Washington University in St. Louis
Washington University Open Scholarship

All Theses and Dissertations (ETDs)

January 2009

Accurate Docking is Achieved by Decoupling Systematic Sampling from Scoring

Jianwen Feng
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/etd

Recommended Citation

Feng, Jianwen, "Accurate Docking is Achieved by Decoupling Systematic Sampling from Scoring" (2009). *All Theses and Dissertations (ETDs)*. 108.
https://openscholarship.wustl.edu/etd/108

This Dissertation is brought to you for free and open access by Washington University Open Scholarship. It has been accepted for inclusion in All Theses and Dissertations (ETDs) by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Arts and Sciences

Division of Biology and Biomedical Sciences

Computational Biology

Department of Biochemistry and Molecular Biophysics

Thesis Examination Committee:
Garland R. Marshall, Chair
Nathan A. Baker
Peter C. Chivers
Jay W. Ponder
David Sept
Gary D. Stormo

ACCURATE DOCKING IS ACHIEVED BY DECOUPLING SYSTEMATIC SAMPLING FROM SCORING

by

Jianwen A. Feng

A dissertation presented to the Graduate School of Arts and Sciences
of Washington University in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2009
Saint Louis, Missouri

copyright by

Jianwen A. Feng

2009

ABSTRACT OF THE DISSERTATION

Accurate Docking is Achieved by Decoupling Systematic Sampling From Scoring

by

Jianwen A. Feng

Doctor of Philosophy in Computational Biology

Washington University in St. Louis, 2009

Research Advisor: Professor Garland R. Marshall

This dissertation discusses two main projects from my thesis work. The first project focuses on the development of a small molecule docking program, SKATE, for drug discovery. The second project focuses on the critical analysis of the thermal stability of a mini-protein, FSD-1.

SKATE is a novel approach to small molecule docking. It removes any inter-dependence between sampling and scoring to improve docking accuracy. SKATE systematically and exhaustively samples a ligand's conformational, rotational and translational degrees of freedom, as constrained by a receptor pocket, to find sterically allowed poses. A total of 266 ligands were re-docked to their respective receptors to assess SKATE's performance. The results show that SKATE was able to sample poses within 2 Å RMSD of the native structure for 97% of the cases. The best performing scoring function was able to rank a pose that is within 2 Å RMSD of the native structure as the top-scoring pose for 83% of the cases. Compared to published data, SKATE has a higher self-docking accuracy rate than or is at least comparable to GOLD, Glide, MolDock and Surflex. The cross-docking accuracy of SKATE was assessed by docking 83 ligands to their respective receptors. The cross-docking results were comparable to those in published methods.

Mini-proteins that contain fewer than 50 amino acids often serve as model systems for studying protein folding because their small size makes long time-scale simulations possible. However, not all mini-proteins are created equal. The stability and structure of FSD-1, a 28-residue mini-protein that adopts the $\beta\beta\alpha$ zinc-finger motif independent of zinc binding, was investigated using circular dichroism (CD), differential scanning calorimetry (DSC), and replica-exchange molecular dynamics (REMD). FSD-1's broad melting transition, similar to that of a helix-to-coil transition, was observed in CD, DSC, and REMD experiments. The N-terminal β-hairpin was found to be flexible. FSD-1's apparent melting temperature of 41 ^{o}C may be a reflection of the melting of its α-helical segment instead of the entire protein. Thus, FSD-1's status as a model system for studying protein folding should be reconsidered despite its attractiveness for being small in size and it was designed to contain essential helix, sheet, and turn secondary structures.

An electronic copy of this dissertation is available online at www.ccb.wustl.edu/~jafeng

Acknowledgments

This dissertation would not have been possible without the help and guidance from generous faculties, colleagues and friends. Thank you, Garland, for providing a stimulating environment for me to mature as a scientist. Having the independence to indulge in my scientific curiosities certainly enriched my graduate training experience.

To Chris Ho, thank you for answering countless questions from molecular modeling, to drug discovery, to algorithm development. Writing SKATE would have been much more difficult without your help. To Christy Taylor, thank you for numerous advice on designing and performing wet lab experiments. To Dan Kuster and Rob Yang, thank you for helpful discussions on running molecular dynamic simulations and performing docking experiments. To Abby Fisher, thank you for teaching me solid-phase peptide synthesis. Thank you, past and present members of the Marshall lab (Gregory Nikiforovich, Sage Arbor, Eric Welsh, Yaniv Barda, Yat Tang, Greg Bourne, Xiaoming Zhang, and Jon Våbenø and the many other students and colleagues who passed through). Thank you, Drs. Jeff Kao and Alex Kozlov for your help in collecting NMR and DSC data.

I appreciate the advice and guidance of my dissertation committee members, Drs. Peter Chivers, Garland Marshall, Jay Ponder, David Sept, Gary Stormo and especially Nathan Baker, who graciously chaired the committee.

I am grateful for financial support from the National Institutes of Health (NIH) and a fellowship for Bioentrepreneurship from the Kauffman Foundation.

Finally, I want to thank my family and friends for their support. To my mother and father, Caichang and Rudong, thank you for being wonderful parents. To my brothers Huan and Tim, thank you for your encouragement and support. To my wife Lily, thank you for your support and inspiration in this journey; I love you!

Jianwen A. Feng

Washington University in Saint Louis
August 2009

Dedicated to my mentors.

Your kind willingness to invest in a naïve student's growth is the soul of humanity.

I will pay it forward!

Contents

Abstract . . . . **ii**

Acknowledgments . . . . **iv**

List of Tables . . . . **x**

List of Figures . . . . **xi**

1 Introduction **1**
- 1.1 Computer-aided drug discovery 1
- 1.2 Mini-proteins 2

2 Drug discovery **4**
- 2.1 Introduction 4
- 2.2 High throughput screening 5
- 2.3 Computational methods in drug discovery 5
- 2.4 Molecular docking 6
  - 2.4.1 Historical perspective 7
  - 2.4.2 Interdependence of sampling and scoring 8
- 2.5 Summary 8

3 SKATE: A docking program **10**
- 3.1 Introduction 10
- 3.2 Overview of docking methodology 11
  - 3.2.1 Sampling 11
  - 3.2.2 Scoring 14
- 3.3 Methods 15
- 3.4 Data Sets 24
  - 3.4.1 Astex/CDCC diverse set 24
  - 3.4.2 Surflex set 25
  - 3.4.3 Vertex set 25
  - 3.4.4 Thymidine kinase set 26
  - 3.4.5 Cyclin dependent kinase 2 set 26
- 3.5 Docking setup 27
- 3.6 Results and discussion 28
  - 3.6.1 Sampling accuracy 28
  - 3.6.2 Analysis of failed sampling cases 35

3.6.3 Scoring accuracy . . . 36
3.6.4 Examples of scoring errors . . . 43
3.6.5 Comparison with other docking programs . . . 45
3.6.6 Cross docking . . . 49
3.6.7 Pitfalls in complex preparation . . . 53
3.6.8 Computational time . . . 56
3.7 Conclusions . . . 56
3.8 Acknowledgment . . . 57

4 SKATE: Potential Improvements . . . 58
4.1 Introduction . . . 58
4.2 Usability improvements . . . 58
4.3 Performance improvements . . . 59
4.4 Estimating Entropy . . . 61
4.5 Summary . . . 61

5 Stability of FSD-1 . . . 62
5.1 Introduction . . . 62
5.2 Materials and methods . . . 65
5.2.1 Peptide synthesis and purification . . . 65
5.2.2 Circular dichroism . . . 66
5.2.3 Differential scanning calorimetry . . . 68
5.2.4 NMR . . . 68
5.2.5 Replica-exchange molecular dynamics . . . 69
5.2.6 Curve fitting . . . 71
5.3 Results and discussion . . . 72
5.3.1 Circular dichroism . . . 72
5.3.2 Differential scanning calorimetry . . . 75
5.3.3 Molecular dynamics simulations . . . 78
5.3.4 Structural analysis . . . 81
5.3.5 Alternative interpretation . . . 85
5.4 Conclusions . . . 86
5.5 Acknowledgments . . . 86

6 Protein Design . . . 87
6.1 Preorganization . . . 87
6.2 Reverse-turn mimetics . . . 89
6.2.1 BTD . . . 89
6.2.2 D-Pro-Gly . . . 90
6.2.3 Aib-Gly . . . 91
6.2.4 R-Nip-S-Nip . . . 91
6.2.5 Dimethyl-L-Proline . . . 92
6.2.6 D-Pro-Pro . . . 94

6.3 Impacts of reverse-turn mimetics on protein stability 95
6.4 Stabilizing FSD-1 99
    6.4.1 Molecular dynamics 99
    6.4.2 Results 100
6.5 Summary 110

Appendix A Abbreviations and Acronyms 113

References 114

Curriculum Vitae 123

List of Tables

3.1 Self docking results 30
3.2 Cross docking results 51

6.1 Number of hydrophobic core contacts 109

List of Figures

3.1 Stematic search of conformational space. 13
3.2 Illustration of aggregates. 14
3.3 Discriminant search. 17
3.4 Overview of SKATE algorithm 20
3.5 Best pose results 29
3.6 Astex/CDCC set scoring results 39
3.7 Surflex set scoring results 40
3.8 Vertex set scoring results. 41
3.9 1JJE scoring error 43
3.10 1OQ5 scoring error 44
3.11 Competitive performance: Vertex set 45
3.12 Competitive performance: Astex/CDCC set 47
3.13 Competitive performance: Surflex set 48
3.14 CDK2 cross docking 54
3.15 Pitfalls in complex preparation 55

5.1 FSD-1 structure 64
5.2 FSD1 NMR. 67
5.3 CD spectra of FSD-1 73
5.4 FSD-1 thermal unfolding monitored by CD 74
5.5 FSD-1 thermal unfolding monitored by DSC 76
5.6 Potential-energy overlap between neighboring REMD replicas 78
5.7 Sampling of temperature space and conformational space 80
5.8 Thermal unfolding monitored by RMSD and RMSF 82
5.9 FSD-1 native contacts 83

6.1 Reverse turn 89
6.2 Bicyclic turn dipeptide 90
6.3 D-Proline-Glycine 90
6.4 α-Aminoisobutyryl-Glycine 91
6.5 Ribonuclease A 93
6.6 D-Proline-Proline 94
6.7 Ribonuclease A turn mimetics 95
6.8 RNase A reverse turn mimetics 97
6.9 Tyr3-Phe12 H-bonds, 300 K 102
6.10 Tyr3-Phe12 H-bonds, 307 K 103

6.11 Inter-strand H-bonds, 300 K 104
6.12 Inter-strand H-bonds, 307 K 105
6.13 FSD-DPP folded structures 106
6.14 FSD-DPP unfolded structures 108
6.15 Hydrophobic Core Contacts, 300K 111
6.16 Hydrophobic Core Contacts, 307K 112

Chapter 1

Introduction

1.1 Computer-aided drug discovery

Application of docking to drug discovery and understanding mini-protein stability are the two main topics discussed in this work. Developing a marketable drug is estimated to take 15 years and a billion dollars. Computer-aided drug discovery (CADD) tools are critical in reducing the time and cost of drug development. The early stages of drug development are where CADD tools can make significant impact in guiding the direction of a therapeutic program.

Chapter 2 introduces the application of high throughput screening (HTS) in drug discovery and how computational tools can be applied to optimize the expensive, random hit-identification strategy of HTS. Cheminformatics tools can filter out problematic compounds that may aggregate, contain known toxic groups, or are non-specific binders. If a crystal structure of the receptor target is available, then virtual screening methods can be applied to eliminate compounds that are less likely to bind. Virtual screening tools like docking programs increase the odds of hit-identification and are complementary to HTS in drug discovery. However, the inter-dependence of sampling and scoring in current docking programs makes it difficult to determine whether a sampling error, or a scoring error, caused a program to fail in validated docking experiments. We have developed a novel docking program, SKATE, that decouples systematic sampling from imperfect scoring. Chapter 3 describes the implementation and results of SKATE.

SKATE was written to prove the concept that systematic sampling improves docking accuracy. It has not been optimized for speed or usability. Chapter 4 discusses improvements to the SKATE docking program that will make it more user-friendly and more likely to be adopted by the drug discovery community.

1.2 Mini-proteins

Mini-proteins that contain fewer than 50 amino acids and fold independently of metal-binding centers or disulfide cross-linking sites are considered model structures for investigating the driving forces behind protein folding. These minimal model systems contain essential features of larger proteins: defined structures, important intramolecular contacts that stabilize the folded state and, in some instances, co-operative folding and unfolding. Their small size makes it feasible to study folding pathways and protein-energy landscapes with long time-scale, molecular-dynamics (MD) simulations.

Chapter 5 provides a detailed analysis of the thermal stability of FSD-1, a 28-residue mini-protein designed to fold into the zinc-finger $\beta\beta\alpha$ motif independent of zinc binding. FSD-1 was an attractive target in simulations studies mostly because of its small size, its sequence consisting of only natural amino acids, and its design has both α-helix and β-sheet secondary structures as well as assumed accessibility of its thermal unfolding transition. However, FSD-1's apparent melting temperature of 42 ℃ and its reported NMR structure have been assumed in subsequent studies without further experimental validation. In Chapter 5, we present a critical analysis of FSD-1's stability by studying its thermal unfolding and solution structure by circular dichroism (CD), differential scanning calorimetry (DSC), replica exchange molecular dynamics (REMD), and NMR spectroscopy. The results suggest an alternative interpretation; the apparent melting temperature reflects a local helix-coil transition and not a protein unfolding transition. FSD-1 is not necessarily a robust mini-protein model system for studying protein folding.

To design more robust model systems, Chapter 6 discusses the impact of pre-organization in protein design and how it can be applied to design mini-proteins that exhibit higher thermal stability. Different semi-rigid reverse-turn mimetics and their impact on protein stability are discussed. A D-proline–Proline turn mimetic was incorporated into FSD-1 and the stability of the resulting chimeric protein was estimated by long timescale molecular dynamics simulations. Simulation results suggest that the Dpro-Pro mimetic could stabilize the β-hairpin in the chimeric protein but further experimental validation is required.

Chapter 2

Drug discovery

2.1 Introduction

The modern drug discovery process for validated biomolecule targets start with the identification of small-molecule hits that modulate a desired function. Receptors and enzymes make up more than 70% of known therapeutic targets[1]. High throughput screening (HTS) of diverse libraries of drug-like compounds is a widely used strategy for identification of hits. Compounds in a HTS library are derived from, but not limited to, combinatorial chemistry, natural products and legacy programs[2]. Hits are active compounds with non-promiscuous binding behavior and they meet certain activity thresholds for given assays. Validated hits are progressed into lead series that are synthetically accessible, exhibit well-defined structure-activity-relationships (SAR), and have good physico-chemical properties (absorption, distribution, metabolism, and excretion (ADME)). To reduce the attrition rates in later, more costly, stages of of drug-development, lead-series must have good *in vitro* affinity and selectivity, but more importantly, they must be optimized for solubility, permeability and metabolic stability. For a review on hit-and-lead generation, see Bleicher et al.[2]. Select compounds in the lead series are further optimized in the lead refinement stage to produce clinical candidates with drug-like properties.

2.2 High throughput screening

High throughput screening is a corner stone technology in identifying hit compounds in the pharmaceutical industry. It is also increasingly being used by academia[3]. Screening libraries of over a million compounds per target is becoming the standard practice in major pharmaceutical companies[4]. HTS, despite its ability to produce enormous amount of information, is not a panacea because the results depend on the composition of the libraries screened. Without a thoughtfully designed compound library, HTS is essentially a very expensive, random hit-identification, strategy. The success rate of finding hits using current HTS technology is approximately 0.1-0.2%. Practitioners are increasingly recognizing that the quality of screening libraries and the accuracy of HTS assays are more important than the number of compounds screened[4]. Using focused libraries for specific therapeutic target classes, kinases for example, will reduce the cost of finding hits. The goal is to perform fewer but high information-content experiments.

2.3 Computational methods in drug discovery

Computational methods play important roles in every stage of the drug development process, from target validation to optimizing lead compounds into clinical candidates.

Often, computer-aided drug discovery makes its largest impact in the early stages. In the case of HTS, cheminformatics tools can be used to filter out poor candidates (ex: compounds containing known toxic subunits) in the compound databases. Filtering tools generally require only 2D information about the molecules and are computationally efficient. Millions of compounds can be processed in a relatively short period of time. Common chemical descriptors are used to filter out compounds that do not meet ADME and toxicity requirements. If specific information about the receptor is available, additional descriptors can be added to further reduce the size of the compound library. The compounds in the resulting smaller library can be prioritized using ligand-based or structured-based virtual screening. If a pharmacophore or structure of the receptor target is available, then 3D screening can be applied to further screen out compounds that are less likely to bind. Molecular docking is a 3D screening technology that finds optimal binding pose(s) of a given ligand. The ligands in a compound library are ranked by their estimated affinity for the receptor. High affinity compounds will then be tested first. Prioritizing the order that compounds in the library are tested will lead to cost savings and faster hit-identification.

2.4 Molecular docking

Small molecule docking programs are used extensively in the pharmaceutical industry and in academia for the discovery of novel lead compounds. A number of docking programs are available as commercial software and from academic labs[5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. Molecular docking programs have three major components: a representation of the system, a sampling algorithm and a scoring function[18]. A docking program must be able to sample near-native poses in order to rank them as top-scoring poses. A pose defines the relative orientation and conformation of a ligand when bound to a receptor.

2.4.1 Historical perspective

Kuntz and colleagues developed one of the earliest docking programs, DOCK[19], to geometrically match the shapes of ligands to the complementary shapes of the binding pocket. The goal was to find small molecules with high degrees of shape complementarity to the receptor binding pocket. The binding pocket was represented by a set of overlapping spheres of varying radii. Each sphere touches the molecular surface at two points. Another set of spheres represented the rigid ligand. Geometrically similar sphere-clusters between the ligand and binding pocket were identified by matching the internal distances of the clusters, subjected to some error limit. The rigid ligand was then transformed by rigid body rotation and translation to fit into the binding site. Top scoring orientations of the ligand poses were selected for subsequent analysis.

In the quest to solve the docking problem, subsequent docking programs by Kuntz's group and others have incorporated features like ligand flexibility, local optimization, receptor flexibility, and advanced scoring functions. However, docking programs have yet to deliver on the promise of predicting binding affinity of compounds, *in silico*, with much consistency or accuracy. One major problem with existing stochastic-based docking programs is that they must couple sampling with imperfect scoring functions. The dependency of sampling on scoring makes it difficult if not impossible to determine whether a sampling or a scoring problem caused a docking program to fail in validation studies.

2.4.2 Interdependence of sampling and scoring

Evolutionary algorithms and other stochastic search methods are a common type of sampling algorithm. They rely on scoring functions to guide their stochastic steps, so the search and scoring processes are necessarily coupled. Scoring functions need to evaluate anywhere from thousands to millions of poses in a docking experiment. To speed up the calculations, the energy functions are simplified so that they can be evaluated quickly. The tradeoff is a less accurate energy function that at best approximates the binding energy of a pose. If a coarse energy function scores a near-native pose poorly, it will be discarded. This problem of false negatives is often the root cause of poor performance. A more rigorous sampling method is systematic search. Implementation of this method can be divided into two subcategories, those that approximate conformational space and those that exhaustively search conformational space. Rigid-body docking of low energy conformers[20], incremental fragment construction[6, 7, 13], and hybrid methods that combine systematic pose generation and stochastic optimization are examples of implementations that approximate a complete systematic search[17, 14]. To the authors' knowledge, only eHiTS[16] and SKATE systematically and exhaustively sample a ligands conformational, rotational and translational degrees of freedom that are constrained by a binding site.

2.5 Summary

Computational tools like docking programs are complementary to HTS in drug discovery. However, the inter-dependence of sampling and scoring in current docking programs makes it difficult to determine whether a sampling error or a scoring error caused a program to fail in a docking experiment. We have implemented a novel docking program, SKATE, which decouples systematic sampling from imperfect scoring.

Chapter 3

SKATE: A docking program

3.1 Introduction

The inter-dependence of sampling and scoring in current docking programs makes it difficult to determine whether a sampling error or a scoring error caused a program to fail in a docking experiment. SKATE is a novel docking program that decouples systematic sampling from imperfect scoring. It employs a rigorous search method to systematically sample conformational, orientation and rotational degrees of freedom of a ligand to find optimal docking poses. A naïve brute force approach, literally rotating each bond, results in combinatorial explosion and becomes computationally intractable. Efficient systematic and exhaustive sampling is achieved by pruning the combinatorial tree using aggregate assembly, discriminant analysis, adaptive sampling, radial sampling and clustering. The resulting sterically allowed poses of a ligand bound to a receptor are then ranked independently with three scoring functions. The docking performance of SKATE is evaluated by three large test sets in terms of self-docking, and two test sets in terms of cross-docking. Compared to state-of-the-art docking programs, SKATE is more accurate.

3.2 Overview of docking methodology

3.2.1 Sampling

Hydrogen-bonding interactions are essential in drug specificity and high affinity binding. SKATE takes advantage of this natural phenomenon by forming all possible hydrogen-bonds between the ligand and the receptor pocket to anchor systematic search. Once a sterically allowed hydrogen-bond is formed between a receptor atom and a ligand atom, SKATE then systematically and exhaustively samples the ligand's torsional degrees of freedom. The simplest systematic approach to find all sterically allowed conformations of a flexible ligand is to rotate each rotatable bond. Assuming a ligand molecule of N atoms with T rotatable bonds, and a receptor pocket of M atoms, if each rotatable bond of the ligand is explored at angular increments of A degrees, there are $\frac{360}{A}$ values to be examined for each T resulting in $\left(\frac{360}{A}\right)^T$ possible conformations to be examined for steric conflict. The 3D coordinates that determine the geometry of a conformation can be generated by applying appropriate transformation matrices to different subsets of atoms. These conformers must be checked for van der Waals (VDW) overlap to eliminate sterically impossible conformations. To a first approximation, there are $\frac{N(N-1)}{2}$ pair-wise distance calculations that must be performed for each conformation. Then $M \times N$ pair-wise distance calculations must be performed between atoms in each conformation and those in the receptor pocket. These distances are checked against the allowed sum of VDW radii for the two atoms involved. The number of VDW comparison V for a single hydrogen-bond formed between the receptor and the ligand is given by

$$V = \left(\frac{360}{A}\right)^T \times \left(\frac{N\,(N-1)}{2} + M \times N\right) \tag{3.1}$$

The rate-limiting step in this brute force approach is the sheer number of VDW comparisons that must be performed in order to find sterically allowed poses. As an example, sampling at torsional increments of 10 degrees for a ligand with 6 rotatable bonds and 50 atoms and a receptor pocket of 1000 atoms will result in 1.4×10^{13} VDW calculations. Assuming there are a combination of 50 possible hydrogen-bonds that can be formed, it would take 22 years to complete this calculation on a modern, single CPU computer that is capable of processing 1 million VDW comparisons per second.

Such a brute force approach to systematic search is inefficient and unnecessary. SKATE implements a number of strategies that truncate the combinatorial explosion. Sterically allowed poses of a ligand as constrained by a receptor pocket are systematically sampled by a step-wise build up of aggregates (Figure 3.1). An aggregate is defined as a set of atoms whose relative positions are invariant to rotational degrees of freedom[21]. A ligand is divided into individual aggregates around internal rotatable bonds (Figure 3.2). An aggregate capable of hydrogen-bonding is transformed by rigid-body translation and rotation to form an energetically favorable hydrogen-bond with the receptor. The geometries of the newly formed hydrogen-bond are determined by a set of hydrogen-bonding geometric parameters. A second aggregate that shares a common rotatable bond with the first aggregate is spliced onto the partial molecule by applying the appropriate transformations. The range of sterically allowed torsions for this rotatable bond is analytically determined by discriminant analysis[21]. Discrete values in the range of allowed torsions are sampled by rotating the second aggregate around the rotatable bond that joins the first and second aggregates. The step-wise assembly of sterically allowed conformations of the ligand within the receptor pocket continues until all aggregates have been added. As

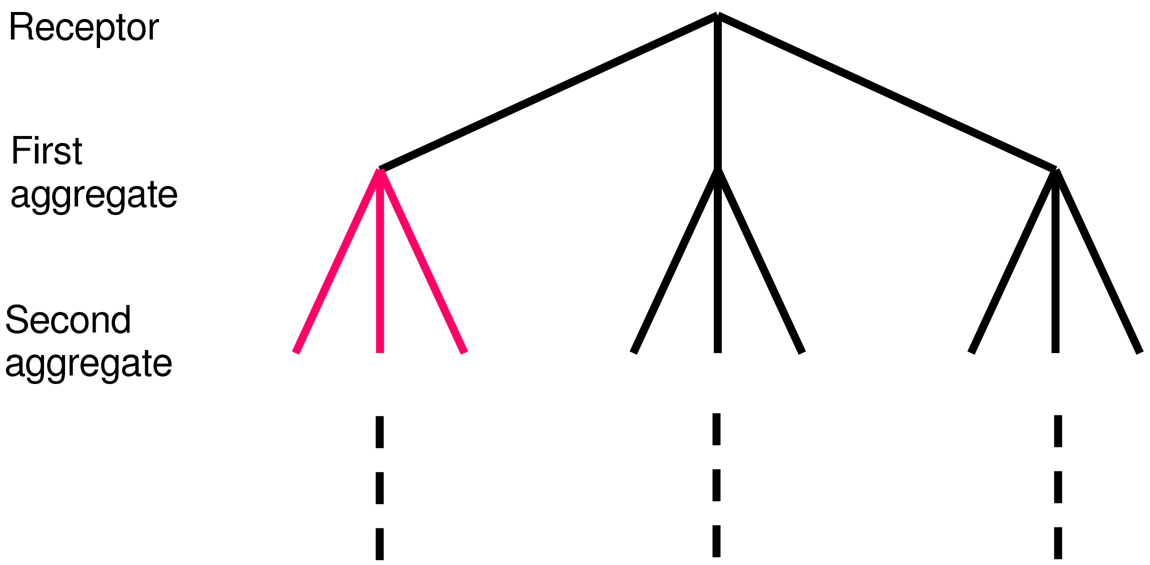


Figure 3.1: The tree structure of systematic search of conformational space for a ligand hydrogen-bonded to a receptor. Vertices of the tree represent ligand aggregates; edges represent discrete torsion values of a ligand's rotatable bonds and a ligand-receptor hydrogen-bond. Red edges represent "pruning" of the search tree by eliminating branches of the tree where the addition of an aggregate is sterically prohibited for any torsion value. Sterically allowed conformations are represented by the tree leaves that are connected by black edges. The first aggregate is hydrogen-bonded to the receptor and the bonding geometries are determined from a set of geometric parameters. At each branch point, a new aggregate may be added to the existing partial conformation if it is sterically allowed (black lines). Each black line represents a torsion value of a rotatable bond where an aggregate is added to the existing partial molecule. The assembly of a sterically allowed conformation continues until aggregates along every branch have been systematically evaluated.

shown in Figure 3.1, the possible conformations of a flexible ligand hydrogen-bonded to a receptor can be represented by a search tree. The tree is anchored by a receptor atom that forms a hydrogen bond with the ligand. SKATE systematically finds sterically allowed ligand poses (tree leaves) by performing a depth-first search of the tree. Systematic search is performed for each possible pairing of hydrogen-bonding atoms between the ligand and receptor. A more detailed explanation of systematic search and discriminant analysis is provided in the methods section.

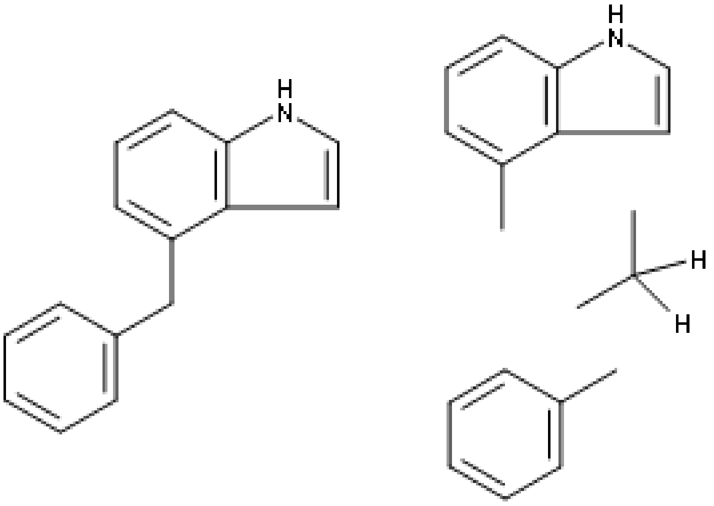


Figure 3.2: A simple molecule (left) is divided into its aggregates (right) by partition at its rotatable bonds.

3.2.2 Scoring

SKATE decouples systematic sampling from scoring. A unique feature of SKATE is that any scoring function may be used to rank the poses generated by SKATE. SKATE itself does not use a scoring function to determine if a pose is low energy. It uses discriminant analysis and incremental build-up to find a set of sterically allowed poses. Those poses are clustered with a heavy atom root-mean-square deviation (RMSD) cutoff of 0.5 Å. In this work, we used energy functions in FRED[20], Rosetta[10] and X-Score[22] to rank or score the poses generated by SKATE. These scoring functions are made available by their respective authors at no charge to academic groups.

FRED or Fast Rigid Exhaustive Docking is a commercial docking program developed by OpenEye Inc. that can also be used to score poses generated by other programs[20]. We used FRED's default consensus scoring function that is an equal-weighted sum of ranks by chemgauss3, PLP, and oechemscore. Chemgauss3 uses smooth Gaussian functions to represent the shape and chemistry of molecules[20]. PLP or Piecewise Linear Potential is a minimal scoring function that includes a steric term and a hydrogen-bonding term, but no electrostatic term[23]. Oechemscore is an OpenEye variant of chemscore, an empirical scoring function[24]. We also examined how FRED scoring is affected by a fast, rigid-body local optimization of SKATE-generated poses prior to scoring.

Rosetta's energy function was originally trained for protein structure prediction and was extended to score protein-ligand interactions[25]. The energy function consists of a weighted sum of force-field-based and knowledge-based terms calculated from the receptor and ligand coordinates. Hydrogen atoms are explicitly treated. The terms include VDW interactions, an implicit solvent model, an explicit orientation-dependent hydrogen-bonding potential, and an electrostatics model. For this work, we used Rosetta's energy function, referred to as Rosetta-Score, to rank poses generated by SKATE. X-Score is an empirical scoring function that treats hydrophobic effect by using three different functions and averaging the results[22]. Each of the three functions includes a VDW interaction term, a hydrogen-bonding term, a hydrophobic effect term, a torsional-entropy penalty, and a regression constant. X-score was trained to reproduce the known binding affinity of 200 protein-ligand complexes.

