LOCAL MINIMA HOPPING ALONG THE PROTEIN ENERGY SURFACE

by

Brian Olson

A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
In Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Computer Science

Committee:

Amarda Shehu
Dr. Amarda Shehu, Thesis Director

Jana Kosecka
Dr. Jana Kosecka, Committee Member

Jyh-Ming Lien
Dr. Jyh-Ming Lien, Committee Member

Sanjeev Setia
Dr. Sanjeev Setia, Chairman, Department of Computer Science

Dr. Lloyd J. Griffiths, Associate Dean for Research and Graduate Studies

Date: 12/2/2011

Fall 2011
George Mason University
Fairfax, VA

Local Minima Hopping along the Protein Energy Surface

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

By

Brian Olson
Bachelor of Science in Engineering
Princeton University, 2005

Director: Dr. Amarda Shehu, Assistant Professor
Department of Computer Science

Fall 2011
George Mason University
Fairfax, VA

Dedication

I dedicate this thesis to my wife Sarah who is the reason I am in graduate school.

Acknowledgments

I would like to thank my advisor Dr. Amarda Shehu for all the help and support she has given me in my graduate studies. She took a chance on me early in both of our careers and it is her guidance that has made me successful in my studies. I would also like to thank the other members of the Shehu lab, especially Kevin Molloy, for collaboration on countless projects. Finally I would like to thank the Hydra cluster for its tireless efforts on my behalf.

This material is based upon work supported by the National Science Foundation under Grant No. 1016995.

Table of Contents

List of Tables

List of Figures

Abstract

LOCAL MINIMA HOPPING ALONG THE PROTEIN ENERGY SURFACE

Brian Olson, M.S.

George Mason University, 2011

Thesis Director: Dr. Amarda Shehu

Modeling of protein molecules in silico for the purpose of elucidating the three-dimensional structure where the protein is biologically active employs the knowledge that the protein conformational space has an underlying funnel-like energy surface. The biologically-active structure, also referred to as the native structure, resides at the basin or global minimum of the energy surface. A common approach among computational methods that seek the protein native structure is to search for local minima in the energy surface, with the hope that one of the local minima corresponds to the global minimum. Typical stochastic search methods, however, fail to explicitly sample local minima. This thesis proposes a novel algorithm to directly sample local minima at a coarse-grained level of detail. The Protein Local Optima Walk (PLOW) algorithm combines a memetic approach from evolutionary computation with cutting-edge structure prediction protocols in computational biophysics. PLOW explores the space of local minima by explicitly projecting each move at the global level to a nearby local minimum. This allows PLOW to jump over local energy barriers and more effectively sample near-native conformations. An additional contribution of this thesis is that the memetic approach in PLOW is applied to FeLTr, a tree-based search framework which ensures geometric diversity of computed conformations through projections of the conformational space. Analysis across a broad range of proteins shows that PLOW and

memetic FeLTr outperform the original FeLTr framework and compare favorably against state-of-the-art ab-initio structure prediction algorithms.

Chapter 1: Introduction

The problem of determining a protein's three-dimensional structure from amino acid sequence alone remains a central challenge in computational structural biology [3]. Proteins play a critical role in countless cellular processes, and their biological functions are largely determined by the three-dimensional structure they adopt under physiologic conditions. The Anfinsen experiments showed that this "native" structure is encoded in the amino-acid sequence [4], and that elucidating a protein's biological function from its amino acids is theoretically possible.

Modern sequencing techniques have led to an exponential growth in the number of known protein sequences, but experimental structure determination methods such as X-ray crystallography and Nuclear Magnetic Resonance (NMR) are time-consuming, expensive, and struggle to keep pace with the incoming data [5]. Development of a computational approach to complement wet-lab efforts will not only be important for elucidating the function of existing proteins, but will also advance the development of synthetically engineered proteins, improve our models of protein ligand docking for drug development, and assist in the prediction of protein-protein interactions in supramolecular assemblies [6–8].

A protein's conformation is the arrangement of its amino acids in space. A chain of amino acids is connected by many molecular bonds, both within and between each amino acid, and it is rotations of these bonds around their axes that give rise to a particular conformation. The amino acids bond end to end to form a common backbone chain, with each amino acid contributing a variable-length side chain that projects from the backbone (see Figure 1.1). A protein's conformation can be represented as a sequence of dihedral bond angles, with each amino acid contributing three backbone dihedral angles and one or more side chain angles. A forward kinematics calculation can recreate the three-dimensional structure represented by a particular conformation from these bond angles. Since even "small" proteins may

contain over 100 amino acids, this representation results in a conformational search space that is both vast and high-dimensional.

Statistical mechanics suggests that the protein conformational space may be represented by a funnel-like energy surface with the native structure at the low-energy basin [9]. The potential energy associated with a conformation is governed by the biophysics of its inter-atomic interactions; conformations with a lower potential energy are thus more likely to form in native conditions. Extensive research has gone into developing physically realistic energy functions to evaluate protein conformations [10]. However, because the computational cost of an energy function is quadratic with respect to the number of atoms, calculating potential energy at the all-atom level of detail is usually prohibitively computationally expensive.

A coarse-grained model of protein conformations is typically employed to reduce the complexity of the search space and lower computational costs. These coarse-grained models employ only two degrees of freedom per amino acid to represent the backbone chain, and typically model the side chains as a static sphere at the center of mass. The energy functions available at the coarse-grained level of detail are significantly less expensive; but the result is that they are semi-empirical, and inaccuracies result in a rugged energy surface. Together the vast size of the conformational search space and the ruggedness of the energy surface make *ab inito* protein native structure prediction a very difficult problem [11–14].

In order to tackle this vast and rugged energy surface, researchers have adopted a two-stage process [1, 2, 6, 15–17]. Stage one is a coarse-grained search for a diverse set of local minima. Stage two is the refinement of local minima at the all-atom level of detail. If a few local minima in the vicinity of the native structure are discovered in stage one, the native structure is more likely to be recovered in stage two [6].

However, the algorithms typically employed in stage one do not explicitly sample local minima. The common approach is to launch many Metropolis Monte Carlo (MMC) or Molecular Dynamics (MD) trajectories to obtain a large number of low-energy decoy conformations. Further analysis of these decoy conformations groups them by geometric similarity in an attempt to reveal explored local minima. A new search framework introduced by our

lab, FeLTr, incorporates this geometric analysis into the stage one search process, but still fails to explicitly model local minima [18, 19]. This thesis addresses this shortcoming by implementing a new coarse-grained conformational search algorithm.

This thesis introduces a novel memetic algorithm, Protein Local Optima Walk (PLOW), that explicitly populates local minima in the coarse-grained energy surface. PLOW, like the Iterated Local Search (ILS) framework, combines a global search with an exploitative local search [20, 21]. The global search allows the algorithm to explore the breadth of the energy surface, biasing sampling towards lower-energy regions, while the local search optimizes each exploration at the global level to the closest low-energy local minimum. PLOW essentially projects the protein conformational space onto a space containing only local minima. The global search is then able to more effectively sample a wide range of near-native conformations from this projected space.

The FeLTr framework combines multiple MMC trajectories into a single efficient tree-based search. This thesis combines multiple PLOW trajectories into a memetic version of FeLTr. In memetic FeLTr each MMC trajectory is replaced by the perturbation and local search functions employed by PLOW. This approach significantly improves the ability of FeLTr to sample near-native conformations. The use of local search to discover local minima provides a straightforward way to incorporate multiple fragment lengths into the FeLTr framework, further improving its sampling ability.

The rest of this chapter provides background on the protein structure prediction problem and covers recent advances in structure prediction protocols. Section 1.3 explains the essential contribution of this work. Chapter 2 describes the PLOW algorithm and the enhanced memetic version of FeLTr. Chapter 3 benchmarks PLOW and memetic FeLTr on 12 diverse proteins, comparing them both to the original FeLTr framework as well as to published results from two other research groups [1, 2].

1.1 Background

1.1.1 Representation of Protein Chains

Proteins are chains of amino acids that vary in length from ten or fewer amino acids in short polypeptides to hundreds of amino acids in complex protein molecules. The arrangement of amino acids in space determines the three-dimensional conformation of a protein. Each amino acid has a backbone, which connect end-to-end to form the protein chain, and a side chain, which projects from the backbone (see Figure 1.1). A protein chain is held together by the molecular bonds between the atoms of each amino acid. Each amino acid's backbone has three backbone dihedral bond angles, ϕ, ψ, and ω, and one or more side chain angles. A protein conformation is then represented by a vector of these bond angles.

The space of possible conformations for a given protein consists of all possible permutations of the vector of dihedral bonds. Since even small proteins can contain hundreds of amino acids, the conformational search space is both vast and high dimensional. In practice, however, a protein's backbone chain may be accurately represented by only the ϕ and ψ angles. This coarse-grained representation significantly contracts the search space and reduces the number of atoms which must be evaluated to calculate the potential energy of a conformation. This thesis models a protein backbone with the N, C_α, C and O atoms, whose positions can be determined from the ϕ and ψ angles through forward kinematics. Side chains are estimated using a single static C_β atom, however, once the backbone of a protein has been discovered, existing all-atom refinement techniques can be used to accurately recreate the side chains [22, 23].

1.1.2 Evaluation of Potential Energy

Protein conformations are evaluated by functions that measure the degree of potential energy present. A particular energy function thus defines the energy surface for the conformational search space, with the native structure residing at the low-energy basin. At the coarse-grained level of detail, available energy functions provide only a rough estimate

