino terminus of H1 forms and docks
Figure 2.15: Folding pathway for CI2.
Figure 2.16: Non-native hairpin formation in protein L. TerItFix produces a 3.2 Å Cα RMSD structure albeit with a non-native register for hairpin 2. The TerItFix (red arrows) and native (green arrows) φ/ψ angles for the 3 turn residues are indicated.
Figure 2.17: Folding pathway for protein L.
Figure 2.18: Folding pathway for α3D.
Figure 2.19: Folding pathway for protein A.
against the other two helices. The lowest Cα RMSD of our best structure is 2.9 Å in fact, slightly better than the 3.1 Å obtained in the Shaw MD simulations [16]. Protein A’s three helices form almost simultaneously within three TerItFix rounds, producing a lowest RMSD structure of 2.9 Å (Figure 2.19). Notably, H2 is kinked at the center in R2, but is corrected by R3.

2.5 Comparison to all atom molecular dynamics simulations

Most proteins considered here highlight TerItFix’s ability to identify interesting features of the folding landscapes. Unlike most homology-free structure prediction algorithms, TerItFix doesn’t use fragments or invoke any prior assumptions about the protein’s secondary structure, while running many orders of magnitude faster than MD simulations (CPU hours compared to CPU weeks). In the absence of major kinetic traps, we expect that TerItFix can predict the native structure for many small proteins. A further test emerges from additional simulations for the set of 12 fast-folding proteins recently investigated by the DE Shaw group using all-atom MD simulations [16]. Using 500 Monte Carlo trajectories per round, we obtain an average Cα-RMSD$_{best}$ of $2.7 \pm 1.2$ Å as compared to $2.0 \pm 1.3$ Å from the MD simulations, with TerItFix producing lower values for 5 of the 12 targets. By increasing the number of Monte Carlo runs to 5000 in the last round, this prediction improves to an average Cα-RMSD$_{best}$ of $2.3 \pm 1.0$ Å with TerItFix producing lower RMSD values for 7 out of 12 targets (Figure 2.20), while the computation still remains within a day for each protein.

While multiple structure prediction methods now are able to predict the tertiary struc-
Figure 2.20: Lowest RMSD structures from TerItFix predictions of 12 fast folding proteins compared to DESRES all-atom molecular dynamics simulations[16]
ture of small proteins to less than 5 Å what makes the all-atom explicit MD simulations significant is the detailed insights one can obtain about the folding mechanism by analyzing the trajectories in atomistic detail. In the original paper by Lindorff, et. al. [16], they quantify the order of structure formation by quantifying how early each amino acid adopts a native-like conformation in their folding pathways. Specifically, they use the expression

\[ d_i = 1 - e^{-0.5 \times MSD_i} \]  

where MSD \(_i\) is the mean square deviation from the native structure for a stretch of five amino acids centered on residue \(i\), and a value of \(d_i\) close to 0 implies a native-like conformation. Thus when the average \(d_i\) is close to 0 in the folding segment of the trajectory, it implies that residue \(i\) becomes native-like early in the folding pathway. Using this approach, one of their major conclusions regarding the folding pathways of the 12 fast-folding proteins was that the order in which an amino acid becomes native-like in the process of folding correlates directly with the propensity of that particular residue to form local native-like structure in the unfolded state (Figure 2.21 Upper plots).

Since TerItFix has the ability to predict the order of structure formation in the folding pathways as well, next we compare the TerItFix predicted pathways for the 12 proteins to those predicted by the DESRES all-atom molecular dynamics simulations. Because these proteins are small and fast-folding compared to the examples discussed in earlier sections, we found that the TerItFix simulations converged within the first 2-3 rounds for the ma-
Figure 2.21: Comparison of DESRES pathways to TerItFix predicted pathways. For DESRES, the closer the average time of formation is to 0, the earlier that residue appears in the folding pathway [16]. For TerItFix, the closer the average TM-score of a residue is to 1, the earlier it appears in the folding pathway.
Figure 2.21: Comparison of DESRES pathways to TerItFix predicted pathways. For DESRES, the closer the average time of formation is to 0, the earlier that residue appears in the folding pathway [16]. For TerItFix, the closer the average TM-score of a residue is to 1, the earlier it appears in the folding pathway.
Figure 2.21: Comparison of DESRES pathways to TerItFix predicted pathways. For DESRES, the closer the average time of formation is to 0, the earlier that residue appears in the folding pathway [16]. For TerItFix, the closer the average TM-score of a residue is to 1, the earlier it appears in the folding pathway.
ority of them. To be able to compare to the DESRES pathways, we needed to quantify how early each amino acid adopts a native-like conformation. While the order of structure formation in the MD simulations is explicitly defined according to the time of formation in a single trajectory, the order of structure formation in TerItFix is defined according to the occurrence of consistent tertiary motifs in low energy structure ensembles across the different rounds. Hence, we define a similar metric, TM<sub>i</sub>, to quantify nativeness for residue, where \( TM_i = 1 \) if RMSD of residue \( i < 3 \) Å from the corresponding residue in the native structure (otherwise \( TM_i = 0 \)) when aligned using TM-score, a metric that is more sensitive to local structure deviation compared to global RMSD. Now, for each residue \( i \), its nativeness in a particular round can be calculated by averaging the TM<sub>i</sub> across the structures in the low energy ensemble from that round. The closer the average TM<sub>i</sub> is to 1, the more native-like residue \( i \) is in that round. In the lower plots in Figure 2.21, we plot this quantity across the various rounds for each of the 11 proteins (Chignolin ignored since no DESRES pathways available). We also plot an average TM<sub>i</sub> for the unfolded state of each of the proteins using the ensemble of the initial structures from first round of TerItFix for each protein.

In most of the cases, the TerItFix pathways appears to be consistent with DESRES pathways in terms order of structure formation quantified by metrics for distance from native structure for individual amino acids described above (Figure 2.21). Qualitatively, the same regions of the sequence appears to form early in both TerItFix and the all-atom DESRES simulations. This provides further validation of the fact that our method is able to identify gross aspects of folding pathways for proteins. We consider this to be a significant result given
the vast disparity in computational time and resources required for the two methods. The
most significant deviation between the TerItFix and DESRES predictions appears to be in the
observed propensities of unfolded state for all the proteins. Particularly, DESRES all-atom
MD simulations appear to consistently have very high structural content in their unfolded
states compared to TerItFix. This could be either be an artifact of the forcefield used or an
issue of definition when partitioning the MD trajectories into folded and unfolded regions.
While studies of structural content in the unfolded state of various proteins under native-
like conditions is an active area of research, at least in the case of λ repressor, the all-atom
simulations appear to grossly overestimate the helical content compared to experimental
data [67], which is closer to TerItFix’s prediction as discussed earlier. A further validation of
our description of the unfolded state comes from that fact that the unfolded state ensembles
generated from our nearest-neighbor dependent coil library appears to reproduce the NMR
residual dipolar coupling patterns for multiple proteins [56].

2.6 Discussion

The aa sequence of a protein codes for its structure as well as the energy landscape that guides
it to that structure. Thus, a fundamental challenge is to identify the basic principles that
enable the prediction of folding pathways and structure from sequence alone. The present
work is notable in the integration of the prediction of both structure and folding pathways,
and in producing agreement with experiments for diverse systems beginning from a realistic
unfolded state and using a computationally rapid model lacking explicit side chains.
The three primary components of protein structure - local backbone propensities, hydrogen bonding and tertiary packing - are combined with the principle of SS to guide the search process by iteratively fixing secondary and tertiary structure. The use of energy functions specifically designed to capture the major stereo-chemical properties (e.g., orientation dependence of pairwise interactions, backbone desolvation, neighbor effects on dihedral preferences) enables the method to describe subtle influences of the primary sequence on the energetic landscape. Previous methods also have used hierarchical approaches to build protein structures, while others, including our own [32], integrate secondary and tertiary structure prediction [48] [49] [50] [51] [52]. A hybrid version of TerItFix utilizing sequence but not structural homology [53] has been validated in CASP8 & 9 and ranks as one of the best groups in the CASP9 refinement category that involves improving template-based models to solve the crystallographic phase problem [22]. Nonetheless, these methods still primarily focus on one aspect, either structure prediction or the folding mechanism.

Our approach departs from Gō-like methods that require knowledge of the native state and invoke the assumption that folding is driven by native interactions on funneled energy landscapes with minimal energetic frustration [5]. While the Gō landscape might describe many features of folding, its predictive power can be limited when non-native interactions are important [75] [11] [21], or when slight changes in the aa sequence can drastically alter the folding properties [12], as occurs for the Im proteins discussed here. Gō variants exist that employ sequence dependence and even all-atom representations [76] [14] [15] [77] [78], but knowledge of the native state is still required.
While the detection of kinetic traps is one success of our method, resolving them remains difficult. To counter this difficulty, we refold the protein in every round starting from an extended conformation but using the information garnered from the previous round in the form of sampling and energetic biases. Because the prior information is implemented as biases, rather than as enforced contacts, both native and non-native contacts can weaken in successive rounds. For example, the native-like contacts between H1-H2 in λ repressor form early, are lost in middle rounds, and then reappear in later rounds. Im7, however, provides an example where the new contacts cannot override the earlier, non-native ones, and the protein becomes trapped in an intermediate state. A signature of a kinetic trap in our simulations is the presence of region(s) whose structural diversity varies within and between rounds. Potentially, the threshold for fixing secondary structure assignments and biasing tertiary contacts can be reduced to drive the escape from the trap.

Another impediment to modeling protein folding is the inherent difficulty of correctly balancing the energies associated with different types of contacts and backbone geometries. Small errors in the energy function, or the lack of explicit hydrogen bonds and backbone \( \phi, \psi \) dihedral angles, can greatly impact the order of structure formation and the location of the TSE on the reaction surface. These issues contribute to the inability of nearly all prior methods to accurately describe the TSE of Protein L [21] and Protein A [66]. The TerItFix algorithm’s central feature of coupling the secondary and tertiary structure by iterative fixing and SS helps identify low energy pathway(s) with the proper order of structure formation. Nevertheless, we experience difficulty identifying the TSE for Protein L and Protein A. Even
though simulations for both these proteins converge within three rounds, ascertaining the TSE is difficult and requires auxiliary information. Our prior $\psi$ analysis studies of four proteins with disparate RCO levels indicate that their TSEs acquire a similar fraction of native topology, $\text{RCO}_{\text{TSE}} \sim 0.7 \cdot \text{RCO}_{\text{Native}}$ [55] [64] [21] [42]. Accordingly, we cluster all structures from the TerItFix simulations whose RCOs are between 60 and 80% of the native value to identify a TSE (Figure 2.22). The major cluster for Protein L has both hairpins folded, in agreement with experiment. But the amino portion of the helix is also folded, which is not observed experimentally [21]. Overestimation of the helical content in Protein L’s TSE is typical of other methods as well [21]. The TerItFix-determined TSE for Protein A has H1 and H3 along with a kinked helix H2. This structure is close to experiment, except that in the experimental studies, the ends of H1 and H3 are frayed and H2 is not kinked.

2.7 Conclusion

We present TerItFix, a holistic approach for predicting pathways and structure that couples basic principles of protein chemistry with a realistic and robust search strategy involving sequential stabilization to find low energy folding routes. Central to the TerItFix folding algorithm is the progressive learning and biasing of secondary structure, tertiary contacts, and backbone hydrogen bonding. Information learned in one round of folding simulations is used in the following round. This work demonstrates that the empirical principle of SS can be applied as a computational strategy to predict both pathways and structure. By unifying the determination of folding mechanism and prediction of structure, this work has positive
Figure 2.22: Transition state ensemble for Protein L and Protein A selected by clustering structures with $0.6 < \text{RCO/RCO-Native} < 0.8$
implications for both areas. Because no knowledge about the native state is required, we can predict non-native kinetic traps and structures. Our nature inspired computational search strategy can benefit the prediction of larger proteins, one of the major frontiers of the field. Finally, our work is equally applicable to fast or slow folding proteins and thus provides a suitable alternative for cases that are outside the range of current MD simulations. Moving forward, we plan to use TerItFix predicted steps as an initial path to launch MD simulations, which could then be connected using Markov-state models or network analysis to obtain the energy surface and a more complete description of the kinetics, including timescales and barrier heights.
CHAPTER 3

INVESTIGATIONS OF NON-NATIVE INTERACTIONS IN
FOLDING OF PROTEIN L AND PROTEIN G

Parts of this chapter appear in Yoo, T. Y., Adhikari, A., Xia, Z., Huynh, T., Freed, K. F., Zhou, R., and Sosnick, T. R. (2012) The folding transition state of protein L is extensive with nonnative interactions (and not small and polarized), J. Mol. Biol. 420, 220-234. Experiments described in this chapter were performed by Tae Yeon Yoo and all-atom simulations were performed by Zhen Xia in the IBM group. The results on protein L have been published, but results on protein G have not yet been published.

Two proteins with identical topology but different sequences, protein L and protein G, have served as a model system for numerous folding studies, most of which focus on characterizing the rate-limiting step for folding. Here, we determine the structure of the transition state for Protein L because standard mutational methods identify an inordinately small structure (a single hairpin). Using the relatively new $\psi$ analysis approach that employs metal ion binding sites, we find that the transition state contains the entire four-stranded $\beta$ sheet. One hairpin is non-native, a significant result that agrees with PDB-based backbone sampling and all-atom simulations. The non-native character partially explains the failure of accepted experimental and native-centric computational approaches to adequately describe the transition state. We also provide active computational predictions of the structure of the transition state and possible non-native interactions in protein G, which we plan to test.
using experiments. This study stresses that caution is required even when agreement exists between experiment and theory, thus highlighting the importance of having alternative methods for characterizing transition states.

I present the study on protein L using both computer simulations and experiments first. Towards the end of the chapter, I present active computational predictions for protein G, along with a comparative analysis of the transition states of protein L, protein G and DesG (redesigned protein G) as predicted by TerItFix.

3.1 Protein L

3.1.1 Introduction

The IgG binding domain of protein L (Protein L) contains two hairpins and a central helix and has been a test bed for many experimental and theoretical studies of folding [74, 76, 79, 80, 81, 82, 83, 84, 85, 86]. Mutational $\phi$ analysis experiments indicate that the folding transition state ensemble (TSE) contains only the amino terminal hairpin [74, 79, 80, 81] (Figure 3.1). The TS of a protein with the same $\alpha/\beta$ fold, Protein G, also is assigned by $\phi$ analysis to have a single hairpin, but this hairpin is located at the carboxy terminus [87], a behavior attributed to different properties of the turn sequences [88, 89]. The difference between the TSs of these two proteins is cited as an example where the specific sequence, rather than just the protein’s topology, influences the folding pathway. A variety of computational studies support this view [76, 82, 83, 84, 90].
Despite this broad consensus, we decided to reexamine the folding behavior of Protein L because a TS with only a single hairpin seems inordinately small. A hairpin scarcely defines Protein L’s topology, yet this protein obeys the well-known correlation between folding rate and topology (relative contact order, RCO)[6, 91]. Our studies of three other proteins with disparate RCOs, indicate that their TSEs acquire a similar level of native topology, $RCO_{TSE} \approx 0.7 \times RCO_{Native}$ [42, 44, 55, 64, 92]. If this relationship is generally applicable, it would provide a simple rationalization for the $k_f$-RCO correlation, as well as a constraint for possible TS structures of other proteins. In the case of Protein L, the presence of only a single hairpin in the TSE equates to a RCO fraction of only 25%, and even the inclusion of the helix would increase the RCO only to 40%. In order to achieve an RCO fraction close to 70%, the TSE must minimally include long-range contacts between the amino and carboxy terminal strands. Furthermore, whereas a 1:1 relationship between hydrogen bond content and surface burial is found in the TSs of other proteins [65, 66], the hydrogen bond content of a single hairpin is grossly inadequate to match the surface burial of the highly collapsed TS of Protein L.

Here we employ $\psi$ analysis [93] to characterize the TS structure of Protein L. The $\psi$ methodology employs bihistidine (BiHis) metal ion binding sites and is well suited for determining TS topology because $\psi$ directly identifies inter-residue contacts. The metal ion-induced stabilization of the TS relative to the native state is represented by the $\psi$ value that reflects the proximity of the two histidines in the TS. Data for a multitude of BiHis sites (individually introduced) can be used to generate structural models of the TS analogous to
the use of NOE distance constraints in NMR-based structure determination.

These experiments demonstrate that Protein L’s TS contains the entire four stranded $\beta$ sheet. Although the amino terminal hairpin is native-like, the carboxy hairpin and the long-range interactions between the two hairpins have non-native properties. We conduct simulations of the individual hairpins using our TerItFix folding protocols with a reduced representation, in which the side chains are represented by single C atoms [32, 53], along with all-atom, explicit solvent molecular dynamics (MD) simulations. Without invoking any knowledge of the native state, both methods indicate the carboxy terminal hairpin forms rapidly, but with a non-native turn. We discuss the implications of our findings with regards to TS topology, the accuracy of $\psi$ analysis, and its ability to validate theoretical studies.

3.1.2 Overview of experimental results

$\psi$-analysis [42, 64, 66, 92, 94] proceeds by introducing BiHis metal binding sites at specific positions on the protein’s surface. Upon addition of metal ions, these sites stabilize secondary and tertiary structures because an increase in metal ion concentration stabilizes the interaction between the two histidine partners. The metal ion-induced stabilization of the TS relative to the native state is represented by $\psi$. It reports on the proximity of the two partners in the TS and can be used to generate structural models of the TS analogous to the use of NOE distance constraints in NMR-based structure determination with $\psi$ values of zero or one indicating that the biHis site is absent or native-like in the TS, respectively. Intermediate values arise when the TS recovers only part of the binding-induced stabilization.
in the native state.

