Molecules in Motion: Computing Structural Flexibility

by

Amarda Shehu

RICE UNIVERSITY

Molecules in Motion:

Computing Structural Flexibility

by

Amarda Shehu

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

Approved, Thesis Committee:

Lydia E. Kavraki, Professor, Chair
Computer Science

Cecilia Clementi, Associate Professor
Chemistry

Luay Nakhleh, Assistant Professor
Computer Science

Houston, Texas
July 2008

Abstract

Molecules in Motion: Computing Structural Flexibility

by

Amarda Shehu

Growing databases of protein sequences in the post-genomic era call for computational methods to extract structure and function from a protein sequence. In flexible molecules like proteins, function cannot be reliably extracted from a few structures. The amino-acid chain assumes various spatial arrangements (conformations) to modulate biological function. Characterizing the flexibility of a protein under physiological (native) conditions remains an open problem in computational biology.

This thesis addresses the problem of characterizing the native flexibility of a protein by computing conformations populated under native conditions. Such computation involves locating free-energy minima in a high-dimensional conformational space.

The methods proposed in this thesis search for native conformations using systematically less information from experiment: first employing an experimental structure, then using only a closure constraint in cyclic cysteine-rich peptides, and finally employing only the amino-acid sequence of small- to medium-size proteins.

A novel method is proposed to compute structural fluctuations of a protein around

an experimental structure. The method combines a robotics-inspired exploration of the conformational space with a statistical mechanics formulation. Thermodynamic quantities measured over generated conformations reproduce experimental data of broad time scales on small ($\sim$ 100 amino acids) proteins with non-concerted motions. Capturing concerted motions motivates the development of the next methods.

A second method is proposed that employs a closure constraint to generate native conformations of cyclic cysteine-rich peptides. The method first explores the entire conformational space, then explores in present energy minima until no lower-energy minima emerge. The method captures relevant features of the native state also observed in experiment for $20 - 30$ amino-acid long peptides.

A final method is proposed that implements a similar exploration but for longer proteins and employing only amino-acid sequence. In its first stage, the method explores the entire conformational space at a coarse-grained level of detail. A second stage focuses the exploration to low-energy regions in more detail. All-atom conformational ensembles are obtained for proteins that populate various functional states through large-scale concerted motions. These ensembles capture well the populated functional states of proteins up to 214 amino-acids long.

Acknowledgments

I thank my advisor, Dr. Lydia Kavraki, for her guidance, mentoring, and support, and Dr. Cecilia Clementi for trying hard to turn me into a physicist. Special thanks go to Dr. Luay Nakhleh, Dr. Ronald Goldman, Dr. Mark Moll, Dr. Moshe Vardi, and the Kavraki and Clementi groups for their instructive comments. My gratitude goes to the NAMD, AMBER, and VMD communities, and researchers like John Stone, Ilya Yildirim, Carlos Simmerling, and David Case. I owe special thanks to my husband, Erion Plaku, and the rest of my family for unconditional love and support.

This work was supported by NSF (CHE-0349303, ITR-0205671, CCF-0523908), NIH (GM078988), the Welch Foundation (Norman Hackermann Young Investigator award and C-1570), the Advanced Texas Technology Program (003604-0010-2003), the Sloan Foundation, Rice University Funds, and by a fellowship awarded to Amarda Shehu from the Nanobiology Training Program of the W. M. Keck Center for Computational and Structural Biology of the Gulf Coast Consortia funded by NIH (1 R90 DK71504-01). Equipment was supported in part by NSF (EIA-0216467) and the Rice Computational Research Cluster funded by NSF (CNS-0421109 and CNS-0454333) in a partnership between Rice University, AMD and Cray.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the current age, computer scientists are increasingly reaching across disciplines and employing computing to simulate the behavior of physical systems. Often, the understanding of the underlying physics that determines the evolution in time and space of the system is incomplete. Incomplete understanding gives way to approximations and, inevitably, inherent error. Hence, efficiency of computations is just one concern. Algorithms designed for physical systems have to capture the correct behavior, despite the approximations and inherent error, and hopefully discover novel information that improves understanding of the rules guiding a physical system.

