RICE UNIVERSITY

Sampling Biomolecular Conformations with Spatial and Energetic Constraints

by

Amarda Shehu

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree Master of Science

Approved, Thesis Committee:

Lydia Kavraki
Dr. Lydia E. Kavraki (Chair),
Noah Harding Professor,
Computer Science, Rice University

Ron Goldman
Dr. Ron Goldman,
Professor,
Computer Science, Rice University

Cecilia Clementi
Dr. Cecilia Clementi,
Assistant Professor,
Chemistry, Rice University

Luay Nakhleh
Dr. Luay Nakhleh,
Assistant Professor,
Computer Science, Rice University

Houston, Texas
December, 2004
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
Abstract

Sampling Biomolecular Conformations with Spatial and Energetic Constraints

by

Amarda Shehu

In the past decade it has become clear that structural fluctuations are often times related to biological function. In biomedically relevant molecules, function-related fluctuations are spatially and energetically constrained. Imposing both spatial and energetic constraints is crucial to obtaining an ensemble of structures that best characterizes the relationship between flexibility and function. Current approaches to satisfying spatial constraints in biomolecules have practical limitations on the dimensionality of the system and the way energetic constraints are addressed. By employing cyclic coordinate descent, an algorithm that works in dihedral space, we reduce the dimensionality of the problem and minimize detrimental effects on energetic stability. This work extends cyclic coordinate descent to efficiently satisfy multiple spatial constraints, respect the secondary structure of proteins, and work with reduced backbone protein models. Reduced models allow us to treat large systems that are intractable under all-atom models. In addition, this thesis combines the satisfaction of multiple spatial constraints with conformational sampling and energy minimization techniques to generate spatially constrained biomolecular structures that are energetically stable under physiological conditions.

The experiments in this thesis demonstrate the relevance and robustness of our method on three areas of applications: loop closure, backbone reconstruction, and physical trajectory recovery. Addressing the problem of loop closure, we obtain en-
sembles of spatially constrained conformations whose energy landscape is in agreement with laboratory experimental results on the energetic stability of the proteins at hand. Our experiments on backbone reconstruction agree with results from statistical approaches to this problem, but in addition guarantee the energetic feasibility of the completed models. On a broader context, our approach to backbone reconstruction for proteins provides an example of the applicability of our method to model the effect of aminoacid mutations on protein flexibility and energetic stability. Modeling point mutations is an area of great interest in protein engineering and protein folding studies. Our last set of experiments on physical trajectory recovery show the relevance and implications of our work to drug design by producing energetically feasible conformations where known groups of atoms are constrained to a well-characterized collective motion.

The methods presented in this thesis are not limited to proteins but are applicable to the study of flexibility of any biomolecules where one can define spatial constraints. This work provides a robust starting tool to address the dynamics-function relationship in the post-genomic era.
Acknowledgments

First I would like to thank my advisor, Dr. Lydia Kavraki, for giving me the needed vision and calm confidence for this work. I have learned much of what it takes to be a researcher and a scientist from her. Second, I would like to thank Dr. Cecilia Clementi for stressing the importance of energy considerations for the biological relevance of this work and always being willing to make theoretical chemistry approachable to computer scientists. Her guidance and insights have been invaluable to this work. I owe special thanks to Dr. Ronald Goldman, Dr. Luay Nakhleh, and Dr. Mark Moll for reading this thesis and having the kindness and patience to provide me with instructive comments. I also thank the Kavraki physical and biological computing group and the Clementi chemistry group for being supportive and helpful in this work. Finally, I thank my husband and family for their love and support.

This work has been supported in part by a Grant from the Whitaker Foundation, a Grant from the Texas Advanced Research Program, NSF 0319415, and NSF 0203396. Computational work was done in part on equipment obtained through EIA0216467.
Contents

List of Illustrations viii

1 Introduction 1
  1.1 Problem Statement ........................................ 3
  1.2 Motivation .................................................... 7
  1.3 Contribution .................................................. 8
  1.4 Applications .................................................. 10
    1.4.1 Loop Closure .............................................. 10
    1.4.2 Backbone Reconstruction .................................. 11
    1.4.3 Physical Trajectory Recovery ............................. 12
  1.5 Thesis Overview ............................................. 13

2 Protein Modeling 14
  2.1 Protein Structure and Flexibility ............................ 14
    2.1.1 Proteins as Robotic Manipulators ....................... 18
  2.2 Protein Energy Landscapes ................................... 20
    2.2.1 CHARMM Energy Modeling ................................ 21

3 Related Work 24
  3.1 Inverse Kinematics-based Methods ........................... 24
    3.1.1 Classic Inverse Kinematics Methods ..................... 25
    3.1.2 Inverse Kinematics with Optimization ................... 26
  3.2 Search-based Methods ....................................... 27
    3.2.1 Loop Construction Methods .............................. 27
4 An Algorithm to Generate Ensembles of Spatially and Energetically Constrained Conformations

4.1 Protein Kinematics

4.2 Protein Kinematics with Spatial Constraints

4.2.1 Loop Closure

4.3 Extension: Protein Kinematics with Multiple Spatial Constraints

4.4 Extension: Energy Considerations

4.4.1 Dihedral Energy

4.4.2 Long Range Interactions

4.5 Extension: Sampling Conformations with Spatial and Energetic Constraints

4.6 Extension: Recovering Physical Conformations

5 Experiments

5.1 Loop Closure

5.1.1 Proteins Used and Methodology

5.1.2 Loop Closure Ensemble for 1COA

5.1.3 Loop Closure Ensemble for 1ALC

5.1.4 Loop Closure Ensemble for VLSE1

5.1.5 On the Use of Backbone-only Approximation

5.2 Backbone Reconstruction

5.3 Physical Trajectory Recovery

5.4 Discussion of Experimental Results
6 Discussion

Bibliography
Illustrations

1.1 (a) The desired missing fragment in VlsE would pass through aminoacids 92 and 113, thus connecting PRO92 and THR113. (b) The \( C_\alpha \) atoms of flavodoxin semiquinone depicted in black, defining an arrangement of the aminoacids shown in silver, pose spatial constraints that should be satisfied to obtain coordinates for all atoms, depicted in gray. (c) Residues drawn in red consist of atoms that undergo an opening and closing motion to accommodate and stabilize drug candidates in the cavity of HIV protease.  

2.1 Four consecutive residues in a polypeptide chain  

2.2 (a) A \( \beta \)-turn, depicted in black, connects two \( \beta \)-sheets, drawn in grey as arrows pointed along the protein backbone. The \( \alpha \)-helix is drawn in silver. (b) Another representation of an \( \alpha \)-helix as a cylinder is given. A long loop connects the \( \alpha \)-helix to the \( \beta \)-sheet.  

2.3 (a) Illustration of internal coordinates, where \( b \) refers to bond length, \( \alpha \) to bond angle, and \( \theta \) to dihedral angle. (b) Rotation by the dihedral on the second bond induces spatial motion of the fourth atom and any consecutive atoms down the polypeptide chain.
2.4 (a) Electron sharing between the carboxyl carbon and the amide nitrogen gives the peptide bond a partial double-bond character and as a consequence its rigidity. (b) and (c) demonstrate the two backbone dihedrals $\phi$, $\psi$ on two aminoacids, glycine, and alanine. (d) No sidechain dihedrals for the rigid ring in Proline (e) Tryptophan contains 2 sidechain dihedrals (f) Arginine has the highest number of sidechain dihedrals, 4. .................................................. 19

2.5 A protein can be modeled as an articulated mechanism, where $d_i$ is the length of the bond between atoms $A_{i-1}$ and $A_i$, $\alpha_{i-1}$ the bond angle between $A_{i-2}$, $A_{i-1}$, and $A_i$, $\theta_{i-1}$ the $\psi$ dihedral, and $\theta_i$ the $\phi$ dihedral. ................................................................. 20

2.6 (a) There are multiple global minima in the energy landscape. (b) The native structure is not a strong stability point, resulting in a shallow basin. (c) The native structure is a strong stability point, reflected in the single global minimum. ................................. 22

3.1 (a) The generated fragments’ mobile anchors, drawn in grey, are not attached to the stationary anchors drawn in black. (b) Fragments can be translated so one of the mobile anchors attaches to its corresponding stationary anchor. (c) Fragment is closed when both mobile anchors overlap with the stationary ones. ................................. 25

4.1 Illustration of a protein with three feature atoms depicted. ............. 33

4.2 CCD Schematic - rotation by $\alpha$ around the O axis steers feature atom from current position $M$ to target position $F$. ......................... 33

4.3 Features specified in order along the backbone. When feature atoms are part of sidechains, pseudo-features are new feature atoms added to split a kinematic chain between three feature atoms. .............. 41
4.4 Three schematic conformations (drawn in red, green, and blue) sampled in the self-motion manifold satisfy spatial constraints posed by three feature atoms (drawn in red.)

5.1 (a) and (b) The native loops of 1COA and 1ALC lie between residues VAL53 and ASP64, and LYS51 and THR76, respectively. The C\textsubscript{α} atoms of the terminal residues of these loops are shown as red spheres, while the rest of the proteins are shown with their secondary structure segments in ribbon representation. (c) and (d) The desired missing loop in VLSE1 needs to connect PRO92 and THR113.

5.2 (a) and (b) 1COA loop is depicted as a string of beads, with black and purple beads denoting atoms involved in peptide bonds. The C-terminal loop residue depicted in red is not perfectly oriented with 50 iterations(a) and requires 500 iterations for no error in the orientation(b). (c) A loop closure 1ALC conformation with 5000 iterations shown in blue and superimposed on the native loop in red shows perfect orientation. (d) Depicted in red are the C\textsubscript{α} atoms of the residues that need to be connected. (e) VLSE1 loop closure conformation shown in blue reaches perfect orientation with 5000 maximum iterations.

5.3 (a) There is low geometric variability in the reported 1COA loop closure conformations when the total energy with both backbone and sidechain contributions is considered. (b) The energy landscape of the loop closure conformations is funnel-like.

5.4 (a) There is low geometric variability in the feasible 1ALC loop closure conformations. (b) The energy landscape associated with this ensemble of loop closure conformations is funnel-like.
5.5 (a) Structural fluctuations (RMSD) per residue in the native ensemble of conformations obtained in [VPDK03] (red line) compared with the X-ray B factors (green line). (b) RMSD fluctuations per residue for the conformations obtained in our simulation are in complete agreement with those obtained in [VPDK03]. (c) Superimposed conformations obtained in [VPDK03], with the most variable protein regions drawn in red. (d) Superimposed conformations obtained in our simulation, with the loops drawn in red, strongly resemble those obtained in [VPDK03].

5.6 (a) There is high geometric variability in the feasible VLSE1 loop closure conformations. (b) VLSE1 energy landscape resembles a plateau, consistent with experimental evidence.

5.7 (a) A large scale search reveals a variety of 1COA loop conformations. (b) The potential energy landscape associated with the obtained conformations does not present a single minimum. (c) Clustering shows two major clusters: cluster 1 in red with 523 conformations and cluster 2 in blue with 483 conformations. (d) The similarity matrix reveals that the two clusters are well-defined. (e) Representatives are not all native-like (native loop in cyan). (f) While one of the medians is native-like, the other is very different from the native.

5.8 (a) Loop sidechain atoms (in black) are clashing with the sidechain atoms of the β-sheet (in cyan). (b) Steric clash between atom CD2 of TYR61 and HB of VAL82.

5.9 Dihedral, improper, and electrostatic energy terms are insignificant fluctuations around the native conformation of CI2.
5.10 (a) 1COA closure conformations are clustered around the native. (b) The funnel-like energy landscape of the reconstructed conformations confirms that CI2 has a well-defined native structure. (c) The five most stable conformations lie at most 0.2 Å RMSD away from the native 1COA conformation.

5.11 (a) Superimposed HIV protease conformations are obtained from linear translations of the native conformation along the first principal component. (b) Zooming on the flaps (residues 45-55 drawn in red) reveals their opening/closing motion. (c) Atoms on the flaps are the ones with the largest displacement along the first principal component.

5.12 (a) HIV protease conformations recovered from the linearly translated conformations along the first principal component are shown superimposed. (b) The motion of the flaps (residues 45-55 drawn in red) is shown more detailed. (c) The closing motion of the flaps is more energetically constrained. This can be seen in the fact that only up to 100 conformations, in blue, with step of size 0.1 Å can be recovered. On the other hand, the opening motion of the flaps is less energetically constrained, resulting in 258 conformations (d) The energy landscape of HIV protease demonstrates the symmetry of the open- and close-flap conformations and the overall funnel-like shape of the landscape.
Chapter 1

Introduction

Biological molecules such as proteins, nucleic acids, lipids, and carbohydrates are essential to every known form of life because they are involved in every cellular process in living organisms. Many of these molecules achieve their functionality through complexes resulting from molecular interactions. For example, proteins bind to DNA and RNA to mediate regulation of gene expression, gene transcription, DNA replication, and mRNA intron splicing. In addition, protein-protein interactions play a role in antibody-antigen binding, large scale organismal motion, and cell adhesion. Such complexes require that the participating biomolecules acquire compatible shapes to be geometrically complementary. Many biomolecules are able to achieve particular complementary shapes due to their flexibility.

Many flexible biomolecules may lose the ability to acquire a particular shape thus incapacitating cells from performing an important process. Many proteins and even RNA molecules may misfold into incorrect shapes with dire consequences for their functionality inside cells. For example, diseases such as Jacob-Creutzfeldt and mad cow disease are linked to the misfolding of prion, a protein that folds abnormally and encourages other proteins to become similarly misshapen, affecting their ability to function. On the other hand, certain biomolecules are very robust and can acquire many different shapes while remaining functional. For example, HIV protease, crucial to the replication of the human immunodeficiency virus, remains functional in many different shapes. Inhibiting the activity of this protease is crucial to incapacitating the replication of HIV.

Since biomolecular flexibility is strongly correlated to biomolecular functionality,
flexible biomolecules are often the target of drug design. Designing effective drugs for flexible targets requires an understanding of how biomolecular flexibility is related to the ability of a target molecule to remain functional or be trapped in shapes that lose functionality. Understanding and modeling function-related flexibility is very crucial to designing effective treatments for misfolding-related diseases and effective drugs to inhibit the activity of flexible molecules such as HIV protease. In particular, modeling the flexibility of viral proteins in drug design is crucial to producing effective drugs to bind targets that always change shapes. Therefore, understanding and modeling the relationship between biomolecular flexibility and function is essential for designing new effective drugs.