3.3 Methods

An aggregate is defined as a set of atoms whose relative positions are invariant to rotational degrees of freedom[26]. Atoms in an aggregate could be directly bonded, have a 1-3 relationship defined by a bond angle, be part of a ring system, or have bonds between them conjugated by resonance. Table 3.1 lists the number of rotatable bonds, sampled by SKATE, for the ligands in the three self-docking test sets. Figure 3.2 illustrates how a simple molecule was divided into three aggregates. There are $T + 1$ aggregates and T torsional degrees of freedom in a flexible molecule. Sterically allowed conformations of a ligand are generated by assembling its aggregates. Since the distance between two atoms within an aggregate is constant, it is not necessary to check for VDW clashes between atoms within the same aggregate.

In SKATE, sterically allowed poses of a ligand are constructed in a stepwise fashion by re-assembling the aggregates comprising the ligand. Starting with an initial aggregate that contains an atom that forms a hydrogen bond with a receptor atom, a second aggregate is added via the rotatable bond that joins the two aggregates (Figure 3.1). Some torsion values around this shared rotatable bond will lead to VDW overlaps between atoms in aggregate two and atoms in aggregate one, as well as atoms in the receptor. It is extremely inefficient to assemble two aggregates for a given torsion only to find out that it is a sterically impossible conformation. Discriminant analysis solves this problem by analytically calculating the range of sterically allowed torsions within which two aggregates can be assembled together. The result is that only allowed torsions are sampled. In theory, systematic sampling should find all sterically allowed poses of a ligand. In practice, SKATE discretizes the continuous conformational space and then uses adaptive torsion sampling and radial sampling to ensure sufficient sampling[27].

Discriminant analysis was first applied to systematically search the conformational hyperspace available to a flexible molecule to define three-dimensional quantitative structure-activity relationships (3D-QSAR) and biological receptor mapping[26]. In the construction of a molecule from stepwise addition of aggregates, there are two sets of atoms to consider. First are those in the sterically allowed partial molecule (set A) previously constructed. Second are those in the next aggregate (set B) to be added to the existing partial molecule. Atoms in set B must be checked against those in set A to find torsions that are sterically allowed. Distance constraint equations are used analytically to determine the possible torsion ranges such that a new aggregate can be added without steric overlap between atoms in the new aggregate (set B) and the partial conformation (set A). These equations, derived elsewhere[26], describe the variable distance between any two atoms as a function of a single torsion angle (ω).

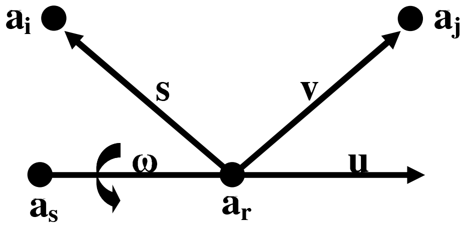


Figure 3.3: The variable distance between a fixed atom a_i and a rotatable atom a_j is a function of a single torsional variable ω. Atoms a_i, a_s and a_r are rigid with respect to each other and they belong to the sterically allowed conformation of a partially docked ligand. Atoms a_s and a_r forms the rotatable bond and determine the rotational axis. $\hat{u}$ is a unit vector along the axis of rotation. The torsional variable of ω is being evaluated by discriminant analysis to determine the range of torsions where atom a_j does not clash with any atoms in the partial conformation.

The square of the interatomic distance between a_j and a_i in Figure 3.3 is given by:

$$d_{ij}^2(\omega) = d_1 + d_2 cos(\omega) + d_3 sin(\omega) \tag{3.2}$$

where coefficients d_1, d_2, and d_3 are defined as follows:

$$d_1 = |\hat{s}|^2 + |\hat{v}|^2 - 2(\hat{s} \cdot \hat{v}_1) \tag{3.3}$$

$$d_2 = -2(\hat{s} \cdot \hat{v}_2) \tag{3.4}$$

$$d_3 = -2(\hat{s} \cdot \hat{v}_3) \tag{3.5}$$

v_1, v_2, and v_3 are the three orthogonal components of the vector v in Figure 3.3 where

$$\hat{v} = a_j - a_r \tag{3.6}$$

$$\hat{v}_3 = \hat{u} \times \hat{v} \tag{3.7}$$

$$\hat{v}_2 = \hat{u} \times \hat{v}_3 \tag{3.8}$$

$$\hat{v}_1 = \hat{v} - \hat{v}_2 \tag{3.9}$$

Equation 3.2 can be rewritten as

$$d_{ij}^2(\omega) = \frac{ax^2 + bx + c}{1 + x^2} \tag{3.10}$$

where

$$a = d_1 - d_2 \tag{3.11}$$

$$b = 2 \times d_3 \tag{3.12}$$

$$c = d_1 + d_2 \tag{3.13}$$

$$x = tan(\frac{\omega}{2}) \tag{3.14}$$

Let c_{ij} be the sum of the VDW radii for atoms i and j, then differential distance function

$$\delta_{ij}^2(\omega) = d_{ij}(\omega) - c_{ij} \tag{3.15}$$

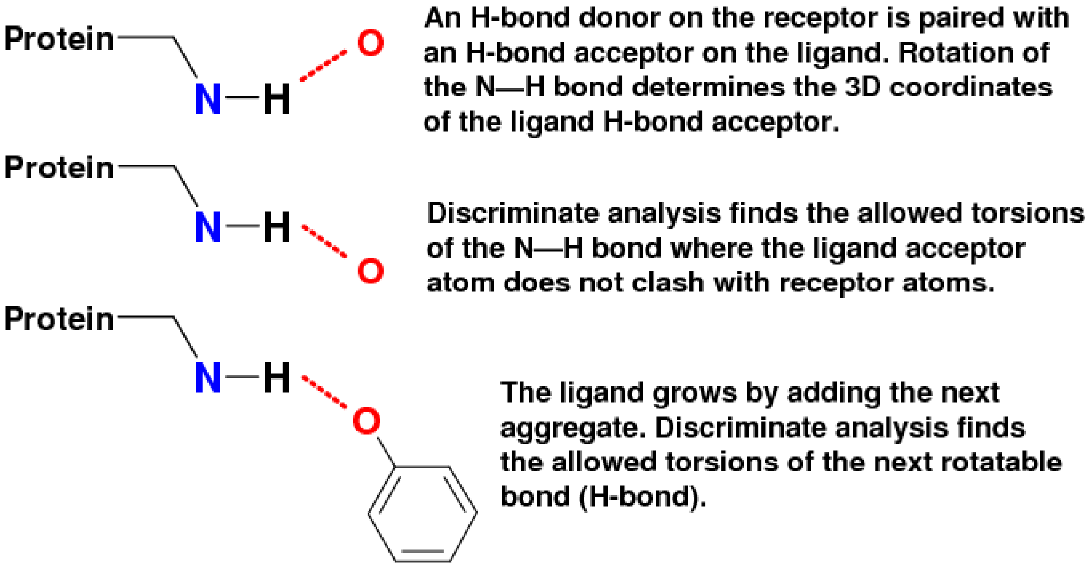
is evaluated to determine whether or not the two atoms are in contact. The differential distance function can be converted to a quadratic form:

$$\begin{aligned}
\delta_{ij}^2(\omega) &= \frac{(a - c_{ij}^2)x^2 + bx + (c - c_{ij}^2)}{1 + x^2} && (3.16)\\
D &= b^2 - 4(a - c_{ij}^2)(c - c_{ij}^2) && (3.17)\\
x &= \frac{-b \pm \sqrt{D}}{2a} && (3.18)\\
\omega &= 2tan^{-1}(x) && (3.19)
\end{aligned}$$

The resulting discriminant D can be used to determine if there is a real or imaginary solution to $\delta_{ij}^2(\omega)$. If $D > 0$, then $\delta_{ij}^2(\omega)$ has real roots and the upper and lower bound values of the torsional range (ω) can be calculated from the above equations. If $D \leq 0$, $\delta_{ij}^2(\omega)$ has complex or real double degenerate roots. For $c - c_{ij}^2 \geq 0$, $\delta_{ij}^2(\omega)$ is positive for all values of ω implying that atom i and atom j never come in contact for any torsional value of ω. For $c - c_{ij}^2 < 0$, $\delta_{ij}^2(\omega)$ is negative for all values of ω and there is no sterically allowed way to add the new aggregate. In this case, the new partial conformation will be discarded and the search branch truncated.

The distance constraint equations minimize the number of pair-wise intramolecular and intermolecular distances that must be evaluated in a systematic search. They prune the search tree by analytically determining torsion ranges that result in sterically allowed partial or complete conformers. The intersection of allowed torsion ranges for every atom pair spanning a rotatable bond results in discontinuous slices of torsion ranges in which a new aggregate is added during the step-wise construction process. The torsional ranges are discretized by adaptive sampling and radial sampling to ensure sufficient sampling[27]. Adaptive sampling, as opposed to uniform sampling, ensures that SKATE does not over-sample or under-sample a torsional range. Radial sampling determines the increment in degrees between two sampled torsions. In SKATE, a rotation of an aggregate around its rotatable bond displaces an atom in the aggregate by a maximum of 0.25 Å.

SKATE pairs an H-bond donor of a receptor with an H-bond acceptor of a ligand, and vice versa, to anchor systematic search. In SKATE, three parameters are used to define a hydrogen bond, the distance between the hydrogen atom and the acceptor atom; the angle formed by the acceptor, hydrogen, and donating atoms; the angle formed by the acceptor base, acceptor, and hydrogen atoms. Figure 3.4 illustrates how SKATE initializes its H-bond pairing and systematic search process. A receptor H-bond donor is paired with a ligand H-bond acceptor. Rotation of the N–H bond



An H-bond donor on the receptor is paired with an H-bond acceptor on the ligand. Rotation of the N—H bond determines the 3D coordinates of the ligand H-bond acceptor.

Discriminate analysis finds the allowed torsions of the N—H bond where the ligand acceptor atom does not clash with receptor atoms.

The ligand grows by adding the next aggregate. Discriminate analysis finds the allowed torsions of the next rotatable bond (H-bond).

Figure 3.4: Docking of a ligand to a receptor by pairing H-bonding partners. Rotatable bonds in the ligand are searched systematically to find allowed torsions that generate a bound pose for further evaluation.

on the receptor determines the 3D coordinate of the ligand acceptor atom. Using discriminant analysis, SKATE quickly determines the allowed torsions of the N–H bond such that the ligand acceptor atom does not clash with receptor atoms. The next bond to be rotated is the H-bond between the receptor and the ligand. This determines the allowed torsions of the H-bond such that ligand atoms in the first aggregate do not clash with the receptor atoms. The remaining aggregates are then systematically searched by recursion. Sterically allowed poses for a given ligand-receptor hydrogen-bond represent leaves of a tree graph where nodes represent aggregates and edges represent discrete torsions of rotatable bonds (Figure 3.1). SKATE travels this tree using a depth first search approach as illustrated by the following pseudo code.

Systematic Search Pseudo Code

```
MAIN()
        DOCK(receptor, ligand)
        SEARCH(receptor, ligand, torsions, agg_idx)
SEARCH(receptor, ligand, torsions, agg_idx)
        UPDATE(ligand, agg_idx)
        VALIDATE(receptor, ligand, torsions, agg_idx)
        for each allowed torsion of aggregate agg_idx
                ROTATE(ligand, torsions)
                if last aggregate
                        RECORD(ligand)
                else
                        SEARCH (receptor, ligand, torsions, agg_idx+1)
                end if
        end for
```

The DOCK procedure transforms the coordinates of a ligand H-bond partner such that it forms a hydrogen bond with a receptor partner. The resulting H-bond geometry is determined by a set of geometric H-bond parameters.

The UPDATE procedure transforms the atoms in aggregate agg_indx to be in the same local coordinates as the previously searched aggregates and partially assembled molecule.

The VALIDATE procedure performs discriminant analysis to find allowed torsions of the rotatable bond that connects aggregate agg_indx with the previously searched aggregates of the ligand. A list of allowed torsions is stored in the torsions data structure.

The ROTATE procedure simply rotates an aggregate to an allowed torsion that was calculated by the VALIDATE procedure.

Due to inherent errors in X-ray structure determination, there are often VDW clashes between ligand and receptor atoms in crystal structures. We employed a VDW scaling factor to reduce the VDW radii of protein and ligand atoms to ensure the reproduction of experimental structures[28]. A general scaling factor of 0.95 is applied to ligand intramolecular interactions. A 1,4 scaling factor of 0.87 is applied to ligand atoms in 1-4 relationships. Intermolecular interactions are scaled by a factor of 0.9 and hydrogen-bond interactions are scaled by a factor of 0.6.

Experimentally determined structures of ligand complexes are used to define binding pockets for the purpose of docking validation. A binding pocket in SKATE is defined as any receptor atom that is within 5 Å of any atoms in a co- crystallized ligand. To prevent SKATE from building ligand poses where the ligand extends into solvent space, a shell of dummy solvent atoms is added to the receptor using Sybyl[29]. These dummy solvent atoms surround the entire surface of the protein and do not occupy the binding pocket. Dummy solvent atoms within 5 Å of atoms of a co-crystallized ligand are discarded.

3.4 Data Sets

SKATE was tested on five data sets in assessing its self-docking and cross- docking performance. Results from four of the data sets can be compared directly to results from published docking programs.

3.4.1 Astex/CDCC diverse set

Hartshorn et al. prepared a set of 85 high-quality and diverse protein-ligand complexes and made them publicly available as a validation set for testing docking performance[30]. Protein targets were selected based on their relevance to drug discovery or agrochemical research. Consequently, only complexes with drug-like ligands were allowed in this set. To ensure complex diversity, no receptor was represented more than once. Furthermore, the ligands contained distinct molecular recognition types. A special focus was placed on selecting very high-quality experimental structures of which the experimental binding mode of the ligands was easily assessed. Protein structures were prepared by removing solvents and small ions. Exceptions were made for water molecules that coordinate a metal ion and for small ions that mimic a cofactor. His, Asn and Gln side-chain placements in the crystal structure that were not consistent with hydrogen-bonding patterns were rotated if such rotations would significantly improve hydrogen-bonding. This is reasonable because crystallographers usually cannot place His, Asn, and Gln side chains with absolute certainty based on electron density alone. This data set was downloaded from the Cambridge Crystallographic Data Centre (http://www.ccdc.cam.ac.uk).

3.4.2 Surflex set

To compile a test set for Surflex, Jain filtered 134 protein-ligand complexes in the GOLD data set by removing complexes that (i) contained ligands with more than 15 rotatable bonds, (ii) were covalently attached to the protein, and (iii) contained obvious errors in structure[5, 13]. The resulting 81 complexes were made available on http://jainlab.ucsf.edu. The protein files in the original GOLD set were prepared by removing water molecules and by adding hydrogen atoms while taking protonation states into account. Exceptions were made to keep water molecules and metal atoms that coordinated ligand binding[5].

3.4.3 Vertex set

Perola et al. prepared a test set of 150 protein-ligand complexes to compare the performances of Glide, GOLD and ICM[31]. These complexes were selected for their relevance to modern drug discovery programs. Ligands were selected for (i) their drug-like properties; (ii) molecular weights between 200 and 600 Daltons; (iii) having between 1 and 12 rotatable bonds; and (iv) structural diversity. The ligands in the Vertex set were prepared by extracting them from their respective PDB files and assigning bond orders and correct protonation states by visual inspection. Protein structures were prepared by removing subunits, ions, solvent and other small molecules not involved in binding. Metal ions and tightly bound water molecules in the ligand binding site were preserved[31]. Hydrogen atoms were added to the protein. The structures of ligand, protein, and co-factor were minimized as a complex for 1,000 steps using Macromodel and the OPLS-AA force field. All heavy atoms were constrained to their original positions during minimization. The structures with optimized hydrogen positions were saved. Of the 150 complexes, 100 are PDB entries and 50 are corporate structures. The files of the 100 PDB complexes are available on the Jain Lab website (http://jainlab.ucsf.edu)[32]. Seven complexes in the Vertex set are also included in either the Astex/CDCC set or the Surflex set (Table 3.1).

3.4.4 Thymidine kinase set

Bissantz et al. tested the virtual screening capability of docking programs by using the crystal structure of HSV-1 thymidine kinase (TK) (PDB ID: 1KIM), 10 known ligands, and 990 randomly chosen decoys[33]. In this work, the 10 known ligands were docked to the 1KIM structure to test SKATE's performance in cross-docking. The structures were prepared as described in Bissantz et al. No optimization of ligand or receptor coordinates was performed.

3.4.5 Cyclin dependent kinase 2 set

Seventy-three known ligands that have been co-crystallized with cyclin dependent kinase 2 (CDK2) were docked to a single high resolution CDK2 structure (PDB ID: 2B54, 1.85 Å)[34]. These ligands occupy the ATP-binding site of CDK2. To prepare the receptor, water molecules and co- crystallized ligands were removed from the 2B54 structure, and hydrogen atoms were added to the receptor using Sybyl 8.1[29]. The ligand structures were extracted from their respective complexes, and were assigned correct bond orders and protonation states by visual inspection. To create reference coordinates of the 73 known ligands, their respective co-crystallized receptors were aligned to the 2B54 structure and the ligands were extracted and saved as mol2 files. The CDK2 data set is available for download at http://www.ccb.wustl.edu/~jafeng.

3.5 Docking setup

The files in the Astex/CDCC, Surflex and Vertex sets were downloaded from their respective websites and used as obtained. Hydrogen atoms were already added to protein and ligand structures by the test sets' respective authors. No further optimization of protein or ligand geometries were performed since it could lead to biased results[35]. For the Vertex set, its author did minimize the ligand and receptor hydrogen atoms while constraining the heavy atoms to their original locations[31]. Ligand and protein files in PDB or MOL formats were converted to the mol2 format and assigned Tripos atom types. The coordinates of these ligand files were used as references when calculating RMSD values.

In this study, the experimentally determined ligand was used to define the binding pocket for the purpose of docking. Any receptor atom that is within 5 Å of an atom in the co-crystallized ligand was considered part of the binding pocket. A list of potential hydrogen-bond donors and acceptors were created by inspecting the atoms in the pocket. SKATE attempted to dock all possible pairings of ligand hydrogen-bond donors with protein hydrogen-bond acceptors, and vice versa. A vast majority of these attempted pairings resulted in immediate search termination because they were not sterically allowed. The resulting sterically allowed poses generated by SKATE were written to a file in the mol2 format to be ranked by scoring functions.

To delete conformational memory of the experimentally determined ligands, SKATE set the torsions of all rotatable bonds to 180 degrees. Experimentally determined bond angles and bond lengths were not modified. The current version of SKATE does not sample ring conformations; instead experimentally determined ring conformations were used.

3.6 Results and discussion

3.6.1 Sampling accuracy

In order to rank a near native pose as the top-scoring pose, a docking program must be able to sample such poses. The inter-dependence of sampling and scoring in current docking programs makes it difficult to determine whether it is a sampling error or a scoring error that caused a program to fail in a test case. SKATE approaches the docking problem by decoupling systematic sampling from scoring. It anchors a search by pairing a ligand hydrogen-bond donor to a receptor hydrogen-bond acceptor and vice versa. For each hydrogen-bond formed, SKATE systematically samples a ligand's torsional degrees of freedom to find poses that sterically fit within a receptor pocket. Figure 3.5 shows the cumulative proportion of best poses, as measured by RMSD to the experimental structure (reference), that were generated by SKATE for the complexes in the Astec/CDCC, Surflex and Vertex self-docking test sets. A pose is considered best if its heavy atom RMSD to the reference structure is the lowest. Table 3.1 lists the RMSD values of the best poses and top-scoring poses for each complex in the three test sets.

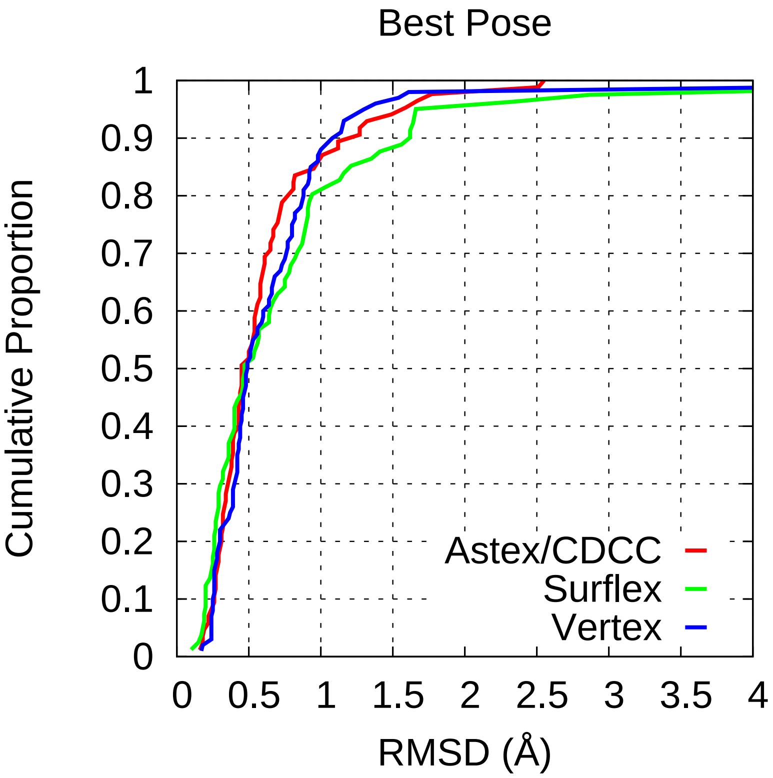


Figure 3.5: Cumulative proportion of best RMSD poses for the Astex/CDCC (red), Surflex (green), and Vertex (blue) sets. There are 85 complexes in the Astex/CDCC set, 81 complexes in the Surflex set and 100 complexes in the Vertex set.

Table 3.1: Results for SKATE on the Astex/CDCC, Surflex and Vertex complexes[1]

Vertex Set				Astex/CDCC Set				Surflex Set			
PDB code	no rot bonds	best pose rmsd	top rank rmsd	PDB code	no rot bonds	best pose rmsd	top rank rmsd	PDB code	no rot bonds	best pose rmsd	top rank rmsd
13gs	3	0.49	0.83	1g9v	5	1.49	1.51	1abe	0	0.45	0.45
1a42	8	0.77	1.41	1gkc	8	1.27	0.95	1acj	1	0.46	0.68
1a4k	5	0.66	1.81	1gm8	4	1.59	2.24	1ack	2	0.53	3.83
1a8t	8	1.00	7.99	1gpk	1	0.27	0.30	1acm	6	0.70	0.65
1afq	9	0.86	8.23	1hnn	1	0.67	0.98	1aco	2	0.27	0.35
1aoe	3	0.48	0.86	1hp0	2	0.45	0.36	1aha	0	0.26	0.18
1atl[2]	8	0.91	1.26	1hq2	1	0.43	0.26	1atl	9	1.35	2.86
1azm	2	0.51	1.28	1hvy	8	1.77	1.64	1baf	4	0.64	0.92
1bnw	5	0.64	5.48	1hwi	9	0.61	1.11	1bbp	9	0.75	0.76
1bqo	6	0.30	0.48	1hww	1	0.22	0.14	1bma	9	1.65	2.47
1br6	3	0.39	1.14	1ia1	2	0.26	0.36	1cbs	0	0.20	0.36
1cet	7	0.93	7.45	1ig3	4	0.44	1.20	1cbx	5	0.29	0.43
1cim	3	0.28	1.09	1j3j	2	0.18	0.30	1com	4	0.46	0.79
1d3p	12	1.14	1.19	1jd0	1	0.72	3.36	1coy	1	0.32	0.51
1d4p	3	0.24	0.60	1jje	7	0.58	7.97	1dbb	1	0.25	0.51
1d6v	7	0.92	2.17	1jla	7	0.70	0.77	1dbj	1	0.32	0.54
1dib	7	0.80	2.88	1k3u	6	0.27	0.29	1dr1	3	0.28	1.48
1dlr	4	0.42	0.64	1ke5	1	0.34	0.29	1dwd	8	1.16	2.97
1efy	3	0.42	1.76	1kzk	9	0.65	0.89	1eap	10	0.82	0.81
1ela	8	0.44	0.68	1l2s	2	0.31	0.51	1epb	0	0.91	0.74
1etr[2]	8	0.46	0.60	1l7f	8	0.33	0.44	1etr	8	0.92	0.93
1ett	6	0.51	0.98	1lpz	6	0.71	1.00	1fen[4]	0	—	—
1eve	6	1.38	1.01	1lrh	2	1.32	1.42	1fkg	9	0.78	1.60
1exa	4	0.25	0.32	1m2z	3	0.19	0.60	1fki	0	0.30	0.34
1ezq	10	0.39	0.71	1meh	7	1.12	1.07	1frp	7	0.26	0.92
1f0r	4	0.40	0.75	1mmv	8	0.81	0.58	1glq	12	1.62	9.14
1f0t	5	0.73	2.57	1mzc	7	1.27	2.26	1hdc	6	1.41	1.61
1f4e	2	0.41	1.09	1n1m	3	0.82	0.57	1hdy	0	0.90	0.74
1f4f	8	0.67	2.23	1n2j	4	0.67	0.47	1hri	9	2.87	10.18
1f4g	11	1.23	1.49	1n2v	3	0.45	1.08	1hsl	4	0.36	0.42
1fcx	4	0.28	0.32	1n46	5	0.43	0.66	1hyt	5	0.65	0.78
1fcz	3	0.26	0.39	1nav	5	0.39	0.73	1lah	6	0.36	0.36
1fjs	8	1.16	2.01	1of1	2	0.32	0.32	1lcp	4	0.56	1.13
1fkg[2]	9	0.64	1.33	1of6	4	0.32	0.64	1ldm	0	0.18	0.39
1fm6	6	0.68	0.74	1opk	2	0.35	0.56	1lic	15	5.05	5.07
1fm9	11	0.48	2.31	1oq5	3	0.37	5.00	1lna	10	0.64	0.74

Continued on Next Page…

Table 3.1 – Continued

Vertex Set				Astex/CDCC Set				Surflex Set			
PDB code	no rot bonds	best pose rmsd	top rank rmsd	PDB code	no rot bonds	best pose rmsd	top rank rmsd	PDB code	no rot bonds	best pose rmsd	top rank rmsd
1frb	5	0.24	0.23	1owe	2	1.01	1.84	1lpm	8	0.89	6.82
1g4o	5	1.04	3.59	1oyt	4	0.40	0.62	1lst	7	0.29	0.21
1gwx	10	1.61	2.19	1p2y	1	1.67	4.87	1mdr	3	0.20	0.47
1h1p	3	0.43	0.43	1p62	3	0.16	0.40	1mrg	0	0.29	0.59
1h1s	4	0.46	0.66	1pmn	6	2.51	6.70	1mrk	3	0.36	0.99
1h9u	3	0.26	0.39	1q1g	3	0.36	0.69	1nco	9	0.91	0.68
1hdq	5	0.98	1.03	1q41	1	0.34	0.54	1phg	3	0.87	4.39
1hfc	10	0.60	0.52	1q4g	3	0.27	0.64	1rds	8	1.03	1.74
1hpv	12	0.88	0.89	1r1h	10	0.43	0.53	1rob	5	0.94	1.41
1htf	13	0.92	2.10	1r55	8	0.98	0.86	1snc	6	0.54	0.80
1i7z	5	0.39	0.48	1r58	9	0.77	0.91	1srj	2	0.40	0.40
1i8z	6	4.82	4.80	1r9o	3	0.52	0.76	1stp	5	0.40	0.79
1if7	7	0.88	5.13	1s19	5	0.38	0.62	1tka	8	1.21	1.46
1iy7	5	0.30	0.64	1s3v	5	0.38	0.73	1tmn	13	0.75	1.48
1jsv	1	0.46	0.39	1sg0	3	0.32	0.49	1tng	2	0.10	0.69
1k1j	8	0.45	1.61	1sj0	6	0.54	0.66	1tni	5	0.57	1.92
1k22	9	0.42	0.49	1sq5	6	0.81	1.68	1tnl	2	0.19	0.40
1k7e	4	0.18	1.00	1sqn	2	0.22	0.27	1trk	9	0.40	0.58
1k7f	5	0.66	1.20	1t40	6	0.65	0.69	1ukz	4	0.20	0.24
1kv1	1	0.24	0.81	1t46	4	0.43	0.38	1ulb	0	0.20	0.46
1kv2	6	0.53	0.55	1t9b	3	0.61	0.47	1wap	4	0.17	0.32
1l2s[3]	1	0.17	0.42	1tow	4	0.56	4.81	2ada	3	0.15	0.19
1l8g	3	0.33	2.13	1tt1	4	0.29	0.73	2ak3	4	0.42	0.55
1lqd	5	0.56	1.00	1tz8	5	0.58	2.29	2cgr	5	1.66	1.80
1m48	7	0.56	0.96	1u1c	6	0.60	1.12	2cht	3	0.67	1.36
1mmb	13	0.82	6.79	1u4d	1	0.28	0.91	2cmd	6	0.57	0.45
1mnc	10	0.80	1.55	1uml	9	0.54	0.62	2ctc	4	0.24	0.41
1mq5	3	0.26	0.41	1unl	6	0.55	0.95	2dbl	6	1.62	1.35
1mq6	4	0.27	0.32	1uou	2	0.73	0.73	2gbp	2	0.19	0.26
1nhu	8	0.44	0.46	1v0p	6	0.50	0.45	2lgs	5	1.13	1.74
1nhv	8	1.15	7.46	1v48	6	0.45	0.35	2phh	2	0.25	0.41
1o86	12	7.97	8.63	1v4s	3	0.29	0.28	2r07	8	1.64	8.96
1ohr	11	0.39	0.46	1vcj	8	0.59	0.84	2sim	5	0.40	1.21
1ppc	8	1.08	1.96	1w1p	0	0.24	0.28	3aah	3	0.79	0.39
1pph	6	0.76	2.03	1w2g	2	0.39	0.56	3cpa	6	0.88	0.92
1qbu	10	0.72	0.77	1x8x	4	0.44	0.78	3hvt	1	2.33	1.62
1qhi	4	0.25	0.40	1xm6	5	0.53	1.23	3ptb	1	0.23	0.42

Continued on Next Page…

Table 3.1 – Continued

Vertex Set				Astex/CDCC Set				Surflex Set			
PDB code	no rot bonds	best pose rmsd	top rank rmsd	PDB code	no rot bonds	best pose rmsd	top rank rmsd	PDB code	no rot bonds	best pose rmsd	top rank rmsd
1ql9	3	0.37	0.37	1xoq	5	0.95	4.03	3tpi	7	0.26	0.26
1qpe	2	0.42	0.55	1xoz	1	0.26	0.28	4cts	2	0.27	0.44
1r09	3	0.75	0.60	1y6b	6	0.39	0.34	4dfr	8	0.84	1.19
1syn	7	0.59	2.48	1ygc	10	2.55	3.85	6abp	1	0.34	0.39
1thl	11	0.82	0.95	1yqy	4	0.18	0.51	6rnt	4	0.38	7.01
1uvs	8	1.30	1.49	1yv3	2	0.30	0.31	6rsa	2	0.47	0.76
1uvt	5	0.49	0.45	1yvf	4	0.58	0.60	7tim	3	0.47	1.13
1ydr	1	0.36	0.36	1ywr	5	0.54	0.45	8gch	8	1.56	2.20
1yds	4	0.44	0.37	1z95	5	0.31	0.34				
1ydt	7	0.87	3.46	2bm2	7	0.45	1.48				
2cgr[2]	5	0.60	0.78	2br1	6	1.12	1.58				
2csn	4	1.54	3.01	2bsm	6	0.50	0.74				
2pcp	2	0.24	0.30								
2qwi	5	0.24	1.01								
3cpa[2]	7	0.45	1.07								
3erk	3	0.25	0.60								
3ert	8	0.48	1.03								
3std	5	0.26	0.26								
3tmn	6	0.47	0.71								
4dfr[2]	8	0.98	1.42								
4std	4	0.26	0.35								
5std	4	0.52	0.32								
5tln	9	0.80	1.06								
7dfr	8	0.77	1.44								
7est	6	0.43	0.82								
830c	7	0.29	0.52								
966c	7	0.30	0.63								

---

[1] Top rank poses were scored by FRED-Opt-Score

[2] Complex is also part of the Surflex set.