This thesis develops methods to conduct biological research in silico. In particular, the methods presented in this thesis aim to describe the behavior of protein molecules. The methods bridge between computer science and life sciences such as biophysics, chemistry, and biology.

Focus in proteins is warranted due to the ubiquitous presence of these molecules in cells. Proteins are central to most cellular processes. They interact with diverse molecules, such as other proteins, DNA, and RNA. For example, proteins bind to

DNA and RNA to mediate regulation of gene expression and transcription, DNA replication, and mRNA intron splicing. Protein-protein interactions play a role in antibody-antigen binding, large-scale organismal motion, and even cell adhesion. This array of interactions allows proteins to participate in important molecular complexes and machines in cells. The emerging disciplines of nanobiology and biomimetics employ proteins as building blocks to engineer novel molecular machines.

Interest in characterizing the behavior of a protein emerges for two main reasons: (i) understanding a protein's biological function may elucidate the role of that protein in a disease and ultimately design drug molecules to activate or inhibit the protein; (ii) a microscopic picture of how a protein associates with other molecules opens up new venues of employing protein molecules to construct molecular machines for therapeutic purposes or for the design of novel functional materials.

1.1 Extracting Protein Function From Native Structure

A long-held assumption in the understanding of proteins was that they were mainly rigid molecules [Sch44] that used geometrical complementarity of molecular shapes (structures) in a lock-and-key scheme to discriminate among molecular associations. Elucidation of the three-dimensional (3D) structure in which a protein carried out its biological function, referred to as the native structure, was held key to understanding what other molecules a protein could recognize and bind.

The post-genomic revolution, with its focus on sequencing organismal genomes

and finding protein-encoding regions, has resulted in rapidly growing databases of protein sequences. More than five million unique protein sequences exist in these databases [WMH04,LRO07]. The usual determination of the protein native structure by experimental techniques, such as X-ray crystallography and Nuclear Magnetic Resonance (NMR), is rapidly falling behind. Experimental techniques have resolved structures for less than 1% of the protein sequences deposited [LRO07], even as the National Institute of Health (NIH) urges faster structural determination through structural genomics projects such as the Protein Structure Initiative (PSI) [oGMS05].

The experimental process of resolving a protein sequence is faster than that of associating a native structure to the sequence. The rise of computational biology proposes computational methods to complement experimental techniques in extracting structural information from a protein sequence [Dod07]. Entire competitions such as the Critical Assessment of Techniques for Protein Structure Prediction (CASP) are devoted to advancing research in this direction [Cen94].

1.2 Extracting Function From Native Flexibility

Most proteins are not rigid but instead exhibit internal motions ranging from local atomic fluctuations to global rearrangements [BKP88, FW94]. Evidence from experiment, theory, and computation suggests that protein motions are often essential for function [PFO99, SDW04, SSB05]. For example, loop and domain movements in the enzyme dihydrofolate reductase are crucial to its catalytic activity [CJO00,SK97].

In flexible molecules such as proteins, function often relates with the ability of the protein to change shape as needed, for instance, to accommodate other molecules for binding [SDW04, SC04]. Proteins employ motions to modulate function and intermolecular associations. It is now widely recognized that functional information for a protein cannot be reliably extracted from one or a few native structures populated by a protein under physiological conditions [KK05, HM05].

The linear chain of amino acids that make up a protein molecule can assume various spatial arrangements in 3D space. In this thesis, these arrangements will be referred to as conformations. Thus, a more accurate description of the state in which a protein carries out its biological function is not a native structure but a native state, possibly comprised of a large ensemble of conformations [SSB05].

Experimental techniques such as X-ray crystallography and NMR can provide few conformations available to a protein under native conditions. NMR experiments additionally report statistical averages over conformations comprising the native state. These averages provide a macroscopic picture of the underlying microscopic ps-ms motions that a protein undergoes under native conditions [Kay05]. No detailed information is available on the actual ensemble of native conformations that lie underneath the measurements.