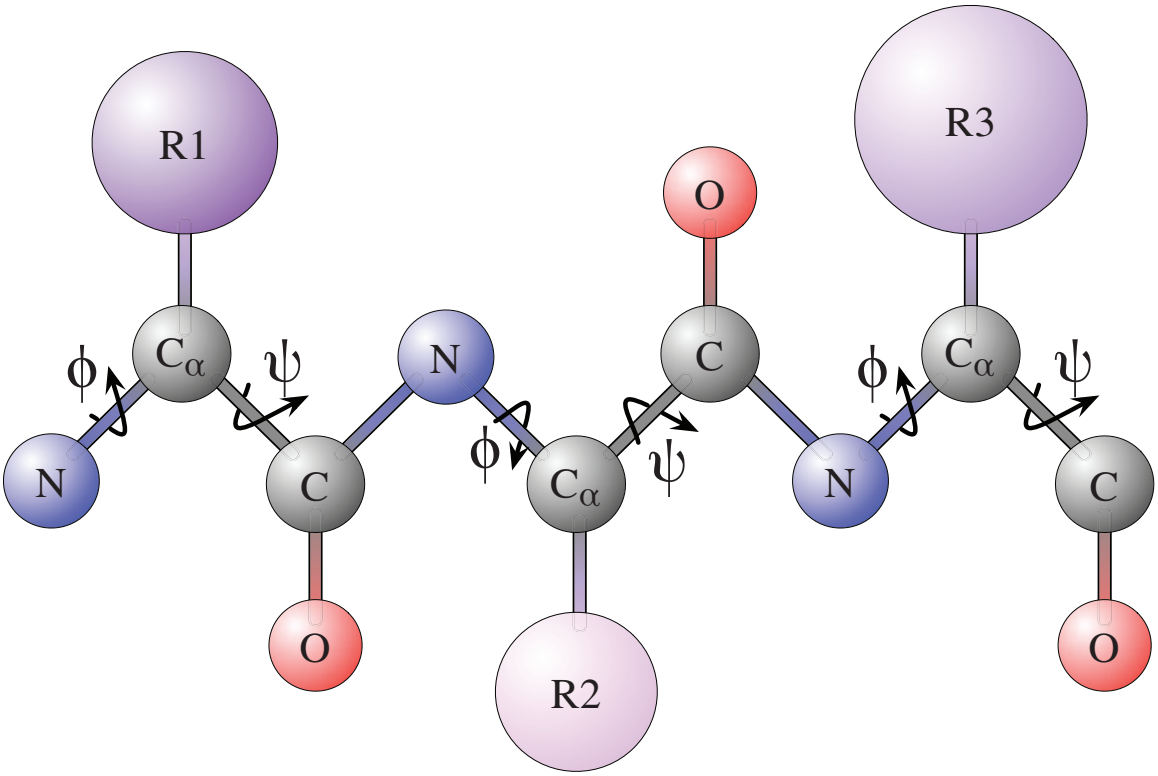


Figure 1.1: A coarse-grained representation of a three amino acid long chain is shown. The side chain for each amino acid is represented as a single "R" group. The ϕ and ψ dihedral torsion angles represent the only degrees of freedom in this model.

of the true potential energy. However, modern coarse-grained energy functions serve as effective objective functions for a stochastic search of the conformational space.

The algorithms presented in this thesis employ a modified version of the Associative Memory Hamiltonian with Water (AMW) energy function [48]. The value of AMW is the sum of the six terms given : $Energy_{AMW} = E_{Lennard-Jones} + E_{H-Bond} + E_{contact} + E_{burial} + E_{water} + E_{Rg}$.

$E_{Lennard-Jones}$ is implemented as the 12-6 Lennard-Jones potential in AMBER9 [49], modified to allow a soft penetration of van der Waals spheres. The E_{H-Bond} term accounts for local and non-local hydrogen bond formation. The terms $E_{contact}$, E_{burial}, and E_{water} allow for non-local contacts, a hydrophobic core, and water-mediated interactions, respectively. The E_{Rg} term measures the difference between the radius of gyration (Rg) of a

conformation and the Rg value predicted for its sequence, given its length [50]. The E_{Rg} term rewards conformations which are more compact, since native-like conformations tend to be compact with a dense and hydrophobic core.

1.1.3 Distance between two Protein Conformations

The ability to compare two different conformations of the same protein sequence is important for measuring the results of a structure prediction algorithm, and many methods have been developed for this purpose [13, 24–26]. This section briefly describes two of the most popular methods: least Root Mean Square Deviation (lRMSD) and Global Distance Test (GDT). Chapter 3 employs both lRMSD and GDT to compare decoy conformations generated during a search to the native structure downloaded from the Protein Data Bank (PDB) [27].

least Root Mean Square Deviation (lRMSD)

The lRMSD between two conformations measures the mean distance in Å between the atoms of the two aligned structures. The two conformations are aligned by center of mass, and a rotation matrix is applied to minimize the Euclidian distance between corresponding atoms in each conformation. The lRMSD is then the RMS distance between corresponding atoms of the aligned structures. The lRMSD can be calculated using every atom or a subset of the backbone atoms. Since the algorithms presented in this thesis model the protein backbone, the backbone atoms N, C_α, C, and O are used to calculate lRMSD.

A protein in native biological conditions is not a static structure, but will fluctuate between an ensemble of native conformations [28, 29]. Therefore, when measuring lRMSD from the native structure, a small level of error is allowed. The level of error is different for each protein. A value of 2Å provides a conservative estimate applicable to the proteins modeled in this thesis.

Global Distance Test (GDT)

The accuracy of lRMSD diminishes as the difference between two conformations increases. A recently proposed method for comparing conformations, GDT, overcomes this limitation [25]. To perform a GDT, two conformations are aligned as in lRMSD. However, rather than calculating the total RMS distance, the GDT measures the number of C_α atoms which are within a threshold distance of each other. Typically the GDT is computed for several threshold values and the average result from each threshold is reported as a percentage of C_α atoms under the threshold. This thesis uses the most common set of thresholds, which is known as GDT_TS: 1Å, 2Å, 4Å, and 8Å. Calculating the optimal alignment for GDT is a much more computationally difficult problem than lRMSD [30]. This thesis employs an approximation of the true optimal alignment as described in [25].

1.2 Related Work

1.2.1 Trajectory-based Exploration

Molecular Dynamics (MD)

MD approaches attempt to simulate the atomic forces at work within a protein molecule by applying the principles of Newtonian physics [31]. An MD simulation calculates the forces exerted by each atom in a protein on every other atom. An MD trajectory simulates a specific period of time, calculating the interatomic forces at each time step and updating the position and momentum of each atom accordingly. MD has the advantage of modeling the actual folding pathway of a protein. However, as the number of amino acids in the target protein grows, the number of atomic interactions which must be computed at each time step increases quadratically. Early milestones in protein structure prediction were achieved using MD approaches [32]. However, given the computational complexity, MD is typically only applied to very small proteins or to carry out fine-grained refinements of existing conformations.

Monte Carlo (MC)

Monte Carlo (MC) based methods sample the conformational space by making a series of modifications or moves to a conformation. Each resulting conformation is evaluated with a potential energy function, and a determination is made as to whether or not to accept or reject the move based on this function. The goal is to drive an MC trajectory towards lower energy conformations which are, in theory, closer to the native structure. The decision of whether or not to accept a move is typically done using the Metropolis criterion [33]; MC methods using the Metropolis criterion are referred to as MMC.

1.2.2 Enhanced Sampling Strategies

Modern structure prediction strategies enhance the sampling ability of trajectory-based exploration methods with parallel execution, varying temperature, and exchanging the seed conformation from which new trajectories are launched. Some of the recent approaches which have been successful are importance sampling, simulated annealing, umbrella sampling, genetic algorithms, replica exchange (also known as parallel tempering), local elevation, activation relaxation, local energy flattening, jump walking, multicanonical ensemble, conformational flooding, Markov state models, discrete time-step MD, and Fragment-based Assembly (FA) [34]. This section outlines several recent approaches which are employed to benchmark the results described in chapter 3.

Fragment-based Assembly (FA)

A modification or move to a protein conformation is the rotation of the backbone around one or more of the dihedral bonds. Extensive research has shown that the use of bond angles found in nature significantly improves sampling of near-native conformations over simply rotating angles uniformly at random [35]. FA replaces the dihedral bond angles with values found in the PDB [27]. For each amino acid in the protein, a library of fragments is defined at the beginning of a search. When a position i in the conformation is to be modified, a corresponding fragment is selected at random from the library. The dihedral angles from

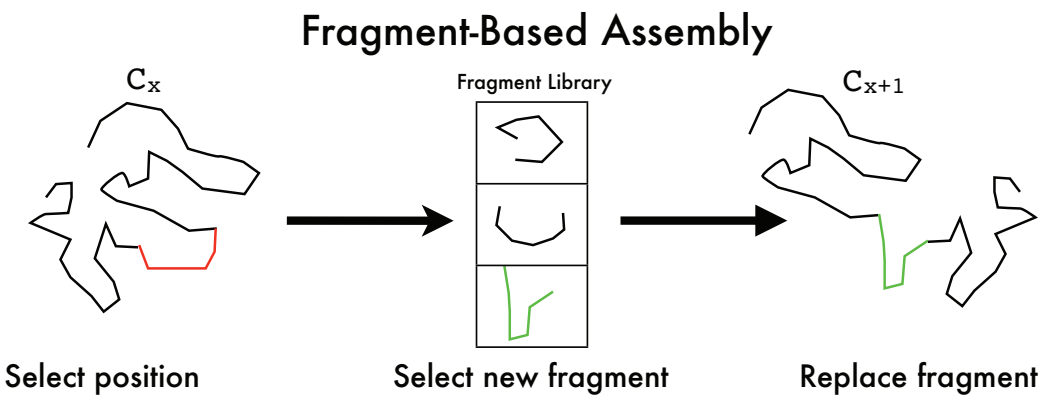


Figure 1.2: A library of fragments taken from the PDB is defined at the beginning of a search. When a position i (shown in red) in the conformation is to be modified, a corresponding fragment is selected at random from the library. The dihedral angles from the conformation are then replaced with those from the fragment, beginning at position i (shown in green).

the conformation are then replaced with those from the fragment, beginning at position i (see Figure 1.2). The use of a subset of dihedral angles greatly contracts the search space and directs sampling towards local structural motifs which have been previously observed in nature. While the goal of a fragment library is to bias search towards structures seen in the PDB, a sufficiently diverse fragment library will also allow for the generation of novel structures. The fragment library used in this thesis is outlined in Chapter 2. The use of FA as the move set in MMC has been shown to greatly improve its sampling ability, so MMC and FA form the basis of most modern protein structure prediction protocols.

MMC-based approaches

Successes in protein structure prediction have given rise to a common template among structure prediction algorithms [1,2,6,15–17]. Many MMC trajectories are run at a coarse-grained level of detail to generate a large sample of low-energy decoy conformations. These decoy conformations are clustered by geometrical similarity in order to highlight centroids that represent a broad range of low-energy local minima [19,36]. These centroids are then

refined at the all-atom level of detail. If the first stage finds enough near-native coarse-grained conformations, the second stage all-atom refinement will, in theory, be able to recover the native structure.

Running many independent MMC trajectories has the advantage of being highly parallelizable, however, there is no guarantee that the independent trajectories will not all converge to the same region of the search space. In order for the all-atom refinement to be successful, the coarse-grained search must sample a broad enough range of local minima that the native structure may be reached from one of them. Brunette and Brock proposed an iterative approach that uses the results of the all-atom refinement as input to a new coarse-grained search [2]. The idea is to use periodic short all-atom refinements to help guide the search at the the coarse-grained level. This method allows the algorithm to dynamically re-apportion computational resources to more promising areas of the energy surface.

The Sosnick group also employs an iterative approach that focuses resources based on an increasingly refined picture of the search space [1]. Their algorithm, as in FA, employs a biased move set of dihedral bond angles consisting of a probability distribution corresponding to frequency in the PDB. The algorithm performs an iterative set of coarse-grained MMC trajectories using these biased move sets. After each iteration, the probabilities for each move set are updated based on their appearance in the search. In practice, this approach allows the algorithm to accurately predict the local secondary structural motifs of the target protein, and thus re-apportion computational resources to the regions of the conformational search space that correspond to the predicted secondary structure.

Both of these approaches use an iterative approach to more efficiently direct independent MMC trajectories. However, neither method directly address the issue of geometric diversity in the results of the coarse-grained search. Furthermore, neither approach explicitly samples local minima in the energy surface; both rely on a post-processing clustering analysis to approximate local minima.