A total of eight biHis sites are individually introduced into Protein L (Figure 3.1). The folding properties of each biHis variant are examined from two separate approaches. First, the guanidinium chloride (GdmCl) dependences of the relaxation rates (chevron analysis) are obtained in the absence and presence of 1 mM zinc or nickel ions at 22°C, pH 7.5. Kinetic data also are taken at dozens of different metal concentrations under strongly folding (\(\sim 0.6\) M GdmCl) and strongly unfolding conditions (\(\sim 3.5-5\) M GdmCl). The changes in stability and activation free energy, and , arise from the difference in metal ion affinities \(K_{eq}\) of the biHis site in the unfolded (U) state, native (N) state, and TS (Eq. 3.1 and 3.2).

\[
\Delta \Delta G_{eq}(\text{[Me}^2+]\text{])} = RT \ln \left(1 + \frac{[\text{Me}^2+]}{K^N}\right) - RT \ln \left(1 + \frac{[\text{Me}^2+]}{K^U}\right)
\] (3.1)

\[
\Delta \Delta G_f(\text{[Me}^2+]\text{])} = RT \ln \left(1 + \frac{[\text{Me}^2+]}{K^{TSE}}\right) - RT \ln \left(1 + \frac{[\text{Me}^2+]}{K^U}\right)
\] (3.2)

These ion-induced changes in free energies are used to define the \(\psi\) value, a parameter analogous to the mutational \(\phi\) value, although \(\psi\) is the instantaneous change as a function of metal ion.

\[
\phi = \frac{\Delta \Delta G_f}{\Delta \Delta G_{eq}}|_{mutation}
\] (3.3)
Figure 3.1: Transition state of protein L according to $\phi$ analysis vs $\psi$ analysis. A) $\phi$ and $\psi$ values mapped onto Protein L. The $\psi$ are obtained from the present study, while the $\phi$ are from ref. [74]. When multiple mutations are studied at a particular site, more emphasis is placed on the X-to-Ala and Ala-to-Gly comparisons for core and helical mutations, respectively. The model of the TS is created by docking the two hairpins using the TerItFix protocols. B) The locations of the biHis sites and $\psi$ values are shown (each site is studied individually).
\[
\phi = \frac{\Delta \Delta G_f}{\Delta \Delta G_{eq} |\text{Me}^{2+}|}
\]  
(3.4)

To remove potential artifacts related to the metal ion binding altering the folding behavior, a \(\psi_0\) is obtained by extrapolating the fitted data to zero metal ion concentration. \(\psi_0\) thus reflects the intrinsic degree of contact formation in the TSE in the absence of metal. This important feature of \(\psi\) analysis lacks a counterpart with \(\phi\) analysis which does not typically account for the effects of even sizable substitutions on the TSE. \(\psi_0\) is calculated from the equation:

\[
\psi_0 = \frac{e^{\Delta \Delta G_f} - 1}{e^{\Delta \Delta G_{eq}} - 1}
\]  
(3.5)

A \(\psi_{\text{chevron}}\) is evaluated from the shifts of the folding and unfolding arms of the chevron plot upon addition of metal ions. The second, independent determination, termed \(\psi_{\text{leffler}}\), is obtained from the fit of a Leffler plot of \(\Delta \Delta G_f\) versus \(\Delta \Delta G_{eq}\) using the kinetic data taken at dozens of metal ion concentrations for both strongly folding and unfolding conditions.

The observed sensitivity of folding rates to metal ion concentration indicates that in the TS, all six biHis sites on the \(\beta\) sheet are metal binding competent (Figure 3.1). The helix is largely absent in the TS, a feature anticipated by its low helical propensity [95] of less than 2% and small hydrophobic content present because all three hydrophobic residues on the helix-sheet interface are alanines rather than larger hydrophobic residues.

As depicted in Figure 3.1, the behavior for sites a and b across the amino hairpin is
consistent with a native-like turn. However, the folding behavior for sites c, d, e and f on the carboxy hairpin and between the two hairpins are indicative of non-native structure. Upon addition of metal ions for sites c and d between the two hairpins, the activation energy for folding is affected more than the equilibrium stability. Consequently, $\psi_0$ greatly exceeds unity, with $\psi_0^{eff} = 1.24 \pm 0.07$ and $3.33 \pm 0.40$ for sites c and d, respectively. The metal ion binding affinities in the TS are 1.5 and 3-fold times tighter than in the native state, respectively. Hence, the two hairpins are in close proximity in the TSE, but the two histidines reside in a non-native configuration where they can bind metal ions more tightly than in the native orientation.

Site f is located at the distal end of the carboxy hairpin and presents folding and unfolding rates that are equally accelerated upon addition of zinc ions. The metal ion binding affinity in the TSE is 2.6-fold stronger than in the native and unfolded states. Thus, the change in equilibrium stability is near zero, and $\psi_0$ is ill-determined. Nevertheless, the stronger metal ion binding affinity in the TS indicates that both histidines form a binding site in the TS, but with a non-native orientation.

A stronger metal ion binding affinity in the unfolded state is observed for site e. This result can be explained by the presence in the denatured state of residual structure with a biHis arrangement that binds zinc ions stronger than the biHis site in the native state. For the site h variant in Zn$^{2+}$, both the folding and unfolding rates mildly decrease, and the change in stability is near zero, $\Delta\Delta G_{eq} = 0.33 \pm 0.20$ kcal mol$^{-1}$. This unexpected downward shift in the chevron plot renders $\psi_0^{chevron}$ ill-determined. Although the Ni$^{2+}$ data
imply the absence of the helix in the TS, the Zn$^{2+}$ data for the site h variant imply that the TSE binds metal ions more weakly than either the unfolded or folded states, e.g., $K_{eq}^{T_{seq}}/K_{eq}^{N} = 1.99 \pm 0.71$. Potentially, this part of the chain adopts a non-helical arrangement in the TS, with an extended backbone geometry in which the histidines are located farther apart than in the unfolded state (where transient helix formation can occur) and in the native state (which is helical).

3.1.3 Hairpin simulations and TS modeling of protein L

To gain further insight into the non-native $\psi$ values, we conduct simulations of the individual hairpins using both our homology-free folding algorithm [32, 53] and standard, explicit solvent MD simulations. Both methods do not invoke any knowledge of the native state, and they concur that the carboxy terminal hairpin forms with a non-native turn geometry. This result is a consequence of the native turn having three consecutive residues with positive backbone $\phi$ dihedral angles (Figure 3.3a).

The TerItFix Monte Carlo simulated annealing (MCSA) simulations represent each side chain with a single C$\beta$ atom. The conformational search space is restricted by iteratively fixing secondary structure assignments of certain portions of the sequence after incorporating the influence of tertiary context, thereby coupling secondary and tertiary structure determinations. The move set involves changes only in a single residue’s $\phi, \psi$ backbone dihedral angles (i.e., not a fragment insertion method). The dihedral angles are derived from a PDB-based coil library lacking helices and sheets, and the angles depend on the amino
acid type of the residue and each flanking residue. The energy function is the sum of our orientational-dependent DOPE-PW statistical potential [32] plus a burial term based on the number of heavy atoms in a 11 Å hemisphere in the direction of the $C\alpha$-$C\beta$ vector for each amino acid type[47].

Energy minimization in each successive round proceeds with an energy function that also includes the distance constraints derived from the average residue-residue contact and hydrogen bond maps derived from the structures with the 25% lowest energies that are generated from the previous round. Hence, information concerning both secondary and tertiary structure learned in prior folding rounds is carried over to the next round. This iterative fixing protocol mimics the sequential stabilization process observed in the folding of real proteins wherein steps represent the building of new portions on top of existing structures [36].

The TerItFix algorithm predicts that the amino hairpin folds to a near-native conformation (Figures 2.17, 3.2). However, the carboxy hairpin folds into a non-native hairpin with a two-residue registry shift in the hydrogen bonding partners (Figure 3.4). This hairpin folds in a single TerItFix round, with 98% of the structures containing hydrogen bonds between the three non-native residues. In contrast, the amino terminal hairpin requires two rounds of folding to obtain a native hairpin conformation. After the first and second rounds, 19% and 81% of the native hydrogen bonds are formed, respectively. Both predicted hairpins are Type 1 $\beta$ turns, which have dihedral angles that correspond to the dominant dihedral angles in the PBD-based sampling library. The non-native character of the carboxy hairpin
emerges because the native $\phi$ angle is positive, is therefore unfavorable for turn residues D53 and K54, and is thus only infrequently sampled during the TerItFix simulations.

Over the course of the microsecond all-atom MD simulations, both hairpins sample a variety of conformations, with an RMSD from the native state for the six turn residues fluctuating between 1-5 Å (Figure 3.3b). However, the amino hairpin has similar RMSDs to the native and the TerItFix-predicted turns, consistent with their joint RMSD being only 1 Å. Further, the MD simulation finds the amino turn within 1 Å of the native turn for about 20% of the trajectory. In contrast, the RMSD of the carboxy hairpin to the native turn remains above 2 Å. For half of the MD trajectory, the RMSD to the TerItFix prediction is closer, $\sim$ 1 Å. Moreover, the dihedral angles of the turn residues from the MD simulation lie in the same basins as the TerItFix predictions. In summary, both computational methods are consistent with the experimental findings of a native-like geometry for the amino hairpin but a non-native geometry for the carboxy hairpin.

Models of the TS are generated using the TerItFix algorithm by docking the native-like amino terminal hairpin against the non-native carboxy hairpin (Figure 3.5). Docking is achieved by introducing an additional interaction term between the terminal strands and employing single $\phi,\psi$ pivot moves of the unstructured residues between the two hairpins. Ten such structures are selected and further refined in a second MCSA round using our “double-crank” local move set that features compensating $\phi,\psi$ counter-rotations of the neighboring residues.

A considerable extent of Protein L’s native state topology is formed in these TS models.
After insertion of side chains using SCWRL [96], their RCO is 73 ± 5% of the native value.

The fraction of buried surface in the TS models, relative the native state and normalized to an unfolded state ensemble [56], is close to the fraction of hydrogen bonds formed in the TS, approximately 55% and 50%, respectively.

![Rate of β hairpin formation in protein L. The carboxy-terminal hairpin forms in fewer simulation rounds than the amino-terminal hairpin. The carboxy hairpin folds in a single round, with 98% of the structures containing hydrogen bonds between the three nonnative amino acid partners. In contrast, the amino-terminal hairpin requires two rounds of folding to obtain a native hairpin conformation. After the first and second rounds, 19% and 81% of the native hydrogen bonds are formed, respectively. The y-axis presents the fraction of the maximum number of hydrogen bonds for the hairpins. The color intensity reflects the native character (0 = nonnative and 1 = all native).](image_url)
Figure 3.3: Simulations of the protein L $\beta$ hairpins. A) TerItFix predictions overlaid on the native hairpins. Ramachandran distributions during the simulations displayed for both TerItFix and all-atom MD simulations. The native (green arrow) and predicted (red arrow) torsional angles are depicted. B) The RMSD values of the six turn residues in the MD simulations to the native and to the TerItFix predicted turns.
Figure 3.4: **Native (Tube) vs predicted (Trace) overlay of the C terminal hairpins for protein L.** The hydrogen bonding partners shift by two amino acids (e.g., THR45) in the predicted hairpin, but the hydrophobicity pattern remains native-like.

### 3.1.4 Discussion

The present study is motivated by the belief that a TSE for Protein L containing only a single hairpin, as suggested by φ analysis \[74, 79, 80, 81\] is unreasonably small because it scarcely defines the protein’s topology or has enough hydrogen-bonded structure to be commensurate with the observed degree of surface burial. The recent experimental ψ analysis has demonstrated that the TSE is extensive, containing the entire β sheet network along with some nonnative structure associated with the carboxy hairpin. According to our simulations, this hairpin adopts a nonnative registry by virtue of the highly unfavorable dihedral angles in the native turn. The same nonnative registry is observed in both the PDB-based backbone sampling TerItFix algorithm and the all-atom simulations. These two methods also concur
on the native-like geometry of the amino hairpin, consistent with the experimental $\psi$ data.

These findings have extensive implications concerning the role of nonnative structure, the relationship between TSE topology and folding rates and the malleability and multiplicity of TSs and whether mutational $\phi$ analysis, which has been the primary method for comparing experiment and theory, is sufficiently reliable that agreement between the two is an adequate validation of both approaches.

We observe six values of $\psi_0$ equal to unity or larger for biHis sites extending across the four $\beta$ strands and two near-zero $\psi_0$ values on the helical sites. This pattern indicates that the TSE contains the entire $\beta$ sheet network but minimal helical structure. The $\psi_0$ values greatly exceed unity for sites situated across strands $\beta_1$-$\beta_4$ and $\beta_3$-$\beta_4$ ($\psi_0$ site d = 3.3 ± 0.4 and $\psi_0$ site f $\gg 1$), a behavior indicating that, in the TS, the biHis site has a geometry with tighter metal ion affinity than the site experiences in the native state. Taking further advantage of $\psi$ analysis’ ability to individually determine the metal ion binding affinities in the unfolded, transition and native states, we identified the presence of residual structure in the denatured state for the turn region of the carboxy-terminal hairpin. We emphasize that $\psi_0$ is the limiting $\psi$ value in the absence of metal ions. Therefore, these properties are intrinsic to the folding behavior of Protein L and are not artifacts induced by metal ion binding.

The absence of the large helix in the TSE is attributable to its sequence having a low average intrinsic helicity: $< 2\%$. In addition, all three hydrophobic residues on the buried helical face, which docks against the $\beta$ sheet, are alanines rather than larger hydrophobic residues.
The alanines’ low hydrophobicity reduces the driving force for helix-sheet association.

Evidence that the amino acid sequence, rather than topology, can control the structure of the TSE has been found in experimental ϕ analysis and computational studies comparing the folding behavior of Protein L and Protein G, two proteins with the same α/β fold [74, 76, 79, 80, 81, 82, 83, 84, 85, 86]. The possibility of different sequences having alternative TSEs with one hairpin or the other formed implies that a hybrid sequence could fold with significant flux going through two structurally disjoint TSEs. However, our present finding that Protein L’s TSE contains both hairpins currently precludes the use of the Protein L/Protein G comparison as evidence either for sequence altering the TSE structure or for the possibility of structurally disjoint TSEs.

Our ψ-based models of the TSE (Figure 3.5) are parsimonious with prior data. The fraction of the surface buried in the TSE models is close to the fraction of hydrogen bonds formed in the TSE. These findings are consistent with kinetic isotope studies that indicate the presence of a commensurate level of surface burial and hydrogen bond formation in the TSE [65, 66]. The underlying principle is that hydrophobic association leads to partial backbone desolvation that can be offset by protein-protein hydrogen bonding.

A considerable extent of Protein L’s topology is formed in the TSE. Our TSE models have $\text{RCO}_{TSE} \approx 0.7 \cdot \text{RCO}_{Native}$, in agreement with our prior ψ studies for ubiquitin, [55, 64] and acyl phosphatase [42, 92] and the B domain of Protein A [42]. Because these four proteins have native RCO values that span the range observed for two state proteins, the 70% value is likely to be generalizable to other proteins that obey the well-known $\text{RCO-k}_f$
trend [6].

Similar relationships emerge from other studies [97] with certain Gō-based models producing a CO at the 60 to 80% level [68, 98, 99]. However, their TSE structures generally correspond to a uniformly “expanded version” of the native structure [98], a finding that seems inconsistent with ψ data (many ψ values are either zero or near unity). The use of data from φ analysis produces an $RCO_{TSE}$ fraction closer to 50% [100], supporting the contention that φ analysis can underreport the structural content of the TSE.

The φ values predicted by Gō models exhibit mixed agreement with experimental φ values for Protein L [76] and some other proteins [8, 9, 10, 101, 102]. The TSE structures for Protein L from the Gō models typically contain just the amino hairpin and the helix [76, 82, 83, 90] but none correctly reproduces the ψ-determined TSE containing only the four strands. The closer agreement between the Gō models and φ analysis may partially be a consequence of their shared native-like biases.

The inability of the Gō-type simulations to correctly predict the four stands in the TSE of Protein L is likely due to both the presence of nonnative interactions and the inherent difficulty of correctly balancing the energies associated with different sets of contacts and backbone geometries. Small errors in the energy function or the lack of explicit hydrogen bonds and backbone φ,ψ dihedral angles can greatly impact the order in which structure forms and the location of the TSE on the reaction surface. These issues contribute to the inability of nearly all methods to accurately describe the TSE structure of a Protein L, as well as of the B domain of Protein A [42].
Figure 3.5: **TSE determined by \( \psi \) analysis is more ordered than that identified by \( \phi \) analysis.** Although the \( \phi \)-determined TS depicted for Acp includes three \( \beta \) strands, only the presence of a hydrophobic core is indicated, and the presence of native-like strand-strand interactions is indeterminate at best [103]. Further, the model shown accounts for the remeasured near-zero \( \phi \) value on F94 that indicates that \( \beta_5 \) is absent in Acp’s TSE. For BdpA, the level of the participation of helices H1 and H3 in the TS is unclear according to \( \phi \) analysis [104]; the model depicted reflects a combination of data from \( \psi \) and \( \phi \) analysis and H/D kinetic isotope effects.
3.1.5 *Comparison to φ analysis*

Mutational φ analysis is the most accepted method for characterizing TSEs, developing models for folding and validating theoretical approaches. [9, 76, 101, 102, 105] However, the present ψ analysis findings of an extensive TSE in Protein L significantly differ with that generated based on φ analysis. The φ analysis method indicates that Protein L’s TSE contains only the amino hairpin [74, 79, 80, 81, 88] (Figure 3.1). Seven sites on this hairpin yield φ > 0.6 (although another five positions have φ below 0.31). Six positions on the carboxy-terminal hairpin produce much lower average φ of 0.13, while the values at two other positions are slightly higher (φ-T48A = 0.26 and φ-V49A = 0.31). Similarly low φ are found on the helical sites.

The primary differences between ψ and φ analyses arise because the former directly probes residue-residue contacts between two known partners, whereas the latter reflects energetic perturbations upon mutation. These perturbations may be the consequence of a combination of factors, including changes in the local side-chain environment and backbone dihedral propensities. In ψ analysis, the binding of increasing concentrations of ions to the biHis site produces a nearly continuous increase in the stability of TSE structures that contain the binding site. Hence, the stability is perturbed yet accomplished in an isosteric and isochemical manner. The resulting series of data can be justifiably combined, and the ψ₀ value can be extracted as devoid of any perturbation due to ion binding. The ability of eliminating the influence of perturbations may be inaccessible to traditional mutation studies where the perturbation can arise from multiple sources, including changes in backbone
propensities as well as indeterminate nonlocal interactions.