Interestingly, biomolecules functioning in living organisms under physiological conditions are able to undergo large-scale structural fluctuations and yet not compromise their functionality. Structural fluctuations due to synchronized movements of atom groups define spatial constraints on the three-dimensional locations of the atoms involved. The ability of biomolecules to undergo function-related structural fluctuations introduces energetic constraints on the flexibility of a biomolecule, due to the fact that the functionality of a structure under physiological conditions is associated with a low energy. Therefore, modeling function-related flexibility for biomolecules demands satisfying both spatial and energetic constraints.

Based on the post-genomic era redefinition of the structure-function to flexibility-function relationship [WD99], the goal of this thesis is to model interesting function-related flexibility through ensembles of biomolecular structures that, while displaying large structural fluctuations, remain energetically stable and can function under physiological conditions. In this context, this thesis explores the problem of sampling accessible spatial arrangements for the atoms of a biomolecule while satisfying spatial constraints on specific atoms and energetic constraints aimed at maintaining energetic stability, i.e, functionality.
1.1 Problem Statement

We refer to the spatial arrangement of a biomolecule's atoms in three-dimensional space as a conformation. The problem of satisfying a spatial constraint on a particular atom \( a \), referred to as a feature atom, can be formulated as: Given (i) a molecule in an initial conformation \( C_{\text{start}} \), and (ii) the target position \( p = (p_x, p_y, p_z) \) of feature atom \( a \), generate a target conformation \( C_{\text{goal}} \) so that feature atom \( a \) in this conformation reaches its target position. We refer to target conformation \( C_{\text{goal}} \) as a spatially constrained or a closure conformation.

In this thesis we also address the satisfaction of energetic constraints. The problem of satisfying an energetic constraint can be posed as: Given energy \( E_{\text{lowest}} \) associated with the most stable conformation, generate a conformation \( C_i \) such that its energy \( E_i \in [E_{\text{lowest}} - \text{cutoff}, E_{\text{lowest}} + \text{cutoff}] \), where cutoff is a parameter that determines the energetic feasibility of conformation \( C_i \).

In this thesis we address the problem of sampling energetically feasible biomolecular conformations that satisfy multiple spatial constraints. We place this problem in the context of three concrete instances, which demonstrate the need for generating not just one feasible closure conformation but an ensemble of spatially constrained conformations with energetic constraints. The first instance originally arises in the context of completing the experimentally determined structure for a biomolecule that exhibits large structural fluctuations under physiological conditions. When faced with biomolecular fragments that can exist in varying stable conformations (referred to as inherently disordered fragments), modeling flexibility becomes more important than reporting one structure out of all the possible stable ones.

In the context of proteins, experimental techniques produce an incomplete view of the structure with mobile fragments missing. As is the case for the lyme disease variable surface antigen (VlsE), shown in Figure 1.1(a), cartesian coordinates for atoms between aminoacids 92 and 113 cannot be determined [JWS03]. Sets of reasonable coordinates for the atoms of this missing fragment need to be automatically gener-
ated. Among different possible conformations for the missing fragment, one needs to consider only those that connect atoms in aminoacids 92 and 113 to the rest of the protein. The connectivity of these aminoacids to the rest of the protein poses spatial constraints on the generated conformations for the missing fragment. In addition, the spatially constrained conformations need to preserve the energetic stability of the protein.

The second instance comes from the need to complete a protein model by reporting the cartesian coordinates of all of its atoms when provided only positions for its $C_\alpha$ atoms. Such a situation often arises from the need of minimalist model simulation techniques to connect their results with other all-atom simulations or experimental results. In order to model large conformational changes in proteins with hundreds of residues, minimalist model simulation techniques are forced to reduce their protein models to $C_\alpha$ atoms only [OLSW97, CJO00, CJO01].

As Figure 1.1(b) indicates for the flavodoxin semiquinone protein, while the coordinates of the $C_\alpha$ atoms vaguely define the position in space of every corresponding aminoacid, it is not clear how one can obtain a detailed all-atom model for the protein. Since the positions of $C_\alpha$ atoms are known, every $C_\alpha$ atom defines spatial constraints that need to be satisfied when exploring stable conformations for the missing atoms.

While the first two instances of our problem arise in the context of modeling large but energetically stable conformational fluctuations where particular atoms are forced to remain in designated spatial locations, the third instance arises from the need to generate stable conformations that can exhibit a hypothesized or well-characterized motion. This instance is particularly important in the context of drug design because it consists of finding molecular conformations where functionally important groups of atoms are constrained to experimentally observed or hypothesized concerted motions. Such is the case for many target proteins and drug candidates, where atoms on the contact site undergo movements so these molecules can reach geometrical complementarity and dock with each other. As Figure 1.1(c) demonstrates for HIV protease,
Figure 1.1: (a) The desired missing fragment in VlsE would pass through aminoacids 92 and 113, thus connecting PRO92 and THR113. (b) The $C_{\alpha}$ atoms of flavodoxin semiquinone depicted in black, defining an arrangement of the aminoacids shown in silver, pose spatial constraints that should be satisfied to obtain coordinates for all atoms, depicted in gray. (c) Residues drawn in red consist of atoms that undergo an opening and closing motion to accommodate and stabilize drug candidates in the cavity of HIV protease.
atoms in contact with potential drug candidates are constrained to a specific motion whereas the rest of the protein is practically immobile. Concerted motion of the atoms around the contact site defines spatial constraints on these atoms that need to be satisfied when exploring molecular conformations that observe the geometric complementarity needed for successful docking.

Sampling molecular conformations reduces to robotic manipulator kinematics, which describes the location of the manipulator's joints in time as a function of its link variables without regard to the forces that cause the motion. In proteins, atoms are the equivalent of manipulator joints, and bonds connecting atoms are the equivalent of manipulator links. In this framework, a conformation is a function of the link variables, also referred to as degrees of freedom (DOFs). Therefore, searching in conformational space translates to the exploration for values to the DOFs. In this context, finding a closure conformation that satisfies $n$ spatial constraints on atoms $a_1, \ldots, a_n$ can be redefined as: Given (i) a molecule in an initial conformation $C_{\text{start}}$, and (ii) the target positions $p_1, \ldots, p_n$ of feature atoms $a_1, \ldots, a_n$, respectively, solve for the DOFs so that these $n$ feature atoms in the generated conformation $C_{\text{goal}}$ reach their respective target positions.

Although the conformational space is of high dimensionality due to the large number of DOFs needed to represent molecular conformations, closure conformations typically lie in a low-dimensional region of this space. Therefore, sampling closure conformations involves exploring a constrained space. Moreover, since the closure conformations we are interested in are energetically constrained, this search is narrowed even further to a small region of the conformational space where closure conformations are also energetically feasible.

In this thesis we address the problem of sampling energetically constrained biomolecular conformations with multiple spatial constraints through spatially constrained manipulator kinematics that selects conformations based on their associated energetic feasibility. The solution we provide in this thesis can be applied to a broad
range of biomolecules and in a broad context of applications, as long as the definition of spatial constraints is inherent to the application. To render the description of our methods and the interpretation of our results clearly and concisely, this thesis focuses on a vital class of biomolecules, proteins that are functional under physiological conditions.

1.2 Motivation

It is well understood in the post-genomic era that the “structure-to-function” relationship does not reveal much about a flexible biomolecule’s function or how this function can be affected by mutations or changes in the environmental conditions. This is the case for a vital class of flexible macromolecules such as proteins. Even though originally proteins were classified as rigid objects [Sch44], it is now known that they are flexible and exhibit motions ranging from local atomic fluctuations to global rearrangements [BKP88, FW94]. Particular motions in proteins, such as loop and domain movements in the enzyme dihydrofolate reductase (DHFR) are crucial to this protein’s catalytic activity [CJO00, SK97]. Therefore, a characterization of protein flexibility is crucial to study the effects of protein motions on function.

The functionality of biomolecules like proteins depends on their energetic stability, which is the result of atomic interactions that can stabilize or destabilize structure. Stabilizing forces bring molecular structure to an equilibrium associated with a low energy. Therefore, any computational exploration of functional structures needs to report conformations that are energetically feasible, thus relevant.

Not all biomolecular fragments exist in one unique stable conformational state. For example, many protein regions such as loops are structurally disordered and can be found in different conformations. Structurally disordered fragments, due to their flexibility, are a challenge for structure determination experimental techniques and for the process of designing drugs for flexible target proteins. Therefore, producing multiple reasonable sets of coordinates for mobile fragments is a better and more
interesting characterization of functional flexibility.

Ensembles of energetically feasible conformations better characterize a biomolecule's function-related fluctuations. In this context, spatially constrained molecular kinematics properly addresses the ability of molecules to adopt different conformations while constraining atoms to specific locations in three-dimensional space. Spatial constraints on atoms are correlated to a biomolecule's specific function, its interaction with other molecules, or its property of maintaining proper functionality.

For a protein, its experimentally determined structure submitted to the Protein Data Bank (PDB) [BWF+00, BKW+77] corresponds to the most stable and functional conformation, also referred to as the native state. Depending on the dynamics of proteins, structural fluctuations around this native state can have effects on the functionality of the protein at different scales. Some proteins characterized as having a stable native state are functionally very sensitive to structural fluctuations around their native state. In this case, energetically stable conformations need to be structurally native-like. For other proteins that can accommodate large-scale fluctuations around the native state with no detrimental effects on function, stable conformations may be structurally very different from the native conformation.

1.3 Contribution

In this thesis, we address the problem of sampling energetically feasible conformations with multiple spatial constraints. Our solution to this problem reveals ensembles of closure conformations whose associated energy landscape is in full agreement with the energetic stability that characterizes different proteins. We reduce the problem of satisfying multiple spatial constraints to that of sampling multiple arbitrary-size fragments that are spatially constrained on both ends. To satisfy spatial constraints on both ends of a fragment, also referred to as fragment completion, we employ state-of-the-art cyclic coordinate descent (CCD) [Lue84, WC91], which iteratively
adjusts one dihedral at a time to move feature atoms toward their target positions. Through CCD, the solution to the fragment completion problem is free of singularities, has no limits on fragment length, does not suffer from any resolution issues, and has no dependence on the particular aminoacid sequence when completing protein fragments. Though previously employed in computational biology under the loop closure framework [CD03, vdBLLD04, LvdBDL04, Lot04], this thesis adapts and extends CCD to handle multiple spatial constraints efficiently.

While working in dihedral space reduces the dimensionality of protein conformations thus rendering protein models more tractable, large proteins with hundreds of residues still remain a challenge. We address this concern by accommodating reduced protein models into spatially and energetically constrained molecular kinematics. Due to an all-atom protein representation demanding large storage, we support a backbone representation of the protein, where sidechain atoms are removed from each aminoacid.

To address energy concerns, we exploit CCD's ability of accommodating additional constraints on each dihedral by imposing boundary conditions on solutions in order to respect any local structure such as protein secondary structure. Imposing such constraints on specific dihedrals allows us to control the energetic stability of conformations while steering these conformations to satisfy spatial constraints. Limiting effective DOFs to dihedrals only, also gives us the benefit of limiting detrimental effects of kinematics to protein stability. We further combine CCD with conformational sampling in dihedral space and conjugate gradient descent minimization of the energy associated with each conformation in order to generate closure conformations that are energetically relevant.

The strongest and main contribution of this thesis is its broad applicability to the study of flexibility for any biomolecules where one can define spatial constraints. For clarity and brevity, the contributions of our work are specifically demonstrated for three different needs in computational and structural protein biology: loop closure,
backbone reconstruction, and recovery of a physical trajectory. Our experimental results on these three applications reveal the robustness and broad relevance of our work in producing closure conformations that energetically confirm stability characteristics for different proteins. Our work, therefore, has implications in model completion, drug design, and the characterization of a protein’s stability through ensembles of spatially constrained yet energetically feasible conformations.

1.4 Applications

The underlying theme of all the applications of our method is that ensembles of conformations are a better dynamic picture of biomolecular structure where many fragments are structurally disordered and can therefore exist in many stable conformations. Though our work can be applied to any flexible biomolecules, in this thesis we address three specific problems on proteins: loop closure, backbone reconstruction, and physical trajectory recovery. We now describe each problem in detail.

1.4.1 Loop Closure

Protein fragments such as loops are very mobile and structurally varying across proteins. Their high mobility makes loop modeling a challenge for experimental methods such as X-ray crystallography [Cow75, GMA+02], nuclear magnetic resonance (NMR) [SH93], and cryogenic electron microscopy (cryoEM) [Rei89]. While X-ray crystallography is affected by the mobility of loops through disorder in the crystal, NMR is limited by protein size and scale of structural fluctuations, and cryoEM is inapplicable due to the poor resolution of generated images. As a result, atomic coordinates for mobile loops will often be missing, as is the case with the PDB reported structure of VlsE, which, as Figure 1.1(a) shows, misses a loop between aminoacids 92 and 113 [JWS03].

Therefore, coordinates for missing loops need to be constructed by automated procedures that explore varying but stable loop conformations. The spatial constraints
that need to be satisfied for sampled loop conformations arise from the aminoacids at each end of the loop, which need to be connected through bonds to the rest of the protein. Conformations with spatial constraints due to the specific need for loop completions are referred to as loop closure conformations in this work. This work provides not just one protein structure with the closed loop, but ensembles of loop closure conformations where the mobility of the loop can be placed in context with the protein's overall stability. This thesis correctly addresses the problem of loop closure by generating stable conformations that are minor structural fluctuations around the native-like conformation for strongly stable proteins and more varying yet energetically feasible loop closure conformations for proteins with an experimentally determined weak stability.

1.4.2 Backbone Reconstruction

Many molecular modeling and simulation methods are forced to reduce all-atom protein models to mainchain \( C_\alpha \) atoms in order to render simulations on large proteins tractable [CM04]. While these simplified models provide a coarse-grained analysis of the protein conformational space, they often need to be brought back to an all-atom representation for a more detailed characterization of regions in conformational space. Therefore, algorithms are needed to reconstruct a protein's backbone and report coordinates for all atoms from the given positions of \( C_\alpha \) atoms. Though this is a well-known problem already addressed [MKS97], current approaches generate low resolution models with no guarantees on their energetic feasibility, and thus their relevance.