1.2.3 Projection-guided Tree-based Exploration

Our previously published FeLTr framework attempts to ensure a geometrically-diverse conformational sampling at the coarse-grained level by employing a geometric projection layer [18, 19]. The algorithm grows a search tree in the conformational space by expanding selected conformations with short MMC trajectories, and maintains a representative ensemble of previously visited conformations in memory. Selection from this ensemble is biased towards low-energy conformations and regions in under-explored areas of the conformational space. FeLTr is thus able to dynamically redirect computational resources at the global level to ensure a degree of geometric diversity in its conformational sampling. This section briefly describes the key components of FeLTr. A detailed description is provided in recent publications [18, 19].

FeLTr explores the protein conformational space with a tree-based search shown in Figure 1.3. Algo. 1 provides pseudo-code for the framework. FeLTr executes on a target protein sequence α and produces an output ensemble Ω_α of low-energy decoy conformations. The search tree is initialized with the extended conformation at the root (Algo. 1, lines 1-2). Each iteration of the search selects a vertex from the tree for expansion via a short MMC trajectory. The result of this MMC trajectory is then added to the search tree as a new vertex. FeLTr employs a two-level selection process to bias selection towards both low-energy and geometrically diverse conformations. This allows FeLTr to combine multiple MMC trajectories into a single search, which is a more effective allocation of computational resources.

Selection of a vertex for expansion is a two step process, starting with selection of an energy level ℓ (Algo. 1, line 4). Each decoy conformation C is projected onto a one-dimensional grid of energy levels with increments of 2 kcal/mol. Energy levels are given a weight $w(\ell) = E_{\text{avg}}(\ell) \cdot E_{\text{avg}}(\ell)$. A level ℓ is then selected at random with probability $w(\ell) / \sum_{\ell' \in \text{Layer}_E}(\ell')$.

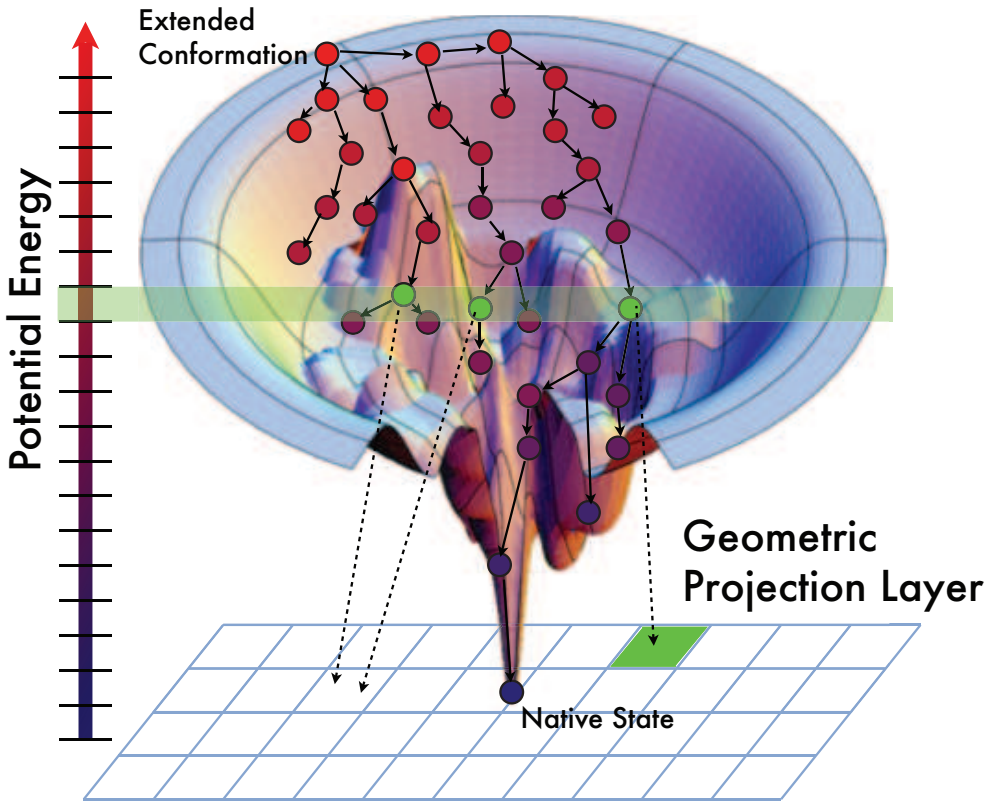


Figure 1.3: The FeLTr search tree is initialized with the extended conformation at the root. Each iteration of the search selects a vertex from the tree for expansion via a short MMC trajectory. The result of this MMC trajectory is then added to the search tree as a new vertex. FeLTr employs a two-level selection process to bias selection towards both low-energy and geometrically diverse conformations. In this example, first the energy level highlighted in green is selected. Then one of the three vertices within that energy level is selected for expansion based on the geometric projection layer.

The second step chooses a geometric cell within the selected ℓ (Algo. 1, line 5). Conformations are projected onto grid cells based on geometric shape using three selected coordinates from the Ultrafast Shape Recognition (USR) method [37,38]. A second weighting function ranks each cell according to the formula $w(cell) = 1.0/[(1.0 + \texttt{nsel}) \cdot \texttt{nconfs}]$. The variable `nsel` represents the number of times a cell has been previously selected, and `nconfs` represents the number of conformations discovered which project into that cell. Finally, a C is selected uniformly at random from the set of conformations which lie in both ℓ and the selected geometric grid cell (Algo. 1, line 6). This selection process allows FeLTr to

bias sampling of the conformational space towards low-energy conformations in unexplored regions of the conformational space.

The selected conformation C is expanded via a short MMC trajectory, resulting in a new conformation C_{new} (Algo. 1, line 7). The trajectory length is $n-2$ moves, where n is the number of amino acids in the target protein. Each MMC move consists of a random trimer fragment replacement using the fragment library described in section 2.3. The energy function used to evaluate each move is a modified version of the AMW function described in section 1.1.2. The resulting C_{new} is then added as a new vertex in the search tree (Algo. 1, line 8) and to the output ensemble Ω_α (Algo. 1, line 9).

Algo. 1 A high-level description of the FeLTr framework is given as pseudo code.

Input: α, amino-acid sequence
Output: ensemble Ω_α of conformations

1: $C_{\text{init}} \leftarrow$ extended coarse-grained conf from α
2: AddConf(C_{init}, Layer$_E$, Layer$_{\text{Proj}}$)
3: **while** Time AND $|\Omega_\alpha|$ do not exceed limits **do**
4: $\quad \ell \leftarrow$ SelectEnergyLevel(Layer$_E$)
5: $\quad$ cell $\leftarrow$ SelectGeomCell(ℓ.Layer$_{\text{Proj}}$.cells)
6: $\quad C \leftarrow$ SelectConf(cell.confs)
7: $\quad C_{\text{new}} \leftarrow$ ExpandConf(C)
8: $\quad$ AddConf(C_{new}, Layer$_E$, Layer$_{\text{Proj}}$)
9: $\quad \Omega_\alpha \leftarrow \Omega_\alpha \cup \{C_{\text{new}}\}$

Recent work shows that the FeLTr framework samples near-native conformations more effectively than MMC-based methods [18, 19, 39, 40]. Like other coarse-grained sampling methods, FeLTr does not explicitly sample local minima, but rather relies on clustering analysis to filter its results down to a subset of conformations which will hopefully correspond to local minima. Cluster centroids, however, are only approximations of true local minima, and analysis shows that promising conformations are frequently discarded during clustering.

1.2.4 Evolutionary Approaches for Exploration

Protein structure prediction has been shown to be NP-hard [41], making metaheuristic and evolutionary computation approaches attractive. Many studies have favorably described the use of evolutionary frameworks for navigating the highly rugged energy surface presented by the conformational search space [42–45]. Techniques adopted from the evolutionary computation community, however, have failed to compete with MMC-based approaches, as they often rely on simplistic representations and energy functions and fail to use widely accepted techniques such as FA.

Work in [46] evaluates the use of a canonical evolutionary framework using realistic physics-based energy functions. This work shows that an *ab initio* evolutionary algorithm can effectively recreate the native structure for a single short protein. The protein modeled, however, is only 5 amino acids in length and thus does not represent a significant computational challenge.

Memetic algorithms combine a global search technique with short local optimizations. This approach allows an algorithm to explicitly probe local minima in a rugged energy surface by projecting each move at the global level to a nearby local minimum. Two studies using lattice models were able to successfully recreate the native structure using a memetic Genetic Algorithm (GA) where the offspring from crossover are refined with either gradient descent or MMC [43, 44]. Lattice models, however, oversimplify protein structure, making them unsuitable for real-world applications.

A subsequent study employs a memetic GA with the physically realistic CHARMM [47] energy function [45]. The authors show that the memetic GA consistently finds conformations of lower energy than either a standard GA or MMC search. However, simply finding the single lowest energy conformation is rarely sufficient to discover the true native structure. Indeed, the memetic GA search often found a lower energy than the native structure taken from the PDB.

Memetic algorithms are especially useful for highly constrained spaces like the protein

energy surface. The protein conformational space contains many regions that are energetically infeasible, and many conformations allowed by a dihedral bond representation result in steric clashes and are thus physically unrealistic. Even a small change to a low-energy conformation can easily result in an infeasible structure with a very high energy. A memetic approach deals with these infeasible regions by efficiently moving a conformation in a constrained region to a nearby unconstrained region of the search space.

Research from the evolutionary computation community suggests that memetic approaches to the protein structure prediction problem hold promise. However, further work is needed to apply these initial studies to real world structure prediction problems. This thesis combines cutting edge stochastic optimization strategies from the evolutionary computation community with established procedures for assembly of coarse-grained structures and analysis of results.

1.3 Contribution of this Work

This thesis explores explicit sampling of local minima along the coarse-grained energy surface. The essential idea is to effectively project the conformational search space onto the sub-space containing only local minima. This dramatically reduces the size of the search space, and analysis shows that restricting the search space in this way does not exclude near-native conformations (section 3.3). This goal is achieved through a memetic approach borrowed from evolutionary computation. The idea is to conduct a two-level search, combing a global search method at the outer level and a local search method at the inner level. Each move made at the global level is projected onto a nearby local minimum by a short local refinement.

The effectiveness of this memetic approach is shown by applying it to MMC. The PLOW algorithm uses an ILS framework to adapt MMC to move over the space containing only local minima. The principles employed in PLOW are then applied to the FeLTr framework to build a powerful memetic algorithm that samples a diverse set of local minima.

This thesis begins to bridge the gap between the advanced optimization and search

algorithms developed by the evolutionary community and state-of-the-art domain-specific solutions developed by the protein structure prediction community. Recent advances in the structure prediction community focus on improving the move sets and energy evaluation functions available for traditional MMC explorations. Evolutionary computation and metaheuristics, on the other hand, offer many rigorously tested approaches to tackling high-dimensional search problems with rugged objective functions. Attempts to apply frameworks like GA to protein structure prediction typically fall short because proven domain specific protocols are not incorporated. Our research objective is to draw from advances in both fields to develop novel approaches that advance our understanding of the protein structure prediction problem as well as generalized optimization problems.