The incomplete picture obtained from experimental techniques calls for computational methods to complement NMR and other wet-lab techniques to characterize the flexibility of a protein under native conditions. Some computational methods

characterize native flexibility by analyzing one or a few native structures to reveal mobile regions, locate hinges, flaps, or provide vectors along which atoms move under native conditions [CB05, NH07]. This is not the approach taken in this thesis. The methods developed here characterize native flexibility by computing the ensemble of conformations populated by a protein under native conditions. This approach allows obtaining a detailed view of the native state in all its conformational diversity. Various measurements then taken over computed conformations as ensemble averages allow analyzing mobility, locating fluctuations, and comparing directly with experiment.

The functionality of a protein depends on its energetic stability, which is the result of atomic interactions that stabilize or destabilize protein structure. Stabilizing forces bring a protein to its native state, associated with the lowest energy [Anf73]. Therefore, obtaining conformations that comprise the native state involves computing conformations that are energetically feasible, i.e., of low energy.

For a naturally-occurring protein, its experimentally-determined structure submitted to the Protein Data Bank (PDB) [BWF$^+$00, BKW$^+$77] provides an average over the conformations making up the native state. The degree to which this average structure captures the native state depends on the dynamics of a protein. Fluctuations around this structure 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 fluctuations around the average structure. In this case, the ensemble of native conformations is homogeneous and fluctuations around the

average structure are insignificant. For other proteins that can accommodate large-scale motions around the average structure with no detrimental effects on function, low-energy conformations may be structurally very different from the average. Indeed, diverse conformations are sometimes employed by molecules to interact with different partner molecules [MK84, Mea88, OD90].

1.3 Characterizing Native Flexibility

Though Anfinsen formulated that the amino-acid (aa) sequence determines a protein's native conformation(s) [Anf73], finding native conformations in silico remains challenging [Dod07]. In fact, the problem of threading a linear chain of amino acids into a lowest-energy conformation is NP-hard [UM93].

The physico-chemical factors regulating the relationship between a protein native conformation and biological function are not fully understood. Moreover, the internal motions that a protein exhibits under native conditions may occur on very different timescales. Sixteen orders of magnitudes (femtoseconds-seconds) in timescales are relevant for protein motions. No method (wet-lab or in silico) can currently span on its own this broad range of timescales. Indeed, many computational methods employ information obtained from experiment to overcome the difficulty of characterizing motions that span multiple timescales [VPDK03, BV04, MC04, LLBD$^+$05].

This thesis addresses proteins in the theoretical energy landscape formulation that the protein native state associated with the global free energy minimum consists of

an ensemble of conformations [OLSW97]. The thesis addresses the problem of characterizing native flexibility by computing ensembles of conformations that comprise the native state.

1.4 Contributions of this Thesis

The methods developed in this thesis search for native conformations by systematically reducing the amount of a priori information assumed to be available from experiment: first using a structure to guide the search for low-energy conformations; then reducing structural information to a closure constraint characterizing the native state; finally, employing only amino-acid sequence.