In this thesis we address such energetic concerns and demonstrate the robustness of our method by providing an alternative approach to backbone reconstruction that guarantees the energetic feasibility of generated all-atom conformations. We sample protein conformations on which every \( C_\alpha \) atom imposes a spatial constraint. We find that imposing energetic constraints on reconstructed backbone conformations yields
all-atom models that are native-like for strongly stable proteins. A demonstration of this can be found on Figure 1.1(b), which shows in grey the cartesian coordinates for the obtained full-atom model of the flavodoxin semiquinone protein. Properly addressing the problem of going from a $C_\alpha$-trace model to an all-atom representation is a crucial part of an hierarchical and "adaptive" modeling procedure through which large biomolecular systems can be studied at inhomogeneous level of detail.

On a broader context our approach to the backbone reconstruction problem can also accommodate and model the effect of mutations on function-related conformational changes. Addressing residue mutations is straightforward through our method, where mutating a residue consists of finding all-atom coordinates for the desired new sidechain around the fixed $C_\alpha$ atom. Modeling mutations is a powerful application of our work as it can be used to direct or confirm laboratory experiments.

1.4.3 Physical Trajectory Recovery

Evidence of a particular motion mechanism for a protein can be revealed through experimental techniques and/or computational searches of protein conformations guided by classic Newtonian mechanics such as molecular dynamic simulations [MBB+98]. Different stable structures for the same protein [LCS+04] raise the question of explaining and reconciling observed structural differences, and therefore, characterizing conformational changes. One way of decoupling the observed motion is through principal component analysis (PCA), which reveals concerted movements of atom groups. This is the case for HIV protease, where linear motion along the highest principal components is characterized as an opening and closing motion [TPK03]. This motion is illustrated in Figure 1.1(c), where the most mobile atoms under the first principal component are depicted in red.

This thesis proposes capturing conformational changes due to an observed motion through spatially and energetically constrained molecular kinematics. It employs a displacement in time analysis to determine which atoms are most affected by the mo-
tion and should therefore be spatially constrained so as to best capture the observed motion. Steering such atoms to obey the motion through the spatially and energetically constrained molecular kinematics laid out in this thesis reveals conformations that are stable and exhibit the designated motion. Spatially constrained molecular kinematics can therefore be employed to propose conformational changes that capture the function-related flexibility of important biomolecules such as target proteins and drug molecules. Modeling such flexibility, currently a challenge for drug design, is therefore crucial for designing successful drugs for drug-resistant and highly flexible target biomolecules such as HIV protease.

Finally, recovering stable conformations where designated atoms undergo a well-defined motion, a problem we refer to as physical trajectory recovery, is crucial to model a characterized motion and determine the existence of energetically feasible conformations that correspond to important large-scale fluctuations associated with or even induced by function.

1.5 Thesis Overview

We start in chapter 2 with a brief background on the biological understanding of protein structure, flexibility, and energy that is employed in the protein model in this thesis. Chapter 3 then summarizes existing techniques on the satisfaction of spatial constraints thus placing the contributions of this thesis in context. Chapter 4 details our methods of sampling conformations with spatial and energetic constraints. In chapter 5 we present three classes of experiments in the area of loop closure, backbone reconstruction, and physical trajectory recovery. Through these experiments we demonstrate the relevance and robustness of our methods over a broad range of applications. Finally, we conclude in chapter 6 with a brief discussion on potential areas of extending this work and implications for computational structural biology.
Chapter 2

Protein Modeling

Capturing relevant conformational changes for biomolecules such as proteins involves having a good understanding of protein structure and flexibility. A well-defined characterization of protein energy is then needed to focus attention to structural fluctuations related to the function of a protein under physiological conditions. Therefore, sampling biologically relevant protein conformations requires that one not oversimplify protein models but capture the most relevant physico-chemical features of these macromolecules.

2.1 Protein Structure and Flexibility

Proteins, involved in virtually any physiological function, though displaying broad functional differences, share the same building blocks, aminoacids. There are 20 naturally occurring aminoacids in living organisms that dictate the shape, chemical properties, and function of proteins. Each aminoacid consists of an alpha carbon (C\(_\alpha\)) to which are attached a hydrogen atom, an amino group NH\(^2+\), a carboxylic group COO\(^-\), and one of 20 different “R” groups, also referred to as residues or sidechains. The R group characterizes an aminoacid. Two aminoacids link to each-other through a peptide bond which is the result of a reaction between the amino group of one aminoacid and the carboxylic group of the other. Aminoacids linked through the peptide bond form a polypeptide, as illustrated in Figure 2.1. Proteins consist of one or more polypeptides.

The one-dimensional sequence of aminoacids, referred to as the primary struc-
Figure 2.1: Four consecutive residues in a polypeptide chain

ture, is believed to determine the three-dimensional structure [Anf73] of proteins, referred to as the tertiary structure. Though possessing different tertiary structures, all proteins share common repeated three-dimensional blocks [Com71] such as α-helices, β-sheets, β-turns, and more that form the secondary structure of a protein. An illustration of some of these building blocks is given in Figure 2.2. While secondary structure elements are well-conserved among proteins, other fragments known as random coils or loops exhibit large structural variability. It is the compact three-dimensional arrangement of secondary structure elements and loops that form the tertiary structure of proteins.

Though originally thought to be rigid [Sch44], proteins are now classified as an important example of flexible macromolecules where conformational changes are strongly correlated with function [BKP88, FW94]. Therefore, a static view of a protein’s structure does not reveal much about its function. In the post-genomic era, it is well accepted that an ensemble of energetically stable protein conformations provides a more realistic characterization of the correlation between flexibility and function [WD99]. A classic example of this is the effect of the loop and domain motions in
the enzyme dihydrofolate reductase to this protein’s catalytic activity [CJO00, SK97]. Therefore, a study of protein flexibility is crucial to understanding its relationship to function.

Protein flexibility can be hierarchically classified into backbone and sidechain flexibility. The three-dimensional protein chain directed from the N-terminus to the C-terminus through the peptide bonds is known as the protein backbone. Motions of the backbone cause large-scale conformational changes in proteins. The aminoacid sidechains attached to the backbone also exhibit flexibility and are mainly responsible for local, smaller-scale, conformational changes.

A protein conformation needs to uniquely refer to a protein’s particular three-dimensional shape. Under the cartesian representation, a conformation for a protein with N atoms is a vector \( \mathbf{\tilde{C}} = \{A_{1x}, A_{1y}, A_{1z}, \ldots, A_{Nx}, A_{Ny}, A_{Nz}\} \), where \( \{A_{ix}, A_{iy}, A_{iz}\} \) refers to the cartesian coordinates for atom \( A_i \). Even though this representation uniquely defines the location of every atom \( A_i \), searching for conformations involves searching in a 3N dimensional space. Cartesian coordinates are redundant DOFs because a conformation can be more succinctly represented through \textit{internal coordinates}. 

Figure 2.2: (a) A \( \beta \)-turn, depicted in black, connects two \( \beta \)-sheets, drawn in grey as arrows pointed along the protein backbone. The \( \alpha \)-helix is drawn in silver. (b) Another representation of an \( \alpha \)-helix as a cylinder is given. A long loop connects the \( \alpha \)-helix to the \( \beta \)-sheet.
Figure 2.3: (a) Illustration of internal coordinates, where $b$ refers to bond length, $\alpha$ to bond angle, and $\theta$ to dihedral angle. (b) Rotation by the dihedral on the second bond induces spatial motion of the fourth atom and any consecutive atoms down the polypeptide chain.

An illustration of internal coordinates is given in Figure 2.3(a), where the bond length refers to the Euclidean distance between two covalently-linked atoms, the bond angle to the angle between two consecutive bonds, and the dihedral angle to the angle between three consecutive bonds. As Figure 2.3(b) shows, the dihedral angle is defined as the angle between the plane $\pi_1$ defined by the first and second bond and the plane $\pi_2$ defined by the second and third bond. Rotation by the dihedral on the second bond changes the position of the fourth atom in space and any consecutive atoms down the polypeptide chain. Internal coordinates reduce the dimensionality to $3N - 6$, because no internal coordinates need to be specified for the first atom in the chain, only one coordinate, bond length, needs to be specified for the second atom in order, and only two coordinates, bond length and bond angle, need to be specified for the third atom in the chain. Therefore, the 6 DOFs saved describe the global position and orientation of the protein.

Less DOFs are actually needed to specify a conformation, as not all bonds are rotatable. For instance, due to the double-bond nature, the peptide bond is rigid, as Figure 2.4(a) shows, forcing the co-planarity of the atoms involved in it and the
dihedral on that bond to a trans conformation of 180°. Although the peptide bond allows no rotation about it, the bonds between the carbon of the carboxyl group and the Cα atom and between the nitrogen of the amino group and the Cα allow rotations by angles labelled \( \phi \) and \( \psi \), respectively, as shown in Figure 2.4(b) and Figure 2.4(c). Aminoacids on the other hand can contribute at most 4 dihedral angles. Aminoacids with short or no sidechains contribute no dihedrals, as Figure 2.4(b) and Figure 2.4(c) indicate. Other sidechains, being bulky, are constrained in their motion and contribute no sidechain dihedrals, as Figure 2.4(d) shows. Aminoacids with long sidechains such as tryptophan or arginine contribute a maximum of 4 sidechain dihedrals, as shown in Figure 2.4(e) and Figure 2.4(f).

In addition, analysis of PDB protein structures reveals that bond lengths and angles do not show much variation among structures [GK97]. Such evidence can be used to assume an ideal or rigid geometry model, where the only DOFs are the dihedral angles. Considering only dihedrals focuses the exploration of conformational space to large and more interesting atomic rearrangements and reduces the dimensionality of this space because at most 6 DOFs - \((\phi, \psi, \chi_1, \chi_2, \chi_3, \chi_4)\) - are needed per residue.

### 2.1.1 Proteins as Robotic Manipulators

A protein can be modeled as a robotic manipulator, as Figure 2.5 shows. Using dihedrals as DOFs makes it possible to model proteins as manipulators with only revolute joints, where each polypeptide chain can be modeled as a kinematic chain. Modeling one polypeptide segment for proteins requires a set of kinematic chains: one for the backbone, and others for the sidechains connected to the backbone. Extending this model to include proteins with many polypeptides adds more backbone and sidechain kinematic chains. Even though manipulator kinematics associates a local frame with every atom, the total number of local frames needed is much less than the number of atoms in the molecule due to two backbone dihedrals \( \phi \) and \( \psi \) per aminoacid and a maximum of four dihedrals per sidechain [ZK02b].
Figure 2.4: (a) Electron sharing between the carboxyl carbon and the amide nitrogen gives the peptide bond a partial double-bond character and as a consequence its rigidity. (b) and (c) demonstrate the two backbone dihedrals $\phi$, $\psi$ on two aminoacids, glycine, and alanine. (d) No sidechain dihedrals for the rigid ring in Proline (e) Tryptophan contains 2 sidechain dihedrals (f) Arginine has the highest number of sidechain dihedrals, 4.
2.2 Protein Energy Landscapes

Conformational changes in proteins relate to favorable and unfavorable atomic interactions, such as interactions due to covalent and hydrogen bonds, electrostatic, ionic, Van Der Waals (VDW), hydrophobic and other weak interactions. The sum of all atomic interactions gives the protein its potential energy. Since different conformational states can possess the same potential energy, one needs to define a measure of chaos/disorder such as entropy $S$, which is directly proportional to the logarithm of the number of different conformational states with the same potential energy. Stable conformations under physiological conditions are a compromise between low potential energy and high entropy, a quantity captured in the definition of free energy $F = E - TS$. Conformations with low free energy are more stable than high free energy conformations. Naturally occurring proteins have been designated by evolution to transition, i.e. fold, to a most-stable native state, associated with a global minimum of the free energy $F$.

The potential surface of a protein is multi-dimensional due to the high dimensionality of conformational space and the intricate network of atomic interactions. Despite its dimensionality, the potential surface is statistically described through
the free energy landscape theory of protein folding [OLSW97, Gru02], which advocates the use of statistical mechanics to organize the multitude of protein conformational states in terms of a minimal number of collective parameters. The statistical formulation of the free energy landscape makes it possible to capture essential features of the free energy surface with only a limited set of parameters [CJO00, CNO00, CJO01, CGO03, COC04].

The modern statistical mechanical picture of protein folding shows a funneled energy landscape where the bottom of the well represents the native state of the protein [Wet73, Anf73]. Though steep, the energy landscape is not a smooth well but rather rugged due to the structural frustration of proteins [FW94, OLSW97]. Any transition of the protein between two conformations corresponding to local minima will create unfavorable interactions that result in barriers. Given the multidimensionality of the energy surface, these barriers are the reason behind the ruggedness of the energy landscape. Despite the ruggedness, proteins are minimally frustrated objects, with an energy landscape that has a smooth overall slope towards the native structure [OLSW97].

There are three main classes of energy surfaces ranging from surfaces with a single well-defined global minimum, corresponding to proteins with a very strong stability point, to surfaces with a few minima, and surfaces with a shallow basin [SW84, BK97, OLSW97], as illustrated in Figure 2.6. It should be noted that actual energy surfaces of proteins may be a combination of the three main cases. In this thesis we only consider the potential energy of protein conformations as measured by the CHARMM forcefield [MBB+98].