Chapter 2: Methods

This chapter describes the algorithms developed to explicitly sample local minima in the protein energy surface. Section 2.1 describes the new PLOW algorithm presented by this thesis. Section 2.2 describes a hybrid algorithm which combines the original FeLTr framework with the local search aspects of PLOW. Section 2.3 briefly describes an enhanced fragment library which is employed as the move set for both FeLTr and PLOW.

2.1 Protein Local Optima Walk (PLOW)

PLOW employs a two layer search process to explore the space of local minima. The outer layer (shown in Algorithm 2) simulates a MMC search at the global level, while the inner layer performs a local hill-climbing search to project each point found in the outer layer onto a nearby local minimum.

PLOW employs a PERTURBATION function to easily overcome local energy barriers by jumping out of its current local minimum, H, to a nearby region of space, H_{new} (Algo. 2, line 5). H_{new} is then projected onto a nearby local minimum by the LOCALSEARCH function (Algo. 2, line 6). The ACCEPTANCECRITERION function decides whether to keep the home base at H or move it to H_{new} (Algo. 2, line 7). A PLOW search trajectory is illustrated in Figure 2.1.

The initial location of the search is determined by the INITIALSELECTION function (Algo. 2, line 3). Jumping from one local minimum to the next proceeds until a specified number of energy function evaluations have occurred (Algo. 2, line 1). Because each call to LOCALSEARCH performs a variable number of evaluations, the $Eval_{count}$ variable is incremented within the LOCALSEARCH function (Algo. 2, line 6).

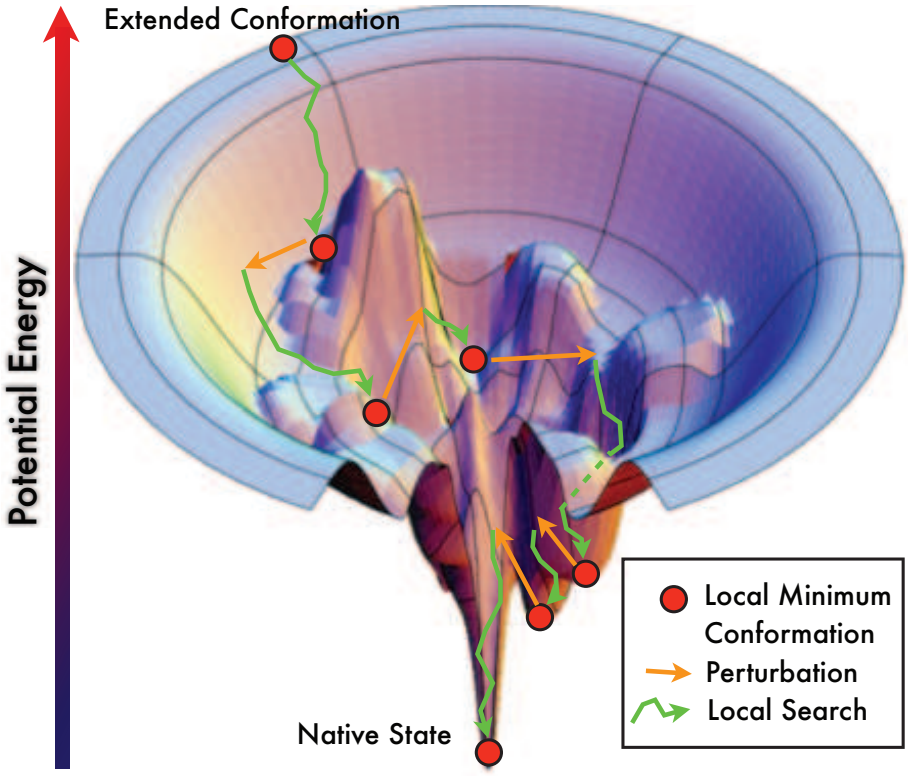


Figure 2.1: PLOW explores the space of local minima in the protein energy surface. The PERTURBATION function can easily overcome local energy barriers by jumping out of the current local minimum to a nearby point in the space. This point is then projected onto a nearby local minimum by the LOCALSEARCH function.

PLOW adapts this general framework into an algorithm suitable for navigating the complex protein energy surface. Sections 2.1.1, 2.1.2, 2.1.3, and 2.1.4 define the new domain-specific implementations of INITIALSELECTION, LOCALSEARCH, PERTURBATION, and ACCEPTANCECRITERION, respectively, employed in PLOW.

2.1.1 Initial Selection

The INITIALSELECTION function initializes H as a fully extended conformation with a very high energy value. H is then projected onto its nearest local minimum using the LOCALSEARCH function, as described in section 2.1.2. While an ILS is typically initialized to a random state, the extended conformation has specific desirable properties and is commonly

Algo. 2 The canonical Iterated Local Search (ILS) framework is shown. This work defines domain-specific implementations of InitialSelection, LocalSearch, Perturbation, and AcceptanceCriterion in sections 2.1.1, 2.1.2, 2.1.3, and 2.1.4 respectively.

Input: Maximum number of energy function evaluations

1: $\text{Eval}_{max} \leftarrow$ *(UserDefined)*
2: $\text{Eval}_{count} \leftarrow$ *0*
3: H $\leftarrow$ InitialSelection()
4: **while** $\text{Eval}_{count} < \text{Eval}_{max}$ **do**
5: $\quad \text{H}_{new} \leftarrow$ Perturbation(H)
6: $\quad \text{H}_{new}$, $\text{Eval}_{count} \leftarrow$ LocalSearch(H_{new}, Eval_{count})
7: $\quad$ H $\leftarrow$ AcceptanceCriterion(H,H_{new})

used as a starting point by the structure prediction community.

2.1.2 Local Search

The LocalSearch function in Algorithm 2 projects a conformation onto a nearby local minimum using a local search incorporating the fragment library and the coarse-grained energy function outlined in sections 2.3 and 1.1.2, respectively. At each iteration, the local search generates a child conformation by performing a single fragment replacement. If the energy of the child conformation is lower than that of its parent, the child conformation replaces its parent; otherwise the child is discarded. This local search process is repeated until k children in a row have been discarded, ostensibly indicating the presence of a local minimum. When this occurs, LocalSearch stops and returns the current conformation. The value of k is set to the length of the target protein. LocalSearch thus encapsulates the precise definition of a local minimum.

PLOW provides a straightforward mechanism for incorporating multiple-length fragments into the search. When fragments of both length 9 and length 3 are employed in the search, the LocalSearch function is repeated in serial for each fragment length. First LocalSearch is run using 9-mers until a local minimum is reached. Then LocalSearch is repeated starting at the local minimum using trimers until a second local minimum is reached.

2.1.3 Perturbation

Perturbation modifies a conformation just enough to jump out of its current local minimum, such that the LocalSearch function is unlikely to return it to the same local minimum. However, if Perturbation makes too drastic a change, then the search is unable to benefit from knowledge of the previous local minimum.

Low-energy conformations tend to be compact and leave little room for movement in their backbone chain without raising their energy. Therefore, even a single random fragment replacement to a structure that is already at a local minimum may disrupt a conformation enough to greatly increase its energy score. Such a perturbed conformation will share nearly all of its local structural features with its parent, but the new conformation will have a much higher energy and a significantly altered overall global structure.

Given a high energy score, the LocalSearch function will be able to easily optimize the perturbed conformation to one of many distinct local minima, leaving little chance that it will return to its previous local minimum. Because most of the local structural features are maintained in the perturbed conformation, LocalSearch will still benefit from previous knowledge of these local structures. For this reason we found that a single trimer fragment replacement serves as an effective Perturbation function.

2.1.4 Acceptance Criterion

After each H_{new} has been projected onto a local minimum by the LocalSearch function, AcceptanceCriterion uses the Metropolis Criterion to decide if the algorithm will move its home base to H_{new} or remain at the current value of H [33]. The algorithm will always move to H_{new} if it is of lower energy than H. If H_{new} has a higher energy than H, then the algorithm will still move to H_{new} with a small probability – 10 kcal/mol jumps in energy occur with a 0.1 probability [19].

2.2 Memetic FeLTr

Memetic FeLTr draws on the strengths of both FeLTr and PLOW. FeLTr provides an efficient implementation of many MMC trajectories with a bias towards geometric diversity and PLOW provides a method for directly sampling a diverse set of local minima. Memetic FeLTr employs the same selection process described in section 1.2.3, projecting sampled conformations with a two-level projection layer employing both potential energy and geometric diversity. In the expansion step, the short MMC trajectory is replaced by a call to PERTUBATION followed by a call to LOCALSEARCH on the selected C to produce a C_{new}. PERTUBATION and LOCALSEARCH are implemented as described in sections 2.1.3 and 2.1.2, respectively.

The result is a tree-based search over the sub-space of local minima with a sampling bias towards unexplored regions of the conformational space. Besides merely sampling local minima, memetic FeLTr gains two advantages with the use of perturbation and local search. PERTUBATION makes it easy for FeLTr to escape its current local minimum, while still retaining local structural features. The LOCALSEARCH function runs for a dynamic number of iterations based on the complexity of the target protein and the depth of the current local minimum, allowing FeLTr to fully explore each local minimum during a tree expansion.

2.3 Fragment Library

The fragment library for a protein sequence α is generated by matching short amino acid sequences of length k from α to corresponding sequences in proteins with known structure. This thesis employs an enhanced fragment library which additionally includes fragments which match based on local structural similarity. A Multiple Sequence Alignment (MSA) finds other known proteins which have sequences similar to α. The PSI-BLAST [51] program analyzes the results of the MSA to produce an alternate sequence of amino acids which can replace α at a given position i. The result is a position-specific profile of α which contains

not only the actual sequence at each position i to $i + k$, but also the set of alternate amino acid k-mers for that position. These alternate sequences are then used to build a list of fragments for the position i in α. Finally, a filtering step is performed to improve the quality of the resulting fragment configurations. PSI-PRED [52] is employed to predict the secondary structure for α. Only fragments which correspond to the predicted secondary structure are included in the final fragment library.

Structure prediction protocols use fragments varying anywhere from 3 to 19 amino acids in length. It is generally accepted that a fragment length of 3 is necessary to make fine adjustments in order to reach a protein's native structure. However, longer fragment lengths allow an algorithm to benefit from larger repeating motifs which are common in the PDB. A common approach uses longer fragment lengths in an initial phase to quickly narrow in on a native-like conformation, followed by one or more phases which employ shorter fragment lengths for more detailed refinement.

This thesis compares each approach using only fragments of length three as a baseline (see Table 3.3). However, the addition of fragments of length 9 has been shown to significantly improve conformational sampling [53]. Therefore, fragments of both length 9 and length 3 are employed in PLOW and memetic FeLTr when comparing them to results published by other research groups (see Table 3.4).