These differences become critical for Protein L for two reasons, the nonnative character of the carboxy hairpin and the exposure of the \( \beta \) sheet’s hydrophobic face in the TSE. The 2-aa register shift in the carboxy hairpin indicated by the simulations results in nonnative contacts along this hairpin and nonnative dihedral angles in the turn. Consequently, the energetic perturbation realized in the TSE likely is smaller than in the less accommodating native state, and \( \phi \) therefore becomes small and mistakenly identifies this hairpin as being absent in the TSE.

The second issue arises because the otherwise buried side chains on the hydrophobic face of the sheet are solvent exposed in the TSE due to the absence of the helix. Consequently, the energetic penalty for the truncation of the side chain for the residues on the inner face of hairpins, for example, imparted by an alanine substitution, is diminished in the TSE relative to the native state, even though the residue is in a hydrogen-bonded \( \beta \) structure. This analysis provides an explanation for the low-to-moderate \( \phi \) values for the native-like amino hairpin [74, 79, 80, 81, 88]. The issue is generally relevant whenever \( \phi \) analysis is applied at any position that is more exposed in the TSE than in the native state.

Overall, these considerations support the contention that \( \phi \) analysis can underestimate or misrepresent the structural content of the TSE [20, 55, 93, 94, 106] due to chain relaxation and accommodation or to nonnative interactions. [107, 108] For example, a residue in fyn SH3 with a helical conformation in the native state adopts a \( \beta \) conformation in the TSE despite having a high canonical \( \phi \) value of 0.7 [109]. A similarly positioned residue in src SH3
also contains a productive nonnative conformation in the TSE [110]. Likewise, in nonnative regions of the cytochrome b562 intermediate, seven high \( \phi \) values are observed \((0.4 \leq \phi \leq 1.0)\) [108]. Conversely, low \( \phi \) values are found in regions of native-like structure in BPTI intermediates [106].

In addition to Protein L, significant underreporting of the TSE’s structural content by \( \phi \) analysis also occurs with acyl phosphatase [103, 111, 112], ubiquitin [20, 94] and the B domain of Protein A [42] (Figure 3.5). The \( \psi \)-determined TSEs for these proteins are extensive and contain persistent native-like tertiary interactions. Unambiguous sites where \( \psi \) is unity indicate that the TSEs of acyl phosphatase and ubiquitin contain a four-stranded \( \beta \) sheet and an \( \alpha \) helix. For these two \( \alpha/\beta \) proteins, \( \psi \) analysis detects the presence of one and two additional long-range \( \beta \) strands than \( \phi \) analysis identifies, respectively. We suspect that underreporting occurs with proteins having a TSE characterized by \( \phi \) analysis as polarized, such as cold shock protein, src SH3 [113] and Protein G [87].

### 3.1.6 Conclusion

We demonstrate that the highly studied TSE state of Protein L is extensive and has nonnative properties that likely arise due to the presence in the native state of backbone dihedral angles that are not highly populated at the earliest stage of folding. The TSE is significantly larger than the one identified by \( \phi \) analysis, a result found in the three other proteins probed by \( \psi \) analysis (Figure 3.5). The difference arises because \( \psi \) directly identifies inter-residue contacts between known partners, while \( \phi \) is native centric and the TSE can be less sensitive to
energetic perturbations than the native state even for structured regions. These observations suggest that identification of a TSE as being diffuse, polarized or an expanded version of the native state based on $\phi$ analysis alone should be reconsidered.

Our results also emphasize that apparent agreement between the $\phi$ values and Gō-based models [102] can produce an overly optimistic view of these methods’ ability to accurately determine TSE structures. Accurate modeling of protein folding remains an ongoing challenge, requiring the proper balancing of numerous factors in a changing contextual environment as the chain folds. $\psi$ analysis should be applied to other proteins to address these outstanding issues, search for other nonnative TSE structures and provide a robust test set for benchmarking simulation, which may lead to better agreement between theory and experiment.
3.2 Protein G

3.2.1 Hairpin simulations: Preliminary Results

Protein G has an identical topology to protein L, but its sequence is completely different. The two proteins, therefore, serve as a model system for studies aimed at examining the interplay between sequence, protein fold and pathway. While the experiments on \( \psi \) analysis are ongoing, I've performed simulations of the protein G hairpins analogous to those in protein L. Figure 3.6 show comparisons of the TerItFix simulations to all atom MD simulations performed on the individual hairpins of protein G. Unlike in protein L, the C terminal hairpin of protein G appears to fold easily into a native-like conformation. However, the N terminal hairpin of protein G can fold into two possible hairpin conformations picked up by clustering the TerItFix predicted hairpin conformations, one with a native-like turn and another with non-native conformation. Comparisons of the all-atom molecular dynamics simulations for the C terminal hairpin overlap in terms of similarity to the native and TerItFix prediction hairpins, unsurprisingly. For the N terminal hairpin, although all-atom simulations are not entirely conclusive and perhaps longer simulations need to be performed, the hairpin appears to be closer to the TerItFix predicted native-like turn compared to the TerItFix predicted non-native turn, as shown by the top plot in Figure 3.6.
**Figure 3.6**: Simulations of protein G hairpins using TerItFix and all-atom molecular dynamics
3.3 Comparative analysis of protein L, protein G and DesG:

**TerItFix predictions**

Most of the TerItFix predictions presented in this thesis, while consistent with experiments, weren’t the basis for new experiments. Besides providing new insights, a crucial validation of any predictive algorithm involves making testable predictions, along with suggestions for new experiments which could test those predictions. In that spirit, I here present some active TerItFix predictions regarding the folding mechanisms of protein L and protein G, which needs to be tested experimentally. I include a third protein as well, DesG, which was a redesigned version of protein G [114] that has previously only been studied using all-atom molecular dynamics simulations [16].

### 3.3.1 TerItFix Observations Summary

Figure 3.7 (top) shows the evolution of secondary structure from the TerItFix predictions (for the whole sequence) of protein L, protein G as well as DesG. For all three proteins, most of the native topology was obtained by round 3. Since most of the structure appear to form between R2 and R3, the structures generated in those rounds which satisfied the criteria: $0.6 < \frac{RCO}{RCO_{Native}}$ were clustered, and the largest two cluster centroids were used as candidates for the transition state structures (Figure 3.7 (bottom)), a heuristic obtained from experimental observations that folding transition states of several proteins have around 70% of the native structure already formed [111].
Figure 3.7: Comparative TerItFix predictions of protein L, protein G and DesG (redesigned protein G)
The comparative TerItFix predictions for the three proteins is summarized below:

- **Protein L.** The C terminal hairpin turn of protein L is non-native, as already explained in the earlier part of this chapter ([Yoo et. al. [21]]). The helix is partially to fully formed but may not be docked well to the rest of the $\beta$ hairpins in the transition state for protein L. Both the proposed TS structures had the C terminal hairpin formed completely (non-native turn) as well as a third strand docked to the C terminal hairpin. However, the N terminal hairpin could be either totally or partially formed according to my predictions.

- **Protein G** has the helix most readily form of the three proteins. The N terminal hairpin of the protein G can have two possible conformations (native as well as not native), consistent with the simulations of the isolated hairpins from the earlier section. The origin of the non-native turn in protein G is discussed later in the chapter. Two distinctive transition state clusters were found: one where both the hairpins are formed and docked in native geometry and another where the N terminal (non-native geometry) hairpin docks to the C term 3rd strand without the C term hairpin fully formed.

- **DesG** has both the two hairpins form early on compared to wildtype protein G. Both hairpins appear to be fairly independently stable and likely form before they dock. The transition state therefore could have multiple docking poses of the two hairpins, not necessarily in the native geometry. The helix was found to be partially to fully formed.
in the transition state.

Summary. The major conclusion from the TerItFix simulations is that both protein L and protein G could have non-native hairpins present in the transition state. While this was verified for protein L using \( \psi \) analysis, similar analysis needs to be performed on protein G. Therefore, based on the TerItFix predictions, one testable hypothesis is that both protein G and L have extensive topology formed in the transition states and the fact the earlier \( \phi \) experiments and native-centric simulations saw polarized transition states for the two proteins might highlight their inability to distinguish non-native interactions, unlike our TerItFix method.

3.3.2 TerItFix recommended experiments

Based on the observations in the earlier section, I recommend the following experiments to test our predictions:

- One of the main active predictions is that the protein G N terminus hairpin can have alternative registries based on native or non-native turn geometries. Hence, \( \psi \) analysis needs to be performed on that hairpin by creating multiple metal binding sites consistent to each of the two registries, and verify whether both the registries can only be observed in the transition state. It is entirely possible however that one of the registries is resolved by the time the protein reaches the transition state, making the task of discrimination harder by standard \( \psi \) analysis.
Another strong prediction from TerItFix is regarding the helicity of the three proteins. TerItFix predicts that the wildtype protein G has the helix most readily formed, which can also be testing by creating multiple metal binding sites across the helix for the three proteins.

3.3.3 Explanation of non native N terminus hairpin in protein G

The N terminus hairpin for protein G in its crystal structure has a canonical Type I turn. TerItFix predicts that an alternate non-native Type I’ turn is possible for the N terminal hairpin where the turn residues are shifted by two residues from the native turn. The Ramachandran propensities of the turn residues in the protein G sequence seem to explain this behavior (Figure 3.8). The sequence near the N terminal turn for protein G is ..ILNGKTL.., where NG is involved in the native Type I turn and KT is involved in the non-native Type I’ turn. In other words, the maxima in the nearest-neighbor dependent Ramachandran distributions for residues “NG” is consistent with a Type I turn and that for residues “KT” is consistent with a Type I’ turn, seen in Figure 3.8.

A further justification comes from a previous study where the N terminal turn residues of the wild-type protein G were redesigned to create the protein, NuG2 [88, 114], where φ analysis indicated that the redesigned N terminus hairpin folded much faster than the C terminal hairpin in the folded pathway. It is worth noting, however that in the wild-type protein G, φ analysis observed the C terminus hairpin folded first and was the only structure formed in the transition state. Interestingly, however, the N terminus turn residues in NuG2
are exactly the same as the TerItFix predicted non-native N terminus turn residues, i.e. residues “..NG..”. These observations taken together provides a strong hypothesis that a hairpin with “NG” as turn residues is both viable and stable and also can form quickly (as seen in NuG2). But, since in the wild-type protein G, the “NG” turn residues leads to a hairpin in a non-native geometry, it is entirely possible that previous experiments using mutational $\phi$ value analysis could not detect it.

Figure 3.8: Both native and non native turns are possible for N terminus hairpin of protein G

![Image of protein structure and turn angles](image)
CHAPTER 4

MODELING LARGE REGIONS IN PROTEIN STRUCTURES


Jian Peng and Jinbo Xu provided the Raptor starting models used in this chapter, while Michael Wilde provided assistance with the Swift scripting language.

Template-based methods for predicting structure provide models for a significant portion of the protein but often contain insertions or chain ends (InsEnds) of indeterminate conformation. The local structure prediction problem entails modeling the InsEnds onto the rest of the protein. A well known limit involves predicting loops of $\leq 12$ residues in crystal structures. InsEnds, however, may contain as many as $\sim 50$ amino acids, and the template-based model of the protein itself may be imperfect. To address these challenges, we present a free modeling method for predicting the local structure of loops and large InsEnds in both crystal structures and template-based models. The approach uses single amino acid torsional angle “pivot” moves of the protein backbone with a C$\beta$ level representation. Nevertheless, our accuracy for loops is comparable to existing methods. We also apply a more stringent test, the blind structure prediction and refinement categories of the CASP9 tournament, where we improve the quality of several homology based models by modeling InsEnds as long as 45 amino acids, sizes generally inaccessible to existing loop prediction methods. Our approach
ranks as one of the best groups in the CASP9 refinement category that involves improving template-based models so that they can function as molecular replacement models to solve the phase problem for crystallographic structure determination.

**ABBREVIATIONS**: amino acid, aa; Monte Carlo Simulated Annealing, MCSA; Ramachandran, Rama, nearest neighbor (NN), Solvent accessible surface area, SASA

### 4.1 Introduction

Homology-based methods use known structures as templates and have proven extremely successful in modeling larger proteins in a computationally efficient fashion. The success of these methods, however, depends on the quality of the alignments between the target sequence and those of the templates [115]. Frequently, the sequence alignments contain gaps that correspond to regions in the sequence where no reliable structural information can be extracted from the templates. These gaps may be insertions or additions at the termini (Figure 4.1). Inevitably, the model assembled from the templates lacks these local regions. In order to model the entire structure, alternative methods are required. The problem of reconstructing local regions in a protein is neither new nor exclusive to homology modeling. Experimentally determined structures from crystallography often contain regions that are difficult to characterize because they are flexible or mobile. Consequently, crystal structures can contain loops that have weak or missing electron density. This issue is particularly significant because protein function is often mediated by loops; for example, loops often act as molecular recognition or binding sites and play a crucial role in executing the protein’s
function [116] [117] [118]. The specificity of protein interactions as mediated by active sites and binding pockets is also a consequence of local protein structure. These issues highlight the need for reliable methods to reconstruct local regions in protein structures.

Three important problems arise in developing methods for predicting local spatial structure. First, the local regions must be modeled subject to the constraints imposed by the rest of the protein structure. E.g., the loop termini must end at the correct anchor positions. Some approaches to this long-standing loop closure problem seek analytical solutions to bond angles that properly position the ends [119] [120] [121]. While exact solutions have been found for short polypeptide segments, no general analytical solution is possible for segments containing more than a few amino acids in proteins. Other robotics-inspired algorithms for loop closure [122] [123] likewise experience decreasing accuracy as the size of the loops increases. Additionally, analytical approaches to the closure problem very often yield solutions that place backbone dihedral angles in disallowed regions of Ramachandran (Rama) space and thus generate sterically forbidden conformations.

Second, irrespective of how the loop closure is performed, a procedure is required for sampling various conformations of the local region. Existing approaches for predicting local regions in protein structures can be broadly categorized into two classes: database and de novo (free modeling) methods [124] [125]. Database methods search for loop fragments that best match the anchor geometries, [124] [126] but these approaches usually are confined to short insertions because of poor database coverage for larger fragments. While these methods tend to be fast, the speed comes at the cost of the greater flexibility in exploring
the conformational space of the loops permitted by free modeling methods. The applicability of these methods is further challenged in the modeling of InsEnds in template-based models, because the regions are likely to correspond to parts of the sequence that are inaccessible to the homology methods. In contrast, de novo methods sample sterically feasible loop conformations that are scored with physics-based or statistical potentials. For example, MODELLER places loop atoms uniformly between the anchor positions and optimizes the atom positions using conjugate gradient and MD with simulated annealing, scoring the loops using a combination of the CHARM22 force field and statistical preferences of the dihedral angles and atom contacts [127]. Other free modeling methods like RAPPER [128] and PLOP [129] build loop fragments by sampling from a dihedral angle library for each residue, beginning from one or both anchors and eventually attempting to close the loop while avoiding steric clashes.

The third challenge is associated with the scoring of various conformations. Because the number of residues whose conformation vary between the different structural candidates is small, accurate energy functions can be crucial for the purpose of guiding the conformational search and scoring the final structures. Both statistical potentials [130] and physics-based force fields [128] have been used as scoring functions in loop modeling. Some methods use statistical potentials only for filtering, while the final ranking employs all atom force fields [131]. Other methods focus on all atom energy functions designed specifically for loop modeling [132] [133]. However, energy functions that are good at guiding the conformational search during the loop building stage might be inadequate for the final ranking of the decoys,
especially in methods where the loop building is performed incrementally and separately from closure.

Until recently, efforts in the study of local protein structure have largely centered on predicting loops in defined crystal structures. However, InsEnds predictions are made in the context of template-based models where the structures for the remainder of the protein may be imperfect, being constructed from one or more crystal structures and by relying on a sequence alignment. As a result of this imperfection, the structure prediction algorithm must be lenient, thereby fundamentally distinguishing this problem from traditional loop modeling. Although both the treatment of loops and InsEnds involve local protein modeling, they can present different sets of challenges. While loops in crystal structures are defined as regions connecting different secondary structure elements, InsEnds are defined as regions devoid of information extracted from sequence alignments. Hence, InsEnd may include regions with complete secondary structure elements. In addition, the length of loops is governed by the structural context, and, consequently, usually contain a limited number of residues. InsEnds, on the other hand, can be of arbitrary lengths. Furthermore, the boundaries of loops are generally well defined whereas the boundaries of InsEnds are determined by the gaps in the alignments. When multiple templates are combined to generate one model, the gap regions may appear with different boundaries in different templates, thereby rendering the InsEnds boundaries ambiguous.

Our method is designed to address these issues. We demonstrate the robustness of our methods by successfully predicting the structures of long loop regions in crystal structures
Figure 4.1: The InsEnds modeling problem. A multiple sequence alignment of a target sequence to template sequences can contain insertion regions at the same location.
as well as providing blind structural predictions of InsEnds in the top homology models from CASP9. We present a fragment free method for local structure prediction.

4.2 Approach

Our approach assumes that the principles that govern the folding of proteins are equally applicable for modeling InsEnds. We have shown that single backbone $\phi, \psi$ pivot moves provide an effective way to sample conformations, provided the moves are contingent upon the identities and conformations of the nearest neighbors (NNs). These moves have been used successfully in the fragment-free de novo prediction of the structures of single domain proteins [32] [53].