2.2.1 CHARMM Energy Modeling

Although the atomic interactions in proteins are well understood, only empirical models exist to model a protein's potential energy. In this work, we employ the CHARMM model, which was first introduced by [MBB+98] in the context of molec-
Figure 2.6: (a) There are multiple global minima in the energy landscape. (b) The native structure is not a strong stability point, resulting in a shallow basin. (c) The native structure is a strong stability point, reflected in the single global minimum.

ular dynamic simulations and continues to be successful for the study biomolecular flexibility [MSXK00, BK90, TNM92]. CHARMM is an all-atom force field, meaning that it considers interactions due to all atoms. Parameters to the CHARMM energy function have been empirically determined from experimental data and supplemented with ab-initio results. The CHARMM energy function has the form:

\[ \sum_{\text{bonds}} K_b (b - b_0)^2 + \quad \text{(Bond Term)} \]
\[ \sum_{\text{UB}} K_{\text{UB}} (S - S_0)^2 + \quad \text{(Urey-Bradley Term)} \]
\[ \sum_{\text{angle}} K_\theta (\theta - \theta_0)^2 + \quad \text{(Bond Angle Term)} \]
\[ \sum_{\text{dihedrals}} K_\chi (1 + \cos(n\chi - \delta)) + \quad \text{(Dihedral Term)} \]
\[ \sum_{\text{impropers}} K_\text{imp} (\phi - \phi_0)^2 + \quad \text{(Improper Term)} \]
\[ \sum_{\text{nonbonded}} \epsilon \left[ \left( \frac{r_{\text{min},ij}}{r_{ij}} \right)^{12} - \left( \frac{r_{\text{min},ij}}{r_{ij}} \right)^6 \right] + \quad \text{(Van Der Waals Term)} \]
\[ \sum_{\text{atoms}} \frac{q_i q_j}{\epsilon_1 r_{ij}} + \quad \text{(Electrostatic Term)} \]

where \( K_b, K_{\text{UB}}, K_\theta, K_\chi, \) and \( K_\text{imp} \) refer to empirically determined constants, \( b \) and \( b_0 \) to the current and equilibrium bond length, respectively, \( S \) and \( S_0 \) to the current and equilibrium Urey-Bradley 1,3-distance, respectively, \( \theta \) and \( \theta_0 \) to the current and equilibrium bond angle, respectively, \( \chi \) to the dihedral angle value, \( n \) to the periodicity, \( \phi \) to the improper angle value, \( \epsilon \) to the Lennard-Jones well depth, \( r_{\text{min},ij} \) to the atomic distance corresponding to the Lennard Jones minimum and \( r_{ij} \) to the observed distance between atoms \( i, j \), \( q_i \) and \( q_j \) to charges of atoms \( i \) and \( j \), and \( \epsilon_1 \)
to the effective dielectric constant.
Chapter 3

Related Work

Methods that can be applied to the satisfaction of spatial constraints for biomolecules fall mainly in three categories: inverse kinematics-based, search-based, and database methods. The difficulty of satisfying spatial constraints on two feature atoms increases with the length of the molecular chain between them, and available techniques are either limited or penalized by this.

3.1 Inverse Kinematics-based Methods

Focusing only on spatial constraints posed by two feature atoms, one can define the problem of satisfying these spatial constraints as closing the gap between the two feature atoms. Generating a fragment or a loop that will close this gap requires that the two mobile terminal loop residues, the mobile anchors connect with the anchors attached to the rest of the molecule, the stationary anchors. Closing a loop requires indeed that only one loop terminal residue be mobile, as one can restore connectivity on one end by simply attaching the other terminal residue to its corresponding stationary anchor, as illustrated in Figure 3.1. This problem is geometrically equivalent to finding values for a manipulator’s link variables so that its joints will be placed at specified positions. In the context of molecules under the ideal geometry representation, this problem translates to finding dihedrals that will steer the feature atoms to their target spatial positions. This problem is also known as inverse kinematics, introduced in the context of manipulators [Cra89].
Figure 3.1: (a) The generated fragments' mobile anchors, drawn in grey, are not attached to the stationary anchors drawn in black. (b) Fragments can be translated so one of the mobile anchors attaches to its corresponding stationary anchor. (c) Fragment is closed when both mobile anchors overlap with the stationary ones.

3.1.1 Classic Inverse Kinematics Methods

It is known that for manipulators with no more than 6 DOFs, there is a finite number of solutions to the inverse kinematics problem. There is, however, no analytical method that can find these solutions for all types of manipulators. For manipulators with only revolute joints, which is the case for biomolecules with idealized geometry, the number of unique solutions is at most 16, when the number of DOFs does not exceed 6 [RR89]. An efficient solution was proposed [MC94] and later applied to the conformational analysis of small molecular chains [MZ94, MZW95]. Methods based on curve approximation were proposed [Chi93] for the inverse kinematics of hyper-redundant robots, where the number of regularly distributed joints is very large.

Specialized solutions to inverse kinematics in biology appeared as early as 1970 [GS70], where fragments of up to 6 DOFs were predicted by solving a set of polynomial equations representing geometric transformations. These equations were applied to building tripeptide loops [GS70] under the ideal geometry assumption. Later work [BK85, PS91, MZW95, WS99] offered efficient analytical solutions for three consecutive residues through spherical geometry and polynomial equations.
Bounding inverse kinematic solutions for chains with no more than 6 DOFs within small intervals was applied in the context of drug design [ZK02a]. A new formulation that extends the domain of solutions to any three residues, not necessarily consecutive and with arbitrary geometry, was recently proposed [CSJD04]. Current work that pushes the dimensionality limit from 6 to 9 DOFs makes use of an efficient subdivision of the solution space [ZWW'04].

3.1.2 Inverse Kinematics with Optimization

Currently only optimization-based solutions are considered appropriate for accommodating chains with an arbitrary number of DOFs. Two well-known optimization-based inverse kinematics solutions that iteratively solve a system of equations until loops are closed are random tweak [FWS'86, SYF'87] and cyclic coordinate descent (CCD) [Lue84, WC91, CD03]. Both methods are based on iteratively changing dihedral DOFs of a fragment/kinematic chain until its terminal atom reaches a target position.

Random tweak relies on the computation of the Jacobian of the feature atom distances from their target positions with respect to the dihedrals which is computationally expensive and even numerically unstable as the Jacobian may loose rank. In addition to not being free from singularities, random tweak does not allow additional constraints on individual residues because modifications to dihedral angles are introduced all at once, with a strong dependence of each dihedral proposed change on all the others. Additional constraints on the dihedrals may result in the unpredictable motion of a feature atom away from rather than toward its target position.

Avoiding the use of the Jacobian, CCD is computationally inexpensive, numerically stable, and free from singularities. This method avoids the dependence of dihedrals on one another by adjusting one DOF at a time, which allows for additional constraints on dihedrals with a predictable motion of feature atoms towards their target positions. First introduced in the context of non-linear programming [Lue84],
this method was applied to robotics [WC91], and later was applied to the loop
closure problem for proteins [CD03, vdBLLD04, LvdBDL04, Lot04]. In particu-
lar, [vdBLLD04, LvdBDL04, Lot04] combined this method with motions in the self-
motion manifold, where local motions do not influence the spatial constraints and can
be used to move towards a local minimum of an objective function.

3.2 Search-based Methods

Search-based methods rely on a generate-and-test paradigm that generates many
fragment conformations and filters out the ones that cannot satisfy the spatial con-
straints. We categorize this class into two subclasses, loop construction methods and
motion-planning based approaches.

3.2.1 Loop Construction Methods

Searching the conformational space for fragments was seen by [BK87] as a way to ex-
tend the dimensionality limitation of loop closure algorithms by analytically solving
for small fragments and enumerating solutions for longer loop fragments. Combinator-
torial approaches [MJ86, BHN88, DS90, BVSD93, Bru93, FSBM94a, PM95, DB00]
discretize the solution space to limit a combinatorial explosion by restricting the set of
($\phi, \psi$) dihedral angles from uniform conformational sampling to distributions biased
towards more populated regions of the ($\phi, \psi$) map.

Other search algorithms propose local moves [EGE95], importance sampling by
local minimization of randomly generated conformations [LS89a, LS89b, LS89c],
or globally minimize energy by mapping a trajectory of local minima [DRP98].
Other energy function optimization-based search algorithms consist of molecular
dynamics simulations [BK90, TNM92, RT93, NHK00], Monte Carlo combined
with molecular dynamics [RF99], biased probability Monte Carlo [AT94, EMCG95,
TZM+97], and Monte Carlo with simulated annealing [HCG92, CHG93, CE93, CE96,
VCD94, FDS00]. Even dynamic programming algorithms [FR92] and genetic algo-
rithms [MJ93] have been applied to the loop closure problem. Other search algorithms consist of bond scaling with relaxation [ZRVD93], multi-copy searches [ZRDK94], and self-consistent mean field optimization [KD95]. All these search methods suffer from practical limitations on the number of DOFs in the fragment and low success rate due to the inherent method used to close a fragment generated through an exploration of conformational space.

3.2.2 Motion-Planning based Approaches

Searching for fragment conformations with spatial constraints has parallels in roadmap-based motion planning for sampling manipulator configurations that capture the connectivity of the configurational space, with a manipulator configuration being the equivalent of a molecular conformation. Therefore, robotic algorithms for planning the motion of closed kinematic chains become relevant for closing gaps. Central to all these algorithms is the use of the probabilistic roadmap [KSLO96] for sampling the configurational space. One way of obtaining closure conformations is by first sampling configurations that ignore spatial constraints and then enforce such constraints through gradient descent [YLK01], where neighbor configurations are connected by chaining local steps generated in the null space of the Jacobian. Another approach integrates spatial constraints in the sampling process by breaking the loop into an active part that is sampled with forward kinematics techniques, and a passive part that is closed through inverse kinematics [HA00]; neighbor configurations are connected by a local planner for the active part while allowing the passive part to follow the motion. An efficient extension of the above method for more DOFs exists [XA03]. Sampling the active part of the chain one DOF at a time ensuring that the active part’s endpoints are always reachable by the passive part of the chain is a natural extension of the above idea [CSL02]. The authors have recently applied their algorithm to the closure of long protein loops [CSRST04].

Motion-planning approaches have the computational advantage of closing frag-
ments of any size. However, these approaches often oversimplify the problem on account of the molecular models used and the energetic stability analysis of the closure conformations. As far as we know, there is no current work on the satisfaction of multiple spatial constraints, and no integration of spatial constraints with energetic constraints for a biologically relevant analysis of reported closure conformations.

3.3 Database Methods

Database methods developed in the context of closing loops for partially determined protein structures, first proposed in the context of electron density fitting, by Jones and Thirup [JT86], and similarly addressed in comparative modeling [CLP+87, CCLW89, SK90, TL92, Lev92, TME+93, LS94, MT96, LLL99]. They make use of protein homology, which rests on the assumption that proteins with aminoacid sequence similarity and related function ought to have similar structures as well. This assumption has proven successful to secondary structure prediction [BSST87, SOJB90, JSSB94], and spawned interest in loop modeling, one of the challenges of comparative modeling due to the high structural variability of loops across proteins. Database approaches rest on the assumption that loops already predicted in proteins submitted to the PDB or other protein structure databases provide natural examples to model unknown loops. The increasing number of structures in the PDB [ASPM97] provides for a good set of loops to be selected based on how well they match the missing loop at hand.

It was demonstrated [TL92] that the selected loop fragments from databases did not overlap satisfactorily with the protein termini on all instances. Efforts to improve this overlap combine the database approach with subsequent optimization and ranking based on energetical stability [vVK97, MJ86]. For antibody hyper-variable loops, database methods yield satisfactorily results [CLP+87, CLT+89, MT96, MTR+98] due to these loops forming very specific folds based on key residues. Statistical information, classification, and subsequent optimization coupled with ranking give database
methods the advantage of producing physically reasonable mainchain conformations
extremely fast for short loops [MJ86].

Database methods are limited by the lengths of the loops they can close
[FSBM94b, vVK97] and have practical limitations of 4 or 9 residues for [FSBM94b]
and [vVK97], respectively. Latest database methods [TBHM02] deal with the limitation
of the loop diversity present in the PDB through a divide and conquer approach,
where loops are recursively broken into equal-sized fragments that can be analyti-
cally solved, and then combined to yield loops of arbitrary length. Characteristic
of database methods, the quality of the loop prediction reported by [TBHM02] also
suffers from low RMSD resolution, with top predictions varying from 1.06 Å for three-
residue loops to 3.72 Å for eight-residue loops.
Chapter 4

An Algorithm to Generate Ensembles of Spatially and Energetically Constrained Conformations

After briefly reviewing molecular kinematics and our implementation of the CCD algorithm, we describe in detail our extensions starting first with an adaptation of CCD to minimally disturb protein stability. Then we propose an extension of CCD to the problem of satisfying multiple spatial constraints and present a method that samples conformations that satisfy such constraints. In addition, we propose an algorithm that combines multiple spatial constraints with energetic feasibility to produce ensembles of closure conformations that are energetically stable.

4.1 Protein Kinematics

Representing protein conformations as vectors of dihedral angles only, allows one to employ the modified-Denavit-Hartenberg (mDH) [Cra89, ZK02b] convention to describe the forward kinematics for molecules. The mDH attaches a cartesian coordinate system $F_i$ to each atom $A_i$, and defines the relative location of consecutive frames by a homogenous transformation matrix:

$$
T_i = \begin{pmatrix}
\cos(\theta_i) & -\sin(\theta_i) & 0 & 0 \\
\sin(\theta_i) \cos(\alpha_{i-1}) & \cos(\theta_i) \cos(\alpha_{i-1}) & -\sin(\alpha_{i-1}) & -\sin(\alpha_{i-1}) d_i \\
\sin(\theta_i) \sin(\alpha_{i-1}) & \cos(\theta_i) \sin(\alpha_{i-1}) & \cos(\alpha_{i-1}) & \cos(\alpha_{i-1}) d_i \\
0 & 0 & 0 & 1
\end{pmatrix}
$$

(4.1)

where $d_i$ represents the length of the bond between atoms $A_{i-1}$ and $A_i$, $\alpha_{i-1}$ the bond angle between $A_{i-2}$, $A_{i-1}$, and $A_i$, $\theta_i$ the dihedral angle formed by four consecutively
bonded atoms $A_{i-2}$, $A_{i-1}$, $A_i$, and $A_{i+1}$.

The position of any atom in the molecule can be determined by chaining matrices $T_i$ as defined in Equation 4.1. Not considering rigid motion of the molecule, given a sequence of bonds $b_j, b_{j-1}, \ldots b_1$ representing the chain from a particular atom $A_i$ to a designated anchor atom $A_0$ whose cartesian coordinates are known, the cartesian coordinates of atom $a$ can be computed by Equation 4.2:

$$[x_i \ y_i \ z_i \ 1]^t = T_j T_2 \cdots T_i [x_0 \ y_0 \ z_0 \ 1]^t$$  \hspace{1cm} (4.2)

### 4.2 Protein Kinematics with Spatial Constraints

Steering feature atoms to satisfy spatial constraints requires proposing dihedral angles by whose rotation one can obtain conformations where these atoms lie in a $\epsilon$-neighborhood of the specified positions. Figure 4.1 illustrates a protein with three feature atoms. Given an initial protein conformation $C_s$, a set of goal spatial locations \{F_1, F_2, \ldots, F_n\}, for a set of $n$ atoms currently spatially located at \{M_1, M_2, \ldots, M_n\}, and a tolerance parameter $\epsilon$, our goal is to find a closure conformation $C_g$ where the designated atoms' current locations lie in an $\epsilon$-neighborhood of the target locations, i.e. $|F_i - M_i|^2 \leq \epsilon^2$.