Chapter 3: Results

This chapter presents the results of running experiments on the PLOW and FeLTr algorithms described in Chapter 2. Section 3.1 lists the proteins targeted in this study and section 3.2 describes the experimental procedure employed. Section 3.3 analyzes the correspondence between the native structure and coarse-grained local minima. Section 3.4 then compares the PLOW algorithm to the previous implementation of FeLTr. Sections 3.5 and 3.6 show how FeLTr is improved by incorporating the memetic approach in PLOW. Section 3.7 compares results obtained by PLOW and memetic FeLTr using multiple fragment lengths to two other state-of-the-art structure prediction algorithms. Section 3.8 analyzes the effectiveness of the perturbation function employed by memetic FeLTr. Finally, section 3.9 showcases the result of an all-atom refinement on the best structures discovered by PLOW.

3.1 Protein Systems of Study

Table 3.1 lists the 12 protein systems investigated in this thesis. These proteins range in length from 61 to 123 amino acids and cover a range of α and β fold topologies. These proteins were selected from studies performed by other research groups so that results could be compared not only to previous work, but also to other state-of-the-art structure prediction protocols.

Table 3.1: Protein systems targeted in this study are listed. Length, fold topology, and the percentage of amino acids which form α helices and β sheets are given.

	PDB ID	length	fold	% α	% β
1	1ail	70	α	84	0
2	1aoy	78	α/β	42	14
3	1cc5	83	α	37	0
4	1csp	67	β	0	28
5	1dtdB	61	α/β	10	44
6	1fwp	69	α/β	17	23
7	1hhp	99	β	0	49
8	1sap	66	α/β	21	32
9	1wapA	68	β	0	63
10	2ezk	93	α	70	0
11	2h5nD	123	α/β	74	5
12	2hg6	106	α/β	25	19

3.2 Experiments and Measurements

All experiments are run for 10,000,000 coarse-grained energy function evaluations. Over 90% of the CPU time consumed by all of the algorithms described is spent calculating potential energy. Furthermore, the computational cost of an energy function is directly related to the length of the protein. Therefore, setting the number of calls to the energy function constant ensures a fair comparison across all algorithms and on a variety of protein lengths. Since all algorithms use the same energy function, this also masks any differences in implementation efficiency between each algorithm. In practice, each experiment takes about two to four days of CPU user time on a 2.66 GHz Opteron processor, depending on the length of the target protein.

The lRMSD and GDT_TS results reported are calculated by comparing the native structure to each conformation in the output ensemble Ω_α. For all variants of FeLTr, Ω_α consists of the conformations added to the search tree. For PLOW, Ω_α consists of every local minima

discovered during the search.

3.3 Analysis of Local Minima

If an ideal energy function were available to score each conformation, the native structure would lie at the global energy minimum. However, it is known that coarse-grained energy functions, like the one used in this study, contain significant inaccuracies at lower-energy levels. Nonetheless, it is expected that the native structure will lie near some local minimum in the energy surface, if not the global minimum. This assumption is particularly important in this study, as we restrict the search space to only local minima. Therefore, if this assumption does not hold, then it will be impossible for PLOW or memetic FeLTr to reach the native structure.

To test this assumption, repeated hill-climbing searches are run starting from the native structure downloaded from the PDB. Column 5 of Table 3.2 shows the distance between the native structure and the nearest local minimum discovered by the hill-climber. For 11 out of the 12 proteins investigated in this study, this distance is less than 3Å lRMSD and a distance of 3Å lRMSD can typically be overcome by an all-atom refinement in a later stage [54]. This validates the assumption that a search method can achieve near-native conformations while only considering the subset of conformations which reside at local minima. Of note is the fact that the single case where a local minimum is not found within 3Å of the native structure is also the only case in which the previous implementation of FeLTr outperforms either PLOW or memetic FeLTr. This suggests that it is flaws in the chosen energy function that cause the new memetic algorithm to fail in the case of 1aoy.

3.4 Analysis of PLOW

Table 3.3 compares the results from PLOW to the previous implementation of the FeLTr framework. PLOW is able to find a structure more than 0.5Å lRMSD closer to the native structure than FeLTr for 9 out of the 12 target proteins. In the cases of 1aoy, 1csp, and

Table 3.2: The lRMSD between the native structure and the closest local minimum found when performing multiple greedy local searches starting from the native structure is given.

	PDB ID	Length	Fold	lRMSD of nearest local minimum to native (Å)
1	1ail	70	α	2.5
2	1aoy	78	α/β	3.9
3	1cc5	83	α	1.5
4	1csp	67	β	1.8
5	1dtdB	61	α/β	1.3
6	1fwp	69	α/β	0.4
7	1hhp	99	β	2.2
8	1sap	66	α/β	2.9
9	1wapA	68	β	1.5
10	2ezk	93	α	2.9
11	2h5nD	123	α/β	1.7
12	2hg6	106	α/β	2.5

1fwp both algorithms, on average, find equivalent structures. In the case of 1ail, PLOW actually finds a structure below 3Å lRMSD, which is close enough to the native structure that the difference can be overcome in an all-atom refinement [54]. For 2ezk and 2h5nD, which represent two of the longer proteins, PLOW is not only able to find the lowest average lRMSD, but also a minimum value of 4.2Å and 6.1Å , respectively. These results suggest that the explicit sampling of local minima in PLOW is able to significantly improve sampling of near-native conformations.

3.5 Analysis of Memetic FeLTr

Similar to PLOW, the memetic version of FeLTr (Mem-FeLTr) outperforms the original implementation of FeLTr by more than 0.5Å on all but four proteins (see Table 3.3). FeLTr only outperforms Mem-FeLTr in the single case of 1aoy and both algorithms find equivalent

structures, on average, in the cases of 1wapA, 1fwp, and 1csp. These results indicate that addition of perturbation and local search significantly improve memetic FeLTr's ability to sample near-native conformations. Furthermore, memetic FeLTr produces an output ensemble Ω_α several times smaller than that of FeLTr for the same number of energy function evaluations, significantly reducing the number of conformations which must be refined at all-atom detail. Table 3.3 shows that the more simplistic approach taken by PLOW is able, on average, to sample conformations closer to the native structure than the memetic version of FeLTr. However, in specific cases (1hhp, 1wapA), the best structure discovered by Mem-FeLTr across multiple runs is significantly closer to the native structure than the best structure discovered by PLOW. This suggests that further analysis is needed to best combine the strengths of both approaches.

3.6 Analysis of Extended FeLTr

The length of local search trajectories employed by both PLOW and memetic FeLTr are determined dynamically and tend to be several times longer, on average, than the fixed length MMC trajectories employed in the original implementation of FeLTr. In order to rule out the possibility that PLOW and memetic FeLTr are merely benefiting from longer local searches, an additional experiment is conducted to compare FeLTr and memetic FeLTr more directly. Ext-FeLTr uses the FeLTr algorithm described in section 1.2.3 with the length of each MMC search trajectory extended to the average PLOW search length given in Table 3.3, column 5. Table 3.3 shows that Ext-FeLTr performs slightly better, on average, than the original FeLTr algorithm, however PLOW and memetic FeLTr still outperform Ext-FeLTr for 9 out of 12 target proteins.

3.7 Comparison to Other State-Of-The-Art Methods

Here the results obtained by PLOW and memetic FeLTr are compared to those obtained by the Sosnick and Brock research groups [1,2]. For these experiments, fragments both length

Table 3.3: The lowest lRMSD from the native structure achieved is shown for both PLOW and FeLTr. The lRMSDs given are the average of three runs, with the minimum of the three runs shown in parentheses. Results for FeLTr are given for the previous implementation (FeLTr) as well as for memetic FeLTr (Mem-FeLTr) and Extended FeLTr (Ext-FeLTr). Column 5 shows the average number of iterations of each PLOW LocalSearch function. Ext-FeLTr represents the FeLTr framework using the value from column 5 as its MMC search length.

				avg local	avg (min) lowest lRMSD to native in Å			
	PDBID	len	fold	search len	PLOW	Ext-FeLTr	FeLTr	Mem-FeLTr
1	1ail	70	α	237	2.7(2.3)	4.0(3.4)	4.7(4.5)	3.5(3.3)
2	1aoy	78	α/β	258	5.4(5.2)	5.9(5.2)	5.1(4.6)	5.8(5.2)
3	1cc5	83	α	274	5.5(5.1)	6.0(4.9)	6.4(6.2)	5.5(5.4)
4	1csp	67	β	193	6.4(6.3)	7.2(6.6)	6.4(6.0)	6.7(5.9)
5	1dtdB	61	α/β	160	7.1(6.9)	7.5(7.0)	7.7(7.6)	7.1(6.9)
6	1fwp	69	α/β	210	6.5(6.3)	7.2(6.8)	6.8(6.4)	6.5(6.2)
7	1hhp	99	β	306	10.4(10.1)	11.0(9.7)	11.1(10.0)	9.9(9.3)
8	1sap	66	α/β	211	6.5(6.0)	7.2(6.8)	7.1(6.5)	6.5(5.9)
9	1wapA	68	β	199	7.2(6.7)	7.4(6.5)	7.8(7.3)	7.5(5.9)
10	2ezk	93	α	293	4.6(4.2)	5.9(4.7)	6.4(6.0)	5.0(4.4)
11	2h5nD	123	α/β	482	7.0(6.1)	8.8(8.3)	9.0(8.5)	8.3(7.8)
12	2hg6	106	α/β	376	8.9(8.1)	9.8(9.0)	10.1(9.6)	9.2(8.7)

9 and 3 are employed during the local search as described in section 2.1.2. Previous work shows that using fragments of length 9 followed by length 3 is able to significantly improve the results of FeLTr. However, in the previous implementation of FeLTr, there is no clear way to switch fragment lengths [53]. Table 3.4 shows that PLOW and memetic FeLTr are able to find significantly lower lRMSDs than [1] for 6 out of 8 target proteins. In the case of [2], PLOW and memetic FeLTr are able to find a higher GDT_TS score for two out the four proteins. For 2h5nD, which is the longest of the target proteins, PLOW's best value of 55% GDT_TS represents a significant improvement over the method used in [2], suggesting that the local-search based approach does provide a significant advantage in some cases. For 1csp and 1hhp the Rosetta-based approach employed in [2] was able to find structures

with significantly higher GDT_TS scores. This suggests that the use of all-atom detail may have a significant impact, especially on β sheet proteins.

Table 3.4: The best GDT_TS and lRMSD from the native structure achieved is shown for Mem-FeLTr and PLOW using fragments of both length 9 and length 3. Results from PLOW and Mem-FeLTr are compared to published results from the Sosnick [1] and Brock [2] research groups.