Our local structure prediction method generates random local conformations using the same single pivot $\phi, \psi$ move set as for our global structure prediction scheme (Figure 4.2). The interaction energy is calculated both within the local regions and between the local region and the rest of the protein. The total energy is used to guide the conformational search, an approach that differs from many methods in which the loop fragment is constructed one residue at a time while simultaneously trying to satisfy the loop closure constraint at the end. In contrast to some existing methods that separate loop building and closure into two subsequent stages, our approach integrates the two into a single simulated annealing Monte Carlo (MCSA) scheme, thus retaining the tertiary context of the entire protein during the simulation while attempting to rapidly find the best local conformation. This tertiary context can be critical for identifying crucial loop-protein interactions, thus greatly reducing the search
space. The algorithm is designed to handle multiple loops in the same MCSA trajectory. Hence, when two loops are close enough to interact, they are modeled simultaneously.

The conformational search proceeds through MCSA scheme (described in detail in the Methods section) that is guided by a combination of the pairwise additive, orientation dependent statistical potential DOPE-PW, along with a harmonic ligation energy term to close the loop (Figure 4.3). The relative weight of the ligation energy increases during the MCSA to enforce loop closure. Explicit side chains are absent during the sampling stage of the simulation since the DOPE-PW statistical potential and backbone torsional move set implicitly incorporates sufficient information concerning the side groups [32]. Final conformations are scored using a combination of structural clustering and accessible surface area of the hydrophobic residues to select the best solutions. The standard deviation in the positions of a given loop residue in a cluster (i.e., the tightness of the cluster) provides a metric for assessing the local quality of the predictions for the loop.

### 4.3 Methods

All backbone heavy atoms are explicitly treated, whereas the side chains are represented by single Cβ atoms [32] [53]. The backbone bond lengths and angles are fixed at their ideal values, and only backbone torsional angles \( \phi, \psi \) are sampled during the simulation. Loop closure is achieved by ligating the free ends of the loops to the beginning of the subsequent chain with a harmonic constraint whose strength increases as \( 1/\text{Temperature} \) during the MCSA procedure (Figure 4.3).
Figure 4.2: The InsEnds local protein structure prediction algorithm
4.3.1 Ramachandran Map (Pivot) Move Set and Sampling

The study uses our approach for sampling single residue $\phi, \psi$ backbone torsional angles [53]. A distribution of $\phi, \psi$ angles is generated from a high resolution library of PDB structures for each amino acid (aa), conditional on the identity of the flanking amino acids. These NN dependent torsional angle distributions are pre-calculated for all 20 aas, resulting in 8000 total Rama Maps that are divided into 50x50 bins. During each Monte Carlo step, a selected residue’s $\phi, \psi$ angles are changed. Besides the identity of the NN, the Rama Maps can also be restricted according to secondary structure of the aa and its NNs. The data presented in the paper, however, are obtained without the imposition of this restriction, thereby enabling the exploration of all regions of torsional space allowed for a given amino acid based on its neighbor’s identity. The only exception to this is the CASP8 target T0464, where 5 of the 24 residues were restricted to helical angles as the PSIPRED program [134] predicted them to be helical with high confidence.
4.3.2 Energy functions

The conformational search is guided through the simulation by an energy function that is a combination of the pairwise, orientation-dependent statistical potential DOPE-PW 20 and a harmonic ligation term for the closure of the loop:

\[
E = E_{DOPE-PW} + \frac{T_k}{T}[(D - D_0)^2 + (L - L_0)^2]
\]  

(4.1)

where \( T \) is the simulation temperature, \( D/D_0 \) are the current/initial distances between the two anchor points, and \( L/L_0 \) are the distances between the free end and the anchor point at the site of the cut. The ligation term becomes stronger as the simulated annealing temperature decreases. The initial temperature of the simulations is set to 100, and \( T_k \) is chosen such that the contributions from the DOPE-PW and ligation energies become comparable by the end of the simulation.

The interactions in DOPE-PW are parameterized based on the observed distance distributions in the PDB, contingent on neighbors, amino acid identities, secondary structures, and side chain orientations. DOPE-PW has been demonstrated to perform well in guiding the conformational search during prediction of the structure of small proteins. The DOPE-PW term initially dominates the total energy and provides greater freedom for the conformational search, thereby aiding in properly orienting the loop with respect to the rest of the structure.
4.3.3 Scoring

Once the set of final conformations are generated from the MCSA simulations, the best candidate among this set of conformations is chosen using a combination of quantities computed from clustering, DOPE-PW energies, and solvent accessibility. Clustering based on the Cα RMSD provides a very effective means to identify dominant conformations. Hierarchical clustering proceeds with a distance cutoff of 5 Å using the minimum distance method with the Cluster module in Biopython [60]. Trials with distance cutoffs of 4 Å and 6 Å do not significantly alter the results. Clustering is used only when the largest cluster contains at least 5% of the total structures. The clusters are ranked as detailed in the Results section, while the best individual structures are selected according to the sum of the DOPE-PW energy and the solvent accessible surface area (SASA).

Loop regions reside mostly on the protein surface, and thus solvent interactions can be crucial determinants of loop structures. Hence, most successful loop scoring schemes include some approximate measure for the extent of solvation as part of the scoring function [128] [129] [131]. While the DOPE-PW energy function accurately describes the preferred orientations of the side chains of both hydrophilic and hydrophobic residues as being directed away and toward solvent, respectively, the interactions are still assumed to be pairwise additive between Cα-Cβ bond vectors and thus do not explicitly treat the solvent accessibility. Since explicit side chains are absent during the sampling stage, the program SCWRL 4.0 [96] is used to add side chains to enable calculating the SASA using a rapid approximation with a water radius of 1.4 Å [135]. The SASAs of each residue are assigned into hydrophobic and
hydrophilic components, and the structure that minimizes the hydrophobic ASA and maximizes the hydrophilic ASA is presumed to have the best ASA score. For this purpose, the structures are ranked using both the hydrophobic and hydrophilic ASAs, and the combined rank is taken as the net ASA score.

4.3.4 MCSA Simulation Procedure

The initial torsional angles of the InsEnds are randomly chosen so that no prior information is retained regarding its conformation, while the rest of the protein structure is kept fixed. 700-1000 independent MCSA trajectories are run using the energy functions described above. Each step of the MCSA trajectory involves selection of a random amino acid in the InsEnds whose torsional angle is modified according to the pre-generated NN dependent Rama Map for that amino acid. This results in a new InsEnds conformation whose energy is evaluated, and the conformation is either accepted or rejected based on the Metropolis criteria at that temperature using the energy functions described above. The temperature is updated every 500 Monte Carlo steps, using a polynomial time cooling schedule [57]. The simulation protocol has been implemented in a C library, called the Protein Library, and the input/output is handled using the PDB tools from the Biopython package.

4.3.5 Parallel Scripting

The InsEnds algorithm has been implemented for high throughput structure prediction using the parallel scripting language, Swift [136]. Swift enabled the algorithm to be expressed in a
high-level logical manner independent of any specific computing resources. Swift automatically parallelized the independent invocations of the lower level protein structure manipulation programs, which were written in Python and C. It further provided the flexibility of running on multiple, different, parallel architectures by automating job scheduling and error handling, and logged the provenance of all data objects produced.

4.4 Results and Discussion

Three different modeling scenarios are considered. First, we address the traditional loop modeling problem in crystal structures where the structure surrounding the loop is known. We next address InsEnds modeling as applied in the CASP9 blind prediction competition, where the InsEnds may be as large as 45 aa regions in template-based models generated by Xu’s RAPTOR-X algorithm [137]. The third scenario is for the CASP9 refinement category in which the InsEnds algorithm is applied to the best structure from the server predictions and where the starting model and boundaries are specified by the organizers.

4.4.1 Loops in crystal structures

In order to demonstrate the applicability to larger loops in crystal structures, 26 loops of lengths 8 to 12 have been randomly selected from standard loop benchmarking studies [129]. Loop boundaries for each target are taken as previously specified, and the loops are modeled using our method. Figure 4.5 illustrates the process of selecting the top 5 predictions, and Tables 4.1 and 4.2 presents the best and the remaining four top predictions. After the
predictions are clustered according to the RMSD between the loop structures, the largest five clusters are ranked using a linear combination of the Z-scores for the cluster tightness (RMSD between structures in the cluster), size, and average DOPE-PW energy, defined as $Z_t$, $Z_s$, and $Z_E$, respectively,

\[
\text{ClusterScore} = Z_s - Z_d - Z_t, \quad Z_i = \frac{X_i - <X_i>}{\sigma_i},
\]

(4.2)

where the Z-score for the property of structure $X_i$ is $Z_i = (X_i - <X_i>) / \sigma_i$, and $<X_i>$ and $\sigma_i$ are the mean and standard deviation, respectively. After ranking the clusters, one representative from each cluster is selected using a combination of the DOPE-PW energy and the SASA to obtain the top 5 predictions. Although DOPE-PW is very successful in guiding the protein backbone into a proper conformation based on the orientation of the $\text{C} \alpha - \text{C} \beta$ vectors, it is unable to resolve the details of solvation at an atomistic level because it is parameterized only at the $\text{C} \beta$ level. Hence, explicit SASA calculations are necessary to properly account for the solvation energy.

As discussed in the Methods section, the SASA scores are determined from a combined ranking of the hydrophilic and hydrophobic ASAs. Similarly, the structures are ranked using the DOPE-PW energy function as well. The structure with the lowest total DOPE-PW + SASA rank in a given cluster is taken as the predicted structure from that cluster. Models are discarded when the distance between the free end and the anchor point fails to return to within 1.5 Å of the initial distance. If the largest cluster contains less than 5% of the structures, the scoring for the top 5 candidates uses only the sum of DOPE-PW and SASA
scores As shown in Figure 4.4, the inclusion of SASA to the DOPE-PW energy improves the selection of the top structure in most cases compared to simply using DOPE-PW energy to select the top structure.

A residue-specific deviation quantifies the local confidence score of the prediction for each residue individually in each of the top 5 predictions. The local confidence scores are illustrated by color and thickness in Figure 4.5. The thicker (redder) portions in the predicted local region correspond to residues displaying the largest deviation within the cluster.

The simulations for loops of length 12 and 8-11 residues generate conformations with global loop RMSDs of 2.76 Å and 1.93 Å respectively, where the RMSD is calculated for the loop residues after aligning the structures without the loop regions. These results can be compared to Table II of Lee et al. [122] which presents the minimum backbone RMSDs found using different existing loop sampling protocols. Simulations for 12 and 8 residue loops in crystal structures with the cyclic coordinate descent (CCD) protocol [123] generates minimum RMSDs of 3.05 Å and 1.59 Å the CJSD [122] method obtains 2.34 Å and 1.01 Å the self organizing algorithm using an alternating scheme of pairwise distance adjustments (SOS) [24] yields 2.29 Å and 1.19 Å and the FALC [122] scheme finds 1.84 Å and 0.78 Å respectively.

The average RMSDs of our top ranked predicted loops are 3.98 Å and 3.13 Å for loops with 12 and 8-11 amino acids respectively. These results are comparable to those from two different methods, RAPPER [128] and FALC [122], ranked by DFIRE [130] as listed in Table IV in Lee et al. [122], where the average RMSDs of the top ranked 12 residue loop decoys
Figure 4.4: Inclusion of SASA to pairwise DOPE-PW statistical potential improves decoy selection
Figure 4.5: Selection of the top 5 loop predictions for 1xnb. After clustering, the largest 5 clusters are ranked based on Z-scores with respect to cluster tightness, size, and average DOPE-PW energy. Once ranked, a selection is made from each of the 5 clusters using DOPE-PW + SASA.
Table 4.1: Prediction of 8-12 residue loops in crystal structures - Local RMSD (Align Loop, RMSD of Loop)

<table>
<thead>
<tr>
<th>Target</th>
<th>Loop Length</th>
<th>Best</th>
<th>Pred1</th>
<th>Pred2</th>
<th>Pred3</th>
<th>Pred4</th>
<th>Pred5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1rcf</td>
<td>12</td>
<td>1.81</td>
<td><strong>2.67</strong></td>
<td>3.39</td>
<td>3.27</td>
<td>3.38</td>
<td>2.79</td>
</tr>
<tr>
<td>1thw</td>
<td>12</td>
<td>1.84</td>
<td>4.61</td>
<td><strong>2.4</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2cpl</td>
<td>12</td>
<td>1.09</td>
<td><strong>1.09</strong></td>
<td>4.14</td>
<td>4.02</td>
<td>4.72</td>
<td>4.37</td>
</tr>
<tr>
<td>1cyo</td>
<td>12</td>
<td>0.75</td>
<td>1.97</td>
<td>3.46</td>
<td><strong>1.67</strong></td>
<td>3.94</td>
<td>5.11</td>
</tr>
<tr>
<td>1hfe*</td>
<td>12</td>
<td>1.91</td>
<td><strong>2.08</strong></td>
<td>2.42</td>
<td>3.38</td>
<td>3.06</td>
<td>3.39</td>
</tr>
<tr>
<td>1onc</td>
<td>12</td>
<td>2.19</td>
<td>3.24</td>
<td><strong>2.19</strong></td>
<td>2.95</td>
<td>2.94</td>
<td>2.99</td>
</tr>
<tr>
<td>1pmy</td>
<td>12</td>
<td>0.57</td>
<td>2.47</td>
<td><strong>1.79</strong></td>
<td>2.28</td>
<td>2.64</td>
<td>3.79</td>
</tr>
<tr>
<td>1rro</td>
<td>12</td>
<td>1.84</td>
<td><strong>2.65</strong></td>
<td>2.98</td>
<td>3.79</td>
<td>2.83</td>
<td>2.82</td>
</tr>
<tr>
<td>1scs</td>
<td>12</td>
<td>2.27</td>
<td><strong>3.48</strong></td>
<td>5.2</td>
<td>4.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1bkf</td>
<td>12</td>
<td>1.55</td>
<td>2.26</td>
<td>2.55</td>
<td><strong>2.9</strong></td>
<td>2.69</td>
<td>2.35</td>
</tr>
<tr>
<td>2tgi</td>
<td>12</td>
<td>2.44</td>
<td><strong>2.85</strong></td>
<td>3.35</td>
<td>3.35</td>
<td>3.48</td>
<td>3.22</td>
</tr>
<tr>
<td>1eco</td>
<td>12</td>
<td>0.52</td>
<td><strong>3.43</strong></td>
<td>3.85</td>
<td>5.09</td>
<td>4.16</td>
<td>4.12</td>
</tr>
<tr>
<td>1msc</td>
<td>12</td>
<td>1.96</td>
<td>3.64</td>
<td>5.08</td>
<td><strong>3.57</strong></td>
<td>3.96</td>
<td>5.17</td>
</tr>
<tr>
<td>1acf</td>
<td>11</td>
<td>1.8</td>
<td><strong>2.05</strong></td>
<td>3.59</td>
<td>3.52</td>
<td>3.88</td>
<td>2.22</td>
</tr>
<tr>
<td>1eid</td>
<td>11</td>
<td>1.16</td>
<td><strong>1.17</strong></td>
<td>2.19</td>
<td>1.65</td>
<td>3.34</td>
<td>3.32</td>
</tr>
<tr>
<td>1noa</td>
<td>11</td>
<td>1.68</td>
<td>4.14</td>
<td>4.35</td>
<td>4.7</td>
<td><strong>3.43</strong></td>
<td>4.35</td>
</tr>
<tr>
<td>1plc*</td>
<td>11</td>
<td>1.8</td>
<td>2.05</td>
<td>2</td>
<td><strong>1.8</strong></td>
<td>4.59</td>
<td>1.84</td>
</tr>
<tr>
<td>1xnb</td>
<td>11</td>
<td>0.91</td>
<td><strong>1.62</strong></td>
<td>2.67</td>
<td>3.3</td>
<td>2.72</td>
<td>3.17</td>
</tr>
<tr>
<td>4ilb</td>
<td>11</td>
<td>1.17</td>
<td><strong>1.13</strong></td>
<td>3.72</td>
<td>3.64</td>
<td>3.11</td>
<td></td>
</tr>
<tr>
<td>8dfr*</td>
<td>11</td>
<td>1.71</td>
<td><strong>1.71</strong></td>
<td>2.04</td>
<td>3.2</td>
<td>3.85</td>
<td>2.96</td>
</tr>
<tr>
<td>1aaj</td>
<td>11</td>
<td>1.71</td>
<td><strong>2.5</strong></td>
<td>2.69</td>
<td>2.81</td>
<td>3.14</td>
<td>2.62</td>
</tr>
<tr>
<td>5p21</td>
<td>10</td>
<td>1.74</td>
<td>2.78</td>
<td>2.99</td>
<td>2.47</td>
<td><strong>2.67</strong></td>
<td><strong>2.38</strong></td>
</tr>
<tr>
<td>5fx2</td>
<td>10</td>
<td>1.93</td>
<td>2.79</td>
<td>3.42</td>
<td><strong>2.3</strong></td>
<td>3.77</td>
<td>3.92</td>
</tr>
<tr>
<td>1cbs</td>
<td>8</td>
<td>0.74</td>
<td>3.65</td>
<td><strong>0.74</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1xnb</td>
<td>8</td>
<td>0.28</td>
<td>1.49</td>
<td>1.21</td>
<td><strong>0.53</strong></td>
<td>1.16</td>
<td>1.13</td>
</tr>
<tr>
<td>1poa</td>
<td>8</td>
<td>0.44</td>
<td><strong>0.54</strong></td>
<td>0.97</td>
<td>1.01</td>
<td>3.94</td>
<td>2.71</td>
</tr>
</tbody>
</table>

* refers to cases where the top cluster contained less than 5% of the total structures. In those cases, the top 5 predictions are selected using DOPEPW+SASA instead; units are in ÅRMSD values in bold indicate the best out of the top 5 predictions.
Table 4.2: Prediction of 8-12 residue loops in crystal structures - Global RMSD (Align rest, RMSD of loop)