The method we employ in this thesis for calculating the rotations needed to steer one feature atom towards its specified position, CCD, was first introduced by [Lue84, WC91] and then reintroduced in structure biology for the problem of loop closure by [CD03]. As illustrated in Figure 4.2, let the current position of the only designated feature atom be $M$ and its target position be $F$. Given a dihedral bond on the kinematic chain from a given anchor atom to the designated feature atom, CCD finds a value by which to rotate the given bond so that the feature atom's updated position gets closer to its target position. For the algorithm below, let us assume that our protein is a simple polypeptide chain, with its end atom being $A_j$. To partially update the protein from the anchor atom $A_i$ to its end atom $A_j$ so that $A_j$ is steered towards a target position, we employ our STEER algorithm.
Figure 4.1: Illustration of a protein with three feature atoms depicted.

Figure 4.2: CCD Schematic - rotation by $\alpha$ around the $O$ axis steers feature atom from current position $M$ to target position $F$. 
Algorithm 1 STEER \((C_s, A_i, A_j, M_j, F_j, \epsilon, n_{\text{max}})\)

**Input:**

- \(C_s\): current protein conformation
- \(A_i\): anchor atom
- \(A_j\): feature atom
- \(M_j = [M_{jx}, M_{jy}, M_{jz}]\): current position of feature atom \(A_j\)
- \(F_j = [F_{jx}, F_{jy}, F_{jz}]\): target position of feature atom \(A_j\)
- \(\epsilon\): tolerance closure parameter
- \(n_{\text{max}}\): maximum number of iterations

**Output:**

Target closure conformation \(C_g\)

1: \(n \leftarrow 0\)

2: while \(|M_j - F|^2 \geq \epsilon^2\) and \(n \leq n_{\text{max}}\) do

3: \(B \leftarrow\) dihedral bonds in kinematic chain from \(A_i\) to \(A_j\)

4: \(B_\sigma \leftarrow\) permutation of \(B\)

5: repeat

6: \(b_{\text{curr}} \leftarrow\) pop \(B_\sigma\)

7: \(\theta_{\text{curr}}\) dihedral in \(C_s\) corresponding to \(b_{\text{curr}}\)

8: \(\alpha \leftarrow\) CCD\((M, F, b_{\text{curr}})\)

9: \(\theta_{\text{curr}} \leftarrow \theta_{\text{curr}} + \alpha\)

10: \(A_k \leftarrow\) ancestor of \(A_i\) on bond \(b_{\text{curr}}\)

11: for all \(A_i\) in the kinematic chain past \(A_k\) do

12: \(M_i \leftarrow T_i \cdot M_k\) \hspace{1cm} //Forward kinematics

13: end for

14: if \(n \geq n_{\text{max}}\) or \(|M - F|^2 \leq \epsilon^2\) then

15: TERMINATE

16: end if

17: until \(B_\sigma\) is empty

18: \(n \leftarrow n + 1\)

19: end while
The STEER algorithm first depends on the definition of a kinematic chain between the anchor atom $A_i$ and the feature atom $A_j$. The rotatable bonds on that kinematic chain are the ones that can be used to rotate and therefore steer the feature atom towards its target position. The algorithm proceeds down the rotatable bonds of this kinematic chain, finding maximum angle values by which to rotate the current dihedral bond so that the feature atom gets as close as possible to its target position. Once such a value is computed, through forward kinematics, we rotate the rest of the kinematic chain and update the position of the feature atom. The algorithm proceeds iteratively through all the dihedral bonds of the kinematic chain until the feature atom lies in an $\epsilon$-neighborhood of its target position or we run out of patience, a measure of whether the satisfaction of the spatial constraint is possible at all.

We now describe in detail our implementation of the CCD mathematical routine that reports the angle $\alpha$ by which to rotate the current dihedral bond $b_{\text{curr}}$ so as to get the feature atom as close to the target position as possible. As schematically illustrated in Figure 4.2, the rotation axis $O$ is given by the direction of the bond $b_{\text{curr}}$ corresponding to the dihedral angle $\theta_{\text{curr}}$ that will be modified so that the overlap between the moving position $M$ and the target position $F$ is below the tolerance parameter $\epsilon$. The overlap between the current and desired position is denoted by $S = |FM|^2$, where $FM = OM - OF$. Let's denote $OM = \bar{r}$, $OF = \bar{f}$, the norm of $\bar{r}$ and $\bar{f}$ by $r$ and $f$, respectively, the unit vector of $\bar{r}$ by $\hat{r}$, the unit vector along the rotation axis by $\hat{\theta}$, and the axis orthogonal to $\hat{r}$ and $\hat{\theta}$ by $\hat{s}$. Then we can write:

$$O\bar{M} = r \cos(\theta) \hat{r} + r \sin(\theta) \hat{s}$$  \hspace{1cm} (4.3)

Then it follows that

$$FM = r \cos(\theta) \hat{r} + r \sin(\theta) \hat{s} - \bar{f} \equiv d$$  \hspace{1cm} (4.4)

Calculating the measure of overlap $S$ we obtain:

$$|d|^2 = r^2 + f^2 - 2r \cos(\theta) (\bar{f} \cdot \hat{r}) - 2r \sin(\theta) (\bar{f} \cdot \hat{s})$$  \hspace{1cm} (4.5)
Minimizing the error measure $S$ amounts to looking for solutions of $\theta$ that correspond to critical points of $S$:

$$\frac{dS}{d\theta} = \frac{d(|d_t|^2)}{d\theta} = 2r \sin(\theta) (\hat{\dot{r}} \cdot \hat{s}) - 2r \cos(\theta) (\hat{\dot{r}} \cdot \hat{s})$$

(4.6)

Solving for $\theta$ yields:

$$\tan(\alpha) = \frac{(\hat{\dot{r}} \cdot \hat{s})r}{(\hat{\dot{r}} \cdot \hat{r})r}$$

(4.7)

If we solve Equation 4.7 for $\alpha$, we obtain two different solutions $\pi$ radians apart from each other, corresponding to local minima and maxima of function $S(\theta)$. To distinguish between the two, one needs to look at the second derivative of $S$ with respect to $\theta$. Therefore, since we want a solution for $\theta$ that minimizes $S$, the correct solution is the one that produces a positive value of the second derivative of $S$. Even though such an analysis will yield the right solution for $\theta$, it is possible to obtain the right solution for $\theta$ in a much simpler way, without needing to compute the second derivative.

Note from Equation 4.5 that if we denote $2r$ ($\hat{\dot{r}} \cdot \hat{r}$) by $b$, $2r$ ($\hat{\dot{r}} \cdot \hat{s}$) by $c$, and $r^2 + f^2$ by $a$, it follows that:

$$S = a - b \cos(\theta) - c \sin(\theta)$$

(4.8)

According to Equations 4.8 and 4.7, it is easy to see that $\tan(\alpha) = \frac{c}{b}$. To see why when $\theta = \alpha$ $S$ reaches its minimum, requires some more manipulation. Multiplying Equation 4.8 by $\frac{\sqrt{b^2 + c^2}}{\sqrt{b^2 + c^2}}$ and defining $\cos(\alpha) = \frac{b}{\sqrt{b^2 + c^2}}$ and $\sin(\alpha) = \frac{c}{\sqrt{b^2 + c^2}}$, we can write:

$$S = a - \sqrt{b^2 + c^2} \cos(\theta - \alpha)$$

(4.9)

With $S \geq 0$, it is easy to see that, when $\theta = \alpha$, since $\cos(\theta - \alpha)$ reaches its maximum value 1, $S$ reaches its minimum.

**Design Choices** Note that the STEER algorithm adjusts one dihedral at a time to move the end-effector toward the target. Because of this, the order of which dihedral bonds are rotated first does not affect the steering of the feature. However, different
permutations $B_\sigma$ of $B$ may produce different conformations for the kinematic chain from $A_i$ to $A_i$. Even though the flexibility of the algorithm to work with any permutations $B_\sigma$ may seem a desirable property, the effect of the choice of permutations is not known until the new conformations are generated. Although the identity permutation taxes more the dihedral bonds that lie at the beginning of the kinematic chain, after experimenting with different probabilistic permutations we conclude that a permutation which is less invasive to the protein's energetic constraints remains elusive.

Since CCD involves solving one equation in one unknown for each dihedral, the STEER algorithm is computationally very fast and runs in time linear to the number of dihedrals on the kinematic chain from $A_i$ to $A_j$. Unlike classic inverse-kinematics solutions that use Jacobian matrices [FWS+86, SYF+87] or general numerical approaches, CCD is free of singularities and does not depend on initial guesses for solutions. Compared to inverse kinematics techniques with optimization that suffer from high computational times, CCD is computationally fast. Unlike other methods such as random tweak, CCD gives predictable behavior and suffers from no anomalies when additional constraints are added to the dihedrals. Such properties make CCD very appealing to the satisfaction of spatial constraints in the context of protein kinematics with energetic constraints.

4.2.1 Loop Closure

The loop closure problem is a special case of satisfying spatial constraints since it specifies two mobile anchors, aminoacids that need to attach to the rest of the protein at the stationary anchors. Let us order the mobile anchors along the backbone from the $N$-terminus of the protein to the $C$-terminus of the protein, with loop termini $A_j$ following loop termini $A_i$. Since a closed loop has to attach its mobile anchors to the stationary anchors, $A_i$ is really no feature aminoacid. Closing the loop then, involves solving for dihedrals in the kinematic chain from $A_i$ to $A_j$ so that $A_j$ attaches to the
rest of the protein.

Our algorithm STEER is not well-suited to the loop closure problem, where the end-effector and the target positions are not single points, but three backbone atoms N, Cα, and C of the C-terminal residue of the loop that need to be superimposed on the corresponding atoms of the fixed C-terminal anchor of the protein. Note that these three atoms define a plane that needs to be steered towards its target in the protein. Modifying our STEER algorithm to accomplish this involves simply solving three simultaneous constraints in the CCD technique so that any θ solutions are the ones that minimize the error

\[ S = |F_N \vec{M}_N|^2 + |F_{Cα} \vec{M}_{Cα}|^2 + |F_C \vec{M}_C|^2. \]

Being a greedy method, STEER fails in rare cases of short highly-extended loops. However, as loops get longer, the convergency rate increases. Extending CCD to the satisfaction of multiple features lying anywhere in the protein is not as straightforward as simply increasing the number of constraints for CCD to satisfy. Minimizing distance errors for all feature atoms at once results in problems such as slow convergence and even deadlock situations. What follows is a more detailed description of such problems and our solution to the satisfaction of multiple spatial constraints.

4.3 Extension: Protein Kinematics with Multiple Spatial Constraints

Satisfying multiple constraints can be seen as trying to close multiple loops. Trying to iteratively satisfy all feature atoms at once is not the correct extension of STEER to multiple spatial constraints. The reason is that when the features atoms are far apart along the chain, trying to minimize all deviations from goal positions at once results in a very slow convergence. An angle that brings a feature closer to its target may bring another one farther away from its own (transformation matrices accumulate and therefore rotation by a dihedral affects multiple features along the polypeptide
chain). Thus, a deadlock situation can be very possible under such a naive extension, as no further optimization can be possible while the designated feature atoms are still far away from their target positions.

The correct approach to the satisfaction of multiple constraints is a serial satisfaction of one spatial constraint at a time. This approach guarantees that features that have already been steered close enough to their target positions will not be moved away from these positions when moving down the polypeptide chain to satisfy more spatial constraints. This serial satisfaction of spatial constraints demands an ordering of the features along the backbone and a definition of all the dihedrals on non-overlapping loops that need to be closed.

We order features along the protein backbone from the $N$ to the $C$ terminus. Given a sequence $a$, $b$, $c$ of feature atoms where $b$ lies on a sidechain and $a$ and $c$ lie on the backbone, the loops that will be produced are $a \rightarrow b$ and $d \rightarrow c$ where $d$ is the $C_\alpha$ atom of the aminoacid where $b$ sits, as shown in Figure 4.3. When trying to steer atom $b$ towards its goal position, assuming that atom $a$ has already been steered, the position of atom $d$ will change due it being a part of the loop $a \rightarrow b$. We define a pseudo-feature such as $d$, which by convention we deem to be the $C_\alpha$ atom of the aminoacid at hand. After closing loop $a \rightarrow b$, since the position of atom $d$ is updated correctly, we can close loop $d \rightarrow c$. In this way, all feature atoms are steered towards their target positions sequentially, as their order determines.

Under such an order, extending our STEER algorithm to MULTIPLESTEER is straightforward, as shown in Algorithm 2.

### 4.4 Extension: Energy Considerations

While the structure integrity of the closure conformations is automatically maintained under the idealized geometry model, maintaining structure stability is more challenging due to the fact that our steering algorithms have no inherent considerations of the energy. Even though working in dihedral space minimizes detrimental
Algorithm 2 MULTIPLESTEER \((C_s, A_0, A, F, \epsilon, n_{\text{max}})\)

Input:

\(C_s\): current protein conformation

\(A_0\): anchor atom denoting beginning of protein backbone

\(A = \{A_1, \ldots, A_n\}\): feature atoms

\(F = \{F_1, \ldots, F_n\}\): target positions for feature atoms \(A\)

\(\epsilon\): tolerance closure parameter

\(n_{\text{max}}\): maximum number of iterations

Output:

Target closure conformation \(C_g\)

1: \(n \leftarrow 0\)

2: \(A_{\text{ordered}} \leftarrow\) ordered atoms from \(A\)

3: \(M_{\text{ordered}} \leftarrow\) positions of \(A_{\text{ordered}}\) in \(C_s\)

4: \(F_{\text{ordered}} \leftarrow\) target positions of \(A_{\text{ordered}}\)

5: for all \(A_j\) in \(A_{\text{ordered}}\) do

6: \hspace{1em} if \(A_j\) is first feature then

7: \hspace{2em} \(A_i \leftarrow A_0\)

8: \hspace{1em} else

9: \hspace{2em} \(A_i \leftarrow A_{i-1}\) \hspace{1em} //Feature \(A_{j-1}\) already steered

10: end if

11: \(M_j \leftarrow M_{\text{ordered}}[A_j]\)

12: \(F_j \leftarrow F_{\text{ordered}}[A_j]\)

13: \text{STEER}(C_s, A_i, A_j, M_j, F_j, \epsilon, n_{\text{max}})

14: end for
effects on the energy, rotations by dihedrals cause large-scale motions of the protein, potentially causing the most unfavorable of interactions, steric clashes. Taking into consideration energetic constraints besides spatial constraints requires modeling energy explicitly and understanding the implications of all atomic interactions. Since our steering algorithms works in dihedral space, the only worrying energy terms according to our CHARMM modeling remain: Dihedral energy term, VDW energy term (steric clashes), and the electrostatic energy term.