				PLOW		Mem-FeLTr		Sosnick	Brock
	PDBID	len	fold	lRMSD	GDT_TS	lRMSD	GDT_TS	lRMSD	GDT_TS
1	1ail	70	α	1.8(1.4)	82(88)	2.3(1.6)	77(81)	5.4	NA
2	1aoy	78	α/β	4.9(4.2)	62(65)	5.3(5.1)	62(63)	5.7	NA
3	1cc5	83	α	5.7(5.3)	46(49)	5.5(5.4)	46(50)	6.5	NA
4	1csp	67	β	5.9(5.7)	44(47)	6.1(5.5)	46(49)	NA	91
5	1dtdB	61	α/β	6.9(6.7)	40(41)	6.8(6.7)	43(47)	6.5	NA
6	1fwp	69	α/β	5.9(5.7)	49(52)	6.2(6.0)	46(49)	8.1	NA
7	1hhp	99	β	9.7(8.7)	29(30)	9.7(9.5)	27(28)	NA	84
8	1sap	66	α/β	6.4(5.6)	46(47)	5.8(5.1)	47(49)	4.6	NA
9	1wapA	68	β	7.2(7.0)	38(39)	7.2(7.2)	41(41)	8.0	NA
10	2ezk	93	α	4.0(3.9)	63(65)	4.0(3.3)	66(66)	5.5	NA
11	2h5nD	123	α/β	6.4(6.3)	48(55)	8.0(7.7)	41(47)	NA	33
12	2hg6	106	α/β	8.5(7.8)	29(30)	8.4(8.4)	30(30)	NA	22

3.8 Perturbation Analysis

The perturbation function described in section 2.1.3 modifies a conformation C (which is already at a local minimum) to $C_{perturb}$ such that the local search (see section 2.1.2) is unlikely to return $C_{perturb}$ to the same local minimum as C. Table 3.5 analyzes the difference between a consecutive C and C_{new} created by the perturbation function followed by a local search for memetic FeLTr using fragment lengths of three. If the difference between C and C_{new} is less than 2Å lRMSD, then it can be assumed that C_{new} returned to the same local minimum as C. For all of the target proteins, the percentage of cases where consecutive

Table 3.5: The median distance between consecutive local minima C and C_{new} is given in column 6. Column 5 represents the percent of C_{new}'s which are within 2Å lRMSD of C and thus deemed to have returned to the same local minima as C.

				lRMSD between C and C_{new}	
	PDB ID	Length	Fold	% < 2Å	median (in Å)
1	1ail	70	α	26	5.9
2	1aoy	78	α/β	20	7.3
3	1cc5	83	α	21	7.4
4	1csp	67	β	7	8.3
5	1dtdB	61	α/β	7	8.2
6	1fwp	69	α/β	12	8.1
7	1hhp	99	β	5	10.6
8	1sap	66	α/β	17	7.5
9	1wapA	68	β	7	8.6
10	2ezk	93	α	18	6.3
11	2h5nD	123	α/β	14	10.9
12	2hg6	106	α/β	15	9.9

C and C_{new}'s are within 2Å lRMSD of each other is less than 27%, with most cases under 20%. This suggests that the perturbation function is effective at jumping the search out of the current local minimum.

In addition to escaping the current local minimum, an effective perturbation function should move to a nearby region of the conformational space. If the move is too large, then the perturbation function simply devolves into random restart. Figure 3.1 illustrates the correlation between the lowest lRMSD from native achieved in Table 3.3 column 9 and the median distance between consecutive local minima C and C_{new} in Table 3.5, column 6. The correlation in Figure 3.1 is near linear, with a larger lRMSD from native corresponding to a larger median local minima distance. This suggests that the proteins where PLOW and Mem-FeLTr were able to find low lRMSD from native conformations were cases in which the perturbation function not only was able to jump out of the current local minimum, but also did not move the search too far away in finding the next local minimum.

Figure 3.2 illustrates the distribution of lRMSDs between consecutive C and C_{new}'s for two of the target proteins. The area of the curve shaded in red represents the portion for which C_{new} is within 2Å of C and thus deemed to have returned to its previous local minimum. In Figure 3.2a, 2EZK is an example of target where memetic FeLTr was effective at finding near-native conformations. In this case the distribution contains a large area of short to medium distance moves from 2 to 8Å lRMSD. In contrast, Figure 3.2b, 1HHP, is an example of a target where memetic FeLTr was not able to find any conformations near the native structure. Correspondingly, the distribution of consecutive local minima distances contains much higher distances, with most of the area under the curve above 8Å lRMSD. This suggests that in the case of 1HHP, the perturbation function is approaching random restart.

3.9 All-Atom Refinement

PLOW and FeLTr are designed as the first, coarse-grained, stage in a multi-stage refinement process. The computational resources required to complete a full stage-two all-atom refinement are prohibitive for this thesis. Therefore, an all-atom refinement is performed using the Rosetta Relax program on the best structure generated by PLOW from the results in Table 3.4 [55]. The result of each all-atom refinement is shown in Figures 3.3 and 3.4. The predicted structure is given in red and superimposed over the native structure downloaded from the PDB [27]. Visual inspection indicates that PLOW accurately predicts the native structure for 1ail and approaches the native structure for the other targets with a significant percentage of α helices. In most cases, however, PLOW struggles to accurately predict the formation of β sheets.

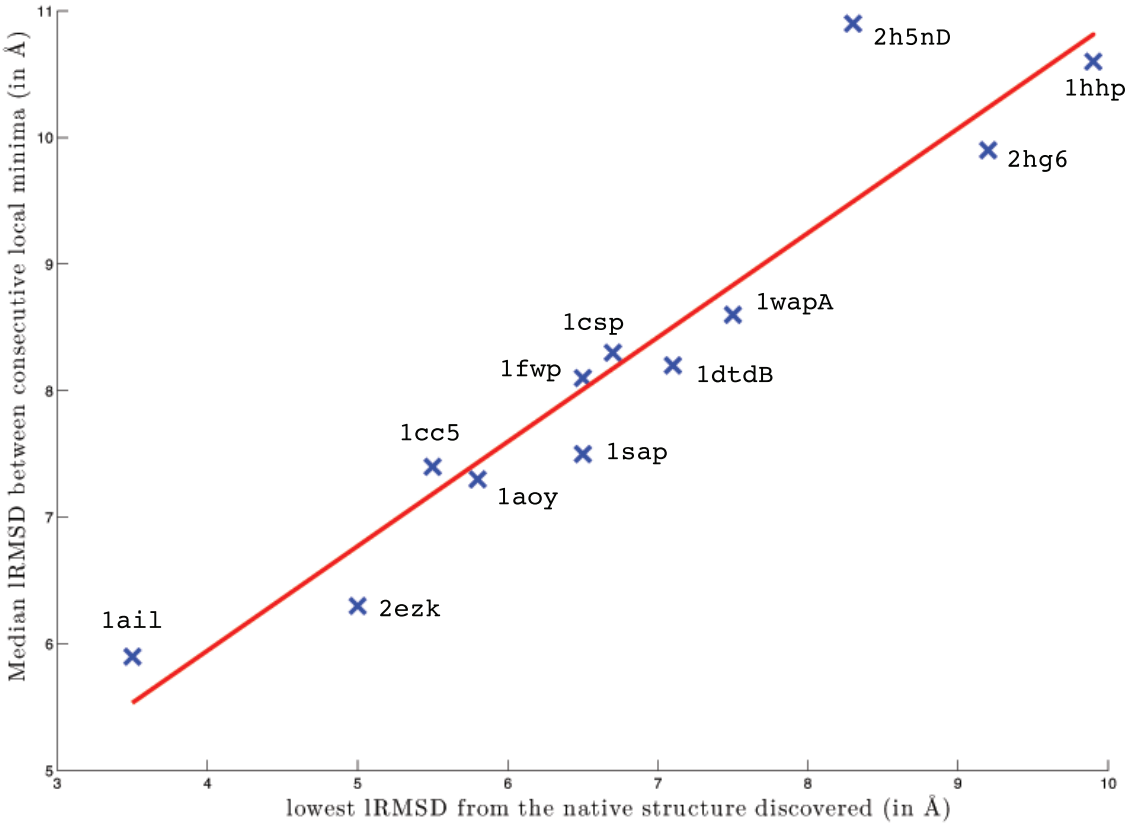


Figure 3.1: The correlation between the lowest lRMSD from the native structure discovered and the median lRMSD between two local minima conformations, C and C_{new} is shown. The strong linear correlation suggests that the efficacy of the perturbation function is directly related to the efficacy of the search.

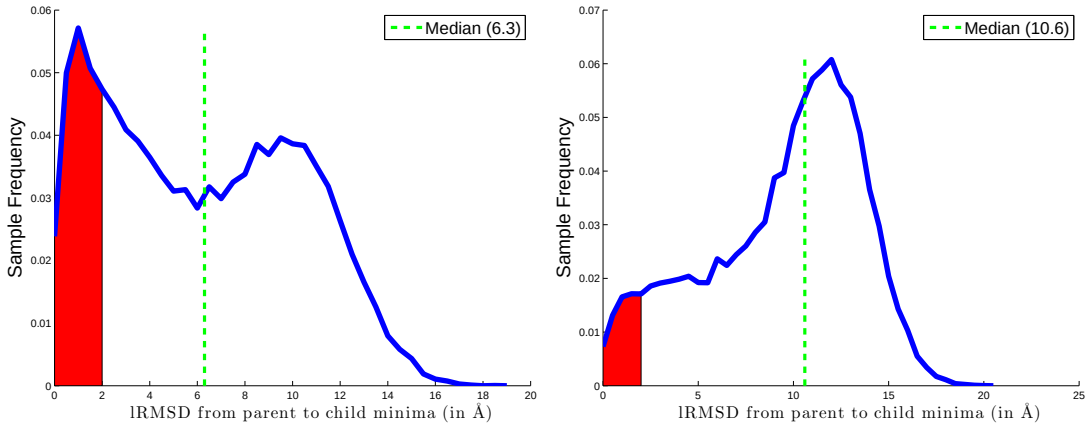
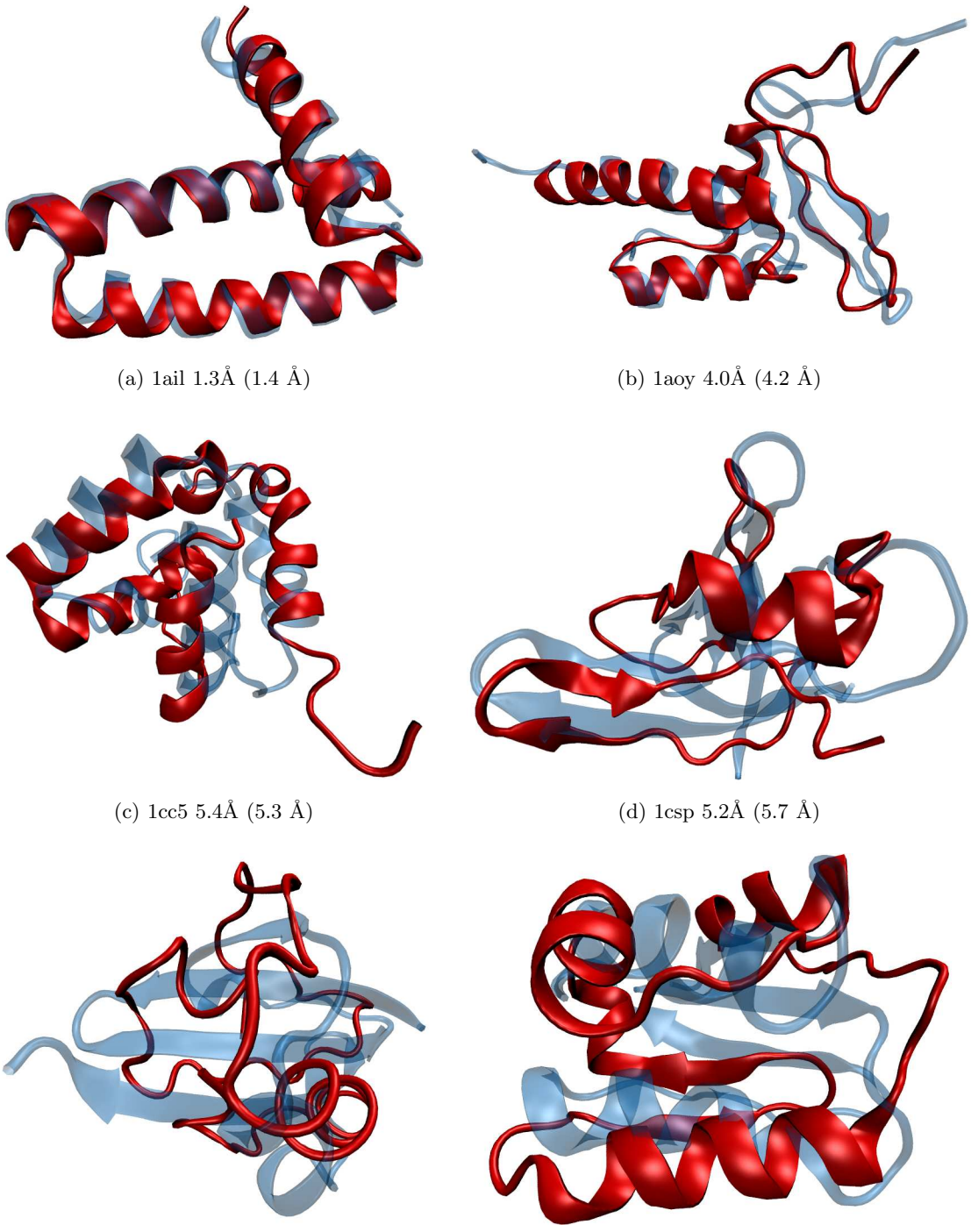