<table>
<thead>
<tr>
<th>Target</th>
<th>Loop Length</th>
<th>Best</th>
<th>Pred1</th>
<th>Pred2</th>
<th>Pred3</th>
<th>Pred4</th>
<th>Pred5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1rcf</td>
<td>12</td>
<td>2.61</td>
<td>3.79</td>
<td>5.29</td>
<td>4.63</td>
<td>4.64</td>
<td>4.53</td>
</tr>
<tr>
<td>1thw</td>
<td>12</td>
<td>4.53</td>
<td>5.94</td>
<td>5.87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2cpl</td>
<td>12</td>
<td>2.43</td>
<td>2.43</td>
<td>6.34</td>
<td>6.94</td>
<td>6.84</td>
<td>8.37</td>
</tr>
<tr>
<td>1cyo</td>
<td>12</td>
<td>1.23</td>
<td>2.47</td>
<td>4.95</td>
<td>4.09</td>
<td>5.15</td>
<td>7.03</td>
</tr>
<tr>
<td>1hfe*</td>
<td>12</td>
<td>3.69</td>
<td>3.69</td>
<td>4.58</td>
<td>5.6</td>
<td>4.29</td>
<td>7.01</td>
</tr>
<tr>
<td>1onc</td>
<td>12</td>
<td>2.91</td>
<td>3.83</td>
<td>5.26</td>
<td>5.36</td>
<td>5.66</td>
<td>5.22</td>
</tr>
<tr>
<td>1pmk</td>
<td>12</td>
<td>1.24</td>
<td>2.97</td>
<td>3.13</td>
<td>2.98</td>
<td>4.07</td>
<td>7.02</td>
</tr>
<tr>
<td>1rro</td>
<td>12</td>
<td>4.63</td>
<td>6.94</td>
<td>7.78</td>
<td>8.44</td>
<td>6.52</td>
<td>7.34</td>
</tr>
<tr>
<td>1scs</td>
<td>12</td>
<td>3.1</td>
<td>4.22</td>
<td>7.87</td>
<td>6.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1bkf</td>
<td>12</td>
<td>2.48</td>
<td>3.37</td>
<td>4.82</td>
<td>6.82</td>
<td>4.85</td>
<td>3.98</td>
</tr>
<tr>
<td>2tgi</td>
<td>12</td>
<td>3.64</td>
<td>4.19</td>
<td>5.32</td>
<td>5.6</td>
<td>4.66</td>
<td>4.33</td>
</tr>
<tr>
<td>1eco</td>
<td>12</td>
<td>0.85</td>
<td>2.65</td>
<td>3.01</td>
<td>1.99</td>
<td>3.45</td>
<td>3.09</td>
</tr>
<tr>
<td>1msc</td>
<td>12</td>
<td>2.57</td>
<td>5.5</td>
<td>7.3</td>
<td>11.9</td>
<td>12.6</td>
<td>7.9</td>
</tr>
<tr>
<td>1acf</td>
<td>11</td>
<td>2.34</td>
<td>2.86</td>
<td>4.21</td>
<td>4.09</td>
<td>4.41</td>
<td>3.19</td>
</tr>
<tr>
<td>1cid</td>
<td>11</td>
<td>1.39</td>
<td>1.83</td>
<td>3.16</td>
<td>2.17</td>
<td>5.47</td>
<td>4.83</td>
</tr>
<tr>
<td>1noa</td>
<td>11</td>
<td>2.59</td>
<td>7.22</td>
<td>6.7</td>
<td>8.3</td>
<td>6.51</td>
<td>7.7</td>
</tr>
<tr>
<td>1plc*</td>
<td>11</td>
<td>3.42</td>
<td>3.42</td>
<td>4.89</td>
<td>4.08</td>
<td>8.35</td>
<td>3.57</td>
</tr>
<tr>
<td>1plc*</td>
<td>11</td>
<td>3.42</td>
<td>3.42</td>
<td>4.89</td>
<td>4.08</td>
<td>8.35</td>
<td>3.57</td>
</tr>
<tr>
<td>1plc</td>
<td>11</td>
<td>3.42</td>
<td>3.42</td>
<td>4.89</td>
<td>4.08</td>
<td>8.35</td>
<td>3.57</td>
</tr>
<tr>
<td>1xnb</td>
<td>11</td>
<td>1.39</td>
<td>2.41</td>
<td>3.42</td>
<td>4.45</td>
<td>3.67</td>
<td>4.6</td>
</tr>
<tr>
<td>411b</td>
<td>11</td>
<td>2.03</td>
<td>2.03</td>
<td>9.71</td>
<td>10.3</td>
<td>10.83</td>
<td></td>
</tr>
<tr>
<td>8dfr*</td>
<td>11</td>
<td>2.89</td>
<td>2.89</td>
<td>4.79</td>
<td>6.19</td>
<td>6.37</td>
<td>6.39</td>
</tr>
<tr>
<td>1ajj</td>
<td>11</td>
<td>2.79</td>
<td>3.59</td>
<td>4.66</td>
<td>5.9</td>
<td>7.08</td>
<td>6.03</td>
</tr>
<tr>
<td>5p21</td>
<td>10</td>
<td>2.62</td>
<td>4.27</td>
<td>3.87</td>
<td>3.24</td>
<td>3.94</td>
<td>3.71</td>
</tr>
<tr>
<td>5fx2</td>
<td>10</td>
<td>2.29</td>
<td>3.23</td>
<td>3.55</td>
<td>2.6</td>
<td>5.32</td>
<td>4.7</td>
</tr>
<tr>
<td>1cbs</td>
<td>8</td>
<td>1.74</td>
<td>5.41</td>
<td>1.74</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1xnb</td>
<td>8</td>
<td>0.77</td>
<td>3.53</td>
<td>1.83</td>
<td>0.85</td>
<td>1.86</td>
<td>1.94</td>
</tr>
<tr>
<td>1poe</td>
<td>8</td>
<td>0.76</td>
<td>1.23</td>
<td>2.1</td>
<td>1.59</td>
<td>5.56</td>
<td>4.84</td>
</tr>
</tbody>
</table>
for RAPPER and FALC are 4.32 Å and 3.84 Å respectively. Rossi et al. [138] compares four different commercial loop modeling packages - Prime (Schrdinger, LLC), MODELLER (Accelrys Software, Inc.), ICM (Molsoft, LLC) and Sybyl (Tripos, Inc.) which obtain RMSD values ranging from 3 to 5 Å for loops with 10-12 amino acids. Our performance is comparable to these methods.

We also compared our results to a recent atomic level loop modeling study which had sub-angstrom level accuracy [139]. Although our Cβ level modeling certainly limits us in terms of obtaining sub-angstrom models, we still are able to obtain better or comparable models for some of the same benchmark proteins compared to the high resolution KIC protocol (Supplementary Table 1 in Ref. [139]). For instance, our top predicted model for 1hfc of 3.69 Å outperforms KIC’s 8.2 Å prediction for the same loop. Similarly, for other targets 4i1b, 1msc, 1cyo and 1pmy from our benchmark set in Table 1a, our top predictions of 2.03 Å 5.5 Å 2.47 Å and 2.97 Å are better or comparable to the high resolution Kinematic Closure method’s top predictions of 3.8 Å 3.2 Å 5.2 Å and 2.6 Å respectively for the same 10-12 amino acid loops. The results demonstrate that a Cβ level representation of the protein chain without a costly analytical closure constraint is sufficient to achieve accuracy comparable to existing methods for relatively long loops in the context of crystal structures.

### 4.4.2 Ends in crystal structures.

Another challenge involves modeling the termini of protein structures, which has seen only limited study [140] [141]. Unlike loops, end regions require no loop closure. To demonstrate
that our method is also applicable to end regions, we refolded the termini for six proteins (Tables 4.3 4.4). In each of the case, twenty residues in the C terminal end of the native proteins were first randomized while the rest of the structure was kept fixed. Starting from these pseudo-random structures, the end residues were sampled and clustered using the loop modeling protocol. Because no loop closure is required, the termini were folded using only the DOPE-PW energy function. In 3/6 cases (1af7,1o2f, 1r69), the best and the predicted structures had a global RMSD of under 3 Å . In all of the cases, the best local RMSD was under 5 Å . Although direct comparisons were not available for the same proteins, there results were comparable to another method [141] for refolding of terminal secondary structures where the average RMSDs of 4.6 Å and 2.0 Å were obtained for 10-23 residue ends after three minimizations using DFIRE and dDFIRE energy functions. We selected the last 20 residues in all the proteins for modeling irrespective of where the secondary structure boundaries lie. This protocol better mimics the situation encountered in authentic template based modeling where the number of unknown residues that need modeling are determined by the gaps in sequence alignment and often no reliable information is available about secondary structure type or boundaries.

Table 4.3: Prediction of 20 end residues in crystal structures - Local RMSD (Align ends, RMSD of ends)

<table>
<thead>
<tr>
<th>Target</th>
<th>Type</th>
<th>Best</th>
<th>Pred1</th>
<th>Pred2</th>
<th>Pred3</th>
<th>Pred4</th>
<th>Pred5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1af7</td>
<td>α</td>
<td>1.33</td>
<td>2.36</td>
<td><strong>2.21</strong></td>
<td>3.74</td>
<td>3.8</td>
<td>3.07</td>
</tr>
<tr>
<td>1o2f</td>
<td>αβ</td>
<td>1.76</td>
<td><strong>2.01</strong></td>
<td>3.26</td>
<td>3.06</td>
<td>3.07</td>
<td>4.72</td>
</tr>
<tr>
<td>1mkj</td>
<td>αβ</td>
<td>3.47</td>
<td>3.82</td>
<td>4.6</td>
<td><strong>3.71</strong></td>
<td>4.3</td>
<td>4.15</td>
</tr>
<tr>
<td>1b7y</td>
<td>α</td>
<td>4.06</td>
<td>5.08</td>
<td>4.95</td>
<td><strong>4.69</strong></td>
<td>5.19</td>
<td>5.11</td>
</tr>
<tr>
<td>1r69</td>
<td>α</td>
<td>2.17</td>
<td><strong>2.5</strong></td>
<td>3.28</td>
<td>7.46</td>
<td>8.04</td>
<td>8.08</td>
</tr>
<tr>
<td>1tif</td>
<td>αβ</td>
<td>4.85</td>
<td><strong>5.5</strong></td>
<td>8.75</td>
<td>9.38</td>
<td>9.19</td>
<td>9.68</td>
</tr>
</tbody>
</table>
Table 4.4: Prediction of 20 end residues in crystal structures - Global RMSD (Align rest, RMSD of ends)

<table>
<thead>
<tr>
<th>Target</th>
<th>Type</th>
<th>Best</th>
<th>Pred1</th>
<th>Pred2</th>
<th>Pred3</th>
<th>Pred4</th>
<th>Pred5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1af7</td>
<td>α</td>
<td>2.24</td>
<td>2.24</td>
<td>3.8</td>
<td>4.36</td>
<td>5.58</td>
<td>5.42</td>
</tr>
<tr>
<td>1o2f</td>
<td>αβ</td>
<td>2.37</td>
<td>2.8</td>
<td>5.94</td>
<td>4.04</td>
<td>4.29</td>
<td>11.9</td>
</tr>
<tr>
<td>1mky</td>
<td>αβ</td>
<td>4.11</td>
<td>5.68</td>
<td>5.58</td>
<td>7.78</td>
<td>9.66</td>
<td>8.72</td>
</tr>
<tr>
<td>1b72</td>
<td>α</td>
<td>5.2</td>
<td>5.8</td>
<td>6.71</td>
<td>5.6</td>
<td>6.75</td>
<td>6.15</td>
</tr>
<tr>
<td>1r69</td>
<td>α</td>
<td>2.72</td>
<td>2.99</td>
<td>5.04</td>
<td>8.35</td>
<td>9.6</td>
<td>8.34</td>
</tr>
<tr>
<td>1tif</td>
<td>αβ</td>
<td>6.31</td>
<td>7.02</td>
<td>10.6</td>
<td>12.1</td>
<td>11.5</td>
<td>11.6</td>
</tr>
</tbody>
</table>

4.4.3 CASP9 blind InsEnds predictions.

Methods designed for predicting the structure of internal loops may be inappropriate for termini of proteins because the energy functions and sampling generally used for loop modeling assume both ends are fixed. Furthermore, InsEnds can encompass whole secondary structure elements. The existing loop modeling methods have been benchmarked for loops in crystal structures where the remaining structure and loop boundaries are known. The situation for homology modeling, however, is more complex, being highly dependent on the quality of the sequence alignments, template identification, and boundary determination. Consequently, the starting point for InsEnds modeling is imperfect and inexact.

The biannual CASP experiments present a unique platform for testing new and benchmarking developed methods through blind predictions. Our participation in CASP9 as MidwayFolding (groups TS435 and TS477) focused on testing our local structure prediction method and on improving poorly predicted local regions in template-based models. Our analysis begins with models generated by the program RAPTOR-X, which utilizes homol-
ogy to identify template structures appropriate to a target sequence through sophisticated sequence/structure alignments. The templates are processed by MODELER to generate our starting model. We also use the sequence alignments of RAPTOR-X to identify the InsEnds regions in the models. Five entries may be submitted to CASP9 for each target, and Figure 4.6 displays the best of the 5 blind models submitted to CASP9 for each target.

The CASP9 targets serve as examples to illustrate several strengths of our method. Several of the insertion regions contain secondary structure elements in the targets. The target T0464 from CASP8 presents a case where the insertion region is a helix, which our method predicts correctly, improving the model’s RMSD from 9.6 Å to 4.5 Å as exhibited in Figure 4.6A. Another target, T0623 has a 25 residue gap in a region that is, in fact, a hairpin that is correctly predicted by our method as well (8.2 Å RMSD improved to 6.3 Å). The largest InsEnds contains 45 residues (T0585), and the RAPTOR-X+MODELER programs describe them as a large loop. Our method correctly identifies that the missing region corresponds to three helices that pack into the protein core, thereby improving the model substantially from 15.1 Å to 9.1 Å overall RMSD as depicted in Figure 4.6H. The target TR606 presents an example where the local modeling is performed for both termini simultaneously to form a pair of beta strands, thereby improving the overall RMSD from 4.9 Å to 3.8 Å for the target as a result of modeling the ends (Figure 4.6G).

Other CASP targets contain InsEnds that are loops connecting different secondary structures. For instance, the targets T0520, T0594 and T0612 yield initial models with loops containing as many as 17 residues (identified from the gap boundaries in the sequence align-
Figure 4.6: **CASP9 InsEnds blind predictions.** Numbers indicate improvement from MODELER (Red) to our model (blue), as compared to the native structure (green) after modeling the regions enclosed by the boxes. RMSD changes are for the whole structure.
ments). Use of our InsEnds protocol for these three loop regions improves the overall RMSDs from 3.2 to 2.6 Å, 2.2 to 1.7 Å and 7.3 to 6.6 Å for T0520, 594 and 612, respectively (Figure 4.6C-D). The demonstration that we successfully model various types of InsEnds with the same protocol without any prior knowledge of whether they are loops or contain secondary structure elements highlights the robustness of the method.

### 4.4.4 Blind Prediction of refinement targets in CASP9

In the refinement category in the CASP experiment, the judges select the best of all submitted (template-based) models from all participating groups. The local regions that deviate most from the native structure are identified to the predictors as the refinement targets. From our perspective, the refinement category is distinct because the starting model is guaranteed to be the best of the all CASP server models rather than one of RAPTOR-X’s model and because the boundaries for InsEnds are specified based on where the server model differs from the native structure (as identified by the organizers) rather than from RAPTOR-X’s sequence alignment.

On average for the 12 refinement targets, the 24 different refinement methods in CASP8 yielded no net improvement over the starting models [142]. Table 4.5 lists the RMSD as well as the Global Distance Test (GDT) changes from the starting models along with the ranking of our method with respect to all the other refinement methods. Our method proceeds by first initializing the InsEnds regions to a completely random conformation so that no structural information about the InsEnds is retained from the starting model.
Table 4.5: Blind InsEnds prediction of refinement targets in CASP9

<table>
<thead>
<tr>
<th>CASP9 refinement target</th>
<th>GDT starting</th>
<th>RMSD starting</th>
<th>GDT Midway-Folding</th>
<th>RMSD Midway-Folding</th>
<th>Rank of Midway-Folding</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR569</td>
<td>73.1</td>
<td>3.01</td>
<td>76.58</td>
<td>2.249</td>
<td>4/121</td>
</tr>
<tr>
<td>TR568</td>
<td>53.35</td>
<td>6.963</td>
<td>56.7</td>
<td>5.108</td>
<td>6/127</td>
</tr>
<tr>
<td>TR517</td>
<td>71.38</td>
<td>4.646</td>
<td>72.17</td>
<td>4.638</td>
<td>11/119</td>
</tr>
<tr>
<td>TR622</td>
<td>67.42</td>
<td>7.47</td>
<td>69.47</td>
<td>5.773</td>
<td>12/120</td>
</tr>
<tr>
<td>TR606</td>
<td>71.95</td>
<td>4.85</td>
<td>72.56</td>
<td>3.915</td>
<td>19/128</td>
</tr>
<tr>
<td>TR594</td>
<td>87.32</td>
<td>1.805</td>
<td>86.07</td>
<td>1.957</td>
<td>23/134</td>
</tr>
<tr>
<td>TR567</td>
<td>78.34</td>
<td>3.435</td>
<td>78.52</td>
<td>3.46</td>
<td>28/107</td>
</tr>
<tr>
<td>TR557</td>
<td>67.6</td>
<td>4.074</td>
<td>68.2</td>
<td>3.74</td>
<td>28/118</td>
</tr>
<tr>
<td>TR614</td>
<td>75.21</td>
<td>4.1</td>
<td>67.36</td>
<td>4.895</td>
<td>43/118</td>
</tr>
<tr>
<td>TR624</td>
<td>54.71</td>
<td>5.529</td>
<td>52.9</td>
<td>5.577</td>
<td>47/122</td>
</tr>
<tr>
<td>TR592</td>
<td>91.43</td>
<td>1.204</td>
<td>82.38</td>
<td>3.415</td>
<td>111/131</td>
</tr>
<tr>
<td>TR576</td>
<td>48.91</td>
<td>10.926</td>
<td>Ignored since initial GDT l.t. 50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The numbers reported are the GDT and RMSD values from the CASP9 website. The values in bold indicate targets with an improvement in the GDT score from the starting model.