[3] Complex is also part of the Astex/CDCC set.

[4] The ligand in complex 1fen does not have any atom that is capable of hydrogen bonding.

For an RMSD threshold of 2 Å, sampling accuracy rates are 98%, 95%, and 98% for the Astex/CDCC, Surflex and Vertex sets, respectively. For an RMSD threshold of 1 Å, the respective sampling accuracy rates are 86%, 80%, and 88% for the three self-docking sets. Highly accurate ligand poses that approximate the native pose below 1 Å RMSD are a prerequisite to improving solutions to the scoring problem[25]. For all but a few test cases in the three test sets, SKATE was able to sample poses that were within 2 Å. The highly accurate sampling of SKATE can be attributed to the systematic sampling algorithm. It is essential for a docking program to sample near-native poses in order to give scoring functions the opportunity to rank them as top-scoring poses.

Perola et al. evaluated some of the most advanced docking programs (ICM, Glide and GOLD) in docking 150 drug-like ligands to their respective receptors[31]. To measure the sampling performance of these three programs, the RMSDs between the closest of top 20 docking poses (nearest native) and the corresponding crystal structure for each complex were reported. Glide identified a docked pose, among the top 20, that was within 2.0 Å of the experimental structure in 79% of the cases, versus 77% by GOLD, and 67% by ICM. The corresponding performance of SKATE sampling coupled with FRED-Score ranking was 91%. This study was limited to docking only 100 PDB entries out of the 150 complexes because the rest were confidential corporate structures.

Systematic sampling in SKATE never repeatedly samples the same point in conformational space. In practice, two conformations can be clustered when their only difference is a 10 degree torsional variation in a terminal rotatable bond. To speed up sampling, SKATE implements heuristics to further reduce conformational space. As shown in Figure 3.1 the possible conformations of a ligand that is hydrogen-bonded to a receptor can be represented by a search tree. The edges, nodes, and leaves of the tree represent torsion values of rotatable bonds, aggregates and sterically allowed poses, respectively. SKATE traverses this tree using a depth-first search approach. Upon reaching a leaf of the tree by traversing down a branch from the root, a sterically allowed conformation is found. SKATE determines if it is necessary to travel down a branch of the tree by checking if the partial ligand constructed thus far is similar to a ligand pose that was already found from visiting previous tree branches. If the RMSD between atoms in a partial ligand and the corresponding atoms in a previously generated pose is less than 0.3 Å, then SKATE terminates the search of the current branch. If the search were to continue, the resulting poses would be very similar to the previously generated poses and would be discarded by clustering.

Discriminant analysis determines the range of torsions that are sterically allowed for a rotatable bond. The allowed range of torsions is discretized and converted into a list of torsions to be sampled. Not all conformers assembled from this list of values will be low in energy. Ligands in the receptor-bound state rarely adopt strained conformations where the torsions of rotatable bonds deviate significantly from the energy minima of the +gauche, -gauche and anti rotations. SKATE truncates conformational space by limiting allowed torsions to be within 30 degrees of +gauche, -gauche and anti torsions for rotatable bonds that (i) are not terminated by oxygen or sulfur atoms, and (ii) contain atoms that are bonded to fewer than four heavy atoms. For terminal aggregates of a ligand, only the three torsions that are nearest to +gauche, -gauche and anti values are sampled for rotatable bonds that meet the above criteria (i) and (ii). SKATE uses a combination of 180 geometric parameters to predict potential hydrogen-bonding interactions between a ligand acceptor and receptor donor, and vice versa. The parameters that represent the most common geometries are tried first. SKATE skips the remaining parameters if a pose is found. These heuristics that reduce the search space and speed up performance are optional and can be enabled or disabled by the user.

3.6.2 Analysis of failed sampling cases

SKATE was able to sample a pose that is within 2 Å RMSD of the reference structure for 98%, 95%, and 98% of the test cases in the Astex/CDCC, Surflex, and Vertex data sets, respectively. Two of the ligands in the 85 complexes Astex/CDCC set barely missed the 2 Å RMSD threshold; their RMSD values were 2.51 and 2.55. SKATE was unable to sample a pose that was within 2 Å RMSD of the native structure for test cases 1O86 and 1I8Z in the Vertex set. In 1O86, lisinopril, an anti-hypertension drug, is bound to the human angiotensin converting enzyme (ACE). There are 12 rotatable bonds in lisinopril. The ACE active site consists of a zinc coordinated narrow center flanked by two large hydrophobic pockets. Poses found by SKATE occupied either one of the pockets exclusively but were not able to bridge the two. To correctly dock lisinopril to ACE, a docking program must sample a pose where the carboxyl group of lisinopril correctly coordinates the zinc atom and still fits sterically into a very narrow channel. For test case 1I8Z, the ligand also coordinates a zinc atom, but SKATE failed to generate a pose that captured this interaction. For the Surflex set, SKATE failed to find near-native poses for 1FEN, 1HRI, 1LIC, and 3HVT. The 1FEN ligand does not have any hydrogen-bonding atoms and SKATE could not anchor its search since it could not form a hydrogen bond between the ligand and the receptor. For 1HRI, the ligand does not form a hydrogen bond with the receptor. SKATE sampled a pose (RMSD = 2.87 Å) where the ligand did form a hydrogen bond with the receptor but its orientation was inverted. Similarly, the 3HVT ligand does not form a hydrogen bond with its receptor and the best pose RMSD value was 2.33 Å. The 1LIC ligand is a simple alkyl chain molecule that has 15 rotatable bonds; it is a poor candidate for testing docking programs because it does not represent drug- or lead-like compounds and should not have been included in the Surflex test set. For the Astex/CDCC and the Vertex sets, SKATE sampled near-native poses (2.0 Å) for 98% of the test cases. Ligands in the Astex/CDCC were selected for unambiguous fitting to experimental electron density. Protons in the Vertex set were optimized to alleviate poor steric contacts. The likelihood of intermolecular penetration of VDW surfaces in these two test sets is lower because of high structural resolution in one case and proton optimization in another.

PDB structures are static models that best fit the available electron density data. Errors in lower resolution structures may result in poor modeling of small molecule ligands. This could lead to poor intermolecular steric contacts and even incorrect fitting of the electron density[36]. It is important to keep this in mind when assessing a docking program's ability to reproduce experimentally determined ligand poses.

3.6.3 Scoring accuracy

SKATE focuses on the systematic sampling of sterically allowed poses of a ligand where its search space is constrained by a binding pocket. It does not provide a scoring function to rank order the generated poses per se, but takes advantage of the many published scoring functions' ability to re-rank docked poses. In this paper, we presented data from using X-Score, Rosetta, and FRED energy functions to rank SKATE-generated poses. X-Score is an empirical scoring function that estimates the hydrophobic effect by using three different functions and averaging the results[22].

Rosetta's energy function was originally trained for protein-structure prediction and was extended to score protein-ligand interactions[10]. In this paper, Rosetta's energy function will be referred to as Rosetta-Score. FRED[20] itself is a docking program, but could also be used to rank previously generated poses with a consensus scoring function that consists of chemgauss3, PLP, and oechemscore. It will be referred to as FRED-Score.

We also evaluated whether rigid-body, local optimization of SKATE-generated poses would improve overall docking performance. SKATE allows some VDW penetration by scaling atomic VDW radii within the systematic sampling algorithm (see methods section for details). FRED's consensus scoring function could rank a near native pose poorly due to poor contacts. Prior to scoring, poses were optimized by performing a fast, small-scale, rigid-body translations (0.75 Å) or rotations (0.5 Å), a total of 72 systematic transformations, using FRED[20]. The optimized pose was selected by using the PLP scoring function. The receptor atoms were fixed throughout the optimization process. We emphasize that we only used the rigid-body, local optimization feature of FRED, not its full-fledged docking capabilities. The process of optimizing and scoring with FRED will be referred to as FRED-Opt-Score.

The results of using X-score, Rosetta-Score, FRED-Score, and FRED-Opt-Score to rank SKATE-generated poses for the Astex/CDCC test set are shown in Figure 3.6. For an RMSD threshold of 2.0 Å, the success rates were 87%, 85%, 73% and 66% for FRED-Opt-Score, FRED-Score, Rosetta-Score and X-Score, respectively. SKATE coupled with FRED-Opt-Score ranking performed particularly well in identifying poses that were less than 1 Å RMSD as the best pose for the Astex/CDCC set. Its accuracy rate was 72%. This is very encouraging because only 86% of the test cases had a pose that was less than 1 Å RMSD. Taking that into account, the scoring accuracy rate is 84% for ranking a pose that is within 1 Å RMSD from the native structure. This could partly be attributed to the high quality of the x-ray structures comprising the Astex/CDCC set.

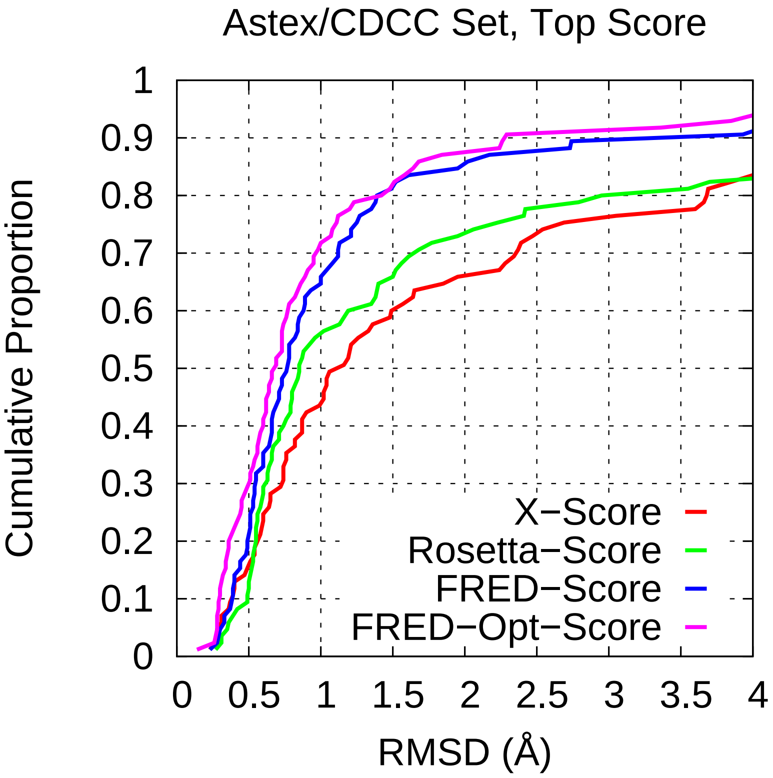


Figure 3.6: Cumulative proportion of top scoring RMSD for 85 complexes in the Astex/CDCC set

Similar to the results in the Astex/CDCC set, FRED-Opt-Score performed best in identifying poses that were within 2 Å RMSD as the best pose for the Surflex set. FRED-Opt-Score's accuracy rate was 84% (Figure 3.7). FRED-Score, X-Score and Rosetta-Score's accuracy rates were 75%, 64% and 52%, respectively.

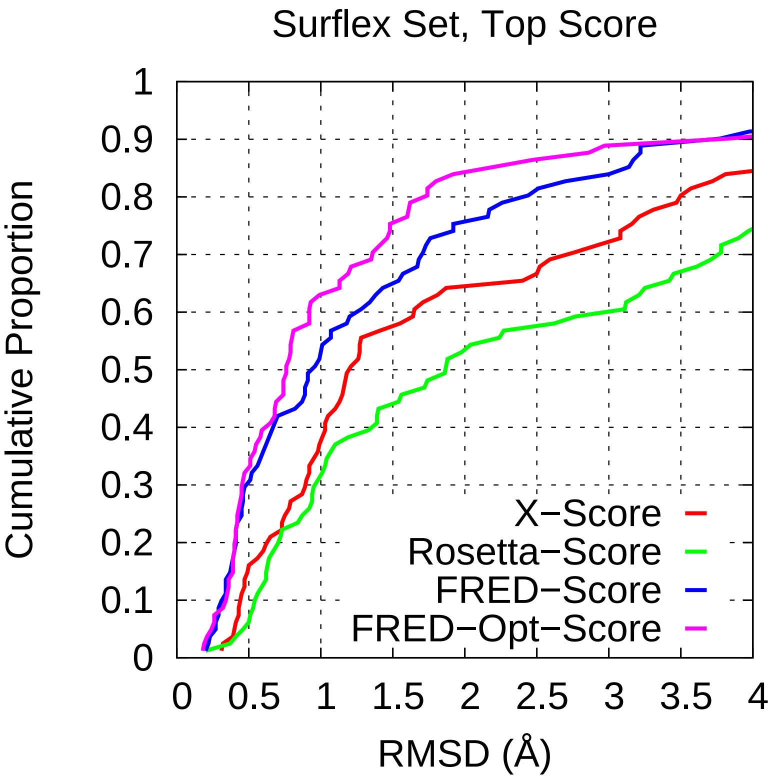


Figure 3.7: Cumulative proportion of top scoring RMSD for 81 complexes in the Surflex set

The results for the Vertex test set are shown in Figure 3.8. For an RMSD threshold of 2.0 Å, the success rates were 77%, 73%, 70% and 69% for FRED-Opt-Score, FRED-Score, Rosetta-Score and X-Score, respectively. For an RMSD threshold of 1 Å, the scoring accuracy of FRED-Opt-Score was 53% and of FRED-Score was 50%. FRED-Score and FRED-Opt-Score performance were comparable in identifying a pose that is within 1 Å RMSD as the best pose.

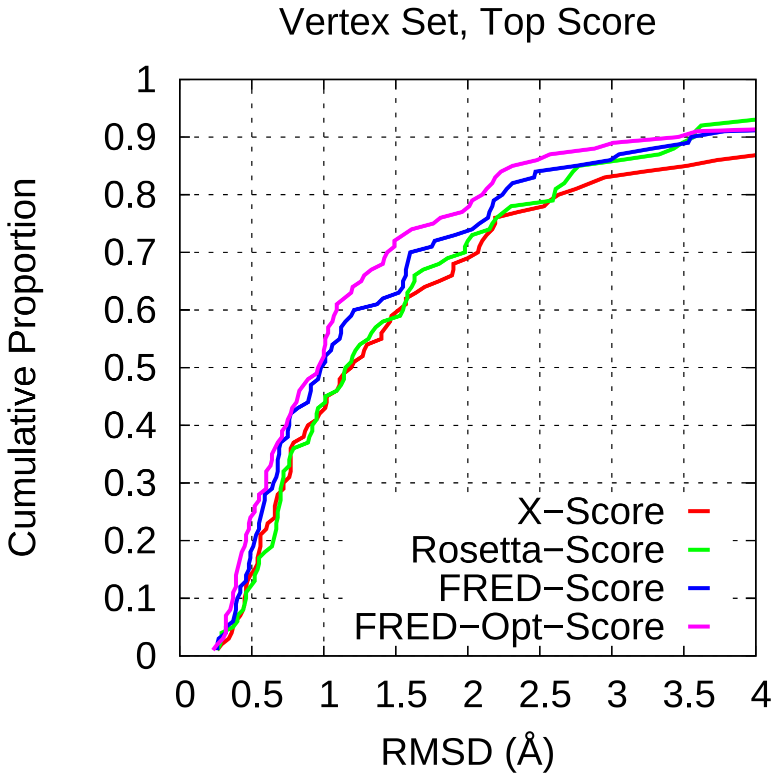


Figure 3.8: Cumulative proportion of top scoring RMSD for 100 complexes in the Vertex set

Existing literature that evaluates docking program performances usually focuses on overall docking results such as the fraction of correctly predicted protein- bound conformations[31, 37]. However, this kind of comparison is not conducive to pin-pointing the cause of the poor performance, i.e. whether it is attributable to poor sampling, inaccurate scoring or both, thereby making it difficult to isolate and fix problem areas. In this work, the same set of high quality SKATE- generated poses was ranked by FRED-Score, X-Score, and Rosetta-Score. Because the sampling and the scoring are separated, it allows for a fair comparison of the scoring function performances[38]. Although comparing scoring performance is not the main purpose of this work, it is still valuable to discuss the results. In all three self-docking test sets, FRED-Score was the most accurate scoring function (Figures 3.6, 3.7, 3.8). FRED-Score summed the individual ranks by chemgauss3, PLP, and oechemscore to produce a consensus rank. This rank- by-rank strategy was also employed by Wang et al. in a study evaluating consensus scoring functions[38]. They showed that combining results from three complementary scoring functions improved the recognition of near-native poses (2.0 Å) as best poses. Coincidentally or not, FRED-Score and one of the best consensus functions in Wang et al. both included the PLP scoring function. Rosetta-Score is an extension of Rosetta's energy function which was designed for in silico protein structure prediction. It may not have been optimally parametrized to score protein-ligand interactions. X-score was successful in ranking a pose that is within 2 Å of the experimental conformation in the range of 64% to 69% for the three test sets. This is consistent with a 66% success rate observed by Wang et al. in evaluating X-score on a 100 complexes test set[38].

Generally, rigid-body local optimization of SKATE-generated poses improved FRED scoring. At RMSD thresholds between 1 Å and 2 Å, optimization followed by FRED scoring improved accuracy by up to nine percentage points. Poses with RMSD values under 2 Å are often considered near-native but some may contain poor contacts that cause a scoring function to rank them poorly. A quick rigid-body local optimization or minimization of those poses alleviated those poor contacts and resulted in better scores. PLP was the scoring function used in the rigid-body optimization of SKATE poses. Despite its simplicity, PLP has been shown to be one of the top performing scoring functions and is incorporated in multiple docking programs[15, 23, 38]. Results from using oechemscore or chemgauss3 as the scoring function for optimization were similar to those from using PLP.

3.6.4 Examples of scoring errors

The best poses for the 1JJE and 1OQ5 complexes in the Astex/CDCC were 0.52 Å and 0.37 Å, respectively. However, the RMSD of the top-scoring pose, ranked by FRED-Opt-Score, for 1JJE was 7.97 Å and that for 1OQ5 was 5.00 Å. Upon closer inspection of the 1JJE poses, we found the shapes of the top scoring pose and the native pose were essentially superimposable. The middle parts of the two poses overlap very well but the two ring systems on the ligand were placed in opposite orientations in the top-scoring pose (Figure 3.9). Due to the symmetric nature of this ligand, this was a challenging case for scoring, because a small difference in 3D docked shape may be flipped to yield a large apparent RMSD.

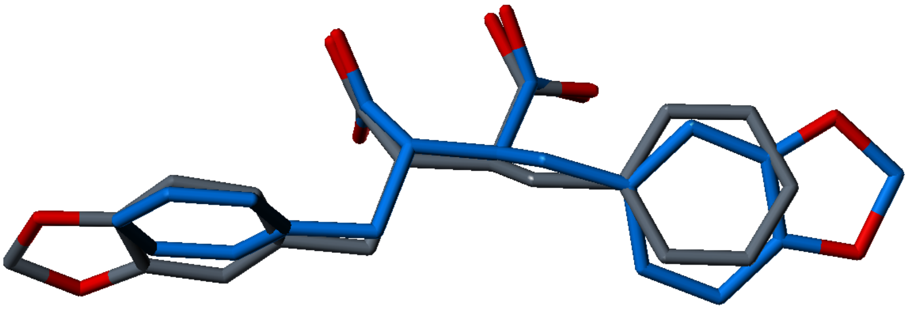


Figure 3.9: The RMSD between the 1JJE native pose (blue) and top-scoring pose (gray) of the ligand was 7.97 Å. The two ring systems of the top scoring pose were oriented in opposite directions.

FRED-Score, X-Score and Rosetta-Score also failed to rank a near-native pose as the top-scoring pose. 1OQ5 is another example where the shapes of the top-scoring pose overlapped well with the native pose (Figure 3.10). A phenyl group was swapped with a trichloromethyl group in the top-scoring pose. FRED-Score, X-Score and Rosetta-Score also failed to rank a near-native pose as the top-scoring pose.

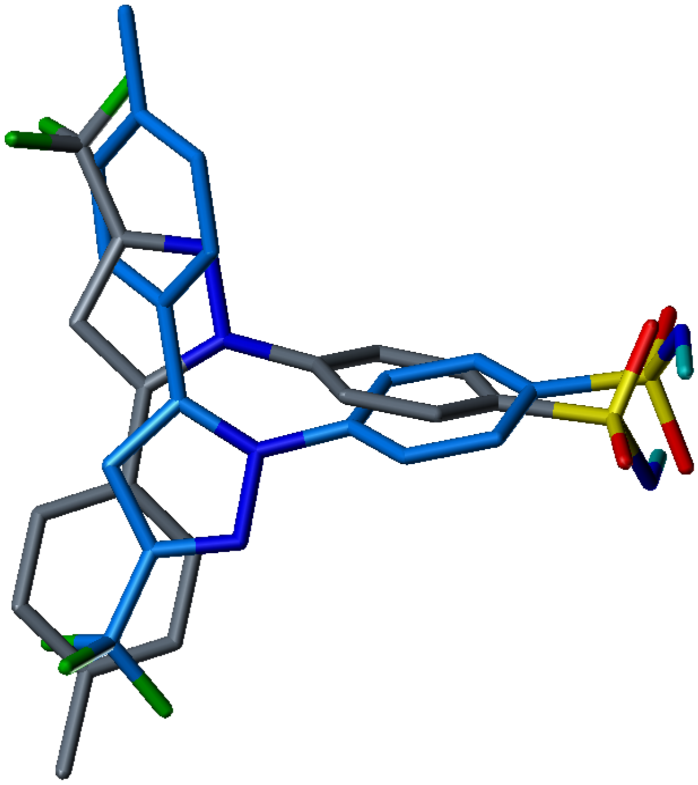


Figure 3.10: The RMSD between the 1OQ5 native pose (blue) and top-scoring pose (gray) was 5.00 Å. A phenyl group was swapped with a trichloromethyl group in the top-scoring pose.

3.6.5 Comparison with other docking programs

Perola et al.[31] prepared a test set of 150 protein-ligand complexes to compare the performances of Glide, GOLD and ICM. Of the 100 publicly available PDB structures, Glide correctly identified a docked pose that was within 2.0 Å RMSD of the experimental structure in 59% of the cases, versus 48% by GOLD (Figure 3.11). The success rate of ICM with this subset of 100 PDB structures was not available, but its success rate with the entire 150 complexes was 45%. Jain docked the same 100 PDB

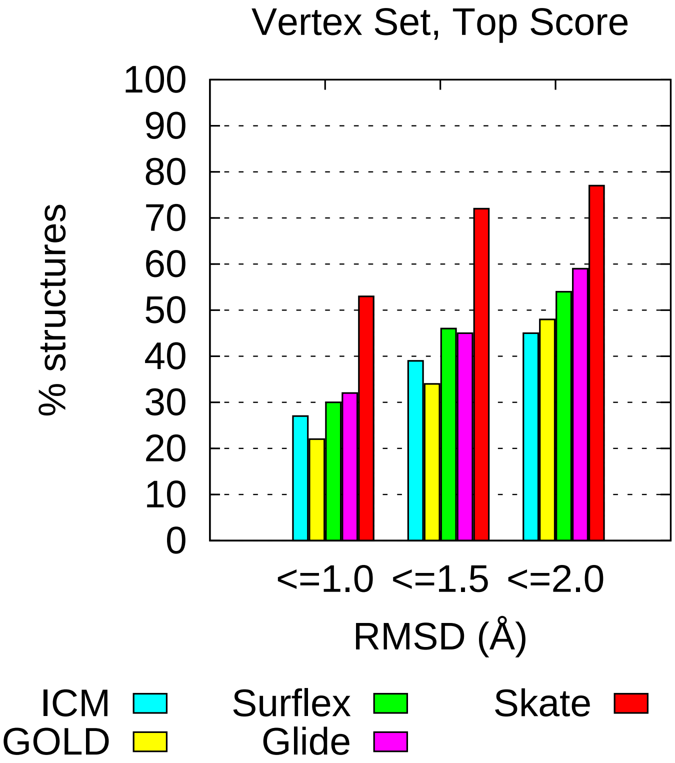


Figure 3.11: Distribution of the RMSD values between the top-ranked docking poses and the corresponding crystal structures in the Vertex set. RMSD values were calculated on the coordinates of the heavy atoms of the ligands. X-axis: RMSD cutoffs; Y-axis: percentage of top-ranked docking poses within a given RMSD cutoff from the crystallographic pose.

complexes using Surflex and its success rate was 54%[32]. SKATE's systematic sampling coupled with FRED-Opt-Score ranking was successful in identifying a pose that was within 2.0 Å RMSD of the native structure as the best pose for 77% of the cases. This represented a 18 percentage point improvement over Glide, the best performing docking program as tested by Perola et al. In this comparison, all docking programs used the same coordinates for the proteins and ligands. Perola et al. prepared the complexes by adding protons to both the bound ligand and the protein and optimized those proton coordinates while constraining the heavy atoms in their original positions. One reason for SKATE's improved results was improved sampling. SKATE sampled poses that were within 1.0 Å RMSD for 88% of the complexes. Ninety-six percent of the complexes had at least one pose that was within 1.5 Å RMSD of the native conformation (Figure 3.5). Perola et al. and Jain analyzed the sampling efficiency of ICM, Glide, GOLD and Surflex by calculating the RMSD values of the best pose among the top 20 results returned by the respective docking program. Their sampling accuracy rates were between 37% and 67% for a cutoff of 1.0 Å RMSD, and between 65% and 74% for a cutoff of 1.5 Å RMSD. Improved sampling by SKATE contributed to the overall higher success rates for the Vertex set.

Docking results for the Astex/CDCC set are available for RosettaLigand[39] and GOLD[5]. RosettaLigand is part of the Rosetta suite of programs. RosettaLigand modified Rosetta's energy function to guide its stochastic search and to rank the resulting poses. This is the same energy function as Rosetta-Score, one of the three energy functions that we used to rank SKATE-generated poses. The docking accuracy of RosettaLigand[39] was 58% for an RMSD threshold of 2.0 Å. SKATE coupled with Rosetta-Score achieved a higher success rate of 73% (Figure 3.6). For comparison, SKATE coupled with FRED-Opt-Score achieved an even higher success rate of

87% (Figure 3.12). It appears that a limiting factor in RosettaLigand's accuracy is

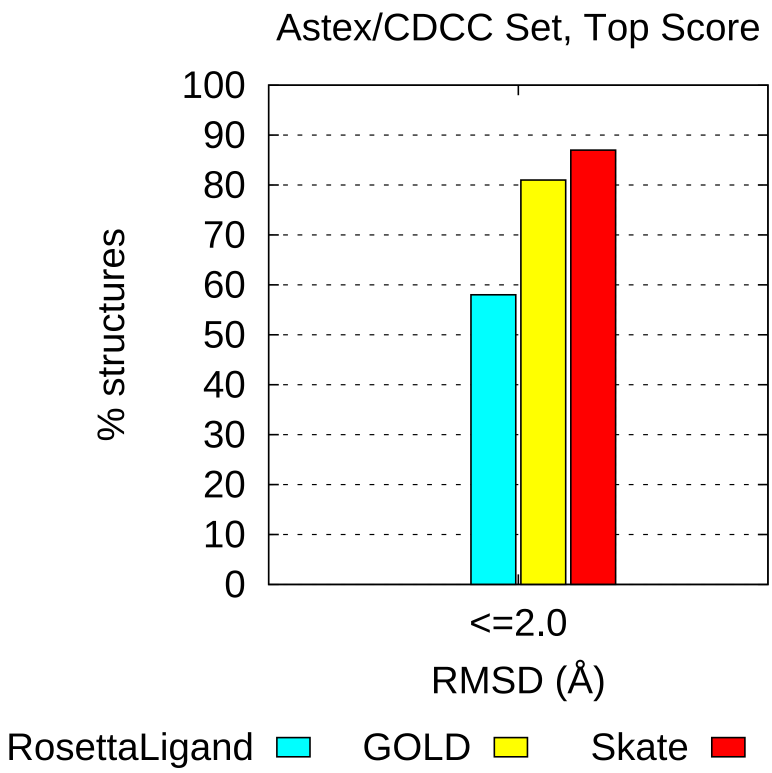


Figure 3.12: Distribution of the RMSD values between the top-ranked docking poses and the corresponding crystal structures in the Astex set. RMSD values were calculated on the coordinates of the heavy atoms of the ligands. X-axis: RMSD cutoffs; Y-axis: percentage of top-ranked docking poses within a given RMSD cutoff from the crystallographic pose.

its scoring function. Rosetta's energy function was optimized for in silico protein-structure prediction but was only recently extended to flexible ligand docking. Using an empirical scoring function that has been shown to work well in ligand docking might improve RosettaLigand's success rate. The success rate for GOLD[30] was 81% (Figure 3.12); which was 6% lower than SKATE.