Figure 3.2: The distribution of lRMSDs between two consecutive local minima conformations, C and C_{new}, generated by performing a perturbation followed by a local search on C to achieve C_{new}. The area shaded in red represents the cases where the distance between C and C_{new} is less than 2Å lRMSD and thus it is deemed that the local search returned C_{new} to the same local minima as C.



(a) 1ail 1.3Å (1.4 Å)

(b) 1aoy 4.0Å (4.2 Å)

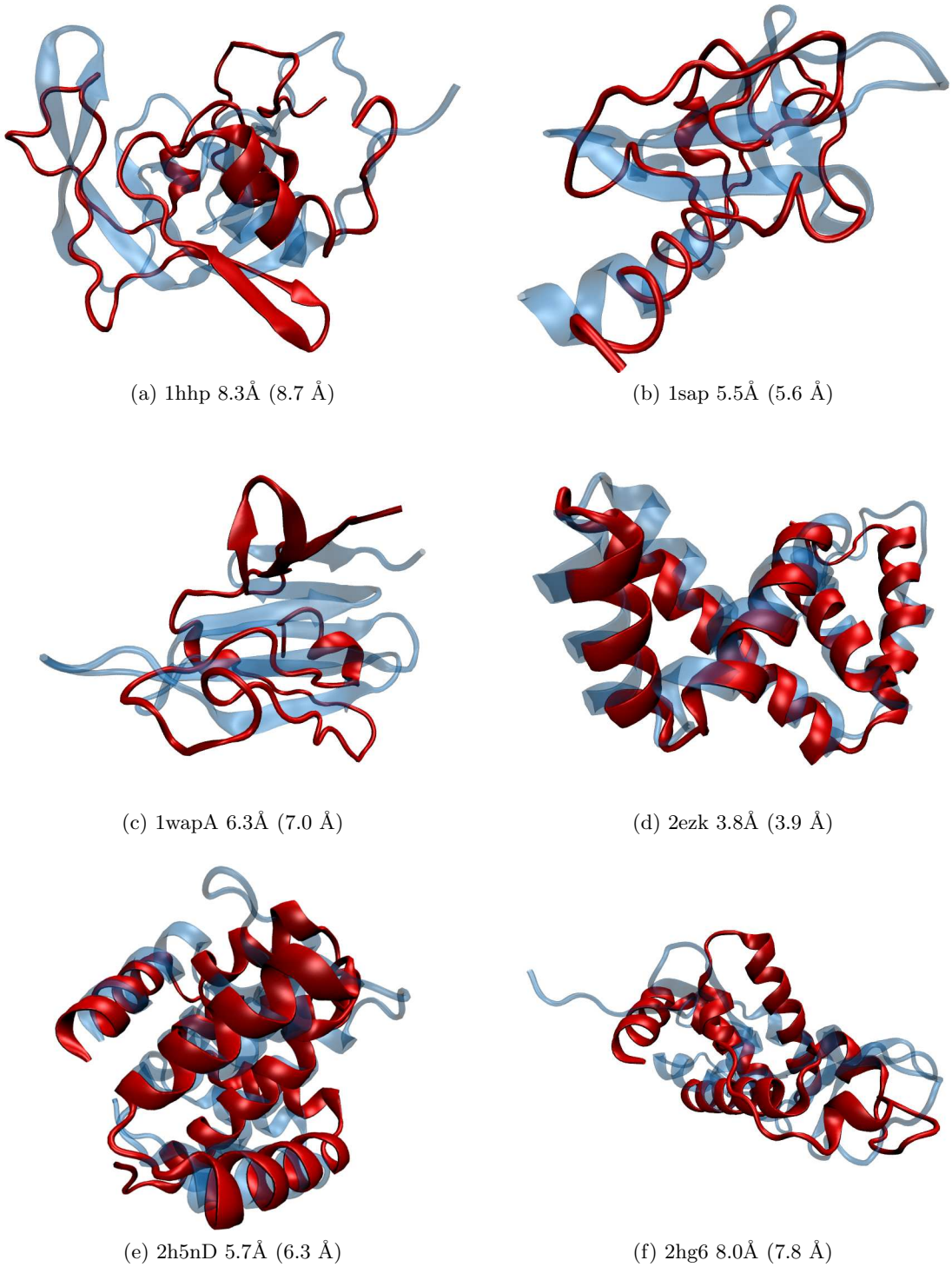
(c) 1cc5 5.4Å (5.3 Å)

(d) 1csp 5.2Å (5.7 Å)

(e) 1dtdB 6.7Å (6.7 Å)

(f) 1fwp 5.6Å (5.7 Å)

Figure 3.3: The best structure produced by PLOW after an all-atom refinement (red) is super-imposed over the native structure downloaded form the PDB (transparent blue). The refined lRMSD from the native structure is given, with the unrefined lRMSD in parentheses.



(a) 1hhp 8.3Å (8.7 Å)

(b) 1sap 5.5Å (5.6 Å)

(c) 1wapA 6.3Å (7.0 Å)

(d) 2ezk 3.8Å (3.9 Å)

(e) 2h5nD 5.7Å (6.3 Å)

(f) 2hg6 8.0Å (7.8 Å)

Figure 3.4: The best structure produced by PLOW after an all-atom refinement (red) is super-imposed over the native structure downloaded form the PDB (transparent blue). The refined lRMSD from the native structure is given, with the unrefined lRMSD in parentheses.

Chapter 4: Conclusions

This thesis proposes two novel algorithms for effectively sampling near-native local minima from a coarse-grained energy surface. PLOW combines the Iterated Local Search (ILS) approach employed by the evolutionary community with a state-of-the-art energy function and fragment library developed by the structure prediction community. The perturbation and local search approach in PLOW is applied to the FeLTr framework to sample a diverse set of low-energy conformations. PLOW and memetic FeLTr work by effectively projecting the search space onto the sub-space of local energy minima. By traversing only these local minima, PLOW and memetic FeLTr more effectively sample conformations near a protein's native structure. Both PLOW and memetic FeLTr outperform earlier work on a diverse set of target proteins [18]. When adapted to use multiple fragment lengths, PLOW and memetic FeLTr are able to accurately recreate medium length α helical proteins, and both perform favorably when compared to published results from other research groups [1, 2].

The efficacy of PLOW and memetic FeLTr is highly correlated to the ability of the perturbation function to make medium-distance jumps in the conformational space. In cases where the algorithms performed poorly, it was found that the perturbation function was more likely to make large moves and thus approach a random restart. Future efforts will focus on adaptive perturbation functions that are able to dynamically adjust perturbation distance to maintain an optimal local minima distance.

In general, memetic FeLTr performs similarly to PLOW. However, in a few select cases the simpler approach taken in PLOW is able to reach local minima with significantly lower lRMSDs from the native structure. This suggests that the existing FeLTr framework needs to be further adapted to take better advantage of the new memetic approach. Selection based on the energy projection layer will be evaluated to accommodate the presence of only local minima in the search tree.

The success of memetic methods illustrates the benefit of examining established methods from other fields that also deal with complex high-dimensional search spaces. The protein conformational space presents unique challenges that go beyond a standard stochastic optimization problem. Combining the theoretical findings from the evolutionary computation community with domain-specific protein structure knowledge can result in new approaches that blend the specialties of both sets of experts. Our research will continue to draw from both fields to develop novel approaches which advance our understanding of both the protein structure prediction problem as well as generalized optimization problems.

Bibliography

Bibliography

[1] J. DeBartolo, A. Colubri, A. K. Jha, J. E. Fitzgerald, K. F. Freed, and T. R. Sosnick, "Mimicking the folding pathway to improve homology-free protein structure prediction," *Proc. Natl. Acad. Sci. USA*, vol. 106, no. 10, pp. 3734–3739, 2009.

[2] T. J. Brunette and O. Brock, "Guiding conformation space search with an all-atom energy potential," *Proteins: Struct. Funct. Bioinf.*, vol. 73, no. 4, pp. 958–972, 2009.

[3] K. A. Dill, B. Ozkan, M. S. Shell, and T. R. Weikl, "The protein folding problem," *Annu. Rev Biophys.*, vol. 37, pp. 289–316, 2008.

[4] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973.

[5] D. Lee, O. Redfern, and C. Orengo, "Predicting protein function from sequence and structure," *Nat. Rev. Mol. Cell Biol.*, vol. 8, no. 12, pp. 995–1005, 2007.

[6] P. Bradley, K. M. S. Misura, and D. Baker, "Toward high-resolution de novo structure prediction for small proteins," *Science*, vol. 309, no. 5742, pp. 1868–1871, 2005.