Unlike the RMSD which relies on a single alignment, the GDT scores reflect the structural similarities at different distance cutoffs and therefore are generally better at assessing improvements in local regions. We have attempted 11 targets for refinement in CASP9 (Figure 4.7) and improve the GDT scores for 7 of them. Among all groups participating in CASP9 refinement, 4 out of our 11 predictions (targets TR517, TR568, TR569 and TR517) fall in the top 10% of all submissions, and 8 out of the 11 reside in the top 25% of all submissions, thereby outperforming several of the all atom refinement methods. The improvements are achieved for targets with a wide range of starting GDTs. The GDT/RMSD for TR569 improves from 73.1/3.01 Å to 76.58/2.24 Å and our method ranks 4th out of the 121 total submissions for this target. The starting values for TR568 are lower at 53.35/6.39 Å and we improve them to 56.7/5.1 Å with an overall ranking of 6th out of 127 submissions for
the target. Our method performs much worse than the rest of the methods for one target, TR592, presumably because the starting structure is already extremely good (91.2/1.2 Å) that our Cβ level representation is inadequate and an all atom side chain representation is required to improve the model further. Moreover, we have not refined the side chains in any of the cases, something that probably would have improved the results even more.

Figure 4.7 displays all of our predictions for the refinement targets in CASP9. The figure illustrates how well the model aligns to the native structure before refinement (initial) and after refinement (after) when superposed using the LGA program [143]. In several cases, the improvements introduced into the local region also help to align the remainder of the protein. For example, in TR614, even though the actual regions modeled are an insertion from 33-50 and the C terminal residues 106-121, the local alignment of the N terminal residues improved over the starting model as indicated with blue in the LGA alignment for TR614 in Figure 4.7.

4.4.5 Molecular Replacement Results for CASP9 refinement targets

One of the CASP9 refinement metrics assesses how well the predicted models reproduce the experimental data [144]. Recently, models generated by the structure prediction methods have been inserted into the molecular replacement likelihood algorithms for X-ray crystallographic refinement to solve the phase problem [145] [146]. The assessors for CASP9 refinement judged the quality of each submitted model in this regard by calculating the Z-score of the best orientation of the model in the unit cell of the crystal compared to placing it in a set
Figure 4.7: **InsEnds predictions for CASP9 refinement targets.** Difference between CA/CA distance across the sequence of the initial (starting) /native and final (refined using InsEnds method)/native after superposition using sequence-dependent LGA protocol. Official data from CASP9 official website (http://predictioncenter.org/casp9/). For each target, the arrows indicate the regions where the InsEnds modeling has been performed. The blue to green color change designates regions where the InsEnds modeling improves upon the given target based on LGA superposition to native structures.
of random orientations. Only models with Z-scores above 6 were considered good enough to solve the phase problem. Table 3 in Ref. [144] summarizes how often various groups improve the Z-score of the targets from likely unrefinable (lt 7) to likely refinable (gt 7). Our method performs as well or better than all the other groups in this test, with positive results in 2 out of 3 cases attempted. Since our approach employs a backbone + Cβ model with the side chains either missing or added simply using SCWRL4.0 with no further refinement, some of our submitted models were discarded in the analysis by assessors. Regardless, the fact that our method ranks at the top in the molecular replacement test proves its real value in X-ray crystal structure refinement.

In contrast to most other methods that expend considerable computing resources on including all-atom interactions, our method lacks explicit side chain atoms. This difference highlights the distinction between the refinement of crystal structures and template-based models. The all atom refinement of crystal structures benefits from having high resolution information for the rest of the structure whereas homology models are usually far from perfect. It is unclear whether the expensive modeling of all the atoms in an imperfect environment is a computationally efficient strategy. In contrast, the first step of our approach is designed to obtain the proper backbone structure and orientation for the local region using a coarse level of modeling that is less sensitive to the atomic level details for the rest of the homology model. Once the coarse level model is obtained for the local region, side chains may be added, and more detailed all-atom refinement can proceed.
4.4.6 Global InsEnds RMSD vs. local InsEnds RMSD

RMSDs are calculated in three ways to quantify the quality of the modeling of local InsEnds regions,

- InsEnds RMSD: Align the loop and calculate the RMSD of only the InsEnds region.
- Global InsEnds RMSD: Align all the residues besides the InsEnds, and then calculate the RMSD of the InsEnds region.
- Global structure RMSD: Optimally align all the residues in the protein and calculate the RMSD of the full chain.

The local InsEnds RMSD is a measure of how well the InsEnds region itself is modeled, and the global InsEnds RMSD provides a measure of how well the modeled InsEnds is oriented with respect to the rest of the protein. The global InsEnds RMSD is the ideal measure of loop quality when predicting loops in crystal structures because the only difference between the native structure and the model can appear in the loop region. In contrast, InsEnds modeling of homology models begins from inexact structures, and, therefore, assessing the refinements requires accounting for the RMSD of the rest of the structure (besides the InsEnds) with respect to the native structure. If the starting homology model deviates significantly from the native structure, the alignment of the non-InsEnds region necessarily must skew the anchor regions, and therefore the global InsEnds RMSD would not provide as a good a metric for reporting the accuracy of InsEnds modeling than either the local InsEnds RMSD or the overall RMSD of the structure.
This utility of the different RMSDs is illustrated for six targets from CASP8 for which the initial RAPTOR models have variable RMSDs to the native structures. The 11-12 residue InsEnds regions in those models are chosen for (post-dictum) prediction using our method (Tables 4.6, 4.7 and 4.8). Not surprisingly, the global InsEnds RMSD is highly dependent on the quality of the initial model (i.e., the RMSD of all but the InsEnds region in the initial model). For target T0478D1, the RMSD of the non-InsEnds region in the starting model is 8.07 Å the best local InsEnds RMSD decreases from 2.9 to 1.58 Å whereas the best global InsEnds RMSD decreases from 12.2 to 8.4 Å

Table 4.6: Prediction of InsEnds in CASP8 structures (post-diction) - Local InsEnds RMSD (align InsEnds, RMSD of InsEnds)

<table>
<thead>
<tr>
<th>Target</th>
<th>InsEnds Length</th>
<th>RMSD of non-InsEnds region</th>
<th>Initial RAPTOR RMSD</th>
<th>InsEnds Best RMSD</th>
<th>InsEnds Predicted RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>T0431D1</td>
<td>11</td>
<td>5.21</td>
<td>0.22</td>
<td>0.52</td>
<td>0.91</td>
</tr>
<tr>
<td>T0456D2</td>
<td>12</td>
<td>2.09</td>
<td>0.49</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>T0478D1</td>
<td>12</td>
<td>8.07</td>
<td>2.9</td>
<td>1.58</td>
<td>3.05</td>
</tr>
<tr>
<td>T0443D1</td>
<td>12</td>
<td>3.42</td>
<td>4.04</td>
<td>3.7</td>
<td>3.48</td>
</tr>
<tr>
<td>T0411D1</td>
<td>11</td>
<td>2.74</td>
<td>3.53</td>
<td>1.85</td>
<td>3.53</td>
</tr>
<tr>
<td>T0479D1</td>
<td>11</td>
<td>1.54</td>
<td>0.75</td>
<td>0.48</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Target T0411D1 has the non-InsEnds RMSD of the starting model much closer to native structure at 2.74 Å and our local InsEnds RMSD improves from 3.53 to 1.85 Å similar to the local InsEnds RMSD improvement in T0478D1(2.9 Å to 1.58 Å. However, the global InsEnds RMSD for this target improves from 10.2 Å to 2.78 Å which is much more remarkable than the global InsEnds RMSD in T0478D1 (12.2 Å to 8.4 Å. The difference can be attributed T0411D1’s starting model having the non-InsEnds region much closer to the native structure
Table 4.7: Prediction of InsEnds in CASP8 structures (post-diction) - Global InsEnds RMSD (align rest, RMSD of InsEnds)

<table>
<thead>
<tr>
<th>Target</th>
<th>InsEnds Length</th>
<th>RMSD of non-InsEnds region</th>
<th>Initial RAP-TOR RMSD</th>
<th>InsEnds Best RMSD</th>
<th>InsEnds Predicted RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>T0431D1</td>
<td>11</td>
<td>5.21</td>
<td>5.34</td>
<td>4.14</td>
<td>4.7</td>
</tr>
<tr>
<td>T0456D2</td>
<td>12</td>
<td>2.09</td>
<td>1.44</td>
<td>1.22</td>
<td>2.03</td>
</tr>
<tr>
<td>T0478D1</td>
<td>12</td>
<td>8.07</td>
<td>12.2</td>
<td>8.4</td>
<td>11.8</td>
</tr>
<tr>
<td>T0443D1</td>
<td>12</td>
<td>3.42</td>
<td>11.13</td>
<td>7.3</td>
<td>7.67</td>
</tr>
<tr>
<td>T0411D1</td>
<td>11</td>
<td>2.74</td>
<td>10.2</td>
<td>2.78</td>
<td>4.98</td>
</tr>
<tr>
<td>T0479D1</td>
<td>11</td>
<td>1.54</td>
<td>2.3</td>
<td>1.59</td>
<td>1.69</td>
</tr>
</tbody>
</table>

Table 4.8: Prediction of InsEnds in CASP8 structures (post-diction) - Global RMSD (align all, RMSD of all)

<table>
<thead>
<tr>
<th>Target</th>
<th>InsEnds Length</th>
<th>RMSD of non-InsEnds region</th>
<th>Initial RAP-TOR RMSD</th>
<th>InsEnds Best RMSD</th>
<th>InsEnds Predicted RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>T0431D1</td>
<td>11</td>
<td>5.21</td>
<td>4.76</td>
<td>4.04</td>
<td>4.81</td>
</tr>
<tr>
<td>T0456D2</td>
<td>12</td>
<td>2.09</td>
<td>2.04</td>
<td>2.12</td>
<td>2.16</td>
</tr>
<tr>
<td>T0478D1</td>
<td>12</td>
<td>8.07</td>
<td>8.26</td>
<td>8.21</td>
<td>8.31</td>
</tr>
<tr>
<td>T0443D1</td>
<td>12</td>
<td>3.42</td>
<td>5.33</td>
<td>4.42</td>
<td>4.62</td>
</tr>
<tr>
<td>T0411D1</td>
<td>11</td>
<td>2.74</td>
<td>3.87</td>
<td>2.85</td>
<td>3.11</td>
</tr>
<tr>
<td>T0479D1</td>
<td>11</td>
<td>1.54</td>
<td>1.62</td>
<td>1.59</td>
<td>1.59</td>
</tr>
</tbody>
</table>
as compared to T0478D1. Figure 4.8 illustrates this behavior and indicates that the local InsEnds RMSD remains relatively unaffected, whereas the global InsEnds RMSD for the same targets are quite severely affected by the RMSD of the remaining region. The successes of the modeling also support our previous contention from protein structure predictions that the neighbor dependent $\phi,\psi$ capture local interactions reasonably well [32].

![Graph showing Global versus local RMSD](image)

**Figure 4.8: Global versus local RMSD.** The RMSD of non InsEnds region is plotted against the global InsEnds RMSD (red) and local InsEnds RMSD (blue) for six CASP8 targets. The global InsEnds RMSD is affected severely by the quality of the homology model.
4.4.7 Applications to protein folding simulations

Although loop modeling is often called the mini-folding problem, traditional approaches to loop modeling do not consider the folding mechanism when predicting loops. Our method on the other hand, views local modeling in a fashion that fits naturally into the larger problem of protein folding.

Experimental studies indicate that proteins fold through sequential stabilization of tertiary structure elements or foldons [55] [147] [148] [36]. Often, long range contacts form early in the folding pathway and produce intermediate species where some entrained local regions are not yet folded. Hence, a computational scheme designed to predict structure by mimicking the natural stepwise fashion of folding pathways should encounter the problem of folding inside of loops.

Our InsEnds algorithm is well suited to address this problem because the undetermined local regions in the structure that arise during the folding pathway can correspond either to distinct secondary structures, loops, or to combinations thereof. As a proof of principle, we test our method by predicting native structures of possible intermediates in the pathways for folding two proteins, ubiquitin and barnase. The late folding intermediate in ubiquitin lacks the $3_{10}$ helix and the $\beta5$ strand while the rest of the structure is well formed [55] [44] (Figure 4.9B). Starting from a native-like structure for the intermediate, the InsEnds algorithm is used to fold the 18 residues insertion. The InsEnds refinement procedure successfully recovers the native structure to a global RMSD of 1.6Å (Figure 4.9C). This illustrates an example where the local region is neither a loop nor a continuous secondary structure. Nevertheless,
we still obtain the right topology, essentially completing the last step of the folding pathway to predict to the native structure.

Barnase is a 108 residue protein that is atypical for a small protein because it contains 3 distinct hydrophobic cores. The two hairpin loops depicted in Figure 4.9D are crucial to the structure because they are involved in formation of the protein’s cores, and, therefore, the correct prediction of the loops is essential for the prediction of the global structure. Experiments indicate that loop 2 is the last structure to form in the folding pathway 36. Our InsEnds method is applied to fold both the 10 and 15 residue loops in barnase (Figure 4.9E,F). Our best predictions in both cases lie in the top clusters, and the best global RMSDs are 2.03 and 1.27 Å for loops 1 and 2, respectively.

The problem of folding inside of loops highlights two aspects of our method. The first is that our approach treats local structure prediction similarly to global structure prediction by mimicking the natural protein folding mechanism. The second aspect is the demonstration that given the correct boundaries, our method is able to reconstruct the local structures irrespective of whether the local regions are well defined secondary structures or loops.

4.4.8 Simultaneous Folding of multiple InsEnds

One crucial feature of our approach is the ability to simultaneously modeling multiple local regions. When the regions are interacting, simultaneous modeling can be essential because the context provided by one local region may be important in guiding the other into place. A good example is the CASP target TR606, where the InsEnds correspond to the two termini
Figure 4.9: InsEnds algorithm applied to protein folding pathways. A) The $\beta_5$ and $3_{10}$ helix in ubiquitin are that last structures to form in the pathway. Their structures are depicted as disordered in B) the model of the folding intermediate, and C) predicted using the InsEnds algorithm. D) Barnase native structure highlighting the two hairpin loops that are part of two different cores, and E) and F) predictions of the loops using InsEnds algorithm, respectively.
that form a hydrogen-bonded pair of $\beta$ strands. The initial template model fails to identify the ends as strands, and, therefore, the ends were wrongly placed. Accurate modeling requires that they be folded simultaneously. Guided only by the orientationally dependent DOPE-PW energy function, we have modeled the free termini and correctly predicted the pair of strands in our top submission (Figure 4.6G).

### 4.4.9 Protein Structure prediction pipeline

Here our goal is to combine the respective strengths of free modeling with template-based modeling for an integrated structure prediction pipeline. This goal is realized through an automated server, created for CASP9 that integrates the InsEnds, RAPTOR-X and ItFix methods. Given a sequence, the pipeline begins by performing homology modeling using RAPTOR-X. If no templates are identified, the pipeline directs the sequence for free modeling using our existing ItFix algorithm for secondary and tertiary structure prediction. If RAPTOR is able to build a template-based model, the InsEnds are modeled to obtain a final structure. The pipeline has been used for the CASP9 structure predictions of the MidwayFolding group (CASP9 group numbers 435, 477).

### 4.5 Conclusions

Loop modeling has been an on-going challenge in protein structure prediction. With the recent surge in template-based modeling, InsEnds modeling is a relatively new topic in need of novel approaches. Previous methods have focused on loops in the context of crystal
structures and may not be generalizable to imprecise template-based models. InsEnds pose a more complicated situation where the poorly predicted local regions must be modeled without assumptions concerning the accuracy of the rest of the structure or the boundaries and secondary structure of the local regions being modeled. This work presents a novel free modeling method for local protein structure prediction that is applicable for modeling large local regions in both exact and inexact environments, as demonstrated by results both for loops in crystal structures and for InsEnds in template-based models. We consider this result as a step towards the generalization of the local protein structure problem. The work also presents a framework in which free and template-based modeling are integrated towards closing the final gaps in protein structure prediction.
CHAPTER 5

SUPPORT VECTOR MACHINES FOR ENERGY FUNCTION OPTIMIZATION

One of the primary challenges in protein structure prediction is the proper representation of the various energetic contributions that capture the underlying protein physics. Specifically, in the problem of protein structure prediction, the energy function to score the protein conformations is typically represented as a sum of various energetic terms. Balancing the weights of these terms can be crucial to the accuracy of the predictions. In this chapter, I will briefly present some work I’ve done regarding the use of support vector machines, a supervised learning technique, specifically in the context of optimization of the energy functions developed in our group. Rather than reduce the problem into a purely data-based approach to modeling, the objective of this chapter is to highlight the fact that use of machine learning can help extract the maximum information even after all the physics has been described in the form of the various energy terms. The work described here is ongoing and has not been published previously.

Support vector machine(SVMs), originally introduced by Vapnik and coworkers as a training algorithm for classification of linearly separable data [149], have recently become popular in biological applications[150]. Even when the data isn’t linearly separable, SVMs can map the training data into a higher dimensional space using kernel functions and then find a maximal margin hyperplane in that space. In recent years, SVMs have also been
extended for regression where they can be trained to output real values rather than simply classify the data.

By creating a training set of decoys generated using OOPS simulations, I use SVMs to optimize the contributions various energy components in the statistical potential to produce better correlation with native RMSD values. Here I show some preliminary results which indicate that SVMs can be used to better select low RMSD models from a set of decoys. This will be directly beneficial for the selection of low energy structures in the end of each TerItFix round, thereby improving the information that is passed from one round to the next.