Seventy-seven of 81 complexes in the Surflex set were docked by the authors of Glide[14]. Glide's success rate for this subset was 82% for an RMSD threshold of 2.0 Å. The same subset was also docked by the authors of MolDock with a resulting success rate of 87%. For comparison, the success rate of Surflex[13] was 77% and that of SKATE/FRED-Opt-Score was 84% for the entire set of 81 complexes (Figure 3.13). It is hard to directly compare the results of Glide, MolDock, Surflex, and SKATE for several reasons. First, Glide and MolDock's success rates are based on 77 complexes, a subset of the 81 complexes in Surflex. Both Surflex and SKATE's success rates

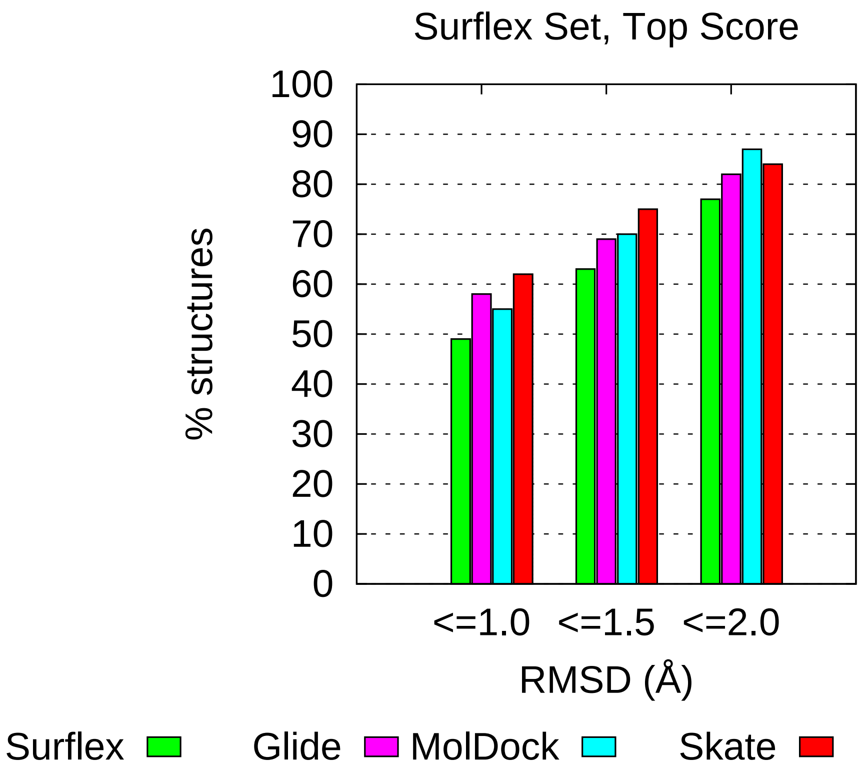


Figure 3.13: Distribution of the RMSD values between the top-ranked docking poses and the corresponding crystal structures in the Surflex set. RMSD values were calculated on the coordinates of the heavy atoms of the ligands. X-axis: RMSD cutoffs; Y-axis: percentage of top-ranked docking poses within a given RMSD cutoff from the crystallographic pose.

are based on the entire 81 complexes in the Surflex set. Second, MolDock basically trained its scoring function on this set of 77 complexes as pointed out by Hawkins et al.[36] Third, Glide calculated RMSD using optimized ligand coordinates instead of experimentally determined coordinates. Glide also used the optimized ligand and protein coordinates in its docking setup. The fact that the same energy function, OPLS/AA, was used in both complex optimization and pose scoring means Glide biased its methods by guaranteeing that the initial coordinates were at a local energy minimum per the OPLS/AA scoring function[32, 14, 16, 36].

Comparing results from different docking programs are not always straightforward[35, 36]. Results depend on, by varying degrees, protein preparation, initial structure of the ligand, docking site volume, and quality and composition of test sets. Generous sharing of protein and ligand files by test set authors has made it easier to do fair comparisons. In this work, we aimed for unbiased comparisons by using the same docking conditions as other docking programs whenever possible. The most visible improvement in docking accuracy is shown in the Vertex set results (Figure 3.11). SKATE results are 18 to 32 percentage points better for three different RMSD thresholds.

3.6.6 Cross docking

The thymidine kinase data set from the comparative paper of Bissantz et al.[33] and the cyclin dependent kinase 2 data set from Yang et al.[34] were used to test the cross-docking performance of SKATE. The TK set was originally used to quantitatively compare the performance of GOLD, DOCK, and FlexX. Data on this set are also available for Glide and Surflex. The TK structure used for docking was the deoxythymidine-bound structure (PDB code 1KIM). Table 3.2 summarizes the top-scoring RMSD values generated by the different docking programs for 10 thymidine kinase ligands.

Table 3.2: Accuracy in Cross Docking of Thymidine Kinase Inhibitors to the 1KIM site

RMSD (Å) of top-scoring pose[1]						
Ligand	SKATE[2]	Glide	DOCK	FlexX	GOLD	Surflex
dT	0.62	0.45	0.82	0.78	0.72	0.74
ahiu	0.67	0.54	1.16	0.88	1.63	0.87
mct	0.56	0.79	7.56	1.11	1.19	0.87
dhbt	1.18	0.68	2.02	3.65	0.93	0.96
idu	0.41	0.35	9.33	1.03	0.77	1.05
hmtt	3.32	2.83	9.62	13.30	2.33	1.78
hpt	4.07	1.58	1.02	4.18	0.49	1.90
acv	3.29	4.22	3.08	2.71	2.74	3.51
gcv	3.34	3.19	3.01	6.07	3.11	3.54
pcv	3.80	3.53	4.10	5.96	3.01	3.84

---

[1] Data for DOCK, FlexX and GOLD are taken from Bissantz et al.[33]; data for Surflex are taken from Jain[13]; data for Glide are taken from Friesner et al [14]

[2] FRED-Score was used to rank the poses generated by SKATE.

The ligand and receptor structures were prepared as described in Bissantz et al. FRED-Score was used to rank the poses generated by SKATE. Five of the 10 ligands were docked to the 1KIM structure with an RMSD of less than 1.2 Å. Another five failed to dock and their RMSD values were between 3 and 4 Å. Of the five failed cases, SKATE generated poses that were less than 2.0 Å RMSD for four ligands. However, neither FRED-Score nor FRED-Opt-Score ranked them as top-scoring poses. The RMSD values of the best pose for ligands hpt, hmtt, gcv, pcv and acv were 0.35 Å, 1.19 Å, 2.11 Å, 1.65 Å, and 1.37 Å, respectively. As pointed out by Friesner et al.[14], ligands acv, gcv and pcv are purine-based ligands and do not fit properly into the pyrimidine-based ligand site. All six docking programs did not sufficiently sample receptor flexibility and therefore failed to dock these three ligands. The cross-docking results by SKATE are comparable to Glide, GOLD, and Surflex.

The CDK2 test set consists of 73 complexes and the ligands were docked to a single CDK2 protein structure (PDB ID 2B54). The resolution of 2B54 is 1.85 Å and is co-crystallized with 6-(3,4-dihydroxybenzyl)-3-ethyl-1-(2,4,6-trichlorophenyl)-1H-prrazolo[3,4-d] pyrimidin-4(5H)-one. The 2B54 structure was selected to be the model receptor because it is the best-resolution structure with no missing residues or side-chain atoms. Two sets of VDW scaling parameters were tested in docking the 73 ligands to 2B54. The default VDW scaling value for intermolecular interactions is 0.9. A second set of parameters allows even more VDW penetration by using a scaling value of 0.8. Reducing VDW radii is a technique docking programs can employ to mimic receptor side-chain flexibility. Admittedly, this is a poor mimicry of receptor flexibility, but nevertheless useful until more advanced features are added to SKATE. To evaluate sampling and scoring accuracy, we used heavy atom RMSD from the native structure. To transform the reference coordinates into the same global coordinates, 72 of the 73 complexes were structurally aligned to 2B54 using pymol[40] and ligands were extracted and saved in the Tripos mol2 format. The sampling results from using the two different VDW scaling parameters are shown in Figure 3.14 (top). More permissive VDW parameters allow more VDW penetration; hence more receptor flexibility resulted in improved sampling. SKATE was able to sample a pose that was within 2 Å RMSD of the native structure for 81% of the ligands (Figure 3.14 top, dotted curve). However, this level of VDW scaling was not accommodated in FRED-Opt-Score. A low RMSD pose will score poorly if there are severe VDW penetrations. The percentage of top-scoring poses as a function of RMSD is shown in Figure 3.14 (bottom). At an RMSD cutoff of 2.0 Å RMSD, the success rate was 38%. Scaling atomic VDW radii by a factor of 0.8 improved sampling but a similar improvement was not achieved in scoring. The percentage of top-scoring ligand poses plotted as a function of RMSD threshold was similar for the two sets of VDW scaling parameters (Figure 3.14 bottom). In terms of overall docking accuracy, there was no significant advantage to using a VDW scaling value of 0.8. Thus, modifications to SKATE to include receptor flexibility are under consideration.

3.6.7 Pitfalls in complex preparation

The Vertex data set was prepared by performing a constrained minimization of the complexes using MacroModel and the OPLS/AA force field[41, 42]. Heavy atoms were constrained to their original position while hydrogen atoms were allowed to optimize. While this alleviated poor contacts between software-added hydrogen atoms, it could lead to artifacts where a hydrogen atom can bend out of plane to relieve steric interactions. Shown in Figure 3.15 is an example where an aromatic hydrogen was

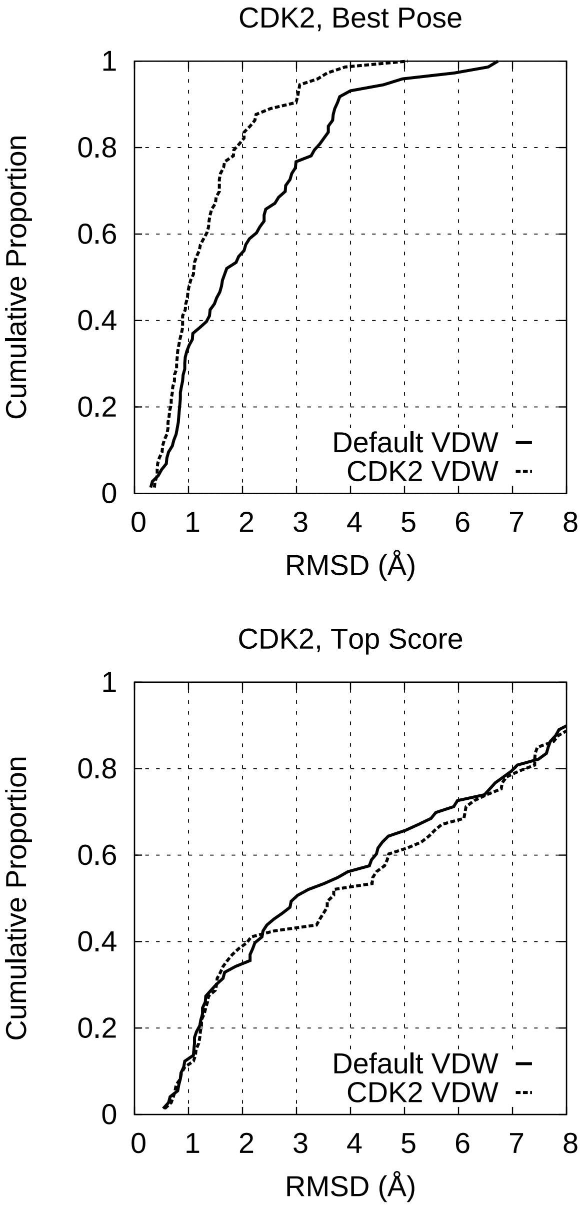


Figure 3.14: Top: Cumulative proportion of best RMSD poses for the CDK2 cross-docking set using two sets of VDW scaling parameters. FRED-Opt-Score was used to rank the poses. Bottom: Cumulative proportion of the RMSD between the top-ranked poses and the native structure. FRED-Opt-Score was used to rank the poses. The default intermolecular VDW scaling value was 0.9. The CDK2 VDW scaling value was 0.8.

bent 25 degrees out of plane during the minimization step. Hawkins et al. pointed out additional pitfalls in complex preparation and x-ray structure quality[36]. However,

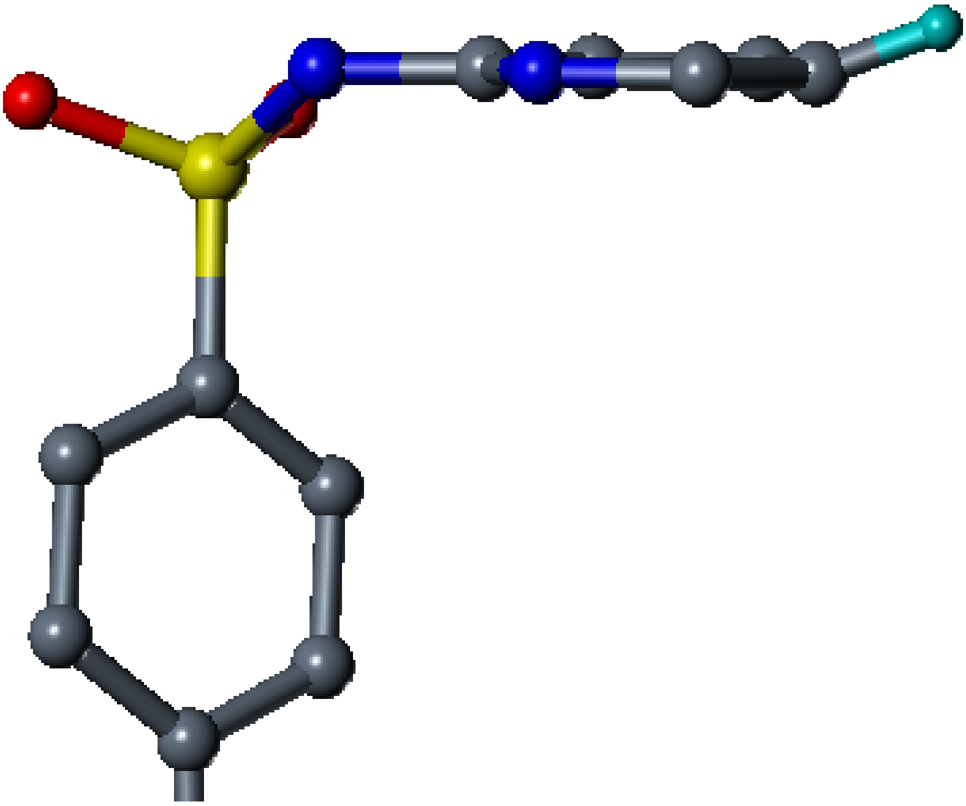


Figure 3.15: An aromatic proton in test case 13GS of the Vertex set was bent out of plane by an optimization step in complex preparation. Heavy atoms in a complex were fixed while protons were allowed to optimize. This proton (cyan) was bent out of plane by 25o to relieve steric overlap with a proton on residue Pro202 of the receptor.

not optimizing software-added hydrogen atoms also has its problems. A proton on the ligand penetrated the VDW surface of a proton atom on a lysine side-chain in complex 1YGC of the Astex/CDCC set. In this case, the poor contacts could have been alleviated by a quick minimization.

3.6.8 Computational time

Using SKATE, the average docking time per ligand-protein hydrogen-bond pair was less than 5 minutes for ligands with 6 or fewer rotatable bonds, and 10 minutes for ligands with 8 rotatable bonds. Total docking time was proportional to the number of possible hydrogen bonds that a ligand can form with the receptor. For the Astex/CDCC set, the median docking time was 42 minutes and the average docking time was 98 minutes on a single CPU (Pentium 4, 2.4 GHz) computer running Linux. SKATE allows simple parallelization by submitting each possible hydrogen-bond pairing to a different CPU in a computing cluster.

SKATE has not been optimized as it is still under development, but it is expected that with some optimization significant reduction in computational time could be achieved. Further speed improvement in SKATE can be made by implementing look-ahead technologies to further prune the combinatorial search tree [21, 27]. Knowledge about distance constraints between pharmacophore points can also be used to prune the search tree. Additional heuristics can be applied to reduce the number of discrete torsions sampled. Speed improvement will make SKATE more amenable to virtual screening applications of large compound libraries.

3.7 Conclusions

We implemented a novel docking concept in SKATE that decouples systematic sampling from scoring to improve overall docking accuracy. SKATE's systematic sampling coupled with FRED's optimization and scoring was more accurate in two large data sets and was equally accurate in a third data set when compared to GOLD, Glide, ICM, MolDock, RosettaLigand, and Surflex. Improved sampling by SKATE resulted in overall higher docking accuracy. Systematic sampling in SKATE was robust as tested by three large self-docking test sets and two cross-docking test sets. The high-quality poses generated by SKATE could be used to train scoring functions to distinguish between near-native and poorly docked poses.

The problem of false negatives is often the root cause of poor performance in docking programs. If a docking program never samples near-native poses, then there is zero chance that a scoring function can rank them as top-scoring poses. Unfortunately, modern docking programs' sampling methods are dependent on scoring functions that at best approximate experimental binding energies. The inter-dependence of sampling and scoring makes it difficult to determine whether a sampling error or a scoring error caused a program to fail in a docking experiment. SKATE breaks this dependence by systematically and exhaustively sampling sterically allowed poses of a ligand that are constrained by a receptor pocket. It is evident from this work that improved sampling contributed significantly to higher docking accuracy.

An executable version of SKATE and the five data sets are available for download from http://www.ccb.wustl.edu/˜jafeng.

3.8 Acknowledgment

I am indebted to Dr. Chris Ho for providing C libraries to read mol2 files and to build data structures. We would like to thank Robert Yang for sharing the CDK2 data set, Dr. Ajay Jain for sharing the Surflex set and Vertex set, and Dr. Marcel Verdonk for sharing the Astex/CDCC set.

Chapter 4

SKATE: Potential Improvements

4.1 Introduction

SKATE was written as a proof-of-concept program and has not been optimized for speed or usability. The following proposed improvements will make it much more user-friendly and more likely to be adopted by the drug discovery community.

4.2 Usability improvements

One of the limitations of SKATE is its dependence on SYBYL to generate a receptor file. This receptor file containes dummy atoms that represente volumes in solvent space that the ligand is prohibited from exploring. The receptor file generation step can be eliminated by representing the disallowed solvent space using grid points. A point in a grid could be marked as part of the protein, as part of the excluded solvent volume, or as part of the binding pocket. A grid point is marked as part of the protein if it is within 1 Å in Cartesian space of any protein atom. A grid point is marked as part of the binding pocket if it is within a user-specified distance of a bound ligand.

The most obvious instance of a bound ligand is the co-crystallized ligand. If a bound ligand is not available, the binding pocket can be defined as a box centered on an user specified point that represents the center of the binding pocket. The remaining points on the grid will be marked as inaccessible (solvent) to the ligand. During the systematic search process, solvent-grid points that are within an aggregate's search space will be checked using discriminant search. If an aggregate is in contact with solvent-grid points, but not protein-grid points, then it may be safe to terminate that branch of the search tree because the ligand is growing out of the binding site and into solvent space.

4.3 Performance improvements

Protons were added to the receptor crystal structures without optimizing potential hydrogen bonds with the ligand. Receptor O-H and N-H vectors could be pointing in a suboptimal or incorrect direction for forming a potential hydrogen bond with the ligand. It would then be necessary to rotate -OH and -NH groups in the receptor to increase the number of near-native poses generated by SKATE. A simple implementation is to make the X-OH and X-NH bonds rotatable where X is the heavy atom directly bonded to O or N. This can be done in SKATE by adding X as a part of the flexible ligand.

The key to making SKATE faster is to terminate a branch of the search tree as early as possible. Implementing look-ahead technology will improve performance by orders of magnitude [27]. The formalism on how to implement look-ahead is described here.

In Chapter 3, equation 3.10 (shown below) was derived from equation 3.2.

$$d_{ij}^2(\omega) = \frac{ax^2 + bx + c}{1 + x^2} \tag{4.1}$$

The two values of x that maximize or minimize $d_{ij}^2(\omega)$ are given by

$$-\left(\frac{d_2 \pm \sqrt{d_2^2 + d_3^2}}{d_3}\right) \tag{4.2}$$

where coefficients d_2, and d_3 are defined as follows (see Figure 3.3 and its subsequent equations for more details)

$$d_2 = -2(\hat{s} \cdot \hat{v}_2) \tag{4.3}$$

$$d_3 = -2(\hat{s} \cdot \hat{v}_3) \tag{4.4}$$

The two values of x may be substituted into equation 4.1 to find the maximum and minimum distances between atoms i and j. If these maximum and minimum distances lie outside the distance constraints for atoms i and j, then the current branch of the search should be terminated. This would further trim the search space.

4.4 Estimating Entropy

The Gibb's free energy of binding is the sum of enthalpic (ΔH) and entropic (ΔS) terms.

$$\Delta G = \Delta H - T\Delta S \tag{4.5}$$

Scoring function can do a good job of estimating the enthalpic energies of a binding pose. However, estimating entropic contributions to binding is still quite primitive. Assigning an energy penalty to each rotatable bond that is "frozen" in the bound state is a common and crude method of calculating entropic penalties. Each docked conformation of a ligand reported by SKATE represents a cluster of poses sharing the same energy minimum. The number of poses in a cluster could be used to estimate the width of that minimum-energy well. Clusters with a large number of members will be more favorable because the entropic penalty of binding is reduced. Clusters with only a few number of members will be less favorable because these poses freeze the ligand into limited conformations, which incurs a large entropic penalty.

4.5 Summary

With the above mentioned improvements, SKATE should contribute significantly to the drug discovery as one of the better docking programs in sampling near-native ligand poses.

Chapter 5

Stability of FSD-1

5.1 Introduction

Mini-proteins that contain fewer than 50 amino acids and fold independently of metal-binding centers or disulfide cross-linking sites are considered model structures for investigating the driving forces behind protein folding. These minimal model systems contain essential features of larger proteins: defined structures, important intramolecular contacts that stabilize the folded state and, in some instances, cooperative folding and unfolding. At the same time, their small size makes it feasible to study folding pathways and protein-energy landscapes with long-time scale, molecular-dynamics (MD) simulations [43] Mini-proteins often serve as benchmarks for validating novel methods in molecular simulations, such as replica-exchange molecular dynamics (REMD) [44, 45, 46]. Insights gained from studying mini-protein folding can be applied to protein-structure prediction, de novo protein design, and the discovery of novel biologics for treating diseases.

The zinc-finger motif consists of an N-terminal β-hairpin and a C-terminal α-helix with the tertiary structure stabilized by a zinc metal center coordinated by two cysteines and two histidines. Mini-proteins designed to fold into the zinc-finger $\beta\beta\alpha$ motif independent of zinc binding are especially interesting because their folded structures contain the helix, sheet, and turn secondary structures of the parent zinc finger. The Imperiali group iteratively designed the 23-residue BBA5 protein to adopt the $\beta\beta\alpha$ motif independent of zinc binding[47, 48]. A D-proline residue at position 4 was essential in stabilizing the β-hairpin in BBA5. The Mayo group used computational methods to design the 28-residue FSD-1 protein that also adopted the $\beta\beta\alpha$ motif independent of zinc binding[49] (Figure 5.1). They started with the backbone coordinates of the zinc-finger protein Zif268, and then selected side-chain rotamers to optimize side-chain/side-chain and backbone/side-chain interactions. The folding pathway, energy landscape, and stability of FSD-1 have been investigated by MD simulations in implicit and explicit solvent and by using improved sampling methods like replica-exchange molecular dynamics (REMD)[50, 51, 52, 53, 54, 55]. These subsequent studies of FSD-1 were conducted mostly because of FSD-1's small size, its sequence consisting of only natural amino acids, and the assumed accessibility of its thermal unfolding transition. However, FSD-1's apparent melting temperature of 42 ℃ and its reported NMR structure[49] have been assumed in previous studies without further experimental validation.

In this work, we present a critical analysis of FSD-1's stability by studying its thermal unfolding and solution structure by circular dichroism (CD), differential scanning calorimetry (DSC), and REMD. Thermodynamic properties such as melting temperature and enthalpy of unfolding were determined by analyzing changes in ellipticity and excess heat capacity, as function of temperature, that were measured by CD and

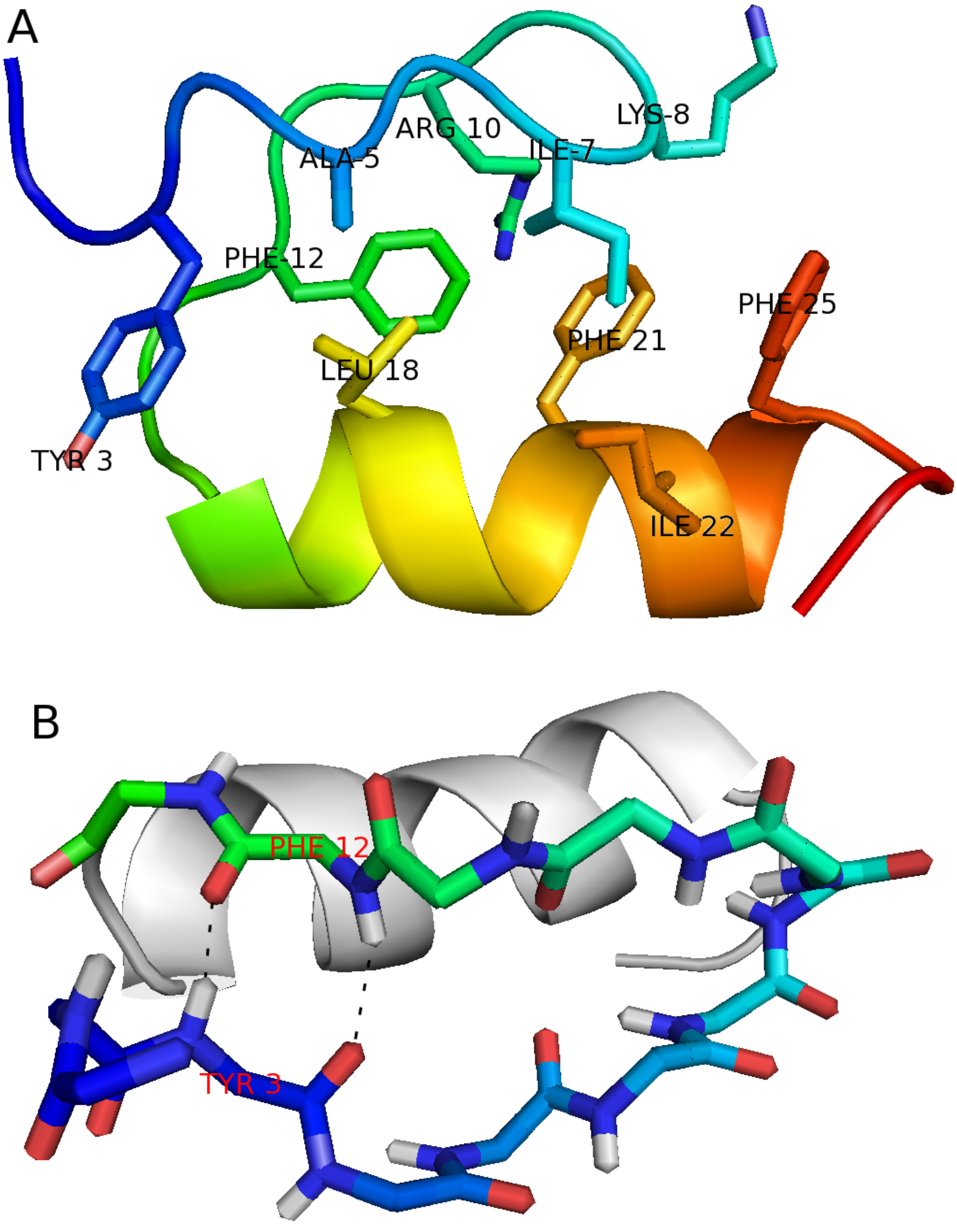


Figure 5.1: Structure of FSD-1 PDB: 1FSD). A: For clarity, side-chains of selected residues are shown. B: The main-chain atoms in the β-hairpin of FSD-1 are shown colors and the hydrogen bonds between Y_3 and F_{12} highlighted by black dashes.The α-helix is shown in light gray. Figures were generated using pymol[40]. Sequence: QQYTAKIKGRTFRNEKELRDFIEKFKGR.

DSC experiments. REMD simulations provided structural details that suggested possible explanations for the unusually broad melting transition of FSD-1. These results suggest an alternative interpretation; the apparent melting temperature is a reflection of a local helix-coil transition and not a protein-unfolding transition. Therefore, FSD-1 may not be a robust model system for studying protein folding.

5.2 Materials and methods

5.2.1 Peptide synthesis and purification

All reagents were obtained from commercial suppliers and used without further purification. FSD-1 was synthesized by solid-phase peptide synthesis using an automated microwave synthesizer, CEM Liberty (Matthews, NC). Fmoc amino acids were used. 2-chlorotrityl resin was preloaded with Fmoc-Arg. The Fmoc groups were deprotected by treatment of 20% piperidine in 0.1 M N-hydroxybenzotriazole (HOBT) in N,N-dimetylformamide (DMF) at 35 W at 75 ℃ for 30 seconds followed by a second treatment at 35 W at 75 ℃ for 3 min. Coupling was achieved with 5 equiv of Fmoc-amino acids, 5 equiv of 2-(1H-benzotriazole-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate (HBTU), and 10 equiv of diisopropylethylamine (DIPEA) at 35 W at 75 ℃ for 5 min. Arginine residues were first coupled at 0 W at 25 ℃ for 25 min then at 20 W at 75 ℃ for 5 min. A second coupling for arginine was performed at 35 W at 75 ℃ for 3 min. FSD-1 was cleaved from the resin by treatment with a mixture of 95% TFA, 2.5% H_2O, and 2.5% triisopropylsilane (TIS) for 2 h at room temperature. After filtration, TFA was removed by evaporation and the crude peptides precipitated with diethylether.

FSD-1 was purified by reversed-phase HLPC. Samples were prepared by dissolving the peptides in a 1:1 mixture of Solvent A (0.05% TFA in H_2O) and Solvent B (0.05% TFA, 10% H_2O in acetonitrile). The eluted samples were then monitored at 220 nm with a Gilson UV/VIS-155. A preparative VyDac C18 column (Cat# 218TP1022) was used with a linear gradient of Solvent A to Solvent B (5%-50% B) over 30 min with flow rate of 15 mL/min. The fraction containing the desired peptide was concentrated and re-purified to greater than 95% purity with a linear gradient of Solvent A to Solvent B (24.5%B to 25.5%B) over 30 min with a flow rate of 15 mL/min. Peptide purity was confirmed by NMR (Figure 5.2). Peptide identity was confirmed by electrospray mass spectrometry on a Waters Quattro micro (Milton, MA). The calculated average $[M+H]^+$ mass was 3489 Da and the observed mass was 3489 Da.

5.2.2 Circular dichroism

CD measurements were performed on a Jasco J-810 (Easton, MD) equipped with a Jasco PTC-424S Peltier temperature controller. Protein concentration was determined by UV-Vis absorbance at 280 nm using a calculated extinction coefficient of 1490 $M^{-1}cm^{-1}$. The protein concentration was 5 μM in 5 mM sodium phosphate buffer at pH 5.0[49]. Spectra were collected prior to thermal unfolding at 4 ℃ and after thermal unfolding at 80 ℃ in a 1 cm quartz cell, averaged over three scans from 260 to 190 nm with 2 s averaging, scanning speed of 20 nm/min and data pitch of 1 nm increments. For thermal unfolding, a thermometer was placed inside the sample cuvette and the sample was constantly stirred. Thermal unfolding was monitored at 218 nm, with averaging time of 15 s, temperature increments of 1 ℃, temperature slope of 30 ℃/h.