4.4.1 Dihedral Energy

Our steering algorithms, being purely geometric routines, not observe the secondary or tertiary structure of proteins. One can try to insert knowledge into the CCD proposed dihedrals by observing an average distribution of $\phi$, $\psi$ dihedrals for every aminoacid as provided from Ramachandran histograms [RRS63] computed from a non-redundant representative set of proteins found in nature. One way we have explored is to define a window of $\phi$, $\psi$ values centered around the dihedral values for $\psi$, $\phi$ proposed by the
CCD technique. Then CCD can traverse this window to choose either one dihedral or a set of dihedrals that are most likely found in nature, i.e., have the highest probability in the Ramachandran histogram, and are not very different from the CCD proposed dihedral. This can be accomplished with the following objective function:

\[ w_1 \cdot \text{abs}([\phi, \psi]_{\text{CCD}} \cdot [\phi, \psi]_{\text{min}}) + w_2 \cdot (1 - P([\phi, \psi]_{\text{min}})) \]

where \( w_1, w_2 \) weigh the importance of each term.

Modifying CCD to pick its suggestions from this window results actually in limiting its ability to satisfy spatial constraints. We also observed that besides impoverishing the satisfaction of spatial constraints, observing the Ramachandran maps did not minimize the energy of the closure conformation in any significant way. The reason for this is that observing the statistical distribution of dihedrals in nature for a specific aminoacid while tends to bound the dihedral energy, it does not affect detrimental long range interactions such as steric clashes. Since the Ramachandran histograms are averages for a specific aminoacid, they cannot model the effect of neighbor aminoacids on the dihedrals, which is important for avoiding steric clashes. As a result, in this thesis we do not limit our steering algorithms by the Ramachandran suggested values beyond bounding its solutions to the interval \([-\pi, \pi]\).

Since Ramachandran histograms reflect the population of \((\phi, \psi)\) values over secondary structure elements, we decide to model the secondary structure of a given protein explicitly, basing our decision on the fact that the steering algorithms should not stretch or distort the secondary structures in any way. Considering \(\alpha\)-helices and \(\beta\)-sheets as segments with limited flexibility actually helps in producing closure conformations whose secondary structure will not deviate from that of the native conformation. We tag every atom with the secondary structure segment it belongs to according to the Kabsch-Sander's algorithm [KS83]. The steering algorithms then bound any solutions reported by CCD for dihedrals that lie on \(\alpha\)-helices or \(\beta\)-sheets to lie below some empirically determined upper bound. In this way, satisfying spatial constraints will not distort the secondary structure segments to a detrimental effect.
on the energy of the closure conformation.

4.4.2 Long Range Interactions

Even the explicit knowledge about the secondary structures will not help in avoiding steric clashes between atoms. Long range interactions due to VDW interactions are the most detrimental in the total energy of a conformation. We observe that there are two ways our steering algorithms can cause steric clashes:

i. Sidechains collide with other sidechains or portions of the coiled backbone.

ii. Backbone portions collide with one another.

Sidechain Collisions Although sidechains with no features on them are not directly affected by our steering algorithms, they experience conformational changes due to the accumulation of transformation matrices. We employ a gradient descent in dihedral space while exploring different sidechain conformations to relieve tensions due to steric collision, a process we refer to as flapping sidechains. Flapping sidechains allows for local motions in dihedral space that while not producing significant conformational changes, minimizes the overall energy and does not disturb features already steered.

Backbone Collisions Even though sidechain flapping may help reduce the overall energy, most of the steric clashes are due to the backbone. While it is important to avoid high energy conformations that geometrically satisfy the features, all implemented energy minimization techniques and secondary structure constraints are not very powerful as they suffer from local minima issues. To deal with the serious issue of steric clashes, we use the fact that depending on starting conformations $C_s$, our steering algorithms can produce different closure conformations $C_g$. Such a feature is very desirable as we may explore the dihedral space for conformations that sat-
isyfy spatial constraints and are energetically stable, and our answer to add energy considerations into the satisfaction of spatial constraints.

4.5 Extension: Sampling Conformations with Spatial and Energetic Constraints

A specification of the feature atoms divides all dihedrals in a molecule into two groups: the first one consists of dihedrals whose values affect the position of the feature atoms and should therefore be bounded through solutions from the CCD routine. The second group consists of dihedrals that can take different values without affecting the feature atoms at all. This second group defines regions in conformational space where conformations, though different, maintain their closure. These regions make the so-called self-motion manifold [ZK02c], which can be explored for conformations that are energetically stable. Therefore, to address energy concerns, we explore regions of the self-motion manifold.

Searching in the self-motion manifold has the potential of generating conformations that are energetically feasible and spatially constrained. Figure 4.4 schematically shows three different closure conformations sampled in the self-motion manifold. Out of all these three conformations, one possesses an α-helix, schematic indication that this conformation is more native-like and, therefore, energetically stable. Exploring the self-motion manifold is the approach we employ to obtain closure conformations that are also energetically feasible. We combine our steering algorithms with sampling of the self-motion manifold and energetic filtering of the obtained closure conformations, so the resulting closure conformations have energies comparable to that of the native structure.

Our MULTIPLESTEERWITHENERGY algorithm, Algorithm 3, samples conformations uniformly in dihedral space, steers such conformations, and rejects the ones that are not feasible. For the loop closure problem, a satisfaction of constraints to \( S = 0 \) degree guarantees a high success rate for closure conformations that are native-like, as
Figure 4.4: Three schematic conformations (drawn in red, green, and blue) sampled in the self-motion manifold satisfy spatial constraints posed by three feature atoms (drawn in red.)

the only mobile portion of the protein is the loop.

4.6 Extension: Recovering Physical Conformations

The MULTIPLESTEER algorithm can also be used to generate physical conformations that exhibit a particular motion vector $v$. The motion vector $v$ can be used to first identify feature atoms that best capture the motion and their spatial constraints. Given the motion vector $v$, we conduct a displacement-in-time analysis that ranks atoms according to their traversed distance along vector $v$ in one time unit. Such an analysis effectively sorts atoms according to a measure of their mobility and so provides a set of feature atoms $A$. Given a current closure conformation $C_s$, the spatial constraints $F$ the next recoverable conformation needs to satisfy can be defined as $F = C_s[A] + \lambda \cdot \vec{v}$, where $\lambda$ denotes step size along vector $\vec{v}$. Once the feature atoms are chosen and the spatial constraints are computed, energetically feasible conformations that exhibit the desired motion $v$ can be generated through our PHYSICALMOTION algorithm, which combines MULTIPLESTEER with a relaxation by a conjugate gradient energy minimization technique [Mez94]. The energy minimization technique we employ in this thesis steers atoms in the direction of the gradient of the CHARMM forcefield to return a conformation whose associated energy is a low minimum in the
Algorithm 3 MULTIPLESTEERWITHENERGY \( (C_{\text{native}}, \text{e.size}, \text{cutoff}, A_0, A, F, \epsilon, n_{\text{max}}) \)

**Input:**

\( C_{\text{native}} \): native conformation  
\( \text{e.size} \): ensemble size  
\( \text{cutoff} \): energy cutoff  
\( A_0 \): anchor atom denoting beginning of protein backbone  
\( A = \{A_1, \ldots, A_n\} \): feature atoms  
\( F = \{F_1, \ldots, F_n\} \): target positions for feature atoms \( A \)  
\( \epsilon \): tolerance closure parameter  
\( n_{\text{max}} \): maximum number of iterations

**Output:**

ensemble of conformations \( C \)

1: native_energy ← energy of \( C_{\text{native}} \)
2: curr.conf ← 1
3: while curr.conf ≤ e.size do
4: sample conformation \( C_s \)
5: MULTIPLESTEER(\( C_s, A_0, A, F, \epsilon, n_{\text{max}} \))
6: curr.energy ← energy of \( C_{\text{native}} \)
7: if \( |\text{curr.energy} - \text{native.energy}| \leq \text{cutoff} \) then
8: curr.conf ← curr.conf + 1
9: end if
10: end while
energy surface.

Algorithm 4 PHYSICALMOTION \((\vec{v}, \lambda, C_s, A_0, A, \epsilon, n_{\text{max}}, \text{opt}_{\text{max}})\)

**Input:**

\(\vec{v}\): principal component vector
\n\(\lambda\): step size along \(\vec{v}\)
\n\(C_s\): native protein conformation
\n\(A_0\): anchor atom denoting beginning of protein backbone
\n\(A = \{A_1, \ldots, A_n\}\): feature atoms
\n\(\epsilon\): tolerance closure parameter
\n\(n_{\text{max}}\): maximum number of iterations
\n\(\text{opt}_{\text{max}}\): maximum number of optimization steps

**Output:**

- Ensemble of closure conformations \(E\)

1: while TRUE do
2: \(F \leftarrow C_s[A] + \lambda \cdot \vec{v}\)
3: \(n_{\text{opt}} \leftarrow 0\)
4: repeat
5: \(C_g \leftarrow \text{MULTIPLESTEER}(C_s, A_0, A, F, \epsilon, n_{\text{max}})\)
6: \(\text{MINIMIZE}(C_g)\) //Conjugate gradient minimization
7: \(n_{\text{opt}} \leftarrow n_{\text{opt}} + 1\)
8: until \(C_g\) is feasible or \(n_{\text{opt}} \geq \text{opt}_{\text{max}}\)
9: \(E \leftarrow E \cup C_g\)
10: end while
Chapter 5

Experiments

This chapter is organized in three sections. We first apply our algorithms to the loop closure problem on loops with lengths well above the limitations of current methods. The second and third sections apply our algorithms to the problem of sampling physical conformations with multiple spatial constraints. In particular, we consider the problem of reconstructing a physico-chemically meaningful protein structure from $C_\alpha$ coordinates only. The last section places our work in the context of drug design, by simulating the motion of atom groups with energetic constraints.

5.1 Loop Closure

Loops are the most mobile and least structurally conserved fragments in proteins. As a consequence, loop modeling remains a challenge for structural biology. The lack of reliable and fast algorithms for modeling long loops [TLM01], i.e., loops of more than five residues, demands the development of new computational approaches. In the following we show that our method performs successfully with long loops. All loop closure simulations are conducted on proteins with loops of length 12, 20, and 26 residues, well beyond the limitations of current loop closure algorithms.

The results presented in this section demonstrate that our method for sampling conformations with spatial and energetic constraints succeeds not only in completing protein models by closing long loops, but more importantly in producing good ensembles of conformations for proteins with flexible or unstructured loops. We demonstrate that for proteins with a known single stable conformation sampling loop closure
conformations with energetic constraints reveals a funnel-like energy landscape. We confirm experimental hypotheses of the existence of multiple minimal for proteins with missing loops by obtaining an energy landscape with a broad flat basin.

5.1.1 Proteins Used and Methodology

Our choice of proteins for the loop closure simulations depends not only on the length of the loop but also on what is experimentally known about the conformational stability of each protein. We choose two proteins each with a strongly stable native state and show that the loop closure conformations we produce are essentially identical to the native conformation. The third protein we choose contains a missing loop. The fact that this missing loop could not be resolved by X-ray crystallography suggests that this part of the protein may be inherently unstructured. We complete the VlsE all-atom model by producing a large ensemble of loop closure conformations, many of them exhibiting comparable energies even if structurally very different. This result confirms the hypothesis of the disordered structure of missing loops. It should be noted that due to the 2.2 Å and 1.70 Å resolution for the first two PDB reported protein structures, we perform a short conjugate gradient energy minimization technique [Mez94] to obtain starting conformations for our experiments. These conformations have potential energies of −1882 kcal/mol and −1288 kcal/mol for chymotrypsin inhibitor 2 and α-lactalbumin, respectively.

1COA Our first protein is the asymmetric unit of chymotrypsin inhibitor 2 (CI2) that serves as a serine-type endopeptidase inhibitor involved in the response to wound-
ing. For our experiments we use the 2.2 Å resolution crystal structure of CI2, (PDB code 1COA), with 64 residues-1057 atoms in total. The native conformation of the protein, shown in Figure 5.1(a), consists of one α-helix and two β-sheets that are connected to each-other through a loop of 12 residues. The 12-residue loop provides a good testing bed for our methods, and its experimentally known structural stability
provides an opportunity to also test the energy landscape associated with the loop closure conformations we generate. In agreement with the known role of this loop for the activity of the protein [LD94, JF94], we obtain an energy landscape that contains a sharp minimum around the native conformation (see Figure 5.3(b)).

1ALC The α-lactalbumin protein is an α/β protein that is involved in the calcium ion binding during lactose synthase. We use an X-ray crystal structure of 1.7 Å resolution (PDB code 1ALC). This protein contains 122 residues and a total of 1930 atoms, as can be seen in Figure 5.1(b). 1ALC contains a loop of 26 residues between an α-helix and a β-sheet. Since this loop is very long and highly constrained, it represents a good test for our loop closure simulations. In agreement with experimental data [VPDK03] we show that the energy landscape contains a sharp minimum around the native conformation (see Figure 5.4(b)). Moreover, we reproduce the experimentally observed peak in the loop mobility around residue 64 [VPDK03] (see Figure 5.5).

VLSE1 The variable surface antigen VlsE of Borrelia burgdorferi, (PDB code VLSE1), is an α protein that is part of the immune system. This protein is found in the cellular membrane of the bacteria borrelia burgdorferi. VLSE1 contains 291 residues and a total of 3938 atoms. The 2.3 Å resolution X-ray crystal structure [JWS03] of VlsE is incomplete due to a missing loop of 20 residues between two α helices. The inability of X-ray crystallography to report cartesian coordinates for this loop is hypothesized to be a consequence of the loop's inherent structural disorder [JWS03]. The available VLSE1 structure can be seen in Figure 5.1(c), with the two terminal residues of the missing loops labelled. We confirm such a hypothesis by obtaining an energy landscape (see Figure 5.6(b)) with a broad plateau.