[7] S. Yin, F. Ding, and N. V. Dokholyan, "Eris: an automated estimator of protein stability," *Nat Methods*, vol. 4, no. 6, pp. 466–467, 2007.

[8] T. Kortemme and D. Baker, "Computational design of protein-protein interactions," *Curr. Opinion Struct. Biol.*, vol. 8, no. 1, pp. 91–97, 2004.

[9] K. A. Dill and H. S. Chan, "From levinthal to pathways to funnels," *Nat. Struct. Biol.*, vol. 4, no. 1, pp. 10–19, 1997.

[10] C. Clementi, "Coarse-grained models of protein folding: Toy-models or predictive tools?" *Curr. Opinion Struct. Biol.*, vol. 18, pp. 10–15, 2008.

[11] Y. Zhang, "Progress and challenges in protein structure prediction," *Curr. Opinion Struct. Biol.*, vol. 18, no. 3, pp. 342–348, 2008.

[12] J. Lee, S. Wu, and Y. Zhang, "Ab initio protein structure prediction," in *Ab Initio Protein Structure Prediction*, D. Rigden, Ed. Springer Science + Business Media B.V., 2009, ch. 1.

[13] M. Ben-David, O. Noivirt-Brik, and A. Paz, "Assessment of CASP8 structure predictions for template free targets," *Proteins: Structure*, Jan 2009.

[14] A. Shehu, "Conformational search for the protein native state," in *Protein Structure Prediction: Method and Algorithms*, H. Rangwala and G. Karypis, Eds. Fairfax, VA: Wiley Book Series on Bioinformatics, 2010, ch. 21.

[15] A. Shehu, L. E. Kavraki, and C. Clementi, "Multiscale characterization of protein conformational ensembles," *Proteins: Struct. Funct. Bioinf.*, vol. 76, no. 4, pp. 837–851, 2009.

[16] R. Bonneau and D. Baker, "De novo prediction of three-dimensional structures for major protein families," *J. Mol. Biol.*, vol. 322, no. 1, pp. 65–78, 2002.

[17] A. Shehu, L. E. Kavraki, and C. Clementi, "Unfolding the fold of cyclic cysteine-rich peptides," *Protein Sci.*, vol. 17, no. 3, pp. 482–493, 2008.

[18] B. Olson, K. Molloy, and A. Shehu, "In search of the protein native state with a probabilistic sampling approach," *J. Bioinf. and Comp. Biol.*, vol. 9, no. 3, pp. 383–398, 2011.

[19] A. Shehu and B. Olson, "Guiding the search for native-like protein conformations with an ab-initio tree-based exploration," *Int. J. Robot. Res.*, vol. 29, no. 8, pp. 1106–11 227, 2010.

[20] Z. Lü and J.-K. Hao, "A critical element-guided perturbation strategy for iterated local search," in *Proceedings of the 9th European Conference on Evolutionary Computation in Combinatorial Optimization*, ser. EvoCOP '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 1–12.

[21] S. Luke, *Essentials of Metaheuristics*. Lulu, 2009, available for free at http://cs.gmu.edu/sean/book/metaheuristics/.

[22] A. A. Canutescu, A. A. Shelenkov, and R. L. Dunbrack Jr., "A graph-theory algorithm for rapid protein side chain prediction," *Protein Sci.*, vol. 12, no. 9, pp. 2001–2014, 2003.

[23] A. P. Heath, L. E. Kavraki, and C. Clementi, "From coarse-grain to all-atom: Towards multiscale analysis of protein landscapes," *Proteins: Struct. Funct. Bioinf.*, vol. 68, no. 3, pp. 646–661, 2007.

[24] M. P. Eastwood, C. Hardin, Z. Luthey-Schulten, and P. G. Wolynes, "Evaluating protein structure-prediction schemes using energy landscape theory," *IBM Journal of Research and Development*, vol. 45, no. 3.4, pp. 475 –497, may 2001.

[25] A. Zemla, "Lga: a method for finding 3d similarities in protein structures," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3370–3374, 2003. [Online]. Available: http://nar.oxfordjournals.org/content/31/13/3370.abstract

[26] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins: Structure, Function, and Bioinformatics*, vol. 68, no. 4, pp. 1020–1020, 2007. [Online]. Available: http://dx.doi.org/10.1002/prot.21643

[27] H. M. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," *Nat. Struct. Biol.*, vol. 10, no. 12, pp. 980–980, 2003.

[28] A. Shehu, C. Clementi, and L. E. Kavraki, "Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations," *Proteins: Struct. Funct. Bioinf.*, vol. 65, no. 1, pp. 164–179, 2006.

[29] S. Wells, S. Menor, B. Hespenheide, and M. F. Thorpe, "Constrained geometric simulation of diffusive motion in proteins," *J. Phys. Biol.*, vol. 2, no. 4, pp. 127–136, 2005.

[30] G. Lancia..., "Protein structure comparison: algorithms and applications," *Mathematical Methods for Protein Structure Analysis ...*, Jan 2004. [Online]. Available: http://www.springerlink.com/index/vcgkrjb5hfcpj73n.pdf

[31] T. Hansson, C. Oostenbrink, and W. F. van Gunsteren, "Molecular dynamics simulations," *Curr. Opinion Struct. Biol.*, vol. 12, no. 2, pp. 190–196, 2002.

[32] Y. Duan and P. Kollman, "Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution," *Science*, pp. 740–744, 1998.

[33] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.*, vol. 21, no. 6, pp. 1087–1092, 1953.

[34] W. F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, H. P. H., M. A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. van der Vegt, and H. B. Yu, "Biomolecular modeling: Goals, problems, perspectives," *Angew. Chem. Int. Ed. Engl.*, vol. 45, no. 25, pp. 4064–4092, 2006.

[35] N. Haspel, C. Tsai, H. Wolfson, and R. Nussinov, "Reducing the computational complexity of protein folding via fragment folding and assembly," *Protein Sci.*, vol. 12, no. 6, pp. 1177–1187, 2003.

[36] A. K. Jain, R. C. Dubes, and C. C. Chen, "Bootstrap techniques for error estimation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 9, no. 55, pp. 628–633, 1987.

[37] P. J. Ballester and G. Richards, "Ultrafast shape recognition to search compound databases for similar molecular shapes," *J. Comput. Chem.*, vol. 28, no. 10, pp. 1711–1723, 2007.

[38] P. J. Ballester, I. Westwood, N. Laurieri, E. Sim, and W. G. Richards, "Prospective virtual screening with ultrafast shape recognition: the identification of novel inhibitors of arylamine n-acetyltransferases," *Journal of The Royal Society Interface*, vol. 7, no. 43, pp. 335–342, 2010. [Online]. Available: http://rsif.royalsocietypublishing.org/content/7/43/335.abstract

[39] B. Olson, K. Molloy, and A. Shehu, "Enhancing sampling of the conformational space near the protein native state," in *BIONETICS: Intl. Conf. on Bio-inspired Models of Network, Information, and Computing Systems*, Boston, MA, December 2010.

[40] A. Shehu, "An ab-initio tree-based exploration to enhance sampling of low-energy protein conformations," in *Robot: Sci. and Sys.*, Seattle, WA, USA, 2009, pp. 241–248.

[41] W. E. Hart and S. Istrail, "Robust proofs of np-hardness for protein folding: General lattices and energy potentials," *J. Comp. Biol.*, vol. 4, no. 1, pp. 1–22, 1997.

[42] D. P. Djurdjevic and M. J. Biggs, "Ab initio protein fold prediction using evolutionary algorithms: Influence of design and control parameters on performance," *J. Comput. Chem.*, vol. 27, no. 11, pp. 1177–1195, 2006.

[43] A. Bazzoli and A. Tettamanzi, *A memetic algorithm for protein structure prediction in a 3D-lattice HP model*, ser. Lecture notes in computer science ; 3005. Berlin: Springer, 2004, pp. 1 – 10.

[44] M. Islam and M. Chetty, *Novel Memetic Algorithm for Protein Structure Prediction*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2009, vol. 5866, pp. 412–421.

[45] J. Keum, E.S., K. Kim, and E. Santos, "Local minima-based exploration for off-lattice protein folding," in *Bioinformatics Conference, 2003. CSB 2003. Proceedings of the 2003 IEEE*, aug. 2003, pp. 615 – 616.

[46] M. Mijajlovic and M. Biggs, "On potential energy models for ea-based ab initio protein structure prediction," *Evolutionary Computation*, Jan 2010. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/evco.2010.18.2.18204

[47] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: a program for macromolecular energy, minimization, and dynamics calculations," *J. Comput. Chem.*, vol. 4, no. 2, pp. 187–217, 1983.

[48] G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes, "Water in protein structure prediction," *Proc. Natl. Acad. Sci. USA*, vol. 101, no. 10, pp. 3352–3357, 2004.

[49] D. A. Case, T. A. Darden, T. E. I. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, K. M. Merz, D. A. Pearlman, M. Crowley, R. C. Walker, W. Zhang, B. Wang, S. Hayik, A. Roitberg, G. Seabra, K. F. Wong, F. Paesani, X. Wu, S. Brozell, V. Tsui, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, P. Beroza, D. H. Mathews, C. Schafmeister, W. S. Ross, and P. A. Kollman, "Amber 9," University of California, San Francisco, 2006.

[50] H. Gong, P. J. Fleming, and G. D. Rose, "Building native protein conformations from highly approximate backbone torsion angles," *Proc. Natl. Acad. Sci. USA*, vol. 102, no. 45, pp. 16 227–16 232, 2005.

[51] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucl. Acids Res.*, vol. 25, no. 17, pp. 3389–33 402, 1997.

[52] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J. Mol. Biol.*, vol. 292, no. 2, pp. 195–202, 1999.

[53] K. Molloy, "Variable-length fragment assembly within a probabilistic protein structure prediction framework," Fairfax, Virginia, 2011.

[54] K. M. Misura and D. Baker, "Progress and challenges in high-resolution refinement of protein structure models," *Proteins: Structure, Function, and Bioinformatics*, vol. 59, no. 1, pp. 15–29, 2005. [Online]. Available: http://dx.doi.org/10.1002/prot.20376

[55] K. W. Kaufmann, G. H. Lemmon, S. L. DeLuca, J. H. Sheehan, and J. Meiler, "Practically useful: What the rosetta protein modeling suite can do for you," *Biochemistry*, vol. 49, no. 14, pp. 2987–2998, 2010, pMID: 20235548. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/bi902153g

Curriculum Vitae

Brian Olson graduated from the Gilman School in Baltimore, MD in 2000. He attended Princeton University from 2000 to 2005, graduating with a Bachelor of Science in Engineering in Computer Science. While at Princeton, he was employed as a software consultant for a small startup venture, Proximities, Inc. Upon graduation, he took a job as a software development engineer at Microsoft. In 2008 Mr. Olson left Microsoft to begin his Ph.D. study at George Mason University. He is currently in his fourth year as a Ph.D. student and will receive a Masters of Computer Science in 2011.