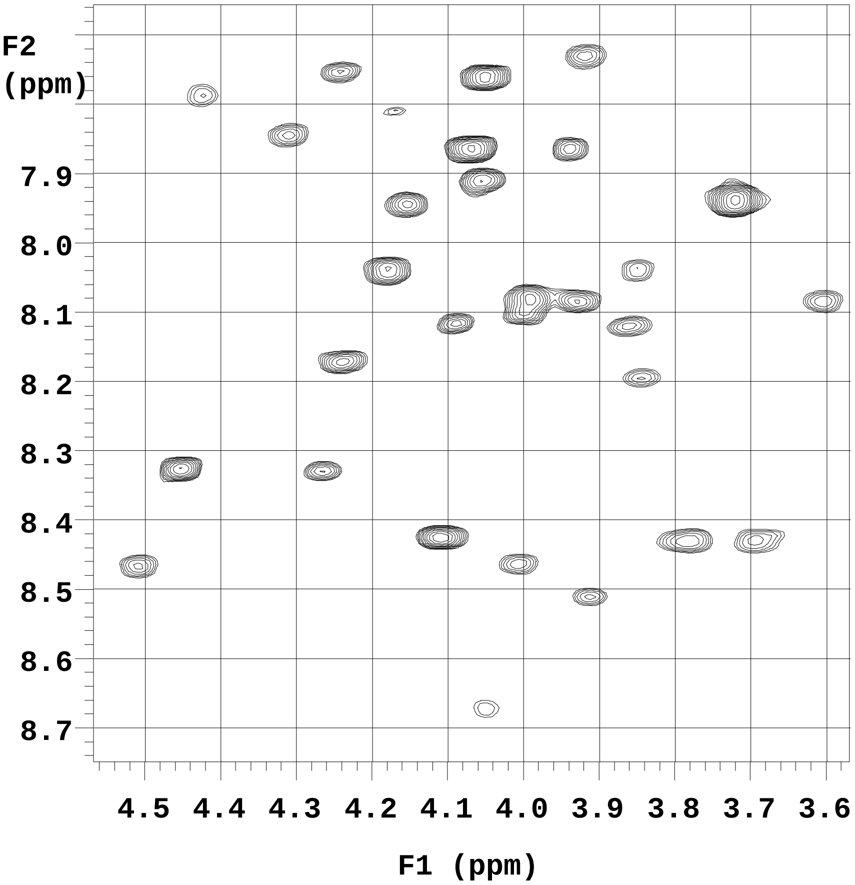


Figure 5.2: NMR spectra of FSD-1 showing the TOCSY Hα-NH fingerprint region.

5.2.3 Differential scanning calorimetry

Differential scanning calorimetry measurements were performed on a VP-DSC microcalorimeter from Microcal (Northamption, MA). Samples were degassed under vacuum for 10 min before they were used for calorimetric analysis. The start and final temperatures were 10 ℃ and 70 ℃, respectively, and the scan rate was 60 ℃/h. A 15 minute pre-scan equilibration was employed. The buffer was 50 mM sodium phosphate (pH 5.0), degassed. The sample cell was pressured to 25 PSI to prevent evaporation. A 0.5 mg/mL protein solution was prepared. Thirteen scans with the sample and buffer cell containing buffer were completed prior to the introduction of protein to the sample cell during a cooling cycle. Reheating runs were repeated to determine the calorimetric reversibility of the thermal-denaturation process. Data analysis was performed using Origin 7.0 and the DSC add-on provided by Microcal.

5.2.4 NMR

NMR spectra were recorded with a Varian Inova-600 (Varian Inc., Palo Alto, CA) spectrometer and the data were processed with VNMR software. NMR samples (˜2 mM) were prepared in H_2O-D_2O (90:10/v:v) with 50 mM sodium phosphate at pH 5.0 (uncorrected glass electrode). All spectra were collected at 7 ℃. The total correlation (TOCSY) spectra were recorded using an MELV-17 mixing sequence of 80 ms flanked by two 2 ms trim pulses. Phase-sensitive 2D spectra were obtained by employing the hypercomplex method. A total of 2 x 256 x 2048 data matrices with 16 scans per t_1 increment were collected. Gaussian and sine-bell apodization functions were used in weighting the t_2 and t_1 dimensions, respectively. After two-dimensional

Fourier transformation, the 2048 x 2048 frequency domain representation was phase and baseline corrected in both dimensions.

The NOESY spectrum resulted in a 2 x 256 x 2049 data matrix with 16 scans per t_1 increment. Spectra were recorded with 250 ms mixing time. The hypercomplex method was used to yield phase-sensitive spectra. The time domain data were zero filled to yield a 2048 x 2048 data matrix and was processed in a similar way as the 2D TOCSY spectrum described above.

5.2.5 Replica-exchange molecular dynamics

Replica exchange molecular dynamics simulations were performed using Gromacs 3.3.1[56]. An energy-minimized structure of FSD-1 (PDB code: 1FSV) was used as the starting structure for the simulations. The termini were charged and the net charge of the protein was plus 5. Five Cl^- ions were added in random locations to neutralize the system. The protein was solvated in a truncated dodecahedron box of TIP4P water where the minimum distance between a protein atom and the edge of the box was 12 Å. The system contained a total of 19,881 atoms. The OPLS-AA/L 2001 force field was used. The system was minimized until the maximum force was less than 100 kJ mol^{-1} nm^{-1}. Sixty-four temperatures were chosen with the average exchange rate of 20%. A one-nanosecond simulation was run to equilibrate the minimized system at each of the sixty-four temperatures. Each trajectory was assigned random initial velocities that were based on their respective temperatures. The NPT ensemble was used, the temperature was coupled to the Berendsen thermostat every 0.1 ps and the pressure was controlled by the Parrinello-Rahman method every 1.0 ps. REMD simulations are often performed at constant volume

(NVT), but constant pressure (NPT) was chosen to avoid extreme-pressure artifacts at higher temperatures[46]. A potential problem of using the NPT ensemble is improper solvation due to lower densities at high temperatures. The box volume at the highest temperature, 445.2 K, was 20% larger than the box volume at the lowest temperature, 226.2 K. This indicates that the protein was solvated in a liquid-like environment at high temperatures. Bond lengths between hydrogen atoms and heavy atoms were constrained with LINCS. Timestep was 2 fs. For each temperature, the temperature-equilibrated system served as the starting coordinates. The resulting 64 structures were used as initial structures in the REMD simulations with attempted exchanges every 1000 time steps (2 ps). Atomic coordinates were recorded every 2 ps for further analysis. 76 ns were simulated for each replica which resulted in 4.8 μs of simulation time. The simulations were run on Teragrid resources[57]. Data from the last 75 ns were used in analysis. To determine the distribution of target temperatures for the replicas, we followed a method described by Sanbonmatsu et al.[45] with minor modifications. The minimized system was equilibrated for 500 ps at temperatures 250, 300, 350, 400, 450, and 500 K. Their average potential energies, U, were calculated and fitted to a linear function (R^2: 0.99).

$$P(exchange) = exp\left[\left(\frac{1}{k_B T_1} - \frac{1}{k_B T_2}\right) \times (U_1 - U_2)\right] \tag{5.1}$$

Equation 5.1 was solved iteratively for the temperature distribution using $P(exchange)$ values of approximately 0.10. k_B was the Boltzmann constant; T_i and U_i were the temperature and potential energy of replica i, respectively. Initial test simulations with the resulting temperature distribution actually resulted in an average exchange rate of 0.20. $P(exchange)$ of 0.20 is recommended to produce ample exchanges between replicas. Temperatures below 273 K were chosen to help establish a baseline.

The resulting temperatures were 262.2, 264.4, 266.6, 268.8, 271.1, 273.4, 275.7, 278.0, 280.3, 282.7, 285.1, 287.5, 289.9, 292.3, 294.8, 297.3, 299.8, 302.3, 304.8, 307.4, 310.0, 312.6, 315.2, 317.9, 320.6, 323.3, 326.0, 328.8, 331.6, 334.4, 337.2, 340.0, 342.9, 345.8, 348.7, 351.6, 354.6, 357.6, 360.6, 363.6, 366.7, 369.8, 372.9, 376.1, 379.3, 382.5, 385.7, 389.0, 392.3, 395.6, 399.0, 402.4, 405.8, 409.2, 412.7, 416.2, 419.7, 423.2, 426.8, 430.4, 434.1, 437.8, 441.5, 445.2.

5.2.6 Curve fitting

Thermal-denaturation curves from CD melting and REMD melting were fit to a two-state model to determine T$_m$ and ΔH$_{vH}$. The following function was used in the fitting procedures[58, 59]:

$$
\begin{aligned}
f(T) &= \frac{y_f + m_f T + (y_u + m_u T) \times K}{1 + K} && (5.2) \\
K &= exp\left(\frac{h}{RT} \times \left(\frac{1}{T_m} - \frac{1}{T}\right)\right) && (5.3)
\end{aligned}
$$

where $f(T)$ was the observed signal, y_f was the folded-baseline intercept and m_f was the corresponding baseline slope, y_u was the unfolded-baseline intercept and m_u was the corresponding baseline slope, T was the temperature (K), h was the van't Hoff enthalpy, R was the gas constant (1.987) and T_m was the temperature (K) in of the transition point. DSC data were fitted to a two-state model using Origin 7.0.

5.3 Results and discussion

5.3.1 Circular dichroism

FSD-1 was synthesized by solid-phase peptide synthesis and purified to greater than 95% purity by reverse-phase HPLC. Its molecular weight $[M+H]^+$ of 3489 Da was confirmed by mass spectrometry. The thermal unfolding of FSD-1 was monitored by CD spectroscopy at 218 nm. The starting temperature was 4 ℃ and the final temperature was 80 ℃. The far-ultraviolet (UV) CD spectra of FSD-1 at 4 ℃, 80 ℃ and 4 ℃ after melting are shown in Figure 5.3A. The CD spectra at 4 ℃ before and after melting overlapped well, which confirms that FSD-1 unfolding is reversible as originally reported by Dahiyat and Mayo[49]. Two minima observed at 207 nm and 220 nm for spectra recorded at 4 ℃ indicated that FSD-1 contains a well-formed α-helical segment[60]. However, the CD spectra provided little information about the formation of a β-hairpin. The CD spectra of a FSD-1 double mutant (I7PK8D-proline) exhibited similar minima at 207 nm and 220 nm (Figure 5.3B), but the double mutant did not contain a stable β-hairpin as determined by NMR (data not shown).

The melting curve of FSD-1 measured at 218 nm was fit to a two-state model (Eq. 5.2) assuming a ΔCp value of zero[58, 59] (Figure 5.4). The melting temperature (T_m) and van't Hoff enthalpy (ΔH_{vH}) were determined from the least-square fit to be 41 ℃ and 18 kcal/mol, respectively. The T_m value reported by Dahiyat and Mayo was 42 ℃[49]. Thermal unfolding of FSD-1 was measured by CD. The mean residue ellipticity at 218 nm, [Θ]218, showed a broad transition with no clearly defined unfolded or folded baselines (Figure 5.4). Lack of a baseline for the fully folded state

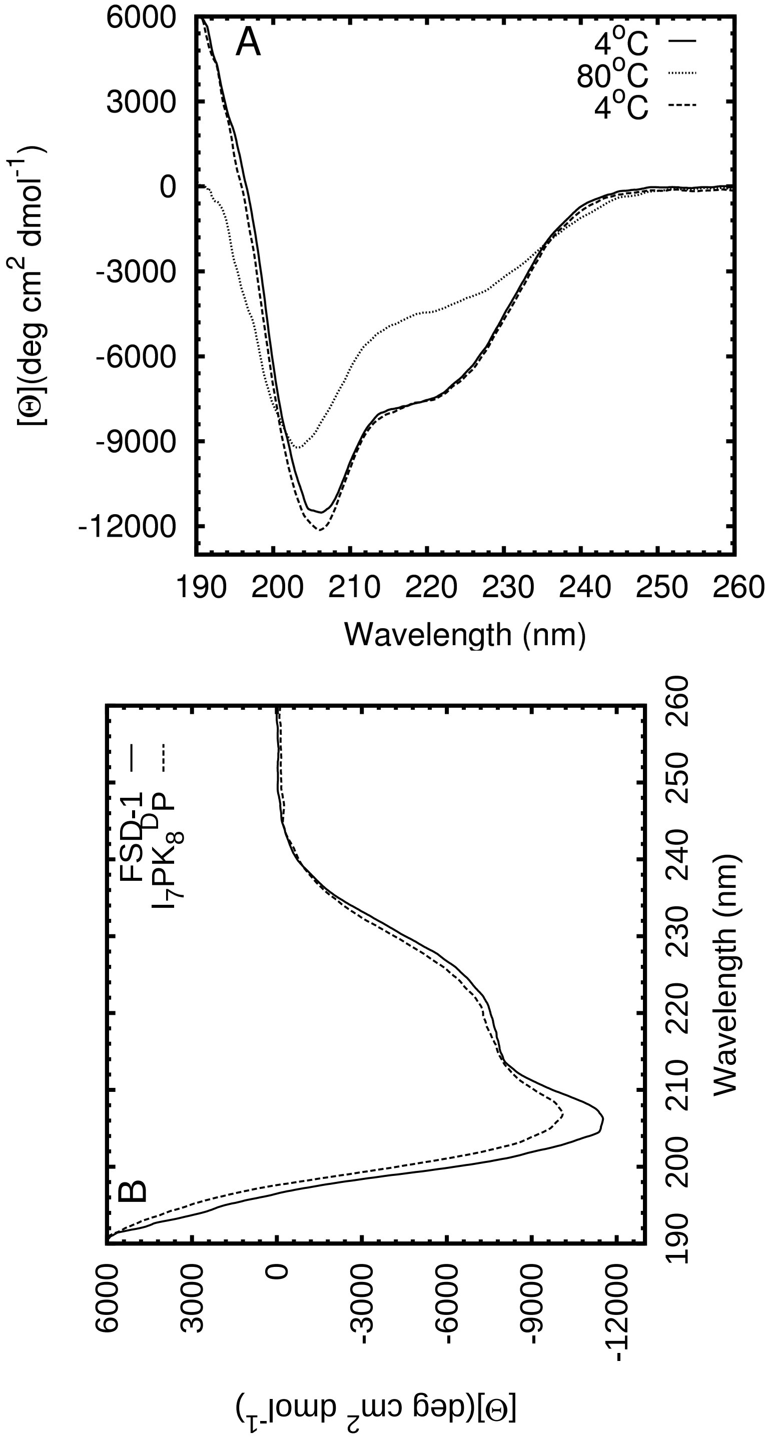


Figure 5.3: A: Far-UV CD spectra of FSD-1 at 4 ℃ and 80 ℃. Spectra were measured at 4 ℃ pre-melting (solid) and post-melting (dotted). B: Spectra of FSD-1 and an unfolded FSD-1 double mutant (I7PK8^{D}P) at 4 ℃. ^{D}P denotes D-Proline.

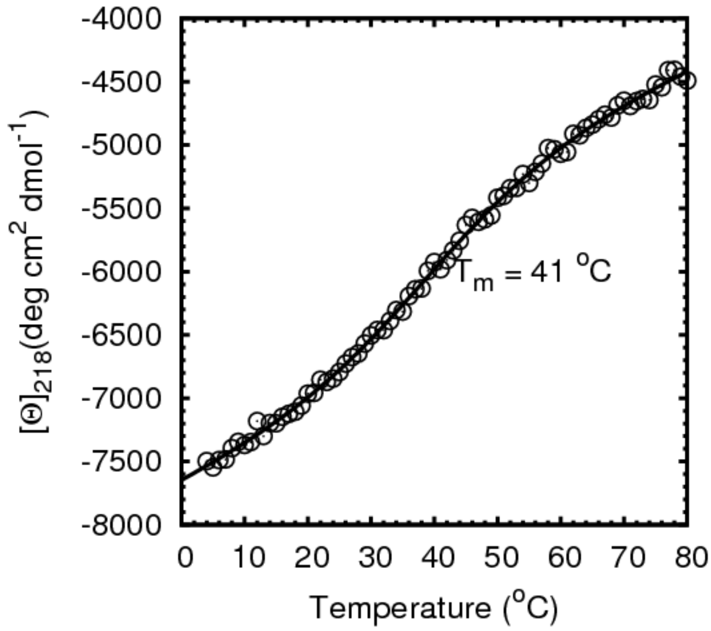


Figure 5.4: Thermal unfolding of FSD-1 monitored by CD at 218 nm. The melting curve was fitted to a two-state model and the resulting T_m was 41 ℃ and ΔH_{vH} was 18 kcal/mol.

indicated that FSD-1 was not well-folded even at 4 ℃. Since FSD-1 consists of only 28 residues, some flexibility was certainly expected, but a well-folded mini-protein should exhibit a better-defined baseline. For instance, the thermal unfolding of a 10-residue mini-protein (CLN025) designed, synthesized, and crystallized by Honda et al. showed a well-defined, folded-state baseline and a T_m of 70 ℃[61]. The broad transition of FSD-1 with undefined baselines was similar to helix unfolding[62, 63, 64]. The broad melting transition observed by CD could be the result of the helix-to-coil transition in the α-helical part of FSD-1, rather than the unfolding of its proposed hydrophobic core between the helix and the hairpin.

5.3.2 Differential scanning calorimetry

DSC showed a broad melting transition for FSD-1 and its unfolding was reversible (Figure 5.5). The broad transition made it difficult to determine the pre- and post-transition baselines necessary for a complete analysis of the calorimetric data. The unfolding of helical peptides also exhibits this behavior[62, 64, 65]. An initial baseline was estimated by drawing a line connecting the heat-capacity values, Cp, at the lowest and highest temperatures. The resulting excess heat-capacity curve obtained by subtracting the baseline was fit to a two-state model while assuming a Cp value of zero. The least-square fit was poor and resulted in high sum of squares-of-residual (SSR) values. To obtain better fits, the baseline was systematically lowered by increments of 25 cal*mol^{-}1 K^{-1}. The estimated Cp baseline resulting in the lowest SSR value for the two-state fit was taken as the best; excess heat capacity values obtained from subtracting this baseline are shown in Figure 5.5. Estimation of baseline by least-squares minimization was similar to that used by Scholtz et al. in determining the baseline for the thermal melting of a 50-residue α-helix[62]. For the two-state fit,

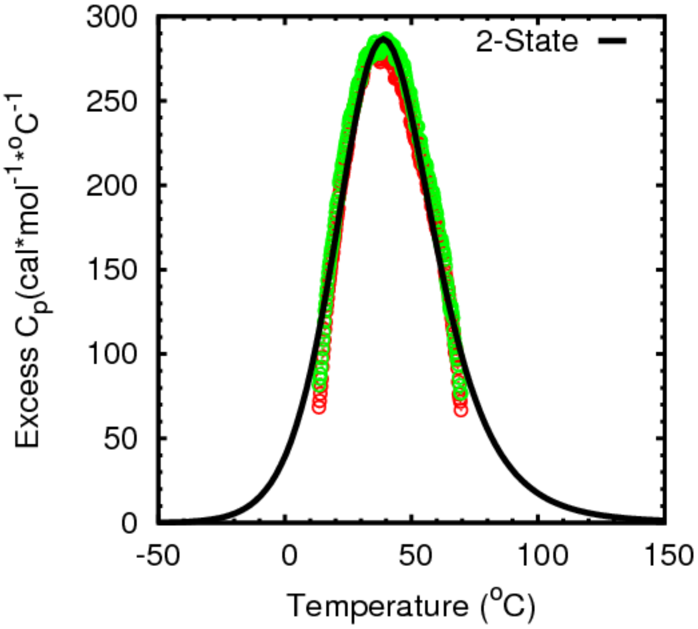


Figure 5.5: DSC melting curve fitted to a two-state model. T_m was determined to be 41 ℃ and ΔH_{cal} was determined to be 15 kcal/mol. Red and green circles represent two back-to-back DSC scans.

T_m was 41 ℃ and ΔH_{cal} was 15 kcal/mol. T_m values calculated from using different baselines were centered at 41 ℃ and varied by less than a degree. However, ΔH_{cal} values varied depending on which baseline was used in calculating excess Cp; the values ranged from 12 to 15 kcal/mol.

The van't Hoff enthalpy determined by CD was 18 kcal/mol, which was 3 kcal/mol higher than the calorimetric enthalpy determined by DSC. The near unity ratio of ΔH_{cal} to ΔH_{vH} suggests that FSD-1 unfolding approximated a two-state transition[66]. FSD-1 unfolding measured by DSC indicated that the unfolding transition began at near -20 ℃ and ended at over 100 ℃. This broad transition is nearly identical to the helix-to-coil transition measured for a 50-residue α-helical peptide, Ac-Y(AEAAKA)$_8$-NH$_2$, by Scholtz et al.[62]. They concluded that Ac-Y(AEAAKA)$_8$-NH$_2$ unfolding was far from being a two-state process because $\Delta H_{cal} \gg \Delta H_{vH}$[62]. GCN4brNC, a 29-residue α-helical peptide with covalently-closed N- and C-terminal loops also exhibited a broad folding-unfolding transition ranging from 5 ℃ to over 80 ℃[64]. The covalent loops between residues 1 and 5 at the N-terminus and between residues 25 and 29 at the C-terminus stabilized this helix. GCN4brNC unfolding was found to closely approximate a two-state transition[64]. Unfolding of the 35-residue subdomain of the villin headpiece was examined by Godoy-Ruiz et al[67]. Its unfolding transition was much narrower, ranging from 40 ℃ to 80 ℃. Its T_m was 65 ℃ and it was reported to fold via a two-state mechanism. The broad unfolding transition of FSD-1 is more like the unfolding of α-helical peptides than that of the more stable 35-residue villin headpiece subdomain.

5.3.3 Molecular dynamics simulations

In CD and DSC melting experiments, measurements such as mean residue ellipticity or excess heat capacity are plotted as a function of temperature to calculate T$_m$. REMD simulations are analogous to thermal unfolding experiments in that a protein is simulated over a range of temperatures and measurements are recorded at each temperature. An advantage of molecular simulations is that atomic details are recorded during the simulation. In REMD, a replica starts at one temperature and exchanges its temperature, based on a Metropolis criterion, with a neighboring replica that has a different temperature[44]. REMD-simulation temperatures were chosen so that potential-energy overlaps between replicas would be consistent across all temperatures and that there would be an optimal exchange rate near 20%[45]. Figure 5.6 shows that the energy overlaps were consistent throughout the temperature range for the last ten ns of the simulation, which was representative of the entire 76-ns simulation. The essence of REMD is high-sampling efficiency achieved by temperature exchanges

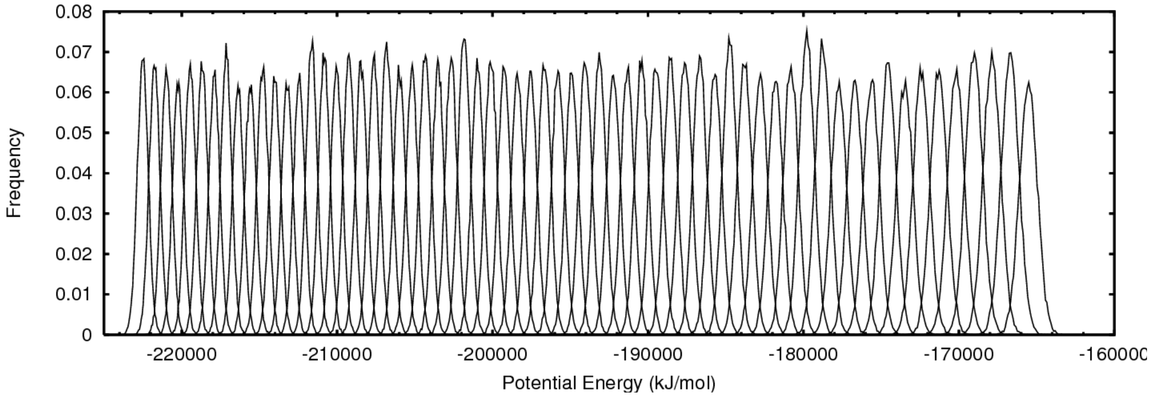


Figure 5.6: Potential-energy overlap between neighboring replicas during the last ten ns of the simulation. Each distribution curve represents the potential-energy distribution at a single temperature. The left-most curve represents the potential energy of the lowest-temperature replica, and the right-most curve represents the potential energy of the highest-temperature replica.

between neighboring replicas. As shown in Figure 5.7A, a wide range of temperatures were sampled by three representative replicas. This indicates the high-quality sampling of the simulation. For example, replica 1 started at 262.2 K. Through a series of temperature exchanges, its temperature reached 445.2 K, the maximum temperature of the simulation, at time point 14.426 ns. Replica 1 then continued to explore a wide range of temperatures throughout the simulation. Figure 5.7B shows backbone root-mean-square deviation (RMSD) between the native protein and trajectory snapshots of the corresponding replicas in Figure 5.7A. A folding event was observed in replica 62 that started at 437.8 K with an unfolded structure (RMSD > 8Å). In the first 40 ns of the simulation, the replica's temperature was limited to the upper half of the allowed temperature space and the protein stayed unfolded. A folding event occurred between 35 and 40 ns of the simulation, concurrent with replica 62 sampling much lower temperatures (Figure 5.7). Unfolding events were observed in replica 1 and replica 31. Replicas 1 and 31 were examples of protein unfolding that is analogous to thermal denaturation. At lower temperatures, the conformations sampled were similar to the native structure, whereas at higher temperatures, unfolded ensembles of conformations were sampled.

Structural properties for each REMD trajectory were analyzed as a function of temperature. Backbone root-mean-square deviation and C-alpha root-mean-square fluctuation (RMSF) were calculated to provide different measures of protein unfolding (Figure 5.8). RMSF is a measure of the average flexibility of an atom with respect to itself. High RMSF values indicate highly flexible atoms. Terminal residues 1, 2 and 26-28 were excluded from RMSF calculations because they were extremely flexible and distorted the overall flexibility of the entire protein. The RMSD and RMSF values were fit to a two-state model (Eq. 5.2) to predict the melting point

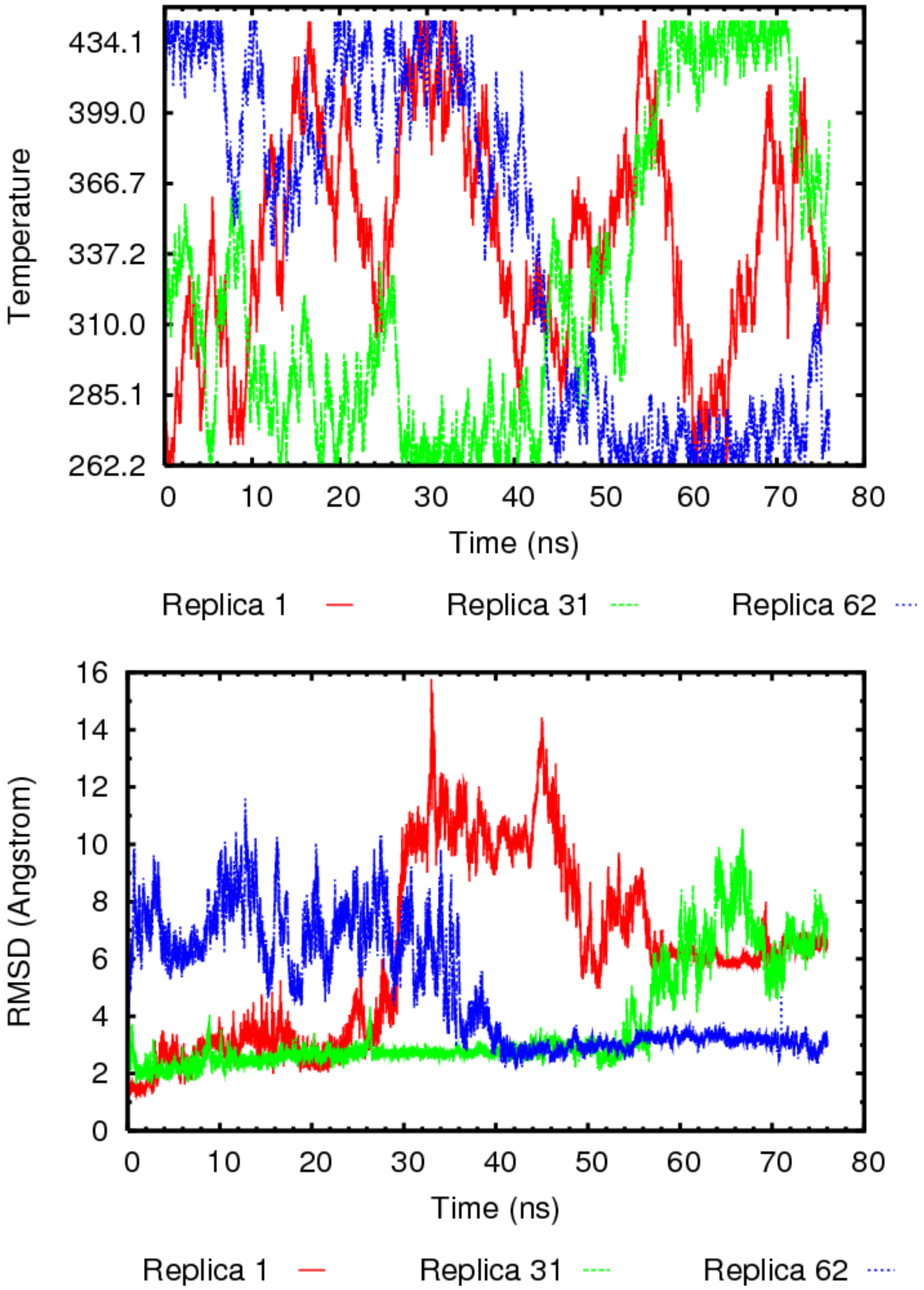


Figure 5.7: A (top): Temperatures sampled by 3 representative replicas, out of 64, during the course of the simulation. Replicas 1, 31 and 62 started at 262.2 K, 337.4 K, and 437.8 K, respectively. B (bottom): RMS deviation of three replicas during the course of the REMD simulation. A folding event is observed in replica 62, and unfolding events are observed in replicas 1 and 31.

of FSD-1. The average predicted T_m was 125 ℃, which was 84 ℃ higher than the experimental T_m - 41 ℃ - determined by CD and DSC. The average FSD-1 T_m predicted by Li et al. was 152 ℃, which was 111 ℃ higher than the experimental T_m[68]. Li et al. used the NVT ensemble instead of the NPT ensemble used in this study. The NVT ensemble tends to stabilize the hydrophobic core at high temperatures[55]. High T_m values were the results of the simulations over-stabilizing proteins at high temperatures[69]. This may be because the force-filed parameters used were originally fit to room-temperature experimental values.