5.1 Support Vector Machines: Theory

In its simplest case, the problem of learning using SVMs can be demonstrated using a typical classification problem. Given a set of training examples \((x_i, y_i)\), where \(i=1,...,n\) (where \(n\) is the total number of training examples) and \(x_i \in \mathbb{R}^m\) (\(x_i\) is a \(m\) dimensional vector) and \(y_i \in (-1,+1)\), the task will be to classify whether new cases from a given test sample of input vectors \(x_j\) belong to one class or the other i.e. \(y_j = -1\) or \(+1\). For naming purposes, \(x_i\) are called co-variates, features or input vectors and \(y_i\) are called response variables or labels. If all the data is linear separable, one can draw a straight line \(f(x) = w^T x + b\), with adjustable weights \(w\) and bias \(b\) to separate the data into two sides, so that for all \(f(x_i) < 0\), \(y_i = -1\) and for all \(f(x_i) > 0\), \(y_i = +1\). However, it is easy to see that several such lines (hyperplanes in higher dimensions) can exist that linearly separate the data into two classes.
Support vector machines (SVMs) tackle this problem by introducing a specific criteria to choose a unique separating hyperplane. Intuitively, a reasonable choice of a hyperplane needs to be robust enough to correctly classify new examples that fall within slight perturbations of the current training set. One way to achieve that would be to choose the separating hyperplane in such a way that it lies as far away as possible from training examples of both classes, thereby allowing maximum room (margin) on either side between itself and the closest examples (vectors) from the two classes. The closest vectors are thus called support vectors and SVMs are also called maximum margin classifiers. This notion of maximizing the margin is central to SVMs and helps in the process of generalization during the learning process.

Mathematically, for the earlier case of vectors and labels, the linear classifier can be expressed as [151]:

$$y(x) = \text{sign}[w^T x + b]$$ \hspace{1cm} (5.1)

When the two classes are linearly separable:

$$w^T x_i + b \geq +1 \; \text{if} \; y_i = +1$$
$$w^T x_i + b \leq -1 \; \text{if} \; y_i = -1$$ \hspace{1cm} (5.2)

which can be combined into a single inequality as

$$y_i \left[ w^T x_i + b \right] \geq 1, \; i = 1, ..., n$$ \hspace{1cm} (5.3)
To obtain a unique hyperplane, Vapnik originally rescaled the problem so that the points closest to the hyperplane satisfy the equation $|\mathbf{w}^T \mathbf{x} + b| = 1$, which gives rise to a margin of $2/||w||$ [151]. Combining the margin with the inequality set in 5.3, one obtains primal form of the SVM optimization problem:

$$\text{Minimize } \frac{||\mathbf{w}||^2}{2}$$

subject to $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$ for $i = 1, 2, ... n$ (5.4)

Within the context of convex optimization theory, the general method to solve the above problem involves formulating the Lagrangian and use the conditions of optimality in the dual space of Lagrange multipliers. The problem can be written using a Lagrangian:

$$L(\mathbf{w}, b, \alpha) = \frac{||\mathbf{w}||^2}{2} - \sum_{i=1}^{n} \alpha_i[y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1]$$ (5.5)

The solution that minimizes the primal problem involves maximizing $\alpha_i$ and minimizing $\mathbf{w}$ and $b$, which gives rise to a saddle point problem, solved by taking derivatives wrt $\mathbf{w}$ and $b$:

$$\frac{d}{db} L(\mathbf{w}, b, \alpha) = 0 \text{ and } \frac{d}{d\mathbf{w}} L(\mathbf{w}, b, \alpha) = 0$$

$$\implies \sum_{i=1}^{n} \alpha_i y_i = 0 \text{ and } \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$ (5.6)

Inserting this back to the Lagrangian (5.5), we get the so called dual problem to calculate the maximized $\alpha_i$. 

146
\[ W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \] \hspace{1cm} (5.8)

where \( \alpha_i \geq 0 \) and \( \sum_{i=1}^{n} \alpha_i y_i = 0 \) \hspace{1cm} (5.9)

Finally, the weight vectors can be calculated using the \( \alpha_i \) as

\[ w = \sum_{i=1}^{n} \alpha_i y_i x_i \] \hspace{1cm} (5.10)

Figure 5.1: Support vector machines maximize the margin to separate the data.

Besides margin maximization, a second crucial feature of SVMs is what is known as kernelization. When the given dataset is not linearly separable, SVMs project the low dimensional training data into a higher dimensional feature space, using specialized functions called kernels. Once the data is projected into the higher dimensional space, SVMs then use the same idea of margin maximization to find a hyperplane to linearly separate the data in that space. The crucial property that allows for such a projection is a simple observation in
the dual formulation called the *kernel trick*.

In Eq 5.8, the problem depends on the input training vectors \( \mathbf{x}_i \) only through the inner product of \( \mathbf{x}_i^T \mathbf{x}_j \). If one were to obtain functions which act through the substitution of the inner products, ie \( \mathbf{x}_i^T \mathbf{x}_j \) to \( K(\mathbf{x}_i, \mathbf{x}_j) \), one can do away with having to calculate the high dimensional inner products and simply use the kernel functions to obtain those values in the original space.

Given a mapping function \( \theta \)

\[
\theta : \mathbb{R}^{orig} \rightarrow \mathbb{R}^{final} \\
\mathbf{x} \rightarrow \theta(\mathbf{x})
\]

which maps the input space of the feature vectors \( \mathbf{x} \), \( \mathbb{R}^{orig} \) into the high-dimensional feature space \( \mathbb{R}^{final} \). Mapping the classification problem into the higher dimensional space \( \mathbb{R}^{final} \) can make it easier to find a hyperplane to separate the data. However, if the dual problem has to be solved in \( \mathbb{R}^{final} \), one will have to calculate the inner products in that high dimensional space, which can be a challenge when \( \mathbb{R}^{final} \) is very large.

One can replace the inner products in a higher dimensional space \( \mathbb{R}^{final} \) by a kernel function \( K \), which is defined in \( \mathbb{R}^{orig} \) using

\[
K(\mathbf{x}_i, \mathbf{x}_j) = \theta(\mathbf{x}_i)^T \theta(\mathbf{x}_j)
\]
where $\theta(x_i)^T \theta(x_j)$ is the inner product in the higher dimensional space, $\mathbb{R}^{final}$. If one is able to find such a function $K$, one can calculate the higher dimensional inner product without even knowing the mapping function $\theta(x)$.

Using what are known as Mercer conditions, one can ensure a validity of a kernel function - which is that $K$ must be symmetric and positive definite.

Thus, using a kernel function the dual (5.8) can be rewritten as:

$$W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1, j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{5.11}$$

where the kernel function is evaluated in the original space of training data, thereby avoiding inner product calculation in the higher dimensional space, yet harnessing the power the higher dimensionality for linear separation.

There are several widely accepted kernel functions that satisfy Mercer’s conditions, and their applicability to particular problems is an area of active research.

1. linear: $K(x_i, x_j) = x_i^T x_j$

2. polynomial of degree $d$: $K(x_i, x_j) = (x_i^T x_j + 1)^d$

3. Gaussian: $K(x_i, x_j) = \exp(-||x_i - x_j||^2) / 2\sigma^2$

4. sigmoid: $K(x_i, x_j) = \tanh(x_i^T x_j + r)$

In recent years, SVMs have been modified for the problem of function estimation or regression as well (SVR). While the basic formulation and central ideas remain the same as
Figure 5.2: Support vector machines map the input data into higher dimensional space to classify linearly inseparable data.
in classification, SVRs allow for estimation of continuous real valued labels instead of the simply binary labels (-1,+1) predicted in classification [152]. As opposed to classification where the margin is measured, SVRs measure the error of approximation. In SVRs, the loss function that quantifies the difference between estimated and real values in classical regression, is reformulated with Vapnik’s \( \epsilon \) insensitivity loss function, where the loss is equal to zero if the difference between the predicted \( f(\mathbf{x}_i, \mathbf{w}) \) and real \( y_i \) is less than \( \epsilon \).

Figures 5.1 and 5.2 visually demonstrates the properties of margin maximization and kernelization discussed above.

### 5.2 Energy function optimization as a SVR problem

I now discuss how the problem of energy function optimization can be formulated in the framework of a support vector regression problem (Figure 5.3). The basic idea is that if the various terms in the energy function is assumed to be additive (each with unique weights), one can use support vector machines to optimize the contributions of the different terms so that it can better rank decoy protein structures that are closer to their native states. Specifically, for a given protein structure decoy, one can first calculate the values of the energy terms individually and associate it with the distance (e.g. RMSD) of the particular decoy to the native structure. Next, a training set is created from a large number of decoy structures so that a list of individual energy terms and the associated RMSD values is available. In this framework, then, the various energy terms serve as the input vectors/features and the RMSD values serve as the labels. One can now use SVM to learn the association between
the different energy terms and the RMSD values and create a model that effectively adjusts
the weights of the different energy terms so that they maximally correlate with the RMSD
values, using the principles described in the earlier section. The fact that the labels are real
values (RMSDs) makes it a support vector regression problem. The SVM model can now
effectively serve as a predictor. For a entirely new test set of decoys not used in the training
set, one can first calculate the energy terms individually. When the SVM model is provided
with a set of values for the energy terms for any given decoy in the test set, it will output a
predicted RMSD value for that decoy based on what it learned from the training set. Hence,
if the SVM was able to create a robust model from the training set, the predicted RMSD
values should correlate well with the actual RMSD values for the decoys in the test set.
There is some precedence of using support vector regression for ranking decoy structures,
where it was shown to improve selection of good decoys [153].

One of the main considerations when using learning methods is the struggle between the
two extremes of overgeneralization and overfitting. An overfitted learning model is a poor
predictor because it is parameterized too specifically to a given training set and therefore
unable to generalize well to new examples. Alternatively, an overgeneralized learning model is
a poor predictor because it is not sensitive to features that are specific to certain examples.
To balance these effects, typically a process of cross-validation is carried out where the
training set is repeatedly and randomly divided up into training and testing sets during
training to optimize the learning process.
Figure 5.3: Support vector regression framework for energy function optimization. Given a list of features (energy terms) and corresponding RMSD values, the first step involves learning from a training set to create a SVM model. The model is then used for prediction of RMSD values given a testing set that only contains the values of the various energy terms (features).
5.3 Results

Next, I present some preliminary results comparing the performance before and after the use of SVM for decoy ranking. For current purposes, I have chosen to optimize the five non-continuous attractive energy terms in our DOPE-PW statistical potential. I chose these terms since they deal with long range interactions and contribute the most to the total DOPE-PW energy. However, it is understood that a more rigorous and updated analysis should potentially include our multibody statistical and desolvation potentials as well.

The five non-continuous DOPE-PW energy terms (input features) are described in detail elsewhere [32]. Briefly, they include the parallel and anti-parallel $\beta$ strand interactions, as well as the non-$\beta$ strand terms for the low, medium and high PW orientational bins. The labels, which are the RMSD values are real valued and range anywhere between 2 Å to 10 Å. In all the results described here, I used the simplest kernel, the linear kernel to train the SVMs. In terms of software, I adapted a publicly available standard SVM library, SVM-Light[154] to perform regression on the DOPE-PW dataset.

5.3.1 Decoy generation

The decoys were generated using folding simulations with the DOPE-PW energy function, using an earlier version of TerItFix [32]. To increase the population of low RMSD decoys where decoy identification becomes more significant, additional decoys were generated by performing refinement of low RMSD decoys using our local doublecrank moves [45]. Around 4000-5000 decoys were generated for each protein, of which 1500 were used for training each
protein and the rest were saved for the testing set. There were a total of 11 proteins used in this study (Table 5.1). It is worth noting that while decoy selection is a standard technique for energy function benchmarking, most decoy sets are biased according to the method used for their generation. For our purposes, since we are primarily interested in ranking the decoys generated using our OOPS tools and statistical potentials, we generated our own decoy set using the tools for testing the SVMs.

5.3.2 Individual Proteins

First, for each of the 11 proteins, I created a unique training set and testing set only for the particular protein. As a result, separate SVMs were trained for each protein individually. While a SVM model created using one protein can not be expected to predict models from another protein, the goal of this exercise was to check whether SVM can improve on the ranking of test decoys for individual proteins compared to simply ranking them using the sum of the five DOPE-PW energy terms.

Figures 5.4, 5.5 and 5.6 show results comparing the ranking of decoys for individual proteins using DOPE-PW energy (left) and the SVM-predicted RMSDs. These clearly show that the correlation with the true RMSD of the decoys improves drastically when the SVM learning model is used compared to simply adding the different energy terms. This is quantified using the Pearson correlation coefficients for each of the two cases for all the proteins, listed in Table 5.1. One protein, 1di2, was particularly challenging, are reflected by the correlation coefficient of 0.37, which the rest were greater than 0.6. The protein 1di2 is slightly
elongated in shape compared to the other globular proteins in the dataset and doesn’t have a very well-defined core. This could, in fact, highlight the inability of DOPE-PW pairwise interactions to properly describe interactions when the protein doesn’t have a single core.

Table 5.1: Pearson Correlation Coefficients of true RMSD values to DOPE-PW energy function as well as SVM predicted RMSD values

<table>
<thead>
<tr>
<th>Protein</th>
<th>Type</th>
<th>Correlation Coefficient to true RMSD of DOPE-PW Energy</th>
<th>SVR predicted RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1af7</td>
<td>α</td>
<td>0.50</td>
<td>0.66</td>
</tr>
<tr>
<td>1b72</td>
<td>α</td>
<td>0.31</td>
<td>0.70</td>
</tr>
<tr>
<td>1csp</td>
<td>β</td>
<td>-0.11</td>
<td>0.86</td>
</tr>
<tr>
<td>1dcj</td>
<td>αβ</td>
<td>0.30</td>
<td>0.94</td>
</tr>
<tr>
<td>1di2</td>
<td>αβ</td>
<td>0.07</td>
<td>0.37</td>
</tr>
<tr>
<td>1mkv</td>
<td>αβ</td>
<td>0.18</td>
<td>0.88</td>
</tr>
<tr>
<td>1o2f</td>
<td>αβ</td>
<td>-0.02</td>
<td>0.74</td>
</tr>
<tr>
<td>1r69</td>
<td>α</td>
<td>0.01</td>
<td>0.82</td>
</tr>
<tr>
<td>1shf</td>
<td>β</td>
<td>0.30</td>
<td>0.78</td>
</tr>
<tr>
<td>1tif</td>
<td>αβ</td>
<td>0.40</td>
<td>0.79</td>
</tr>
<tr>
<td>1tig</td>
<td>αβ</td>
<td>0.20</td>
<td>0.78</td>
</tr>
</tbody>
</table>

5.3.3 Protein Classes

The more important purpose of using SVM in this study was to be able to optimize the statistical potential to more generally improve the ranking of decoys across multiple proteins. Hence, next I create SVMs for a different classes of proteins rather than individual proteins. By combining the training sets for the three different α proteins (1af7, 1b72, 1r69) into one training set, I create a single SVM_α to build a learning model from it. Next, I combine the testing set for those three proteins and predict all their RMSD values using the same SVM_α model. Figure 5.7 (top) shows the correlation between the SVM predictions to the
Figure 5.4: Correlation of true RMSD to the SVM predicted RMSD and DOPE-PW energy in α proteins.
Figure 5.5: Correlation of true RMSD to the SVM predicted RMSD and DOPE-PW energy in β proteins.
Figure 5.6: Correlation of true RMSD to the SVM predicted RMSD and DOPE-PW energy in α/β proteins.
true RMSD values for those test decoys. The correlation coefficient remains relatively high at 0.79. I performed a similar test by grouping all the $\alpha/\beta$ proteins together, and obtain a slightly lower correlation coefficient of 0.56 (Figure 5.7 (bottom)).

These results for the two protein classes show that the learning using SVM is generalizable at least across the different classes for the proteins studied. A thorough test of robustness requires training and testing the SVM framework across a large number of different proteins, which is left for future work.

5.4 Some Comments

As discussed, the results provide only a glimpse of how SVMs can potentially improve the selection of decoys generated using our statistical potentials. By creating an even larger and diverse dataset of decoys, and adding more input features during the training process, a more general SVM predictor can likely be constructed. Furthermore, the weights for the various features obtained from the generalized SVMs could be used to improve the generation of decoys, especially in our TerItFix folding simulations. Additionally, one can envision including decoys from earlier stages of folding for training as well.

Since the design of statistical potentials inherently are biased according to the features chosen for statistical analysis in the databases, often a lot of terms have redundant information. SVMs provides a way of identifying such terms and implicitly correct for their double counting when creating a model from the training set.

A different study using SVR for scoring function optimization found that the most impor-
Figure 5.7: Correlation of true RMSD to the SVM predicted RMSD different protein classes.
tant input feature was a “mutual information” metric that quantified the mutual similarity across the different decoys, similar to that used during clustering [153]. Future work should also probably include a mutual distance metric for OOPS generated decoys as well.

5.5 Conclusions

One of the recurring problems used for benchmarking statistical potentials in protein structure prediction is the proper ranking of decoys with respect to RMSD to the native structures. Since our statistical potential combines multiple features like sequence distance and secondary structure dependence, it is necessary to optimize the relative weights of those various features in the statistical potential. In this chapter, I showed preliminary results that indicate that use of support vector machines can be productive in the task of ranking decoy structures generated using OOPS simulations. Instead of replacing the physical insights by a purely data mining approach, the purpose of this chapter was to show that machine learning techniques like SVMs can help maximally extract that information from the dataset already generated using the physical principles, thereby supplementing them.
CHAPTER 6

FUTURE DIRECTIONS AND CONCLUSION

The previous chapters described a framework for protein structure prediction guided by the process of protein folding. With proper levels of representation, sampling techniques and energy functions, I showed that it is possible to obtain useful information about protein folding pathways, often consistent with various experimental folding studies. Additionally, in some cases, folding simulations can help explain or interpret results from such studies. Notwithstanding the successful predictions in numerous systems, the work described here still remains to be implemented on larger and more complex systems (e.g., multiple core proteins, complicated beta topologies). Although the material described henceforth in this chapter is under development and unpublished, I present some ideas and preliminary results to extend the my work described in this thesis. First, I will discuss some strategies for tackling larger proteins in the TerItFix framework. Next, I will present ideas for obtaining more quantitative data regarding the folding kinetics from TerItFix predicted foldons. Finally, I will discuss some general open challenges and opportunities in the field of protein folding research.