Methodology and Parameters For each of the proteins used we sample 5,000 loop closure conformations through our MULTIPLE STEER WITH ENERGY algorithm.
Figure 5.1: (a) and (b) The native loops of 1COA and 1ALC lie between residues VAL53 and ASP64, and LYS51 and THR76, respectively. The Cα atoms of the terminal residues of these loops are shown as red spheres, while the rest of the proteins are shown with their secondary structure segments in ribbon representation. (c) and (d) The desired missing loop in VLSE1 needs to connect PRO92 and THR113.
We limit the sampled conformations by rejecting any whose energy difference from the native conformation is outside a 600 kcal/mol cutoff. For each of the proteins we compare the sampled loop closure conformations by plotting them over the native conformation to qualitatively see the geometric variability of the sampled loops. We quantify this comparison by plotting the energetic difference of the sampled loops from their geometric difference from the native. We plot the energetic difference as the unitless quantity $\frac{\Delta E}{kRT}$, where $\Delta E$ is the energetic difference from the native in units of kcal/mol, and $kRT$ refers to our 600 kcal/mol cutoff, with $R$ referring to the Boltzmann constant per mol and $T$ to room temperature of 300 Kelvin.

A parameter allowed to vary in our loop closure simulations is the maximum number of iterations needed by the STEER algorithm. This number affects the degree of loop closure, and therefore the orientation of the protein fragments after the loop. Figure 5.2(a) demonstrates that 50 iterations for the loop closure of 1COA are not enough to leave the orientation of the rest of the protein undisturbed. One needs 500 iterations for an undisturbed orientation, as can be seen in Figure 5.2(b). Requiring that there be no error in the orientation of the C-terminal mobile loop anchor guarantees that the conformation for the rest of the protein will be native-like, and therefore, any energy fluctuations will be due to the loop conformations only. We determine that a maximum number of iterations of 5000 is enough for no error in orientation for 1ALC and VLSE1, as can be seen in Figure 5.2(c) and Figure 5.2(d,e) respectively.

5.1.2 Loop Closure Ensemble for 1COA

In our MULTIPLESTEERWITHENERGY algorithm, we sample loop dihedrals uniformly to obtain different loop configurations. Unlike other sampling techniques employed in [CSRST04], there is no need to bound the energy values of these sampled conformations because no error in orientation will restore the portion of the protein that is native-like, thus naturally localizing any energy infeasibility to the loop con-
Figure 5.2: (a) and (b) 1COA loop is depicted as a string of beads, with black and purple beads denoting atoms involved in peptide bonds. The C-terminal loop residue depicted in red is not perfectly oriented with 50 iterations (a) and requires 500 iterations for no error in the orientation (b). (c) A loop closure 1ALC conformation with 5000 iterations shown in blue and superimposed on the native loop in red shows perfect orientation. (d) Depicted in red are the Cα atoms of the residues that need to be connected. (e) VLSE1 loop closure conformation shown in blue reaches perfect orientation with 5000 maximum iterations.
formation only. We limit the sampled conformations by rejecting any whose energy difference from the native conformation is outside the 600 kcal/mol cutoff. This filtering is necessary to reject loops that cause steric clashes. For the case of 1COA, because the loop is not too constrained between the two β-sheets, the rejection rate is not very high.

Sampling 1COA loop closure conformations with energetic constraints reveals loop closure conformations that are mere fluctuations around the native conformation. This result is shown in Figure 5.3(a), where the 1COA loop closure conformations exhibit low variability from the native structure. The measure of conformational similarity can be quantified through the least root mean squared deviation (RMSD), which is the Euclidean distance between two superimposed conformations $C_i, C_j$ (to remove translation and rotation) distributed over all $n$ atoms (Euclidean distance divided by $\sqrt{n}$). Figure 5.3(a) shows that the 1COA loop closure conformations exhibit low variability from the native structure. The energy landscape associated with these conformations is shown in Figure 5.3(b). Clearly, this energy landscape is funnel-like, consistent with the fact that C12 has a well-defined and stable native structure [LD94], where the loop region is important for the stability and activity of the protein [JF94].

5.1.3 Loop Closure Ensemble for 1ALC

Constraining 1ALC loop closure conformations by energy produces an ensemble of conformations that are mere fluctuations around the native structure, consistent with experimental data [VPDK03]. Figure 5.4(a) shows the low geometric variability of all sampled 1ALC loop closure conformations from the native conformation. This result is better illustrated by the potential energy of the conformations vs. their RMSD from the native structure. Figure 5.4(b) shows that the energy landscape associated with the ensemble of 1ALC loop closure conformations is clearly funnel-like.

Figure 5.4(a) reveals that the 1ALC loop is very constrained as it is very long and
Figure 5.3: (a) There is low geometric variability in the reported 1COA loop closure conformations when the total energy with both backbone and sidechain contributions is considered. (b) The energy landscape of the loop closure conformations is funnel-like.
Figure 5.4: (a) There is low geometric variability in the feasible 1ALC loop closure conformations. (b) The energy landscape associated with this ensemble of loop closure conformations is funnel-like.

 goes through stable secondary structure elements while avoiding collisions with them. In particular, the beginning and end of the loop residues of the loop are very constrained, as rotations even by small dihedral angles on these residues accumulate on the long loop and produce large conformational changes. This result is also illustrated in the energy landscape in Figure 5.4(a), where fluctuations by small RMSD from the native produce large energetic fluctuations. The highest structural variability of the sampled loop closure conformations seems to be around residue 64, as Figure 5.4(a) shows.

Such a result is in complete agreement with the study in [VPDK03], where the largest fluctuations of this loop are found to exhibit a peak around residue 64 (see Figure 5.5(a)). Figure 5.5(b) shows the full agreement of our results with those in [VPDK03] by directly computing the RMSD loop fluctuations per residue. Figure 5.5(c) shows the ensemble of conformations obtained in [VPDK03] to correspond
Figure 5.5: (a) Structural fluctuations (RMSD) per residue in the native ensemble of conformations obtained in [VPDK03] (red line) compared with the X-ray B factors (green line). (b) RMSD fluctuations per residue for the conformations obtained in our simulation are in complete agreement with those obtained in [VPDK03]. (c) Superimposed conformations obtained in [VPDK03], with the most variable protein regions drawn in red. (d) Superimposed conformations obtained in our simulation, with the loops drawn in red, strongly resemble those obtained in [VPDK03].
to our results in Figure 5.5(d).

5.1.4 Loop Closure Ensemble for VLSE1

Completing the VLSE1 structure requires finding loop conformations that close a gap of 20 residues. The missing VLSE1 loop is highly mobile, therefore not resolved by X-ray crystallography. It has been shown [JWS03] that even NMR or cryoEM cannot characterize this loop, since VLSE1 is a protein with very low thermodynamic stability for a protein its size, i.e., an ensemble of conformations is needed to characterize its loop.

The high mobility of this loop is mainly due to the fact that, although long, it connects two neighboring α-helices whose Euclidean distance from each-other is much lower than the extended loop length. Therefore, physically meaningful loops need to extend away from the two neighboring α-helices, in order to avoid steric clashes. The STEER algorithm, greedy in nature, needs to be directed by energy, in order to produce loops that first extend out of the helices and then bend back toward them.

Sampling VLSE1 loop closure conformations reveals that there are many geometrically variable VLSE1 loop closure conformations that are relevant at room temperature. This result fully supports the evidence that VLSE1 is a protein with low stability and a highly mobile loop, that cannot be seen through X-ray crystallography. The high conformational heterogeneity of the loop closure conformations can be seen in Figure 5.6(a). The geometric variability of these loop closure conformations can also be seen Figure 5.6(b), where the energy landscape is plotted as a function of the RMSD from the most stable conformation generated by our simulation (with energy -2174 kcal/mol). Clearly, the energy landscape associated with the ensemble of loop closure conformations we generate is plateau-like, supporting the high conformational heterogeneity of the loop at room temperature.

The ensemble of loop closure conformations we produce is a better characterization of VLSE1’s flexibility than a single structure. The conformational heterogeneity of
Figure 5.6: (a) There is high geometric variability in the feasible VLSE1 loop closure conformations. (b) VLSE1 energy landscape resembles a plateau, consistent with experimental evidence.
this protein due the flexibility of its loop can only be captured through our ensemble of stable loop closure conformations. The results shown in Figure 5.6(b) demonstrate the importance and relevance of looking at an ensemble of structures for proteins with inherently disordered fragments.

5.1.5 On the Use of Backbone-only Approximation

Since energy computations are the most time-consuming portion of our MULTIPLESTEERWITHENERGY algorithm, one appealing compromise is to compute only VDW energy terms for the protein backbone, with no consideration for the sidechain contributions to steric clashes. Such a view is very common in computational structure biology since sidechain addition to an already stable backbone is thought to be a simple task [YW04].

In this fashion, we sample 1COA loop closure conformations that avoid backbone steric clashes and assume rigidity of sidechains, as shown in Figure 5.7(a). However, the results of Figure 5.7 are not particularly good. Multiple loop conformations with comparable energy are obtained, although they are structurally very different from the native conformation. The sampled conformations seem to exhibit a two-fold symmetry around the native conformation, that is not supported by experimental data [LD94, JF94]. Projecting the energy values of the obtained 1COA conformations on their RMSD distances from the native conformation reveals a poorly funneled energy landscape in Figure 5.7(b), that does not exhibit a well-defined minimum around the native conformation.

To investigate more the observed symmetry, we cluster the 1COA loop closure conformations according to their RMSD distances from one-another. We use the k-means technique efficiently implemented in [PM99] through a kd-tree of RMSD distances, and use medians rather than means as centroids of the clusters to avoid outliers and un physical centroid conformations. Our clustering confirms the two-fold symmetry, with about half of the loop closure conformations in each one of the two
reported clusters, as can be seen in Figure 5.7(c). Plotting the similarity matrix of the clustered conformations in Figure 5.7(d) reveals that the two clusters are well-defined. Such a clustering indicates that only one of the clusters is centered around the native-conformation, (shown in red in Figure 5.7(c)). Figure 5.7(e) presents a closer look at three representative conformations of each cluster. When comparing the two medians of the clusters in Figure 5.7(f), we also see that one of them resembles the native, whereas the other one is 2.5 Å RMSD away from it.

The experimental evidence of only one well-defined stable loop closure conformation prompted us to investigate the reliability of the commonly used assumption that backbone only modeling provides reasonable structures. Indeed, close investigation of loop closure conformations that are reported as part of the second cluster reveals that these conformations have steric clashes between their sidechains, as shown in Figure 5.8. We conclude that considering only steric clashes due to the backbone of the protein may produce conformations that do not allow physical sidechain fitting. Figure 5.8(a) shows that the loop backbone is too close to the following β-sheet to accommodate sidechains. Figure 5.8(b) provides a closer look into a steric clash between a sidechain carbon atom of loop residue 61 and a sidechain hydrogen atom of β-sheet residue 81.

In order to avoid such steric clashes between sidechains, we cannot ignore sidechain flexibility in the protein model. This conclusion supports similar claims about the appropriate modeling of sidechains [ACSR97]. It is necessary to consider interactions from all atoms of the protein and filter closure conformations by considering VDW interactions over the whole protein. We have confirmed that such a modification keeps the dihedral, electrostatic, and improper energy terms low, as the following figures indicate. Figure 5.9 shows that the dihedral, improper, and electrostatic energy terms of 1000 sampled conformations for 1COA protein are mere fluctuations of the respective energy contributions in the native conformation. However, considering VDW interactions over all atoms of the protein is as expensive as considering
Figure 5.7: (a) A large scale search reveals a variety of 1COA loop conformations. (b) The potential energy landscape associated with the obtained conformations does not present a single minimum. (c) Clustering shows two major clusters: cluster 1 in red with 523 conformations and cluster 2 in blue with 483 conformations. (d) The similarity matrix reveals that the two clusters are well-defined. (e) Representatives are not all native-like (native loop in cyan). (f) While one of the medians is native-like, the other is very different from the native.
the total potential energy. Therefore, all reported conformations in this work are constrained by their total CHARMM energy, which considers both backbone and sidechain contributions.

5.2 Backbone Reconstruction

We have tested our algorithm in the context of structure reconstruction from $C_\alpha$-only coordinates, an essential problem to connect minimalist models to all-atom simulations and experimental data. Addressing this problem involves obtaining cartesian coordinates for all atoms from only $C_\alpha$ coordinates. While database methods that use statistical information for distances between backbone atoms [MKS97] have been quite successful in backbone reconstruction, we can offer a new solution to the same problem with our methods and obtain comparable results, while guaranteeing the energetical feasibility of our ensemble of conformations. This comparison presents a robustness test for our MULTIPLESTEER algorithm.

The existence of cartesian coordinates for the $C_\alpha$ atoms poses natural spatial
Dihedral, Improper, and Electrostatic Energy Fluctuations
for sampled closure conformations

Figure 5.9: Dihedral, improper, and electrostatic energy terms are insignificant fluctuations around the native conformation of CI2.
constraints, with every $C_\alpha$ atom designating a feature. Then, in the context of our sampling with spatial and energetic constraints, the problem of backbone reconstruction consists in obtaining closure conformations where every $C_\alpha$ atom is constrained to specific cartesian coordinates. Even though the number of spatial constraints to be satisfied by our \textsc{MultipleSteerWithEnergy} algorithm is $n$, one constraint per residue, the satisfaction of these constraints is guaranteed by the fact that they come from realistic conformations. Moreover, if the $C_\alpha$ coordinates come from native-like conformations, the obtained closure conformations are also going to be native-like, localizing unfavorable atomic interactions to sidechains mainly. Sampling closure conformations that go through all the specified $C_\alpha$ atoms and bounding their energy yields an ensemble of conformations that are native-like.

We test this concept on the CI2 protein by assuming that the $C_\alpha$ coordinates of the native conformation are the ones we are given as spatial constraints. Since this protein has 64 residues, there are 64 constraints that need to be satisfied. Using our sampling method we obtain multiple closure conformations for 1COA that are energetically feasible. Figure 5.10(a) shows that the reconstructed conformations lie very close the native conformation. To confirm the robustness of our method, we plot the energy of the sampled conformations vs. their RMSD from the native conformation. The resulting funnel-like energy landscape shown in Figure 5.10(b) reiterates the strong thermodynamic stability of CI2. Figure 5.10(c) shows that the most stable closure conformations are native-like.