5.3.4 Structural analysis

Experimental and simulation melting curves showed that FSD-1 exhibited a broad unfolding transition. There were difficulties in establishing a baseline for the folded state of FSD-1 in both simulation and experimental melting curves. In the REMD simulations, there were difficulties even though the lowest temperature was chosen to be -11 ℃ to help establish a baseline for the folded state. These results suggest FSD-1 is only nominally stable even at -11 ℃, which is in agreement with DSC results. To further investigate, we examined hydrogen-bonding patterns and native contacts observed in the ensemble of FSD-1 NMR structures reported by Dahiyat and Mayo[49] (PDB ID: 1FSD) and in the ensemble of trajectory snapshots from the REMD simulation. In the NMR structures, residues 2 to 13 formed a β-hairpin and residues 5 to 10 formed an EbaaagbE reverse turn[70]. The two β-stands in the hairpin were connected by this six-residue loop instead of the more common four residues in a traditional reverse turn. Two hydrogen bonds were formed between the amide and carbonyl groups of Y_3 and F_{12} at 7 ℃ as determined by NMR nuclear Overhauser effects (NOE)[49] (Figure 5.1B). No other hydrogen bonds were observed between

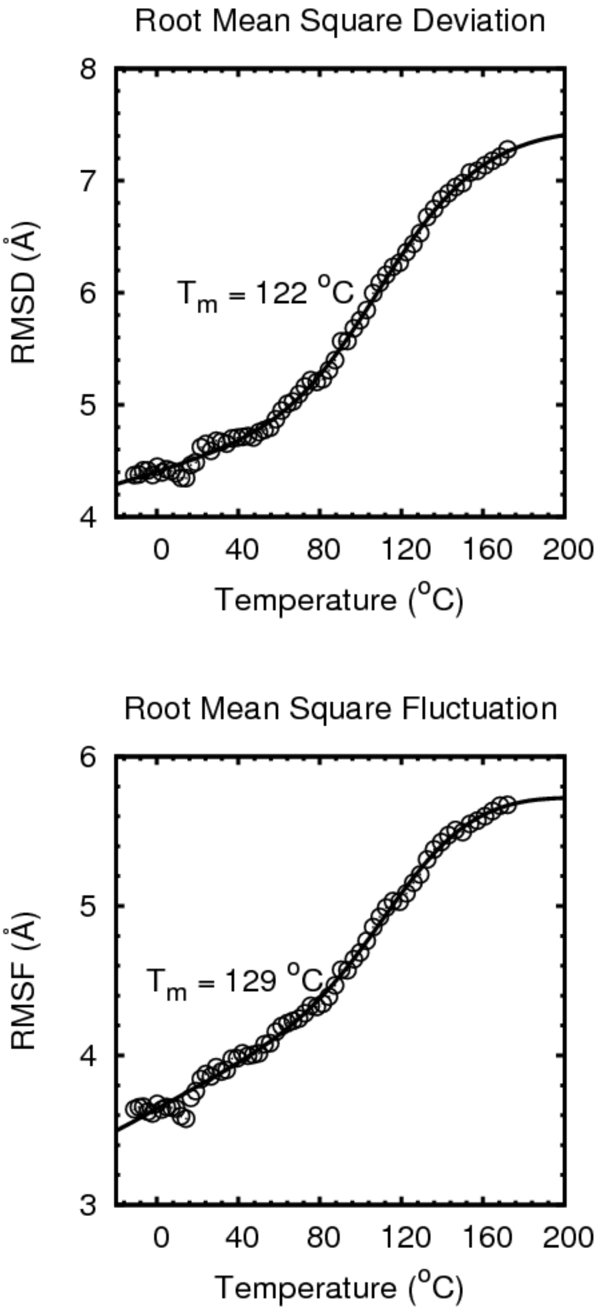


Figure 5.8: Thermal unfolding monitored by RMSD and RMSF. The data were fit to a two-state model. RMSF values were calculated for residues 3 to 25. The fitted melting temperatures T_m are shown in the panels.

main-chain atoms of the hairpin residues. The fact that only two hydrogen bonds were observed in a hairpin of 12 residues indicates that the β-hairpin was minimally stable. For comparison, four inter-strand hydrogen bonds were observed between residues in the 8-residue β-hairpin of BH17[71], a 17-residue synthetic mini-protein containing independent helical and β-hairpin domains. A D-proline residue at position 13 nucleated the β-hairpin of BH17. For all temperatures of the REMD simulation, the highest average number of hydrogen bonds formed between the main-chain atoms of Y_3 and F_{12} was 0.7 (Figure 5.9). This suggested that the limited β-hairpin of FSD-1 was formed only 35% (0.7/2.0) of the time even at low temperatures. The average number of hydrogen bonds formed between main-chain atoms of hairpin residues in strands one (residues 2-6) and two (residues 9-13) were also plotted in Figure 5.9. At most 1.5 hydrogen bonds were formed. The expected number of inter-strand hydrogen

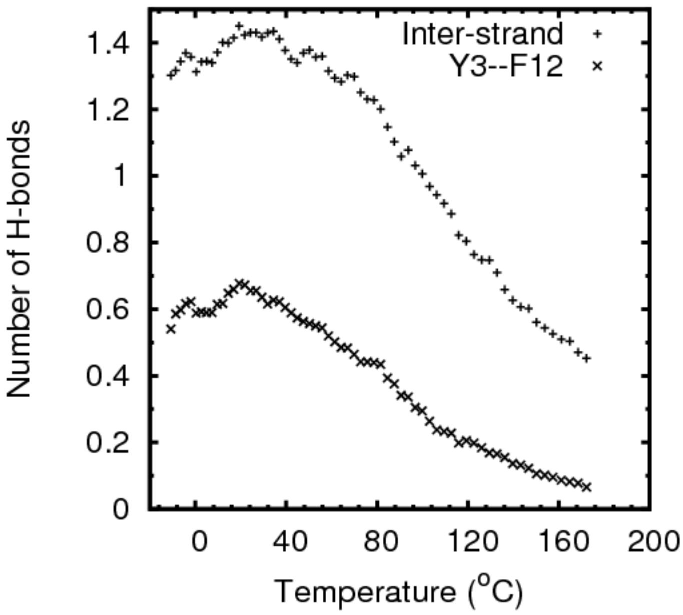


Figure 5.9: Average number of hydrogen bonds formed between main-chain atoms in residues Y_3 and F_{12} (x) or between residues in strand 1 (residues 2-6) and strand 2 (residues 9-13) of the hairpin (+) during the REMD simulation.

bonds for this β-hairpin of 12 residues would be six. Lack of detectable inter-strand hydrogen bonds in FSD-1's β-hairpin contributed significantly to a hypothesis of its overall instability.

The overall fold and topology of FSD-1 as designed depend mostly on the small hydrophobic core formed by residues Y_3, A_5, I_7, K_8, R_{10} and F_{12} in the β-hairpin and residues L_{18}, F_{21}, I_{22} and F_{25} in the α-helix (Figure 5.1A). Two residues were defined as in contact if their side chains have any two heavy atoms that are within 6.0 Å. Using this criteria, 185 contacts were found between residues in the α-helix (18,21,22,25) and those in the β-hairpin (3,5,7,8,10,12) in the ensemble of 41 NMR structures of FSD-1 reported by Dahiyat and Mayo[49]. Contacts within the β-hairpin or α-helix were not considered. The maximum percentage of native hydrophobic-core contacts seen by REMD was 58% at -11 ℃. Li et al. calculated the protein and β-hairpin native contacts for FSD-1 in their REMD simulation[68] to be 60% for the entire protein, and 45% for the β-hairpin. Hydrogen-bond and native-contact information from both Li's simulations and the REMD reported herein both suggest that FSD-1 at low temperatures was very flexible and adopted multiple conformations.

In MD simulations of FSD-1, it was found to be marginally stable at room temperature[50, 53]. The plasticity of the β-hairpin, especially reverse-turn residues 7, 8 and 9 were believed to contribute to the instability of FSD-1[50, 54]. Li et al. observed from their REMD simulations that the C-terminal α-helix was more stable than the β-hairpin by 33 ℃[68]. For the α-helix, the C-terminal helical turn consisting of residues E_{23}, K_{24}, F_{25}, and K_{26} was folded less than 10% of the time in their simulations. In five 200-ns simulations of FSD-1 at 300 K, Lei et al. noted that the N-terminal β-strand (Y_3TAK) were mostly helical instead of forming the native β-strand[53]. They concluded that this was probably due to the high helical propensity of A_5 and K_6 according to the Chou-Fasman scale[72].

5.3.5 Alternative interpretation

Results from CD, DSC, and REMD experiments showed that FSD-1 was only minimally stable even at low temperatures. The FSD-1 thermal-unfolding curves measured by CD and DSC lacked baselines for the folded state, suggesting that FSD-1 adopts multiple conformations. Molecular dynamics simulations provided further evidence of the plasticity of the β-hairpin. The C-terminal residues (26-28) were also very flexible. The changes in ellipticity in the CD unfolding experiment and in heat capacity in the DSC unfolding experiment at low temperatures were likely caused by different dynamics of the β-hairpin and C-terminal residues. The broad melting transition observed by CD and DSC was probably the result of the helix-to-coil transition in the α-helical part of FSD-1, rather than the unfolding of its limited hydrophobic core. The melting temperature of FSD-1 was determined to be 41 ℃ by CD and DSC. Given the minimal stability of the β-hairping, it is quite plausible that the T_m reflects mostly the melting of FSD-1's α-helix, rather than melting of the entire protein. Burial of hydrophobic groups of FSD-1's amphiphilic α-helix by residues in the hairpin region, regardless of whether a hairpin was formed, likely shifted the helix's T_m to 41 ℃. Additional helix stability was gained by the presence of nine charged residues on the hydrophilic side of the 14-residue helical segment.

5.4 Conclusions

We have presented a critical analysis of FSD-1 stability by studying its thermal unfolding and structure by CD, DSC, and REMD. Thermal unfolding experiments and molecular dynamics simulations showed that the unfolding transition started at temperatures much lower than 7 ℃. The plasticity of the β-hairpin contributed significantly to the observed changes in ellipticity in the CD experiment and changes in heat capacity in the DSC experiment. We propose that the apparent melting temperature of FSD-1 – 41 ℃ – primarily reflects the melting of FSD-1's α-helix, not the entire protein. While its small size makes FSD-1 an attractive target for studying protein folding, these results question FSD-1's status as a robust model system of a folded mini-protein.

5.5 Acknowledgments

I want to thank Dr. Alexander Kozlov for assistance in performing DSC experiments and Dr. Jeff Kao for collecting the NMR spectroscopy data. Computing resources for REMD simulations were supported in part by the National Science Foundation through Teragrid resources provided by the Texas Advanced Computing Center at the University of Texas–Austin.

Chapter 6

Protein Design

Protein stability can be enhanced by the incorporation of non-natural amino acids and semi-rigid peptidomimetics to lower the entropic penalty upon protein folding through preorganization. An example is the incorporation of aminoisobutyric acid (Aib, α-methylalanine) into proteins to restrict the ϕ and ψ backbone angles adjacent to Aib to those associated with helix formation. Reverse-turn analogs were introduced into the sequences of HIV protease and ribonuclease A, which enhanced their stability and retained their native enzymatic activity. Therapeutic proteins could be engineered to contain peptidomimetics that survive longer in vivo or retain activity after oral administration. Different reverse-turn analogs and their ability to nucleate β-hairpins are discussed in this chapter.

6.1 Preorganization

Designing a protein sequence that folds into a designed three-dimensional shape is known as the inverse protein-folding problem. In nature, protein sequences are limited to combinations of the naturally occurring 20 amino acids and their post-translational modifications. The incorporation of non-natural amino acids and semi-rigid peptidomimetics provides unique possibilities for designing proteins that adopt a stable predetermined fold, allowing protein engineering to become a reality. Limiting segmental dynamics may be a useful probe of enzyme mechanism and/or specificity and also be of commercial interest in the production of super stable biocatalysts for green chemistry. For example, multiple tons of the proteolytic enzyme subtilisin, engineered to be stable in detergents at alkaline pH and elevated temperatures, are consumed annually in laundry detergents[73].

As the simplest example of preorganization, incorporation of aminoisobutyric acid into proteins restricts the ϕ and ψ backbone angles adjacent to Aib to angles associated with helix formation[74, 75]. It is believed that Aib lowers the entropic penalty of helix formation upon protein folding due to preorganization. By the same principle, incorporating semi-rigid mimetics of α-helices, β-sheets, and reverse turns into a protein would minimize the entropy lost on folding through preorganization, while retaining the interactive surface features that optimize the favorable enthalpic interactions in the folded state. The first examples include the incorporation of reverse-turn analogs into the enzymes HIV protease and ribonuclease A. Chimeric proteins should be thermodynamically more stable because their fold space is limited by semi-rigid mimetics that reduce the entropic penalty upon folding into the desired 3D structure. In addition, semi-rigid mimetics should promote the rate of protein folding by nucleation. Modular secondary structure mimetics can serve as building blocks in the design of ultra-stable, catalytically active chimeric proteins that resist proteolytic degradation and denaturation.

6.2 Reverse-turn mimetics

Reverse-turn mimetics are designed to replace residues (i+1) and (i+2) of a turn with a semi-rigid analog that stabilizes the turn without negatively altering the geometry of the corresponding β-hairpin (Figure 6.1).

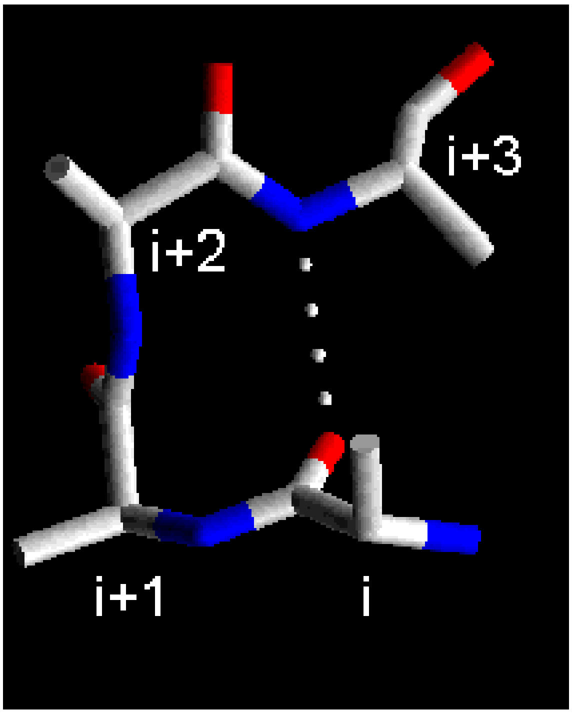


Figure 6.1: Type I reverse turn. Image reproduced from www.swissmodel.expasy.org

6.2.1 BTD

The impact of preorganization on energetics was first illustrated in a chimeric protein that incorporated an unusual bicyclic dipeptide reverse-turn analog (bicyclic turn dipeptide: BTD) into HIV protease. It showed an anticipated increase in fold stability, but of limited amount[76]. The BTD HIV-1 protease was fully active, specific for native ligands, and more resistant to thermal inactivation.

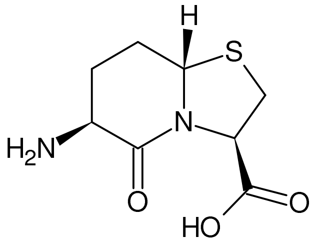


Figure 6.2: Bicyclic turn dipeptide (BTD), a β-turn analog

6.2.2 D-Pro-Gly

The dipeptide D-Proline-L-Glycine (D-Pro-Gly) was found to be superior to L-Asn-Gly for β-hairpin nucleation [77]. A 20-residue, mini-protein containing two D-Pro-Gly dipeptides was found to form a three-strand β-sheet [78]. Replacing one or both of the D-Pro residues with L-Pro resulted in the lost of long-range NOEs that were indicative of sheet formation. In another example of preorganization, Imperiali and co-workers used a D-Pro-Gly dipeptide to induce a type II' β-turn centered about residues 4 and 5 of their BBA series of mini-proteins[48, 47]. Although they did not quantify the energetic contribution of incorporating such a rigid residue, it was essential for the folding of the proteins.

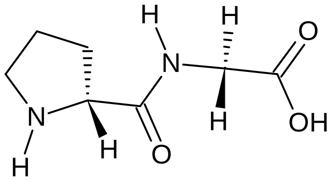


Figure 6.3: D-Pro-Gly turn mimetic

6.2.3 Aib-Gly

The non-stereogenic α-Aminoisobutyryl-Gly (Aib-Gly) dipeptide was found to nucleate type I' β-turn in a 12-residue β-hairpin[79]. α-Aminoisobutyric acid is similar to alanine but in Aib the C-alpha proton was replaced with a methyl group, making it non-sterogenic. An advantage of using the Aib-Gly sequence was the elimination of potential *cis-tran* isomerization of the Xxx–D-Pro peptide bond in sequences containing the D-Pro-Gly dipeptide[79, 47, 48].

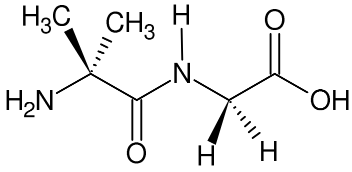


Figure 6.4: Aib-Gly turn mimetic

6.2.4 R-Nip-S-Nip

Nipecotic acid (Nip) is a β-peptide [80, 81] that is similar to proline but it has a six-member ring instead of a five-member ring sidechain. Ribonuclease A is a soluble 124-residue protein that catalyzes the endonucleolytic cleavage of nucleosides[82]. A di-β-peptide, R-Nip-S-Nip (or Nip-D-Nip), was used to nucleate a β-hairpin at residues Asn113-Pro114 which enhanced stability without impacting enzymatic activity(Figure 6.5)[83]. The melting temperature of RNase A containing R-Nip-S-Nip was 1.6 ℃ higher than the wild-type RNase A.

6.2.5 Dimethyl-L-Proline

In further work on the RNase A enzyme, the Raines group mutated Pro114 with 5,5-dimethyl-L-proline and observed a melting temperature increase of 2.8 ℃(Figure 6.5)[84]. In X-ray crystal structures of RNAseA, Asn113-Pro114 forms a *cis*-amide bond. Dimethyl-L-Proline was designed to stabilize the 113-114 cis-amide bond. This mutant (dmP_{114}) did not impact enzymatic activity but the folding rate was accelerated. The effect of adding two dimethyl groups at the 5 position of the proline ring was hypothesized to shift the unfolded state, U_SII, to structures that are more similar to that of the native protein. Alternatively, dmP_{114} could reduce or eliminate slower folding subspecies within U_SII, thus shifting the equilibrium in favor of the faster folding subspecies.

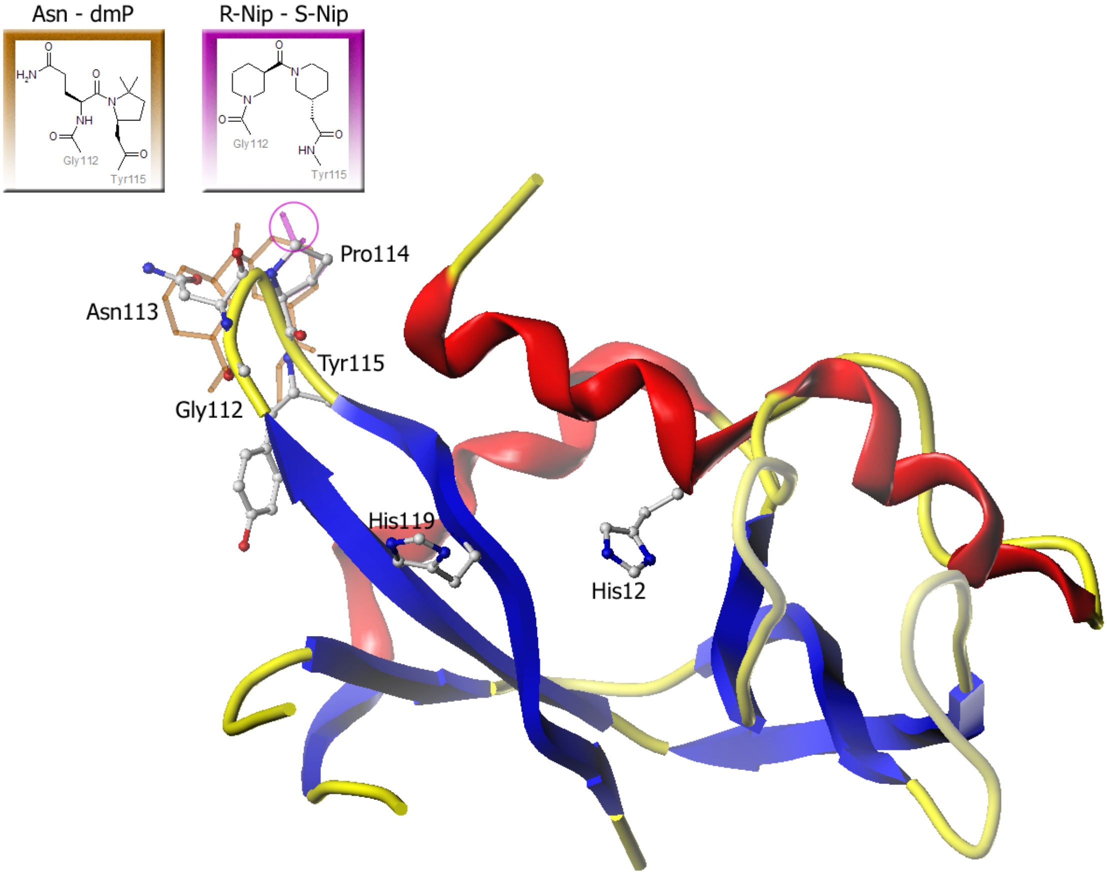


Figure 6.5: RNase A (cartoon representation) has a beta turn at Gly112-Asn113-Pro114-Tyr115 that was chemically modified using expressed protein ligation to generate the chimeric proteins. Both 5,5-dimethylproline (dmP) substitution for Pro114 and R-Nip-S-Nip for Asn113-Pro114 have been studied by the Raines group. The native beta-turn residues are shown in atom-colored CPK representation; the dmP modification (two methyls replacing hydrogens) is shown in magenta, and the R-Nip-S-Nip modification (two six-member β-amino acids) is shown in orange. Figure reproduced from Marshall et al.[85]

6.2.6 D-Pro-Pro

The D-Proline-L-Proline (D-Pro-Pro) dipeptide has been employed as a template by the Robinson group to synthesize cyclic β-hairpin peptide libraries via combinatorial chemistry[86, 87, 88]. Their general strategy was to transplant a hairpin structure from the protein to a D-Pro-Pro template that fixed the conformation of the N- and C-terminal hairpin residues into a β-hairpin geometry. The resulting cyclic peptide maintained its original β-hairpin structure and the N- and C-terminal residues were stapled on to a D-Pro-Pro template. Libraries of β-hairpin mimetics based on the protruding loop (loop III) of human platelet-derived growth factor B (PDGF-B) were synthesized by transplanting residues Glu76 to Ile83, with selected mutations, to a D-Pro-Pro template[88]. Interestingly, it was shown that β-hairpins cyclized by the D-Pro-Pro template could mimic one face of a helix[89, 90]. A 10-residue cyclic β-hairpin served as a scaffold to project key residues of the p53 trans-activation domain to present a surface complementary to the p53 binding pocket of MDM2. A mimetic was optimized to have a IC_{50} value of 140 nM.

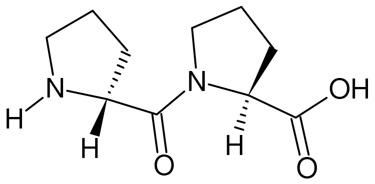


Figure 6.6: D-Pro-Pro turn mimetic

6.3 Impacts of reverse-turn mimetics on protein stability

Ribonuclease A is a soluble 124-residue protein that catalyzes the endonucleolytic cleavage of nucleosides[82]. A di-β-peptide, Nip-D-nip, was used to nucleate a β-hairpin at Asn113-Pro114 that enhanced stability without impacting enzymatic activity[83]. To investigate what was felt to be a minimal effect on the melting temperature (ΔT_m = 1.2±0.3 ℃), the crystal structure of RNAse was minimized, the turn mimetic Nip-D-nip inserted for Asn113-Pro114 and the chimeric structure re-minimized[85]. The two additional methylenes of the two β-amino acids were readily incorporated into the structure by simply extending the hairpin loop with nearly identical torsion angles of the rest of the peptide backbone (Figure 6.7). Thus, no difference was found between the minimum-energy structures of the two structures suggesting that Nip-D-nip did not disrupt the extended β-sheet, and that enthalpic stabilization should be maintained.

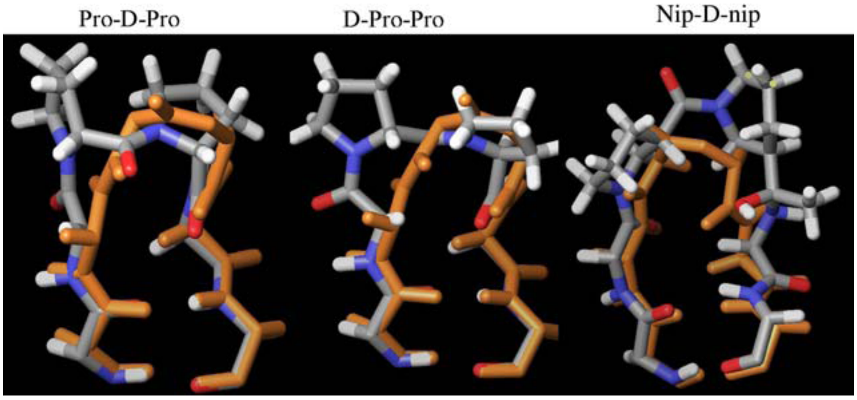


Figure 6.7: RNAseA β-hairpin and its turn mimetics. The native β-hairpin atoms are shown in gold. (Reproduced from Feng et al. [54]

To quantitatively evaluate the relative propensity of reverse turn mimetics to stabilize β-hairpins, Takeuchi and Marshall[91] monitored various parameters during an MD simulation. For example, the relative time that the distance between the two α-carbons of the first and fourth residue of a capped tetrapeptide containing the mimetic was less than 7 Å. To investigate the proclivity of the newer reverse turn mimetics, the native tetrapeptide sequence Gly112- Asn113-Pro114-Tyr115 was capped with acetyl at the N-terminal and with N-methyl amide at the C-terminal. Starting with the native sequence, nine mutants with potential reverse-turn mimetics were generated *in silico*. MacroModel 9.1 was used to run 10-ns MD simulation of the modified peptides in implicit solvent (GB/SA) at 300 K using the OPLS 2005 force field. The distance between the C-alphas of Gly112 and Tyr115 and the distance between the carbonyl oxygen of Gly112 and amide hydrogen of Tyr115 were recorded (Figure 6.8) at each time step, after initial equilibration, similar to the method used by Takeuchi and Marshall[91] to study reverse-turn propensity. The virtual dihedral angle defined by the four C-alpha carbons of the reverse turn, as suggested by Tran et al.[92], of Gly112, Asn113, Pro114, and Tyr115 was also monitored and the results are shown in Figure 6.8. The results of the tetrapeptide simulations were quite revealing. In the top two panels of Figure 6.8, the impacts of the nine different dipeptide substitutions on frequency of observation versus distance between the glycine carbonyl oxygen and the tyrosine amide nitrogen (prevalence of a classic hydrogen bond between the i and i + 3 residue) are plotted. The middle two graphs of Figure 6.8 show the frequency versus distance between the a-carbons of glycine and tyrosine, another measure of the propensity to form conformations resembling reverse turns. In native RNase, the distance between the two α-carbons was less than 7 Å over 80% of the simulation; in the R-Nip-S-Nip chimeric protein, the distance was less than 7 Å for only 10% of the time. While this difference does not directly estimate the amount of

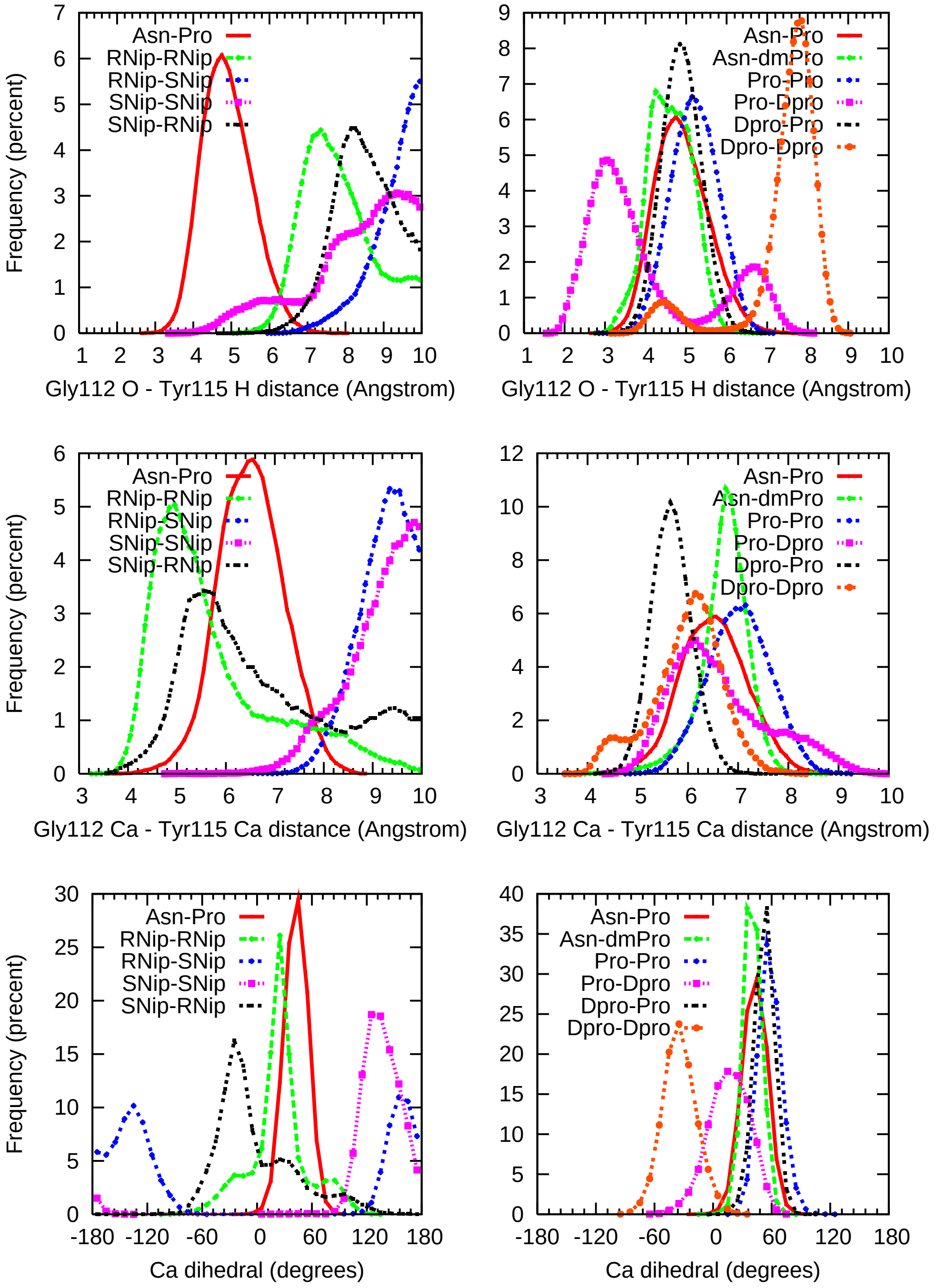


Figure 6.8: Impacts of nine reverse-turn mimetics substituted for dipeptide Asn-Pro segment of acetyl-Gly-Asn-Pro-Tyr-NH-methyl (RNase 112-115) on hydrogen bond distances and virtual dihedral-turn metrics based on MD simulations in implicit solvent.

preorganization in the unfolded RNase versus the chimeric protein, it does indicate that the introduction of two additional methylenes in the backbone of the hairpin loop by R-Nip-S-Nip dramatically increases its inherent flexibility and compromises any anticipated impact of preorganization of the entropy of folding. In contrast, the use of the reverse-turn nucleators, Pro-D-Pro or D-Pro-Pro, enhanced the reverse-turn potential to equal or greater than native Asn113-Pro114 in the simulations, which is consistent with previous estimates of reverse-turn nucleation by Takeuchi and Marshall[91]. It is clear from these graphs that the R-Nip-S-Nip or S-Nip-R-Nip dipeptides do not dynamically stabilize the reverse turn seen with Asn-Pro (red line in all graphs) while Asn-dmP, D-Pro-Pro, and Pro-D-Pro mimic and stabilize the reverse-turn as well as or better then Asn-Pro itself; in fact, the two bottom graphs of Figure 6.8, where the distribution of virtual dihedral angle values between the four α-carbons is shown, further confirm the stabilization of the reverse turn by these three dipeptides. These graphs can be used to give a crude estimate of the entropic consequences of these dipeptide substitutions. From these results, one could predict that the thermal stability of a chimeric RNase with either Pro-D-Pro or D-Pro-Pro replacing Asn113-Pro114 would be greater than that of the chimeric RNase with R-Nip-S-Nip. More sophisticated computations using replica exchange are underway to estimate the changes in melting temperature seen for small chimeric proteins in a model system described later.