6.1 Improvements to TerItFix for prediction of larger proteins

While TerItFix represents advancement in terms of unifying prediction of folding pathways and tertiary structure, improvements are necessary for it to be successfully applied to larger systems. Proteins larger than 100 amino acids typically present a challenge for de novo
prediction methods. One of the main challenges in protein structure prediction is to be able to reliably predict structures of proteins larger than 100 amino acids. As the number of residues increases, the search problem becomes more challenging as well.

I present two preliminary studies, which could potentially increase the accuracy of structure prediction of larger proteins using TerItFix.

### 6.1.1 Multiple levels of local structure biases

In Chapter 2, I discussed simulations where the protein’s dihedral angles are initialized according to their occurrence in the statistical coil library. While this makes the simulations more realistic, it comes at the cost of expensive sampling of all regions of the Ramachandran space in the first round. For larger proteins, however, this strategy can pose a significant challenge. One of the ways to alleviate the problem could be to bias the dihedral angle sampling using information about the secondary structure of the protein. Since TerItFix’s movesets allow each amino acid to be completely flexible in terms of dihedral angle twists (or pivots), any secondary structure biases introduced will still only reflect through preferential sampling of \( \phi, \psi \) angles from the basins corresponding to those secondary structures. In other words, secondary structure motifs are not prefixed in space and will still need to be stabilized by hydrogen bond and tertiary interactions to properly form.

The local sampling bias can be introduced at various levels in the current TerItFix framework.

- Instead of initializing torsional angles from a coil library, angles from the whole PDB
can be used. Although this might increase the bias for helical angles, this level of initialization has previously been shown to obtain low RMSD structures for globular proteins less than 100 residues [32, 53].

• Utilizing the information in growing sequence and structure databases, various algorithms use data mining approaches to predict secondary structure from the primary sequence [134, 155]. These methods are generally more reliable than database driven tertiary structure prediction algorithms. Since programs like PSIPRED usually assign amino acid level confidence scores to their predictions, a reasonable starting secondary structure information can be used for TerItFix by fixing these residues to the predicted secondary structures in the first round. This approach has been used for our submissions in the CASP experiments previously to some success.

• The strongest local bias can be introduced by assuming prior knowledge of the secondary structure of the protein, and using it to generate the Rama sampling distribution for the first round. Although the pathways generated using pre-stabilized secondary structures might not be identical to cases where no secondary structure is assumed, it still can provide useful information about tertiary motifs that are most stabilizing. Figure 6.1 shows an example of how tertiary fixing of structure can incrementally improve the prediction of the tertiary even when the knowledge of the native secondary structure is assumed.

It is also worth noting that most existing de novo structure prediction algorithms [156,
157, 158, 159] use some level of local structure bias in their algorithms to aid prediction of the tertiary structure.

![Image](image.png)

**Figure 6.1:** Tertiary fixing can improve structure prediction even when all secondary structure is fixed. In the example shown, the predicted structure of Top7 with RMSD of 3.5 Å is obtained by iterative rounds of tertiary fixing using native secondary structure, where the population of low RMSD structures increases with each round.

### 6.1.2 Clustering Before Averaging

TerItFix is an iterative algorithm, where successive iterations make use of the information of prior rounds. So, one of the biggest challenges is faced when the first round simulations do not generate reliable contacts to iteratively fix in the next rounds. This can especially be relevant for larger proteins with more higher contact order, where a large number long range tertiary contacts are present in the native structure. One obvious suggestion would be to run longer Monte Carlo simulations for larger proteins to ensure thorough sampling of the conformational space. Similarly, running more independent trajectories can help improve the
chances of picking up larger populations with consistent contacts. Nevertheless, situations arise where the final minimized structures do not have a consensus set of contacts in the low energy population so that the when the contact bias is generated for the next round of TerItFix, all pairwise contact energy terms are homogeneous and weak.

Figure 6.2 (top) demonstrates this for the case of folding RNASE H, a 148 amino acid protein. When the contact maps of the lowest 25% energy structures from the first round are averaged, almost no sets of contacts appear to have a significant consensus score. The next round of TerItFix will thus be almost identical to the first round due to the lack of any significant bias energy. One solution, which I outline in Figure 6.2, will be to first cluster the low energy structures so that subpopulations with common structural motifs can be categorized. Typically, RMSD is used as the distance metric for structural clustering. However, for larger proteins, the final structures in the first round might be of poor quality in terms of global RMSD. Hence, a metric that can pick out the similarities on a more local structural level will be more appropriate, like TM-score [160] or even distance matrices. Once clustered, one can obtain average contact maps for each cluster by using structures from each of the top clusters, as seen in Figure 6.2 (bottom). For RNASE-H, clustering helps differentiate structures with different sets of consistent contacts in the original low energy structures. The average contact maps of each cluster is more populated compared to the original overall contact map, thereby generating different sets of possible biasing energies. A strategy to use these biasing energies from the different clusters for the next round could be to run different number of independent Monte Carlo simulations using each of the biasing
Figure 6.2: Clustering the low energy structures before obtaining average contact maps improves bias signals. In this example shown for the first round TerItFix simulation of RNASE-H, without clustering the consensus contact map is very sparse originally (top). However, once the low energy structures are first clustered, more consensus contacts within each cluster are more visible (bottom).
energies, where the number of Monte Carlo runs are proportional to the size of the cluster corresponding to the particular biasing energy. Depending on how different the various clusters are, one could investigate the possibility of multiple pathways using this approach as well.

6.2 Extending TerItFix to obtain folding kinetics

In Chapter 2, I discussed a general strategy to obtain protein folding pathways using the principle of sequential stabilization. However, the information obtained about the folding pathways is in the form of discrete foldons, represented by the recurring motifs in a folded population of each TerItFix round. In other words, TerItFix in its current form provides a coarse grained view of folding pathway by identifying the obligate species that are present during a stepwise assembly of a protein. However, a rigorous framework for constructing the free energy landscape and obtaining protein folding kinetics is still required. I outline below two possible ways this might be achieved.

6.2.1 Constant Temperature Metropolis Monte Carlo

Monte Carlo simulations have a long history in statistical physics as a standard tool to obtain equilibrium information [161]. The idea that Monte Carlo simulations can be useful in obtaining the relaxation kinetics to the equilibrium distribution is not new either. Given a common free energy landscape and sufficient time separation between major folding events, Monte Carlo simulations can mimic traditional dynamics simulations to produce a coarse-
grained kinetic picture[162]. For example, it was demonstrated that the folding pathways observed in Monte Carlo simulations for a helical hairpin were same as in a Brownian dynamics simulation [163]. Even earlier studies have used local move sets in Monte Carlo simulations of lattice polymer chains to study coil-globule relaxation time [164]. The kinetics of formation of various protein structural motifs like hairpin formation have also been studied using Monte Carlo simulations [165]. There is also precedence for obtaining specific folding events like nucleation in fast folding systems by performing and analyzing ensembles of Monte Carlo runs [166, 167].

In the TerItFix simulations described in Chapter 2, while we also use Monte Carlo simulations to obtain the major steps along the folding pathway, two problems preclude any detailed inferences regarding folding kinetics. First, in TerItFix, the system goes through decrease in temperature during each Monte Carlo run following some annealing schedule, which drives the system out of equilibrium. Second, the major moveset used in the TerItFix are pivot moves which produce drastic movements of the protein chain that are unphysical. These issues need to be addressed before using Monte Carlo scheme to obtain kinetics.

In figure 6.3, I propose a Monte Carlo scheme, termed Seamstress, to use constant temperature Metropolis Monte Carlo (MMC) dynamics after the completion of the TerItFix simulations to obtain detailed kinetics. While the method development is in progress, the basic idea behind the method would be as follows: The main information at the end of each round of TerItFix are the average contact maps, that represent the foldons obtained by the process of stepwise assembly. These are shown for the case of ubiquitin in Figure 6.3. One
can view these foldons as obligate “checkpoints” along the dominant folding pathway for a protein. Hence, starting from the unfolded state, one can run Metropolis Monte Carlo dynamics at some constant temperature and a local moveset, while iteratively biasing the simulation through the sequential contact maps (foldons) obtained from the TerItFix simulation. This would be similar in terms of implementation details to Gō-like Monte Carlo methods, where simulations are biased by native contacts. The major difference, however, would be that unlike Gō methods, no prior knowledge of the native structure is needed or assumed. Instead, the biases can be sequentially applied as defined by the buildup of contacts in TerItFix predicted foldons. Given the foldons are cooperatively folding substructures, one can likely obtain some useful information regarding barriers connecting the various foldons using this method. Moreover, since TerItFix already predicts the foldons, thereby breaking down the larger native-state search problem into smaller subproblems of connecting the different foldons, one can expect the method to be computationally tractable and fast.

In Figure 6.3, using ubiquitin as a test case, I perform constant temperature MMC runs to connect its six foldons by sequentially biasing the dynamics to the subsequent foldons. The local moves were implemented in backbone torsional space using the doublecrank moves described in Chapter 2. The MMC dynamics was run at different constant temperatures. The temperature at which the contact energies based on the subsequent foldon was the lowest was chosen to be the optimal temperature (for ubiquitin was 600 (arbitrary units), Figure 6.4). By clustering the multiple MMC trajectories at the optimal temperature, one can potentially obtain a more detailed picture of the folding pathway. For example, while the
Figure 6.3: Constant temperature biased Metropolis Monte Carlo simulation to connect foldons in ubiquitin
N terminal hairpin is the first motif to form in ubiquitin according to TerItFix, the next step involves formation of the helix as well as docking of the C terminal strand to the N terminal hairpin. By clustering the trajectories from the MMC run connecting the two foldons, it was observed that the C terminal strand docks to the hairpin 80% of the time, compared to the 20% of trajectories where the helix docks against the hairpin, thereby hinting that the hairpin-strand motif is more readily formed compared to the hairpin-helix motif.

Figure 6.4: Obtaining the optimal temperature for Metropolis dynamics

6.2.2 Markov State Models

In recent years, Markov state models (MSMs) have gained popularity as a tool to analyze and integrate independent trajectories to obtain a comprehensive map the molecule’s thermodynamics and kinetics [168, 169, 170, 171]. The basic underlying assumption that MSMs make is that on a long time scale, biomolecular dynamics can be approximated as a Markov process. By discretizing a system’s phase space into regions, MSMs proceed by running
independent short trajectories that visit each of those regions, while simultaneously record-
ing the state-to-state transition probabilities in a Markov matrix [172]. To automatically
determine the various states and ensure equilibration within each state, MSMs generally
perform kinetic clustering where conformations that inter-convert rapidly are grouped into
one state and those that inter-convert slowly are grouped into separate states. One of the
benefits of MSMs is that it is amenable to parallel computation since several independent
short dynamics simulations can be run and combined for a single MSM analysis.

A primary challenge in MSMs however is the choice of pre-defined initial states from
which the unbiased trajectories are launched. Often times in folding, some of the interesting
states might not be known beforehand, in which case MSMs could provide a misleading
picture of the major folding pathways, or worse yet provide a folding picture biased by the
initial states chosen. TerItFix, therefore can serve as an ideal method for generating initial
states for MSM analysis. Depending on how many initial states MSMs require, we can use
the TerItFix predicted foldons, or the constant temperature MMC dynamics trajectories
from the previous section to launch unbiased molecular dynamics simulations, which the
MSMs then can integrate to create a complete thermodynamic and kinetic picture.

6.3 Protein folding : Looking forward

One of the topics of primary interest in the field of protein folding has been the interplay
between the amino acid sequence and the native topology in the context of folding. Par-
ticularly, how much of the energy landscape is dictated by topology and what features of
the energy landscape are obligate given a particular topology? It’ll be instructive to look at various proteins that have been studied over the past years to answer these questions.

Generally, a few mutations do not affect the native structure of most proteins. In fact, there are several examples of proteins with low sequence homology but similar topologies (e.g. CheY/Spo0f/NtrC are all structural homologs with pairwise RMSD of less than 2 Å but sequence identity of 25 %. [173]). Then, there are cases where even though the native state is unchanged, the folding pathway of the proteins can be affected by mutations. The immunity proteins studied in Chapter 2 (Im7, Im9, SIm9) provide a good example of this. Even with two mutations, Im9 switches its folding behavior from two state to three state [12]. The degree and manner in which the folding pathway is affected varies across different systems as well. For example, TIM barrels have conserved folding pathways in terms of the number of intermediates, including the presence of an off-pathway intermediate, but the structure of the intermediates varies across different members of the TIM barrel family [174]. In α spectrin domains R15, R16 and R17, all of which are three helix bundles, but have only around 30 % pairwise sequence identity, the folding rates appear to be different [13, 175]. Remarkably, R15 folds 3000 times faster than its homologues. Additionally, the folding rate constants R16 and R17 show significantly lower viscosity dependence compared to R15.

There have been other clever approaches to study the interplay between sequence and topology in proteins. Repeat proteins are composed of linear tandem repeats that are similar in topology but differ in their sequences. Notch Ankyrin is an example where the overall topology lacks long-range interactions and has been studied experimentally. In the 7 repeat
Many proteins fall somewhere in between on a spectrum that ranges from topology dominated to sequence dominated folding behavior.

- **Low seq homology, same fold – but same pathway?**
  - (CheY/Spo0f/NtrC structural homologs)
  - Same fold, different seq, similar folding pathways
    - (TIM barrel)
  - Same fold, high sequence identity, different folding behavior
    - (Im7/Im9, Ankyrin, Spectrin)
  - Same fold, few mutations, different folding behavior
    - (Im9/SIm9)

- **High seq homology, different fold**
  - (G\text{A}_{95}, G\text{B}_{95})

Figure 6.5: Most proteins fall somewhere in between on a spectrum that ranges from topology dominated to sequence dominated folding behavior.
ankyrin domain, it was found that folding is initiated at repeats 3 to 5. However, when the sequence of the last two repeats were changed to a consensus sequence (without affecting the final fold), the folding was found to be initiated from the C terminus of the protein, demonstrating that local stabilities (sequence effects) can switch folding pathways [176, 177]. This switching also provides some support for the notion of specific pathways directly resulting from specific primary sequence. Another approach to investigate the role of topology involves creating circular permutants of proteins, where the connectivity of various secondary structural elements is changed but the identity of amino acids in the sequence is preserved. Often, the circular permutants will have different folding behavior compared to the original proteins [178].

Perhaps, the most dramatic example of sequence effects on a protein topology can be found in the redesigned heteromorphic proteins, G_A and G_B, which have high sequence identity but completely different folds. Both engineered using small protein domains of a large streptococcal protein G, the fragment G_A is a three helix bundle whereas the fragment G_B has a four β strands + helix topology. Remarkably, recent studies have shown that even with 95 % sequence identity, merely different by three amino acids, these two proteins adopt two different folds [179, 180]. While the native topology certainly influences the folding landscape, this example highlights the importance of sequence in modulating what species can actually be populated in the landscape.

Taken together, the specific examples discussed above highlight the fact that folding behavior of most proteins lie somewhere between a topology-dominated or a sequence-
dominated folding spectrum (Figure 6.5). It is therefore reasonable to assume that evolution uses both the native topology and primary sequence as knobs for modulation and selection of proteins. Experiments and simulations to investigate whether evolution tunes for the folding pathways of various proteins can improve our understanding of protein folding in a biological context. Towards that direction, comparative studies of folding behavior of homologous and evolutionarily related proteins can provide a first step [181, 182, 183, 184].

Figure 6.6: Changes in the primary sequence can affect the folding landscape in different ways

While the traditional view of protein folding states that the amino acid sequence codes for the native structure of proteins, a new picture has been emerging with the recent studies of protein folding, including this current work: The amino acid sequence not only codes for the final structure, but in fact codes for the complete folding landscape, which includes other aspects besides the native structure like intermediates, transition states, kinetic traps,
off-pathway species, etc. Changes in the primary amino acid sequence therefore can alter the protein folding landscape in different ways that include changes in barrier heights, alternative routes to the native state and in some cases, create new kinetic traps (Figure 6.6). A complete understanding of the folding problem therefore must be reflected in the ability of new methods to predict how a specific mutation in a given amino acid sequence will affect the protein’s folding landscape. In my opinion, this goal should motivate the development of next generation of theoretical, computational and experimental protein folding studies, and the work presented in this thesis provides a step forward in that direction.

6.4 Conclusions

The primary motivation behind the work described in this thesis was to identify the general principles of protein folding and encode those into an algorithm for simultaneous prediction of folding pathways and tertiary structure starting from the amino acid sequence. Towards that goal, I described the development of TerItFix, which demonstrably predicted the native structures as well as major details of the folding pathway of several proteins, mostly consistent with corresponding experimental folding studies. Furthermore, the information obtained using TerItFix was shown to be comparable to that obtained from all-atom molecular dynamics simulations which need orders of magnitude greater computational resources. Since the TerItFix simulations don’t require or assume the knowledge of the native structure, it stands out from most contemporary computational studies of protein folding that bias the simulations using native-centric interactions. In fact, this deviation from the native-centric
approach allows TerItFix to better explore non-native interactions, intermediates and kinetic traps in folding landscapes, as demonstrated through the predictions for Im7/Im9, protein L/protein G.

I also demonstrated that the principles that apply for folding the whole sequence can also be used for reconstruction of large local regions in protein structures. By framing the loop and InsEnds modeling problem within the context of the folding problem, I was able to improve local structure prediction in crystal structure loops and insertions in homology models, demonstrated most notably through the blind prediction of refinement targets in CASP9 tournament, where our method ranked at the top based on improvements to homology models for successful X-ray structure determination using molecular replacement test.

These results reaffirm the validity of some of the basic principles we’ve identified which we believe capture the important aspects of protein physics as it relates to protein folding - nearest-neighbor dependent backbone torsional angle sampling and moveset, orientation-dependent pairwise and multibody Cβ-level statistical potentials to capture long range tertiary interactions and the process of sequential stabilization for conformational search. In the spirit of the famous paraphrased quote by Albert Einstein [185] - “Make everything as simple as possible, but no simpler” - the current work demonstrates that when properly formulated, reduced representation models can provide useful insights, and more importantly, predictions regarding both the protein folding mechanism as well as the tertiary structure using only the primary sequence as input.
REFERENCES


181


189