To compare our performance with the statistical approach employed in [MKS97], we apply our method to backbone reconstruction of the list of 14 conformations reported in [MKS97]. For every one of the 14 proteins we use our method to obtain the most energetically stable conformation that satisfies the spatial constraints posed by each $C_\alpha$ atom. Table 5.1 compares our results to those reported in [MKS97]. For each protein we report its size (in terms of number of residues), the RMSD between the $C_\alpha$, N, C, O, and $C_\beta$ atoms of the native and reported conformations, respectively.
Figure 5.10: (a) 1COA closure conformations are clustered around the native. (b) The funnel-like energy landscape of the reconstructed conformations confirms that CI2 has a well-defined native structure. (c) The five most stable conformations lie at most 0.2 Å RMSD away from the native 1COA conformation.
The obtained structure (in cyan) is also shown superimposed on the native (in blue). We also report a backbone average RMSD, and juxtapose our results to those reported in [MKS97].

It is worth noting that the $C_\alpha$ RMSD of the reconstructed structure is not 0 because it depends on the degree of satisfaction of the spatial constraints. The RMSD of the carbonyl oxygen atom is the highest for our reported conformations, as in [MKS97], due to the flexibility of such atoms for structural variations that do not affect energetical feasibility significantly. We could not report RMSD for the $C_\beta$ atom for 3APP, 5CPA, 2CTS, 2PRK, and 1TIM, due to their large number of residues which pose a computational burden on the energy calculations. To sample conformations for such proteins we reduced the respective proteins from a full-atom model to a backbone model. These conformations can be refined even further through our method so as to better satisfy the spatial constraints. However, the goal of this application is not to report better RMSD values than those reported in [MKS97], but to test the robustness of our method and the additional property of our solution, which guarantees the energetic feasibility of the reconstructed conformations.

5.3 Physical Trajectory Recovery

HIV Protease, a protein of the HIV virus, is of great interest to drug discovery because it assists the replication of the HIV virus. Designing drugs that can block the functional site of this protease can essentially inhibit its function, and therefore stop the HIV virus from further replication. As all virus proteins, the HIV protease can accommodate frequent mutations of the aminoacids on its binding site with no detrimental effect to its stability and function. In the contrary, these mutations give a lot of flexibility to the binding site of this protease, making it very challenging to design drugs that can fit in a frequently changing pocket. Drug discovery for the HIV protease remains challenging because current techniques do not model the flexibility of this protein.
Table 5.1: RMSD (Å)

<table>
<thead>
<tr>
<th>Protein</th>
<th>Size</th>
<th>C&lt;sub&gt;α&lt;/sub&gt;</th>
<th>N</th>
<th>C</th>
<th>O</th>
<th>C&lt;sub&gt;β&lt;/sub&gt;</th>
<th>BB</th>
<th>Superimposed</th>
<th>[MKS97]</th>
</tr>
</thead>
<tbody>
<tr>
<td>2ALP</td>
<td>171</td>
<td>0.203</td>
<td>0.256</td>
<td>0.355</td>
<td>0.922</td>
<td>0.390</td>
<td>0.434</td>
<td></td>
<td>0.453</td>
</tr>
<tr>
<td>3APP</td>
<td>323</td>
<td>0.827</td>
<td>0.737</td>
<td>0.901</td>
<td>1.527</td>
<td>x</td>
<td>0.998</td>
<td></td>
<td>0.416</td>
</tr>
<tr>
<td>5CPA</td>
<td>307</td>
<td>0.992</td>
<td>0.822</td>
<td>0.903</td>
<td>1.247</td>
<td>x</td>
<td>0.991</td>
<td></td>
<td>0.480</td>
</tr>
<tr>
<td>1CRN</td>
<td>46</td>
<td>0.398</td>
<td>0.318</td>
<td>0.589</td>
<td>1.032</td>
<td>0.642</td>
<td>0.584</td>
<td></td>
<td>0.408</td>
</tr>
<tr>
<td>1CTF</td>
<td>68</td>
<td>0.109</td>
<td>0.111</td>
<td>0.171</td>
<td>0.433</td>
<td>0.157</td>
<td>0.206</td>
<td></td>
<td>0.461</td>
</tr>
<tr>
<td>2CTS</td>
<td>437</td>
<td>1.026</td>
<td>0.880</td>
<td>0.955</td>
<td>1.303</td>
<td>x</td>
<td>1.041</td>
<td></td>
<td>0.369</td>
</tr>
<tr>
<td>2FOX</td>
<td>138</td>
<td>0.884</td>
<td>0.696</td>
<td>0.805</td>
<td>1.192</td>
<td>1.292</td>
<td>0.894</td>
<td></td>
<td>0.414 0.522</td>
</tr>
<tr>
<td>2MHR</td>
<td>118</td>
<td>0.613</td>
<td>0.605</td>
<td>0.605</td>
<td>0.769</td>
<td>0.621</td>
<td>0.648</td>
<td></td>
<td>0.457</td>
</tr>
<tr>
<td>2PRK</td>
<td>279</td>
<td>0.880</td>
<td>0.729</td>
<td>0.846</td>
<td>1.384</td>
<td>x</td>
<td>0.960</td>
<td></td>
<td>0.358</td>
</tr>
<tr>
<td>6PTI</td>
<td>57</td>
<td>0.090</td>
<td>0.100</td>
<td>0.163</td>
<td>0.432</td>
<td>0.173</td>
<td>0.196</td>
<td></td>
<td>0.381</td>
</tr>
<tr>
<td>1TIM</td>
<td>247</td>
<td>0.853</td>
<td>0.847</td>
<td>0.846</td>
<td>0.846</td>
<td>x</td>
<td>0.848</td>
<td></td>
<td>0.595</td>
</tr>
<tr>
<td>1UBQ</td>
<td>76</td>
<td>0.322</td>
<td>0.334</td>
<td>0.463</td>
<td>0.968</td>
<td>0.586</td>
<td>0.522</td>
<td></td>
<td>0.324</td>
</tr>
<tr>
<td>9WGA</td>
<td>170</td>
<td>0.154</td>
<td>0.148</td>
<td>0.220</td>
<td>0.533</td>
<td>0.491</td>
<td>0.264</td>
<td></td>
<td>0.450</td>
</tr>
<tr>
<td>2WRP</td>
<td>104</td>
<td>0.788</td>
<td>0.652</td>
<td>0.716</td>
<td>0.977</td>
<td>1.171</td>
<td>0.783</td>
<td></td>
<td>0.212</td>
</tr>
</tbody>
</table>
Therefore, the focus of many computational techniques is to classify and model the flexibility of proteins like HIV protease. Principal component analysis [TPK03] of molecular dynamic simulations of HIV protease around its native conformation have established that the first three principal components correspond to physical coordinated motion of the loop regions sealing the binding site.

The motion of HIV protease and other proteins subjected to principal component analysis is simulated by linearly translating all atoms of the protein along the principal components. A trajectory obtained in this fashion is useful to decouple the motion of protein regions. A sample trajectory obtained by linearly translating the HIV protease atoms along the first principal components is shown in Figure 5.11(a). This motion is referred to as the opening/closing of the flaps, most affected by the motion. Zooming on the flaps in Figure 5.11(b) illustrates this motion.

Such trajectories, though informative, are not physical, as atoms linearly translated along principal components are forced to stretch bonds and angles and, therefore, create unfavorable interactions. Although linearly interpolated conformations are not physical, the principal components provide evidence of a particular kind of motion, the opening and closing of the flaps. Through our method, it is possible to combine such evidence with the concern for energetic feasibility to obtain physical conformations that, while undergoing conformational changes described by such a motion, are energetically feasible, i.e., relevant. Such conformations are of great interest to the process of drug design, as they can be used to find candidate drugs that best dock to the HIV protease.

**Recovering HIV Protease Opening of the Flaps** We interpret the motion of the protein as spatial constraints on atoms that best capture the motion. A simple analysis of displacement of each atom along the first principal component reveals in Figure 5.11(c) that most of the motion, as expected, is in the flaps, residues 45 to 55 for the first polypeptide chain, and residues 145 and 155 for the second polypeptide...
Figure 5.11: (a) Superimposed HIV protease conformations are obtained from linear translations of the native conformation along the first principal component. (b) Zooming on the flaps (residues 45-55 drawn in red) reveals their opening/closing motion. (c) Atoms on the flaps are the ones with the largest displacement along the first principal component.
chain. Therefore, through such an analysis we designate 5 atoms from the above-concluded range of residues to be feature atoms. These atoms are features for our PHYSICAL MOTION algorithm that recovers conformations that exhibit conformational changes described by a motion vector. The recovered conformations are shown superimposed in Figure 5.12(a), with zooming on the flaps shown in Figure 5.12(b).

The energetic feasibility of the recovered conformations with respect to the native conformation is demonstrated in Figure 5.12(c). As we can see, the energy values initially drop, only to level for a while, and then rise again. This is consistent with the fact that our recovered conformations go from open flap conformations to native-like and then close flap conformations. We are able to capture more of the conformations around the native, as they are energetically more stable.

Figure 5.12(c) shows in red the open-flap recovered conformations and in blue the close-flap recovered conformations. As seen in Figure 5.12(c), we are able to recover more open-flap conformations due to the fact the the closing flap motion is more constrained as the flaps come too close with the rest of the protein. Note that since we minimize as much as needed to bring the energy of the recovered conformations down to the range \([-kRT, kRT]\), the recovered open-flap conformations surround the energy parabola. A plot of the energy fluctuations with respect to the RMSD fluctuations of the recovered conformations in Figure 5.12(d) reveals that the recovered open-flap and close-flap conformations are symmetric around the native conformation. For the additional open-flap conformations we are able to recover, we again observe the funnel-like shape of the energy landscape.

5.4 Discussion of Experimental Results

In the loop closure application of our work, we reveal ensembles of loop closure conformations that are energetically feasible. Our choice of input data combines proteins with long loops that have to go through extremely difficult compact regions, proteins with well-characterized energy landscape, and proteins with an incomplete structure
Figure 5.12: (a) HIV protease conformations recovered from the linearly translated conformations along the first principal component are shown superimposed. (b) The motion of the flaps (residues 45-55 drawn in red) is shown more detailed. (c) The closing motion of the flaps is more energetically constrained. This can be seen in the fact that only up to 100 conformations, in blue, with step of size 0.1 Å can be recovered. On the other hand, the opening motion of the flaps is less energetically constrained, resulting in 258 conformations (d) The energy landscape of HIV protease demonstrates the symmetry of the open- and close-flap conformations and the overall funnel-like shape of the landscape.
due to a missing loop. For the Cl2 protein with a loop of length 12 residues, we produce ensembles of loop closure conformations and a corresponding energy landscape that is fully consistent with the characterization of this protein as having a strong stable native state [LD94, JF94]. α-lactalbumin, a protein with a very long loop of 26 residues, provides a challenging case with the long loop forced to go through a narrow region between secondary structures. For this protein too, the energy landscape associated with the ensemble of loop closure conformations generated by our method agrees with experimental evidence that α-lactalbumin has a very stable native state. An analysis of the per-residue flexibility of the energetically constrained loop closure conformations for this protein supports the results of a different study [VPDK03] reporting that the most flexible loop residues are those around aminoacid 64 [VPDK03].

This thesis also applies our methods to VlsE, a protein with a missing loop of length 20. We provide an ensemble of conformations for the known aminoacid sequence of the loop and show that their associated energy landscape is plateau-like. This supports wet-lab experimental evidence that VlsE has a very weak stability for a protein its size [JWS03].

Our algorithm's robustness and broad relevance is demonstrated on its application to the backbone reconstruction problem, where we sample ensembles of reconstructed all-atom models from Cα mainchain atoms only. The energy landscape of these conformations for Cl2 again supports the characterization of this protein as being very stable [LD94, JF94]. We apply our algorithm to a list of 14 proteins and obtain comparable results with the statistical approach employed in [MKS97]. In addition, our methods report all-atom models that are energetically stable.

This work also provides a feasibility study to model conformational changes in proteins. We have applied our methods in this context to the physical trajectory recovery of HIV protease. Through our methods, we obtain energetically feasible conformations that undergo collective fluctuations characterized by a motion vector.
Chapter 6

Discussion

This thesis provides a novel approach that combines the satisfaction of multiple spatial constraints with energetic constraints for biomedically relevant molecules. This work supports evidence that an ensemble of energetically constrained conformations properly addresses the relationship between dynamics and function [WD99]. Our methods are applicable to any biomolecule where one can define spatial constraints.

In this thesis, values for dihedrals are uniformly sampled in the interval $[-\pi, \pi]$, which makes our method not specific to proteins, but applicable to any flexible biomolecule. While our approach can accommodate a specialized Ramachandran distribution for proteins, our first attempts reveal that this distribution is not particularly informative, as it models no dependence on neighboring aminoacids, and is not effective in avoiding long-range steric clashes. To our knowledge, limitations of the Ramachandran distribution have been raised recently in the structural and computational biology community [FR04, vdBL04]. Therefore, one of the prospects that remains open for future work is developing a more descriptive distribution of dihedrals that integrates the presence of neighboring aminoacids. To our knowledge, this has never been applied in the context of conformational sampling in dihedral space.

The CCD algorithm has been criticized for placing a large burden on the first few dihedrals. We have experimented with different permutations of the dihedrals to observe their impact on the results. Our initial tests reveal that the identity permutation performs as well as the rest in practice.
Our methods are applicable to protein folding, where conformations can undergo fluctuations constrained by designated residues in close spatial proximity to one another. Our algorithms can be applied to sample such conformations, where the residues in contact with one another are designated as features. Accommodating residues that are hypothesized to be in contact with some probabilistic measure of confidence is a simple straightforward extension to our STEER algorithm. Sampling such conformations is very important to extensively characterize regions of the folding landscape where wet-lab experimental data provides evidence that certain residues or atoms remain close to another. On a broader context, our methods can also be used to study the effects of mutations in the context of the function-related flexibility of molecules. Indeed, by finding cartesian coordinates for all atoms from constrained Cα atoms, our approach readily accommodates any aminoacid, effectively modeling point mutations in a protein.

In this thesis we consider only the CHARMM potential energy and an energetic cutoff of 600 kcal/mol from the native conformation. We are currently considering ways to incorporate solvation energetic and entropic effects that together with configurational entropy will allow us to measure the free energy associated with conformational states. Free energy computations will allow us to consider a statistical mechanics formulation of the feasibility of conformational states under a given set of conditions.
Bibliography


[CD03] A. A. Canutescu and R. L. Dunbrack. Cyclic coordinate descent: A