6.4 Stabilizing FSD-1

In chapter 5, the stability of FSD-1 was critically analyzed using circular dichroism, differential scanning calorimetry, and molecular dynamics. The plasticity of its β-hairpin was found to be the main contributor to the lack of a well-defined folded state of FSD-1. The β-hairpin could be stabilized by introducing a reverse-turn mimetic to nucleate hairpin formation. The Robinson group's experimental data and our prior simulation data showed that D-Pro-Pro is good reverse turn mimetic. The D-Pro-Pro dipeptide was explored as a reverse-turn mimetic that could enhance the stability of FSD-1 by pre-organizing its β-hairpin. Residues 2 to 13 of FSD-1 formed a β-hairpin and residues 5 to 10 formed an EbaaagbE reverse turn[70]. The two β-stands in the hairpin were connected by this six-residue loop instead of the more common four-residue loop in a traditional reverse turn. FSD-EY, a FSD-1 double mutant containing mutations N1E and I7Y, formed a type I' β-turn. FSD-EY was suggested to be slightly more stable than FSD-1, but definitive melting temperatures were not obtained[70].

Residues Ile$_7$ and Lys$_8$ were mutated to the D-Pro-Pro turn mimetic *in silico*. Gly$_9$ was mutated to Ala to reduce backbone flexibility at position 9. To investigate the impact of incorporating D-Pro-Pro dipeptide into FSD-1, ten 100-ns, explicit solvent, simulations of the Dp7P8A9 mutant were performed, five at 300K and five at 307K.

6.4.1 Molecular dynamics

Residues 7, 8, and 9 of an energy minimized structure of FSD-1 (PDB code: 1FSV) were mutated to Dpro7, Pro8, and Ala9 *in silico* using Sybyl. The mutant protein was denoted FSD-DPP. The Gromacs MD software package was used to set up and run the simulations[56]. The termini were charged and the net charge of the protein was plus 4. Four Cl^- ions were added in random locations to neutralize the system. The protein was solvated in a truncated dodecahedron box of TIP4P water where the minimum distance between a protein atom and the edge of the box was 10 Å. The system contained a total of 16,306 atoms. The OPLS-AA/L 2001 force field was used. The system was minimized until the maximum force was less than 100 kJ mol^-1 nm^-1. The system was heated to 300K from 0K using 50K increments followed by a 1-ns equilibration using the NPT ensemble. Production runs of 100-ns were performed using the NVT ensemble at 300K and 307K. Hydrogen bonds were constrained with LINCS. Timestep was 2 fs. Atomic coordinates were recorded every 2ps for further analysis. The simulations were run on Teragrid resources[57]. Tools provided by the Gromacs package and in-house scripts were used to analyze the MD data.

6.4.2 Results

The global fold of FSD-DPP and its local contacts were analyzed to determine the impact of mutating Ile7, Lys8 and Gly9 to D-Pro, Pro and Ala, respectively. In chapter 5, the importance of inter-strand H-bonds in the β-hairpin and native contacts between the hydrophobic core residues were established. We analyzed the MD trajectories focusing on these local contacts.

In the ensemble of FSD-1 NMR structures, the main-chain atoms of residues Tyr3 and Phe12 formed two hydrogen bonds (Figure 5.1). The formation or maintenance of this pair of hydrogen bonds was determined for ten 100-ns MD simulations (five at 300 K and five at 307 K). The simulations started with the presumed “folded” structure of FSD-DPP. It was our hypothesis that incorporating the D-Pro-Pro residues would not introduce steric repulsions that might disrupt the overall hairpin structure. In fact, it was expected to stabilize the hairpin, in turn, maintaining the Tyr3-Phe12 main-chain hydrogen bonds. The number of hydrogen bonds formed between main-chain atoms of Tyr3 and Phe12 was calculated for the ten 100-ns simulations and shown in Figures 6.9 and 6.10. At 300 K, only simulation D maintained the Tyr3-Phe12 hydrogen bonds. The remaining four simulations showed that the Tyr3-Phe12 hydrogen bonds were lost after 15 to 40 nanoseconds of simulations. At 307 K, Simulations B and E maintained the Tyr3-Phe12 hydrogen bonds, where as simulations A, C and D lost these key hydrogen bonds in the first 20 nanoseconds of the simulations. Simulations D at 300 K, B and E at 307 K all showed the breaking and then reformation of both hydrogen bonds. In addition to monitoring specific hydrogen bonds between residues Tyr3 and Phe12, inter-strand main-chain hydrogen bonds between residues 2 to 7 of β-stand 1 and residues 8 to 13 of β-strand 2 were analyzed (Figures 6.11, 6.12). For simulations D at 300 K, B and E at 307 K where the Tyr3-Phe12 H-bonds were maintained, up to two more additional H-bonds were observed. Formation of additional H-bonds suggested zipping of the β-hairpin.

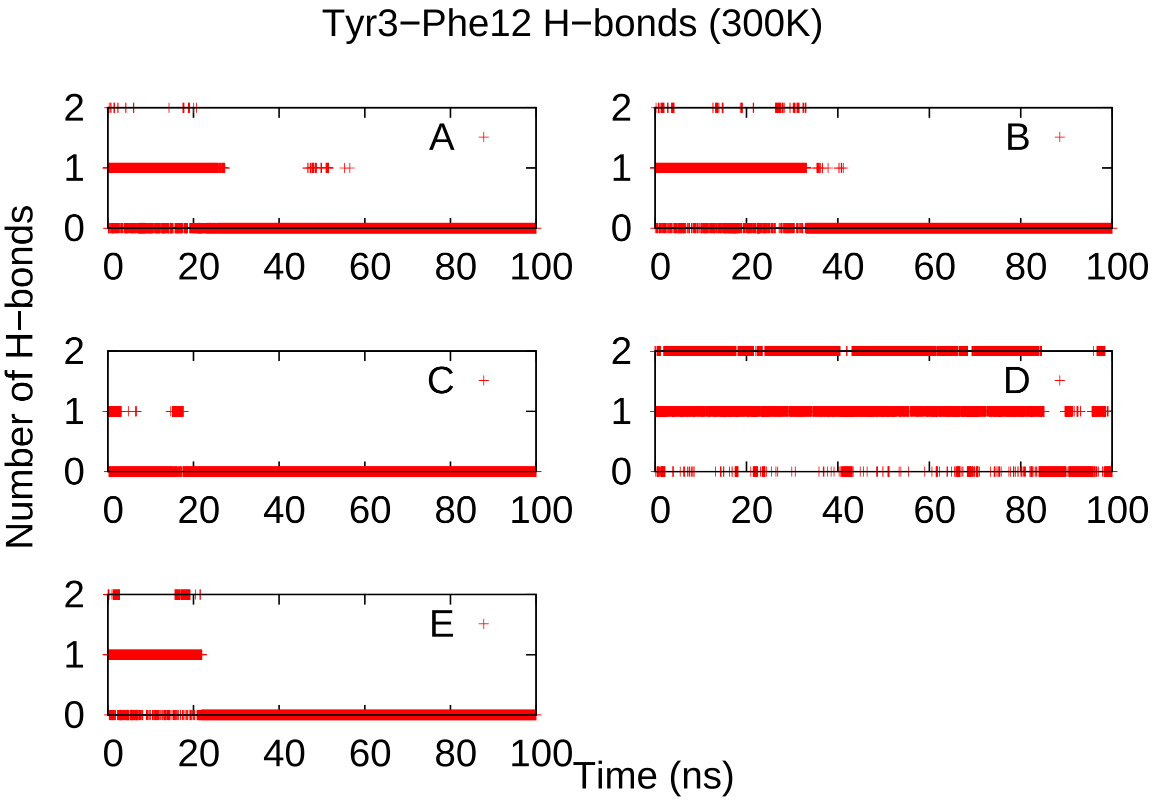


Figure 6.9: Number of hydrogen bonds formed between main-chain atoms of residues Tyr3 and Phe12 for the FSD-DPP mutant. Five independent simulations were performed at 300 K.

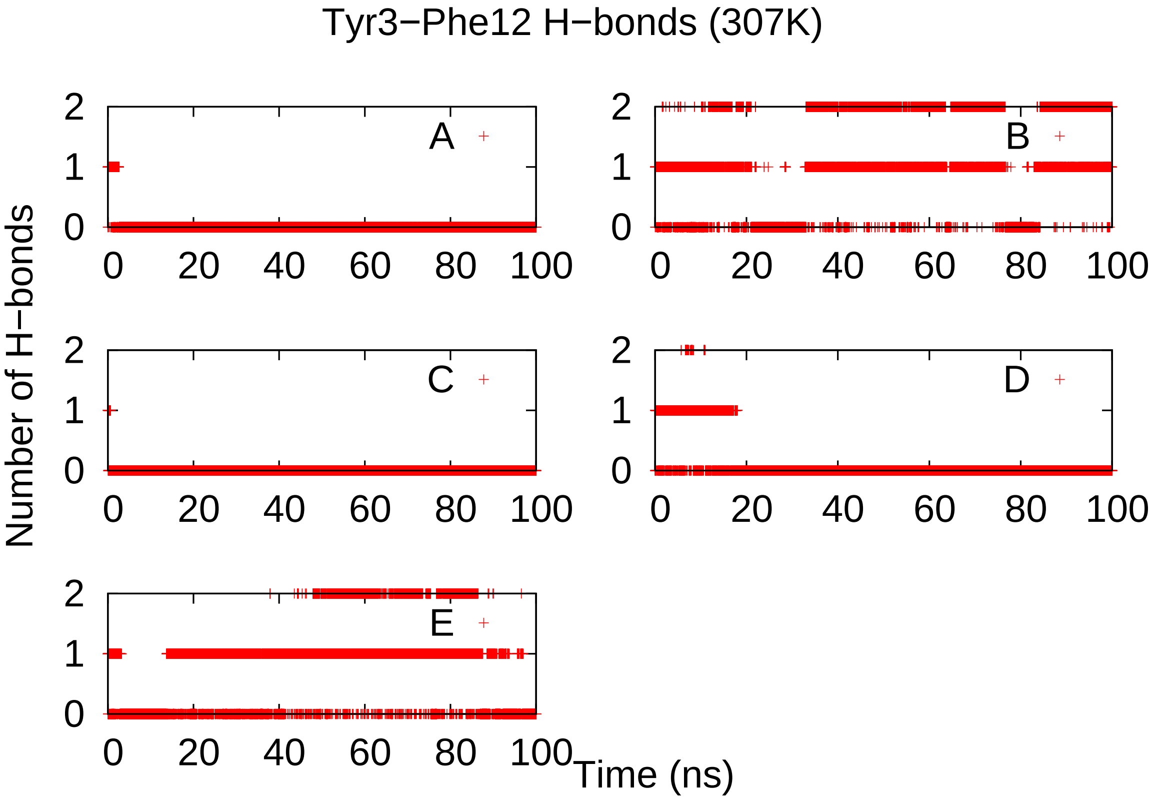


Figure 6.10: Number of hydrogen bonds formed between main-chain atoms of residues Tyr3 and Phe12 for the FSD-DPP mutant. Five independent simulations were performed at 307 K.

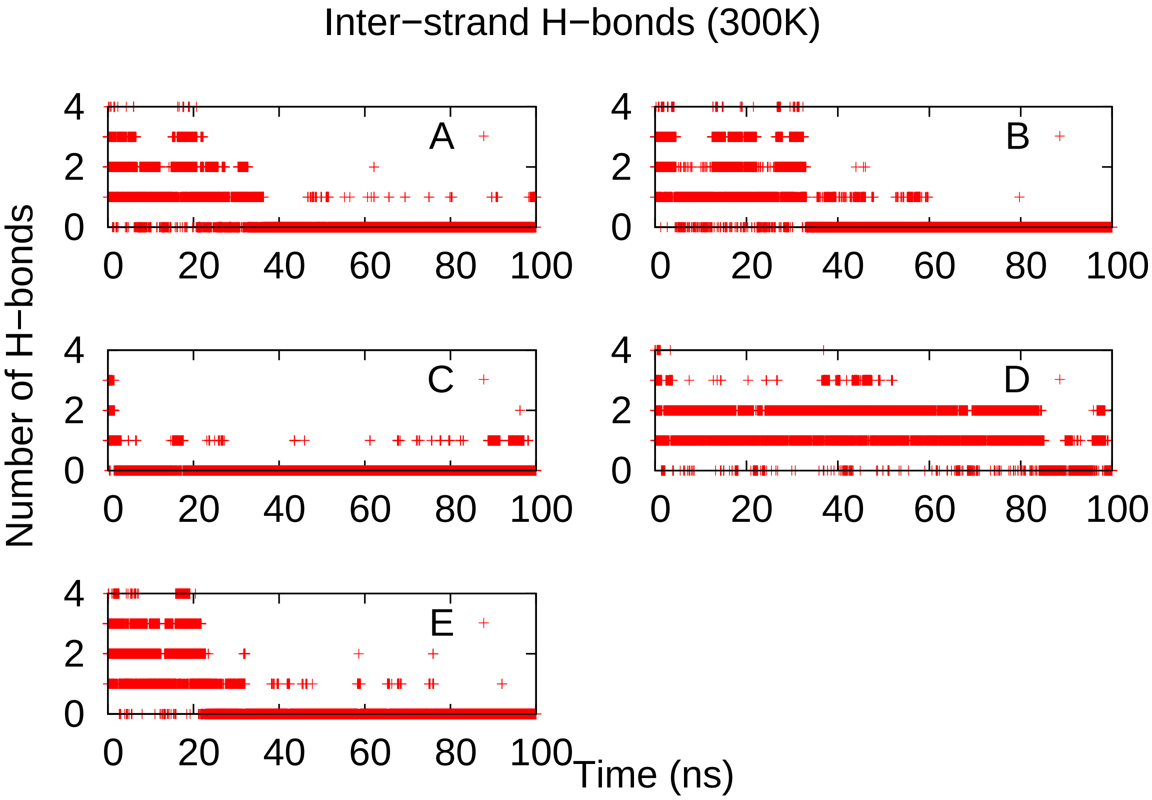


Figure 6.11: Number of hydrogen bonds formed between main-chain atoms of residues in strand one and strand two of the β-hairpin. Five independent simulations were performed at 300 K.

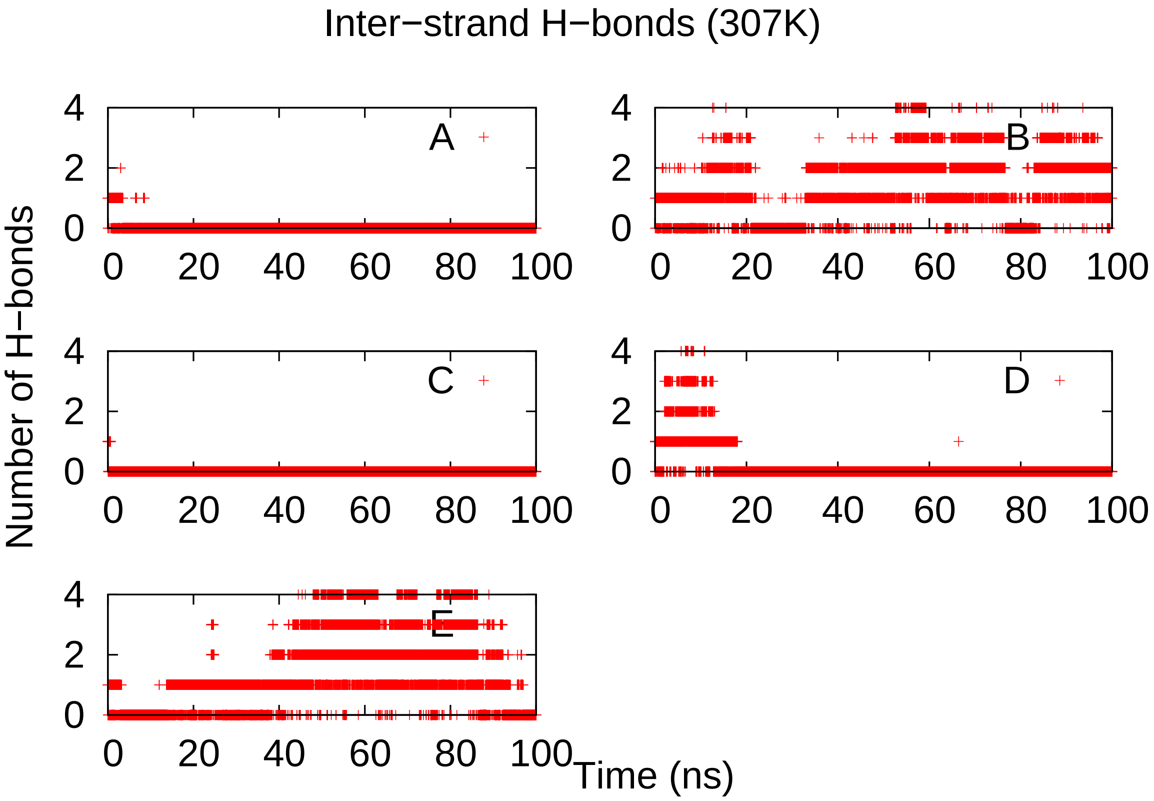
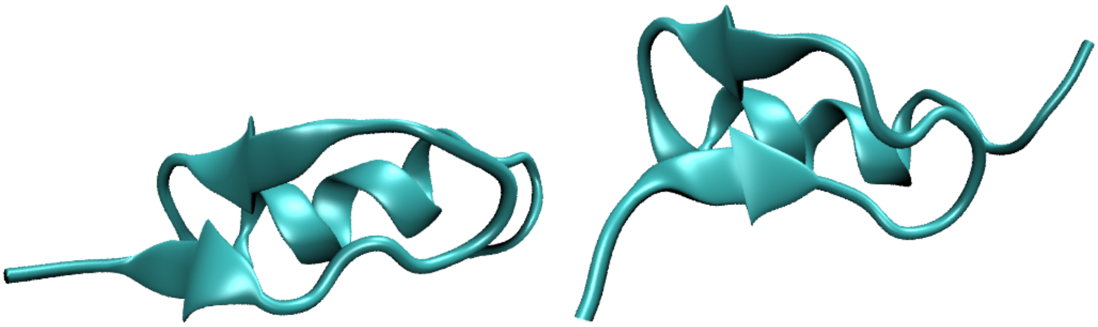
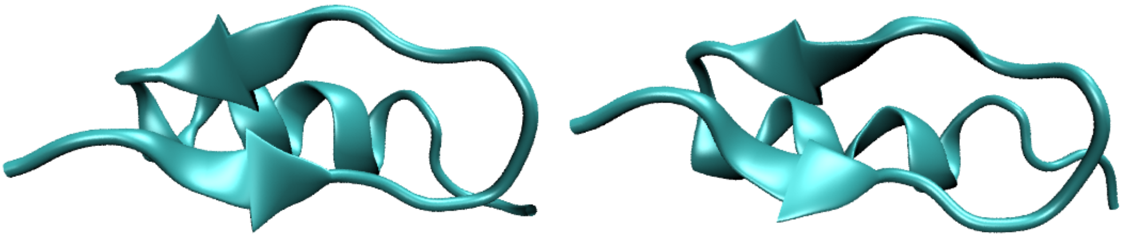


Figure 6.12: Number of hydrogen bonds formed between main-chain atoms of residues in strand one and strand two of the β-hairpin. Five independent simulations were performed at 307 K.

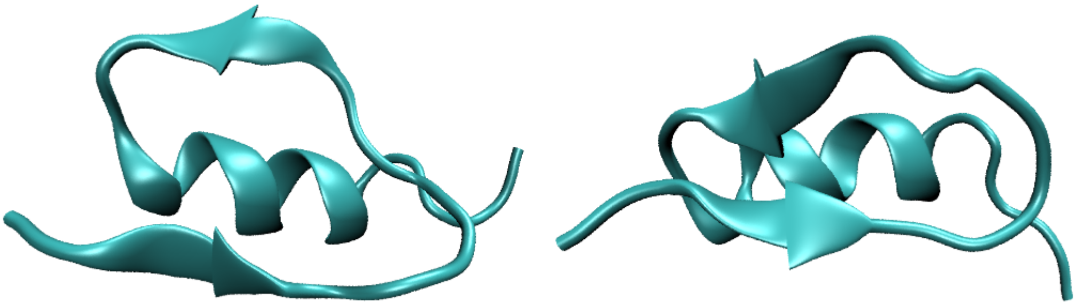
Figure 6.13 shows representative states of various β-hairpin conformations of FSD-DPP from trajectory B at 307K at 20 ns intervals.



(a) 0 ns (b) 20 ns



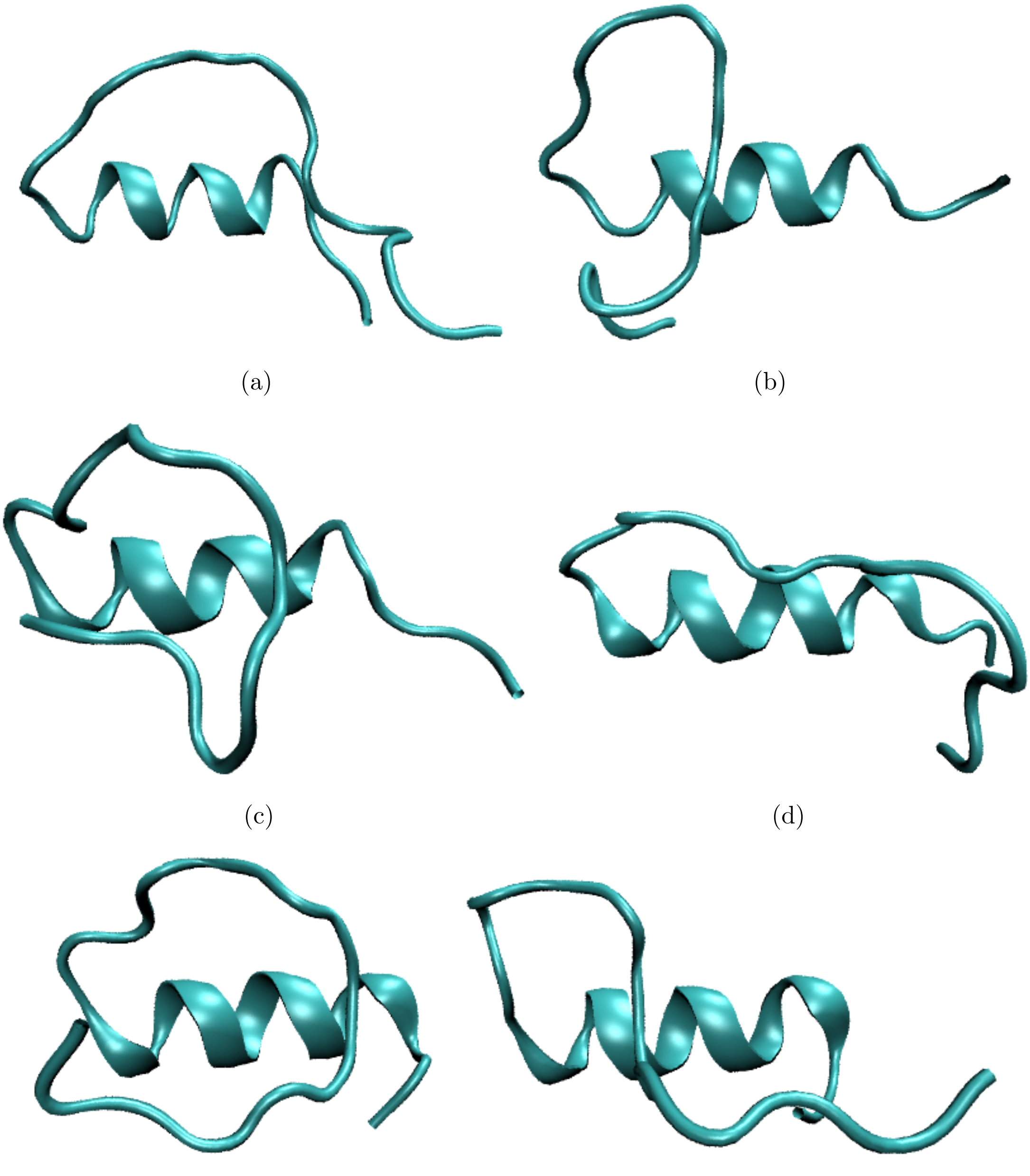
(c) 40 ns (d) 60 ns



(e) 80 ns (f) 100 ns

Figure 6.13: Representative folded structures of FSD-DPP at 0, 20, 40, 60, 80, and 100 nanoseconds of a simulation at 307K. The structures were oriented with the helix shown behind the hairpin.

For the seven simulations that did not maintain the key Tyr3-Phe12 hydrogen bonds, the β-hairpin unfolded as indicated by the lack of inter-strand hydrogen bonds between residues 2-7 and residues 8-13 (Figures 6.9, 6.10, 6.11, 6.12). The more stable α-helix was folded throughout these simulations. Some fraying of the flexible C-terminal turn was observed. Figure 6.14 illustrates various unfolded structures of FSD-DPP. Among the ensemble of unfolded structures, the N-terminal strand of the β-hairpin shows the most flexibility.



(a) (b)

(c) (d)

(e) (f)

Figure 6.14: Representative unfolded structures of FSD-DPP. The structures were oriented with the helix shown in the back.

The amount of atomic contacts between residues in the hydrophobic core was not as useful as hydrogen-bond formation in predicting the stability of the β-hairpin and that of the overall protein. In the NMR ensemble of structures of FSD-1, residues 3, 5, 7, 8, 10, 12 of the hairpin and residues 18, 21, 22, 25 of the helix form a small hydrophobic core. Simulations D at 300K, B and E at 307K indicated that the β-hairpin was not disrupted because the Tyr3-Phe12 H-bonds were maintained. Table 6.1 listed the average number of atomic contacts for ten 100-ns simulations. The average number of contacts for the three simulations that maintained the β-hairpin was 120.4, 101.2, and 120.4, respectively. These values are within the extrema of average contacts from simulations that did not maintain the β-hairpin structure. Figures 6.15 and 6.16 show the average number of contacts as a function of time for the ten 100-ns simulations.

Table 6.1: Average number of hydrophobic core contacts for ten 100-ns simulations

Simulation	Avg. No. of Contacts
300K A	118.1
300K B	126.6
300K C	101.9
300K D	120.4
300K E	110.2
307K A	117.4
307K B	101.2
307K C	88.6
307K D	143.3
307K E	120.4

6.5 Summary

Molecular dynamics simulations suggest that incorporating a turn mimetic into FSD-1 should stabilize its overall fold by nucleating FSD-1's β-hairpin. However, the data was suggestive of nucleation and is not conclusive. Three out of ten simulations showed that FSD-DPP maintained the hydrogen-bonding pattern necessary to form a stable β-hairpin to maintain the overall $\beta\beta\alpha$ fold of FSD-1.

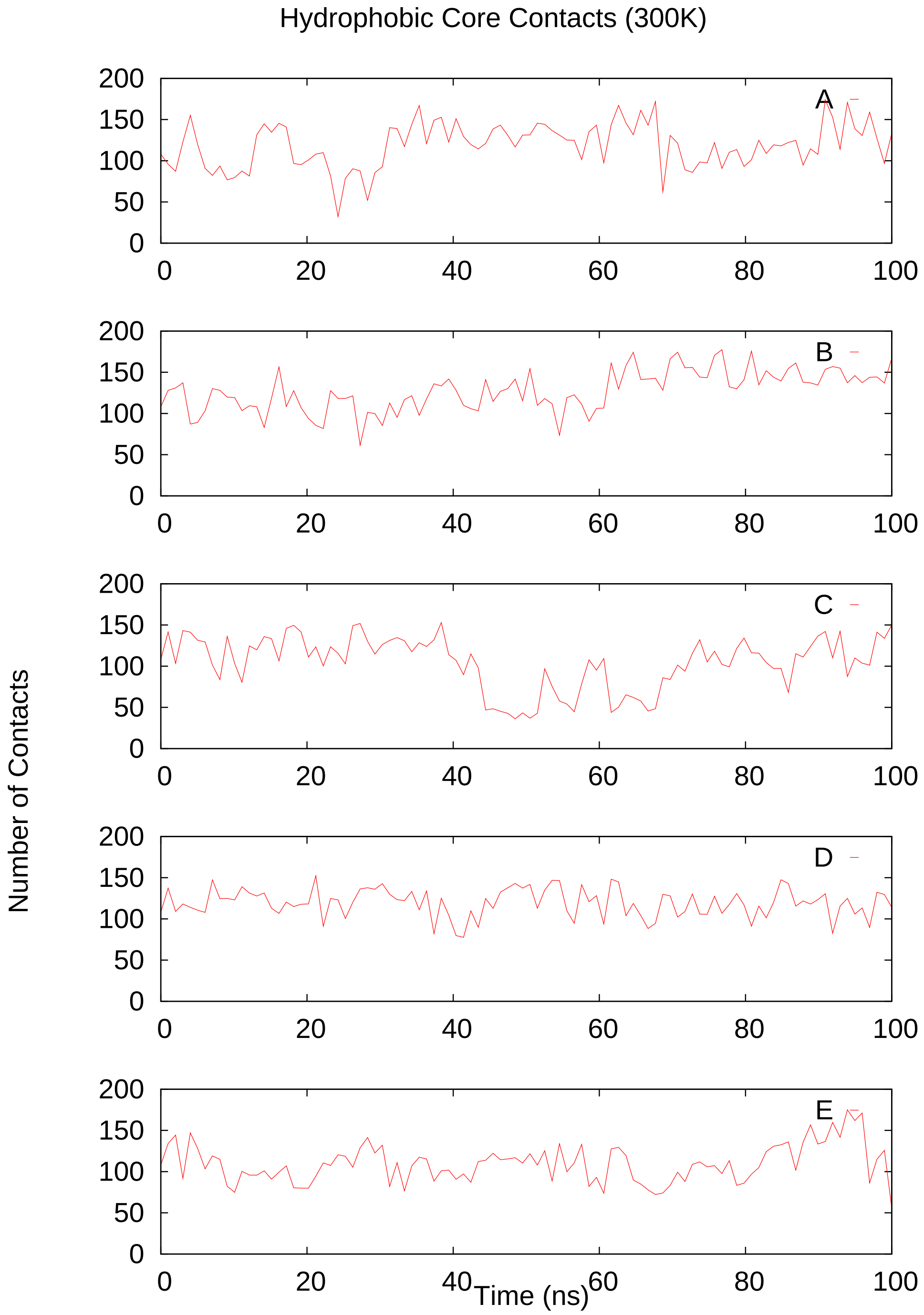


Figure 6.15: Hydrophobic core contacts between side-chain heavy atoms of residues 3, 5, 7, 8, 10, 12 of the hairpin and residues 18, 21, 22, 25 of the helix.

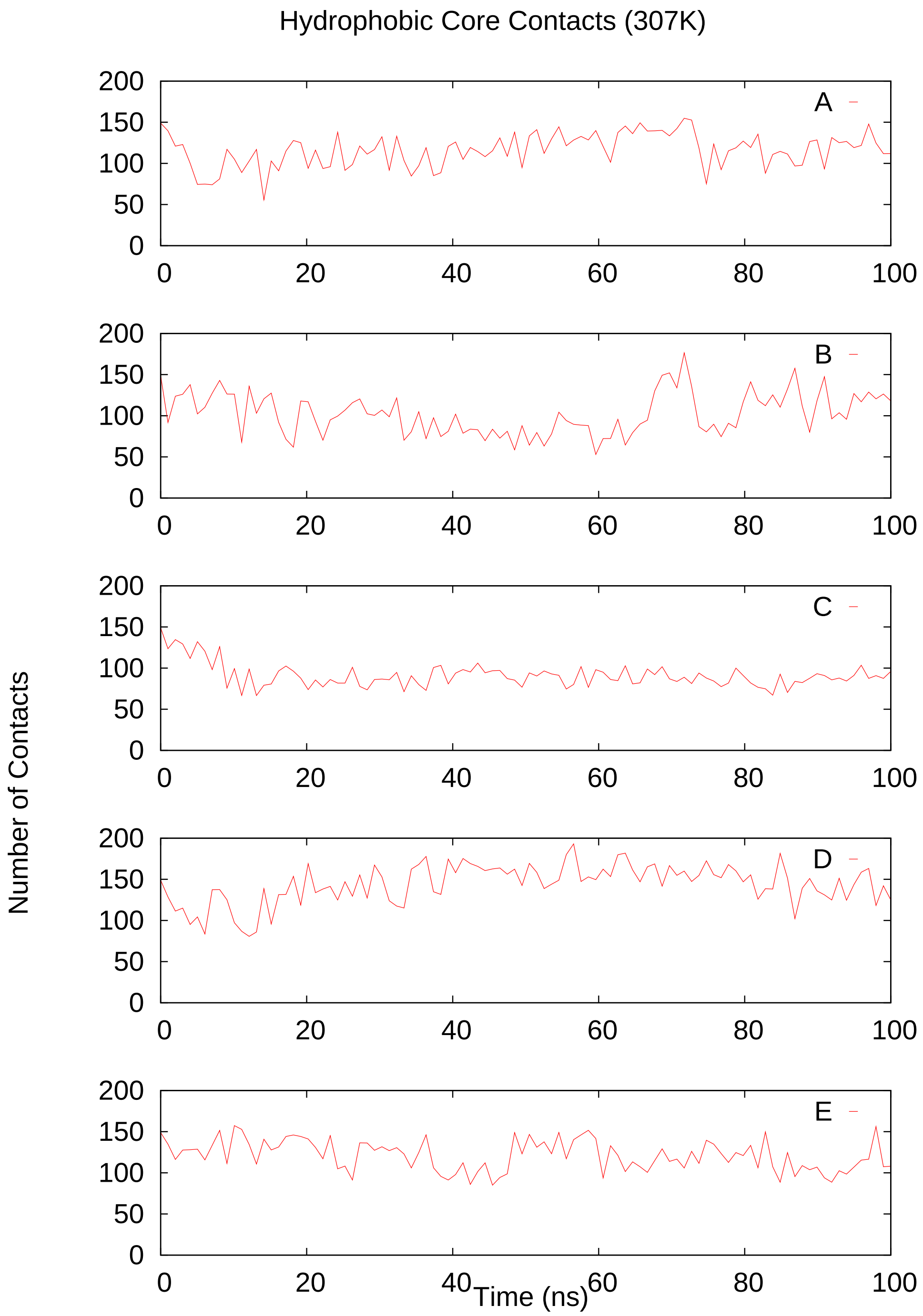


Figure 6.16: Hydrophobic core contacts between side-chain heavy atoms of residues 3, 5, 7, 8, 10, 12 of the hairpin and residues 18, 21, 22, 25 of the helix.

Appendix A

Abbreviations and Acronyms

2D	two-dimensional
3D	three-dimensional
CD	circular dichroism
DSC	differential scanning calorimetry
FSD-1	full sequece design - 1
MD	molecular dynamics
NMR	nuclear magnetic resonance
NOE	nuclear Overhauser effect
NOESY	nuclear Overhauser spectroscopy
REMD	replica exchange molecular dynamics
RMSD	root-mean squared difference
RMSF	root-mean squared fluxuation
TOCSY	total correlation spectroscopy

References

[1] J. Drews, “Drug discovery: a historical perspective,” *Science (New York, N.Y.)*, vol. 287, pp. 1960–1964, Mar. 2000. PMID: 10720314.

[2] K. H. Bleicher, H. Bohm, K. Muller, and A. I. Alanine, “Hit and lead generation: beyond high-throughput screening,” *Nat Rev Drug Discov*, vol. 2, pp. 369–378, May 2003.

[3] J. Kaiser, “MOLECULAR BIOLOGY: Industrial-Style screening meets academic biology,” *Science*, vol. 321, no. 5890, pp. 764–766, 2008.

[4] J. Bajorath, “Integration of virtual and high-throughput screening,” *Nat Rev Drug Discov*, vol. 1, pp. 882–894, Nov. 2002.

[5] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, “Development and validation of a genetic algorithm for flexible docking,” *Journal of molecular biology*, vol. 267, no. 3, pp. 727–48, 1997.

[6] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, “A fast flexible docking method using an incremental construction algorithm,” *Journal of molecular biology*, vol. 261, no. 3, pp. 470–89, 1996.

[7] T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz, “DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases,” *J Comput Aided Mol Des*, vol. 15, no. 5, pp. 411–28, 2001.

[8] R. Abagyan, M. Totrov, and D. Kuznetsov, “Icm - a new method for protein modeling and design - applications to docking and structure prediction from the distorted native conformation,” *Journal of Computational Chemistry*, vol. 15, no. 5, pp. 488–506, 1994.

[9] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson, “Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function,” *J Comput Chem J Comput Chem*, vol. 19, no. 14, pp. 1639–1662, 1998.

[10] J. Meiler and D. Baker, “ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility,” *Proteins*, vol. 65, no. 3, pp. 538–48, 2006.

[11] M. R. McGann, H. R. Almond, A. Nicholls, J. A. Grant, and F. K. Brown, “Gaussian docking functions,” *Biopolymers*, vol. 68, no. 1, pp. 76–90, 2003.

[12] C. M. Venkatachalam, X. Jiang, T. Oldfield, and M. Waldman, “LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites,” *Journal of molecular graphics & modelling*, vol. 21, no. 4, pp. 289–307, 2003.

[13] A. N. Jain, “Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine,” *Journal of medicinal chemistry*, vol. 46, no. 4, pp. 499–511, 2003.

[14] R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, D. E. Shaw, P. Francis, and P. S. Shenkin, “Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy,” *Journal of medicinal chemistry*, vol. 47, no. 7, pp. 1739–49, 2004.

[15] R. Thomsen and M. H. Christensen, “MolDock: a new technique for high-accuracy molecular docking,” *Journal of medicinal chemistry*, vol. 49, no. 11, pp. 3315–21, 2006.

[16] Z. Zsoldos, D. Reid, A. Simon, S. B. Sadjad, and A. P. Johnson, “eHiTS: a new fast, exhaustive flexible ligand docking system,” *Journal of molecular graphics & modelling*, vol. 26, no. 1, pp. 198–212, 2007.

[17] C. McMartin and R. S. Bohacek, “QXP: powerful, rapid computer algorithms for structure-based drug design,” *Journal of computer-aided molecular design*, vol. 11, no. 4, pp. 333–44, 1997.

[18] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov, “Principles of docking: An overview of search algorithms and a guide to scoring functions,” *Proteins*, vol. 47, no. 4, pp. 409–43, 2002.

[19] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin, “A geometric approach to macromolecule-ligand interactions,” *Journal of Molecular Biology*, vol. 161, no. 2, pp. 269–288, 1982.

[20] OpenEye, “FRED.”

[21] R. A. Dammkoehler, S. F. Karasek, E. F. Shands, and G. R. Marshall, “Constrained search of conformational hyperspace,” *Journal of computer-aided molecular design*, vol. 3, no. 1, pp. 3–21, 1989.

[22] R. Wang, L. Lai, and S. Wang, “Further development and validation of empirical scoring functions for structure-based binding affinity prediction,” *Journal of computer-aided molecular design*, vol. 16, no. 1, pp. 11–26, 2002.

[23] D. K. Gehlhaar, G. M. Verkhivker, P. A. Rejto, C. J. Sherman, D. B. Fogel, L. J. Fogel, and S. T. Freer, "Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming," *Chemistry & biology*, vol. 2, no. 5, pp. 317–24, 1995.

[24] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee, "Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes," *Journal of computer-aided molecular design*, vol. 11, no. 5, pp. 425–45, 1997.

[25] H. F. Velec, H. Gohlke, and G. Klebe, "DrugScore(CSD)-knowledge-based scoring function derived from small molecule crystal data with superior recognition rate of near-native ligand poses and better affinity prediction," *Journal of medicinal chemistry*, vol. 48, no. 20, pp. 6296–303, 2005.

[26] I. Motoc, R. Dammkoehler, G. Marshall, and N. Trinajstic, *Three-Dimensional Structure-Activity Relationahips and Biological Receptor Mapping*, pp. 222–251. Chichester: Ellis Horwood, 1986.

[27] D. D. Beusen, E. F. B. Shands, S. F. Karasek, G. R. Marshall, and R. A. Dammkoehler, "Systematic search in conformational analysis," *Theochem-J Mol Struc Theochem-J Mol Struc*, vol. 370, no. 2-3, pp. 157–171, 1996.

[28] H. Iijima, J. B. Dunbar, and G. R. Marshall, "Calibration of effective van der waals atomic contact radii for proteins and peptides," *Proteins*, vol. 2, no. 4, pp. 330–9, 1987.

[29] Tripos, "Sybyl 8.1."

[30] M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. Mooij, P. N. Mortenson, and C. W. Murray, "Diverse, high-quality test set for the validation of protein-ligand docking performance," *Journal of medicinal chemistry*, vol. 50, no. 4, pp. 726–41, 2007.

[31] E. Perola, W. P. Walters, and P. S. Charifson, "A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance," *Proteins*, vol. 56, no. 2, pp. 235–49, 2004.

[32] A. N. Jain, "Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search," *Journal of computer-aided molecular design*, vol. 21, no. 5, pp. 281–306, 2007.

[33] C. Bissantz, G. Folkers, and D. Rognan, "Protein-based virtual screening of chemical databases. 1. evaluation of different docking/scoring combinations," *J Med Chem*, vol. 43, no. 25, pp. 4759–67, 2000.

[34] R. Yang, C. Parnot, and G. Marshall, "OpenScreening on xgrid: an open-access virtual screen web server," *submitted*, 2009.

[35] A. N. Jain, "Bias, reporting, and sharing: computational evaluations of docking methods," *Journal of computer-aided molecular design*, vol. 22, no. 3-4, pp. 201–12, 2008.

[36] P. C. Hawkins, G. L. Warren, A. G. Skillman, and A. Nicholls, "How to do an evaluation: pitfalls and traps," *Journal of computer-aided molecular design*, vol. 22, no. 3-4, pp. 179–90, 2008.

[37] G. L. Warren, C. W. Andrews, A. M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff, and M. S. Head, "A critical assessment of docking programs and scoring functions," *Journal of medicinal chemistry*, vol. 49, no. 20, pp. 5912–31, 2006.

[38] R. Wang, Y. Lu, and S. Wang, "Comparative evaluation of 11 scoring functions for molecular docking," *J Med Chem*, vol. 46, no. 12, pp. 2287–303, 2003.

[39] I. W. Davis and D. Baker, "RosettaLigand docking with full ligand and receptor flexibility," *Journal of molecular biology*, vol. 385, no. 2, pp. 381–92, 2009.

[40] W. Delano, "The PyMOL molecular graphics system."

[41] Schrodinger, "Macromodel."

[42] W. L. Jorgensen and J. Tirado-Rives, "Development of the OPLS-AA force field for organic and biomolecular systems.," *Abstr Pap Am Chem S Abstr Pap Am Chem S*, vol. 216, pp. U696–U696, 1998.

[43] C. D. Snow, E. J. Sorin, Y. M. Rhee, and V. S. Pande, "How well can simulation predict protein folding kinetics and thermodynamics?," *Annual review of biophysics and biomolecular structure*, vol. 34, pp. 43–69, 2005.

[44] Y. Sugita and Y. Okamoto, "Replica-exchange molecular dynamics method for protein folding," *Chemical Physics Letters*, vol. 314, no. 1-2, pp. 141–151, 1999.

[45] K. Y. Sanbonmatsu and A. E. Garcia, "Structure of met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics," *Proteins*, vol. 46, no. 2, pp. 225–34, 2002.

[46] M. M. Seibert, A. Patriksson, B. Hess, and D. van der Spoel, "Reproducible polypeptide folding and structure prediction using molecular dynamics simulations," *J Mol Biol*, vol. 354, no. 1, pp. 173–83, 2005.

[47] M. D. Struthers, R. P. Cheng, and B. Imperiali, “Design of a monomeric 23-residue polypeptide with defined tertiary structure,” *Science*, vol. 271, no. 5247, pp. 342–5, 1996.

[48] M. Struthers, J. J. Ottesen, and B. Imperiali, “Design and NMR analyses of compact, independently folded BBA motifs,” *Fold Des*, vol. 3, no. 2, pp. 95–103, 1998.

[49] B. I. Dahiyat and S. L. Mayo, “De novo protein design: fully automated sequence selection,” *Science*, vol. 278, no. 5335, pp. 82–7, 1997.

[50] H. Lei and Y. Duan, “The role of plastic beta-hairpin and weak hydrophobic core in the stability and unfolding of a full sequence design protein,” *J Chem Phys*, vol. 121, no. 23, pp. 12104–11, 2004.

[51] S. Y. Kim, J. Lee, and J. Lee, “Folding simulations of small proteins,” *Biophysical chemistry*, vol. 115, no. 2-3, pp. 195–200, 2005.

[52] S. Jang, E. Kim, and Y. Pak, “Free energy surfaces of miniproteins with a betabetaalpha motif: replica exchange molecular dynamics simulation with an implicit solvation model,” *Proteins*, vol. 62, no. 3, pp. 663–71, 2006.

[53] H. Lei, S. G. Dastidar, and Y. Duan, “Folding Transition-State and Denatured-State ensembles of FSD-1 from folding and unfolding simulations,” *J Phys Chem B Condens Matter Mater Surf Interfaces Biophys*, vol. 110, no. 43, pp. 22001–8, 2006.

[54] J. Feng, L. Tessler, and G. Marshall, “Chimeric protein engineering,” *International Journal of Peptide Research and Therapeutics*, vol. 13, no. 1, pp. 151–160, 2007.

[55] W. Zhang, C. Wu, and Y. Duan, “Convergence of replica exchange molecular dynamics,” *The Journal of chemical physics*, vol. 123, no. 15, p. 154105, 2005.

[56] D. V. D. Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, “GROMACS: fast, flexible, and free,” *J Comput Chem*, vol. 26, no. 16, pp. 1701–18, 2005.

[57] C. Catlett and L. Grandinetti, *TeraGrid: Analysis of Organization, System Architecture, and Middleware Enabling New Types of Applications, HPC and Grids in Action.* Amsterdam: IOS Press, 2007.

[58] M. M. Santoro and D. W. Bolen, “Unfolding free energy changes determined by the linear extrapolation method. 1. unfolding of phenylmethanesulfonyl alpha-chymotrypsin using different denaturants,” *Biochemistry*, vol. 27, no. 21, pp. 8063–8, 1988.

[59] V. Consalvi, R. Chiaraluce, L. Giangiacomo, R. Scandurra, P. Christova, A. Karshikoff, S. Knapp, and R. Ladenstein, "Thermal unfolding and conformational stability of the recombinant domain II of glutamate dehydrogenase from the hyperthermophile thermotoga maritima," *Protein engineering*, vol. 13, no. 7, pp. 501–7, 2000.

[60] N. J. Greenfield, "Analysis of circular dichroism data," *Methods in enzymology*, vol. 383, pp. 282–317, 2004.

[61] S. Honda, T. Akiba, Y. S. Kato, Y. Sawada, M. Sekijima, M. Ishimura, A. Ooishi, H. Watanabe, T. Odahara, and K. Harata, "Crystal structure of a ten-amino acid protein," *Journal of the American Chemical Society*, vol. 130, no. 46, pp. 15327–31, 2008.

[62] J. M. Scholtz, S. Marqusee, R. L. Baldwin, E. J. York, J. M. Stewart, M. Santoro, and D. W. Bolen, "Calorimetric determination of the enthalpy change for the alpha-helix to coil transition of an alanine peptide in water," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 7, pp. 2854–8, 1991.

[63] J. M. Scholtz, H. Qian, E. J. York, J. M. Stewart, and R. L. Baldwin, "Parameters of helix-coil transition theory for alanine-based peptides of varying chain lengths in water," *Biopolymers*, vol. 31, no. 13, pp. 1463–70, 1991.

[64] J. W. Taylor, N. J. Greenfield, B. Wu, and P. L. Privalov, "A calorimetric study of the folding-unfolding of an alpha-helix with covalently closed n and c-terminal loops," *Journal of molecular biology*, vol. 291, no. 4, pp. 965–76, 1999.

[65] J. M. Richardson and G. I. Makhatadze, "Temperature dependence of the thermodynamics of helix-coil transition," *Journal of molecular biology*, vol. 335, no. 4, pp. 1029–37, 2004.

[66] E. Freire, "Thermal denaturation methods in the study of protein folding," *Methods in enzymology*, vol. 259, pp. 144–68, 1995.

[67] R. Godoy-Ruiz, E. R. Henry, J. Kubelka, J. Hofrichter, V. Munoz, J. M. Sanchez-Ruiz, and W. A. Eaton, "Estimating free-energy barrier heights for an ultrafast folding protein from calorimetric and kinetic data," *The journal of physical chemistry*, vol. 112, no. 19, pp. 5938–49, 2008.

[68] W. Li, J. Zhang, and W. Wang, "Understanding the folding and stability of a zinc finger-based full sequence design protein with replica exchange molecular dynamics simulations," *Proteins*, vol. 67, no. 2, pp. 338–49, 2007.

[69] J. W. Pitera and W. Swope, “Understanding folding and design: replica-exchange simulations of ”Trp-cage” miniproteins,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 13, pp. 7587–92, 2003.

[70] C. A. Sarisky and S. L. Mayo, “The beta-beta-alpha fold: explorations in sequence space,” *J Mol Biol*, vol. 307, no. 5, pp. 1411–8, 2001.

[71] I. L. Karle, C. Das, and P. Balaram, “De novo protein design: crystallographic characterization of a synthetic peptide containing independent helical and hairpin domains,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 7, pp. 3034–7, 2000.

[72] P. Y. Chou and G. D. Fasman, “Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins,” *Biochemistry*, vol. 13, no. 2, pp. 211–22, 1974.

[73] R. Gupta, Q. K. Beg, and P. Lorenz, “Bacterial alkaline proteases: molecular approaches and industrial applications,” *Appl Microbiol Biotechnol*, vol. 59, no. 1, pp. 15–32, 2002.

[74] G. R. Marshall and H. E. Bosshard, “Angiotensin II. studies on the biologically active conformation,” *Circ Res*, vol. 31, no. 9, pp. Suppl 2:143–50, 1972.

[75] G. R. Marshall, E. E. Hodgkin, D. A. Langs, G. D. Smith, J. Zabrocki, and M. T. Leplawy, “Factors governing helical preference of peptides containing multiple alpha,alpha-dialkyl amino acids,” *Proc Natl Acad Sci U S A*, vol. 87, no. 1, pp. 487–91, 1990.

[76] M. Baca, P. F. Alewood, and S. B. Kent, “Structural engineering of the HIV-1 protease molecule with a beta-turn mimic of fixed geometry,” *Protein Sci*, vol. 2, no. 7, pp. 1085–91, 1993.

[77] H. E. Stanger and S. H. Gellman, “Rules for antiparallel -Sheet design: d-Pro-Gly is superior to l-Asn-Gly for -Hairpin nucleation1,” *Journal of the American Chemical Society*, vol. 120, no. 17, pp. 4236–4237, 1998.

[78] H. L. Schenck and S. H. Gellman, “Use of a designed Triple-Stranded antiparallel -Sheet to probe -Sheet cooperativity in aqueous solution,” *Journal of the American Chemical Society*, vol. 120, no. 19, pp. 4869–4870, 1998.

[79] L. R. Masterson, M. A. Etienne, F. Porcelli, G. Barany, R. P. Hammer, and G. Veglia, “Nonstereogenic alpha-aminoisobutyryl-glycyl dipeptidyl unit nucleates type iprime beta-turn in linear peptides in aqueous solution,” *Peptide Science*, vol. 88, no. 5, pp. 746–753, 2007.

[80] D. Seebach, M. Overhand, F. N. M. Khnle, B. Martinoni, L. Oberer, U. Hommel, and H. Widmer, “beta-Peptides: synthesis by <I>Arndt-Eistert</I> homologation with concomitant peptide coupling. structure determination by NMR and CD spectroscopy and by x-ray crystallography. helical secondary structure of a beta-hexapeptide in solution and its stability towards pepsin,” *Helvetica Chimica Acta*, vol. 79, no. 4, pp. 913–941, 1996.

[81] D. H. Appella, L. A. Christianson, I. L. Karle, D. R. Powell, and S. H. Gellman, “-Peptide foldamers: Robust helix formation in a new family of -Amino acid oligomers,” *Journal of the American Chemical Society*, vol. 118, no. 51, pp. 13071–13072, 1996.

[82] R. T. Raines, “Ribonuclease a,” *Chem Rev*, vol. 98, no. 3, pp. 1045–1066, 1998.

[83] U. Arnold, M. P. Hinderaker, B. L. Nilsson, B. R. Huck, S. H. Gellman, and R. T. Raines, “Protein prosthesis: a semisynthetic enzyme with a beta-peptide reverse turn,” *J Am Chem Soc*, vol. 124, no. 29, pp. 8522–3, 2002.

[84] U. Arnold, M. P. Hinderaker, J. Koditz, R. Golbik, R. Ulbrich-Hofmann, and R. T. Raines, “Protein prosthesis: a nonnatural residue accelerates folding and increases stability,” *J Am Chem Soc*, vol. 125, no. 25, pp. 7500–1, 2003.

[85] G. R. Marshall, J. A. Feng, and D. J. Kuster, “Back to the future: ribonuclease a,” *Biopolymers*, vol. 90, no. 3, pp. 259–277, 2008. PMID: 17868092.

[86] J. Spth, F. Stuart, L. Jiang, and J. Robinson, “Stabilization of a <I>beta</I>-Hairpin conformation in a cyclic peptide using the templating effect of a heterochiral diproline unit,” *Helvetica Chimica Acta*, vol. 81, no. 9, pp. 1726–1738, 1998.

[87] M. Favre, K. Moehle, L. Jiang, B. Pfeiffer, and J. A. Robinson, “Structural mimicry of canonical conformations in antibody hypervariable loops using cyclic peptides containing a heterochiral diproline template,” *Journal of the American Chemical Society*, vol. 121, no. 12, pp. 2679–2685, 1999.

[88] L. Jiang, K. Moehle, B. Dhanapal, D. Obrecht, and J. Robinson, “Combinatorial biomimetic chemistry: Parallel synthesis of a small library of <I>beta</I>-Hairpin mimetics based on loop III from human Platelet-Derived growth factor b,” *Helvetica Chimica Acta*, vol. 83, no. 12, pp. 3097–3112, 2000.

[89] R. Fasan, R. L. Dias, K. Moehle, O. Zerbe, J. W. Vrijbloed, D. Obrecht, and J. A. Robinson, “Using a beta-hairpin to mimic an alpha-helix: cyclic peptidomimetic inhibitors of the p53-HDM2 protein-protein interaction,” *Angewandte Chemie (International ed*, vol. 43, no. 16, pp. 2109–12, 2004.

[90] R. Fasan, R. L. Dias, K. Moehle, O. Zerbe, D. Obrecht, P. R. Mittl, M. G. Grutter, and J. A. Robinson, "Structure-activity studies in a family of beta-hairpin protein epitope mimetic inhibitors of the p53-HDM2 protein-protein interaction," *Chembiochem*, vol. 7, no. 3, pp. 515–26, 2006.

[91] Y. Takeuchi and G. R. Marshall, "Conformational analysis of reverse-turn constraints by n-methylation and n-hydroxylation of amide bonds in peptides and non-peptide mimetics," *J Am Chem Soc J Am Chem Soc*, vol. 120, no. 22, pp. 5363–5372, 1998.

[92] T. Tran, J. McKie, W. Meutermans, G. Bourne, P. Andrews, and M. Smythe, "Topological side-chain classification of -turns: Ideal motifs for peptidomimetic development," *Journal of Computer-Aided Molecular Design*, vol. 19, no. 8, pp. 551–566, 2005.

[93] U. Arnold, M. P. Hinderaker, and R. T. Raines, "Semisynthesis of ribonuclease a using intein-mediated protein ligation," *ScientificWorldJournal*, vol. 2, pp. 1823–7, 2002.

[94] H. J. Bohm, "The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure," *J Comput Aided Mol Des*, vol. 8, no. 3, pp. 243–56, 1994.

[95] B. I. Dahiyat and S. L. Mayo, "Protein design automation," *Protein Sci*, vol. 5, no. 5, pp. 895–903, 1996.

[96] S. Fox, S. Farr-Jones, L. Sopchak, A. Boggs, H. W. Nicely, R. Khoury, and M. Biros, "High-Throughput screening: Update on practices and success," *J Biomol Screen*, vol. 11, no. 7, pp. 864–869, 2006.

[97] P. C. Hawkins, A. G. Skillman, and A. Nicholls, "Comparison of shape-matching and docking as virtual screening tools," *Journal of medicinal chemistry*, vol. 50, no. 1, pp. 74–82, 2007.

[98] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen, "Designing a 20-residue protein," *Nat Struct Mol Biol*, vol. 9, pp. 425–430, June 2002.

[99] G. C. K. Roberts, *NMR of macromolecules : a practical approach*. Practical approach series ; 134., Oxford ; New York: IRL Press at Oxford University Press, 1993.

[100] K. Wthrich, *NMR of proteins and nucleic acids*. The George Fisher Baker non-resident lectureship in chemistry at Cornell University, New York: Wiley, 1986.

Curriculum Vitae

Jianwen A. Feng

Born

July 24, 1977
Enping, Guangdong, People's Republic of China

Degrees

2001 - B.S. Biochemistry (with Distinction), *summa cum laude*, The Ohio State University, Columbus, OH
2001 - B.S. Computer and Information Science, *summa cum laude*, The Ohio State University, Columbus, OH
2009 - Ph.D. (anticipated) Computational Biology, Washington University in St. Louis, St. Louis, MO
Thesis Advisor: Garland R. Marshall

Select Honors and Awards

Kauffman Graduate Fellowship in Bioentrepreneurship
NIH Computational Biology Training Grant
Pfizer Summer Undergraduate Research Fellowship
Excellence Scholarship (full tuition at OSU)
Phi Beta Kappa (Academic Honor Society)

Professional Societies

Biophysical Society

Positions and Employment

2001-2003 Software Engineer, WebSphere Application Server Performance, IBM, Rochester, MN
2003-2009 Graduate Student, Dept. of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, MO
2008 Research Intern, Computational Drug Discovery Group, Genentech, South San Francisco, CA

Peer-reviewed Publications

Schweitzer-Stenner R, Gonzlez W, Bourne GT, Feng JA, Marshall GR. "Conformational Manifold of α-Aminoisobutyric Acid (Aib) Containing Alanine-Based Tripeptides in Aqueous Solution Explored by Vibrational Spectroscopy, Electronic Circular Dichroism Spectroscopy, and Molecular Dynamics Simulations." *J Am Chem Soc*, **129**, 13095-13109, 2007.

Feng JA, Tessler LA, Marshall GR. “Chimeric Protein Engineering.” *International Journal of Peptide Research and Therapeutics*, **13**, 151-160, 2007.

Marshall GR, Feng JA, Kuster DJ. “Back to the future: Ribonuclease A.” *Biopolymers*, **90**, 259-77, 2008.

Feng JA, Kao J, Marshall GR. “Critical Assessment of Miniprotein Stability Using Molecular Simulations, Circular Dichroism, Calorimetry and NMR.” Accepted with minor revisions *Biophysical Journal*, 2009.

Manuscripts in Preparation

Feng JA, Marshall GR. “SKATE: Accurate Docking is Achieved by Decoupling Systematic Sampling from Scoring.” Submitted to *J Comp Chem*, 2009.

August 2009

Small Molecule Docking, Feng, Ph.D. 2009