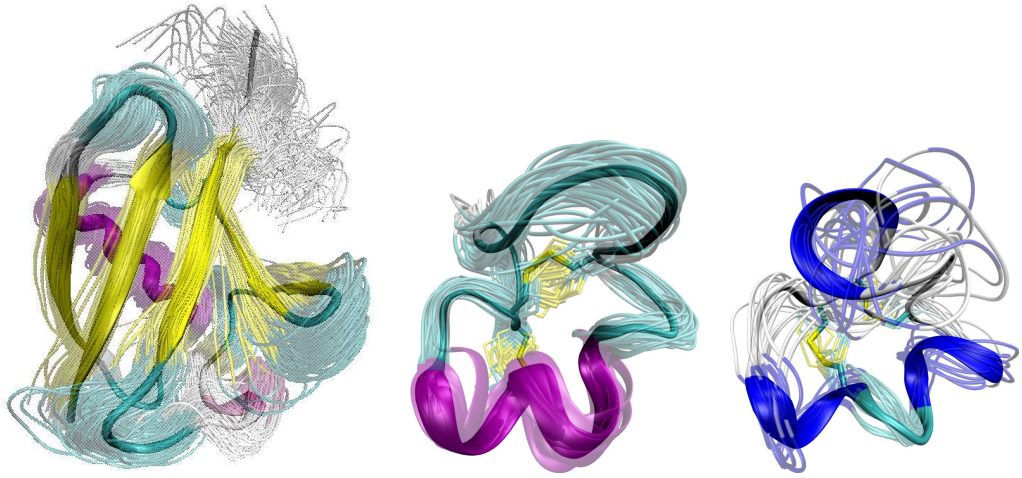
1.4.1 Let Structure Guide the Search for Functional Motions

An experimental structure that represents the native state of a protein is employed as reference to denote the location of the global energy minimum. The search for native conformations focuses in the conformational space around this minimum. The `Protein Ensemble Method` (PEM) presented in chapter 4 exploits locality, the fact that, in proteins with non-concerted motions, global information on flexibility can be obtained locally. The problem of computing fluctuations around a reference structure is divided into independent subproblems of computing fluctuations of consecutive overlapping fragments of the protein chain [SCK06, SCK07].

The core computational unit in PEM is the `Fragment Ensemble Method` (FEM),

described in chapter 3, which computes low-energy conformations of a fragment. FEM exploits analogies between a protein fragment and a kinematically-constrained robotic chain and employs a probabilistic exploration to sample the fragment conformational space. Figure 1.1(a) shows conformations computed for ubiquitin.

Employment of FEM in a local-to-global strategy and a statistical mechanics formulation allow PEM to reproduce NMR data of broad (ns-μs) timescales [SCK06]. PEM accurately captures the native flexibility of proteins of various folds and lengths up to 100 amino acids long [SKC07], as presented in chapter 4.



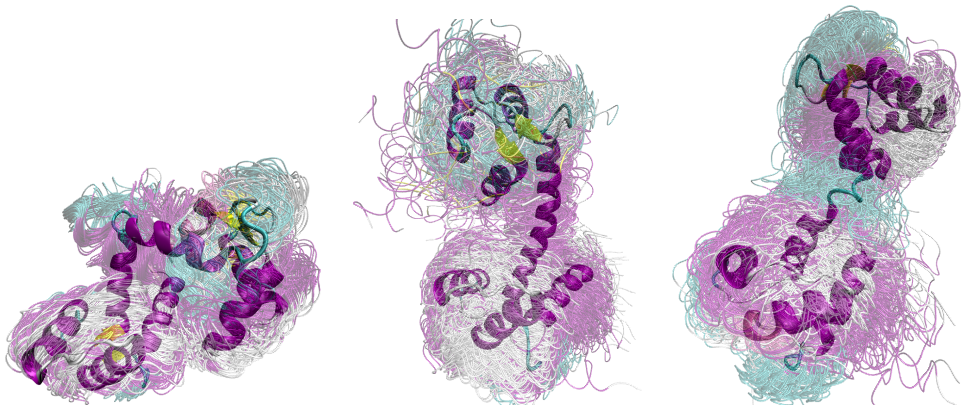
(a) (b) (c)

Fig. 1.1: (a) The ubiquitin experimental structure in opaque is used as reference by PEM. Computed conformations in transparent are populated with high probability under native conditions. (b)-(c) are ensembles obtained by NcCYP when a linker is added for cyclization. (b) reproduces the uncyclized state. (c) emerges from the addition of the linker. Lowest-energy conformations are shown in opaque. The rest of the conformations are in transparent.

1.4.2 Replace Structure with Geometric Constraints

The Native state characterization of cysteine-rich CYclic Peptides (NcCYP) method described in chapter 5 is proposed to compute native conformations of cyclic cysteine-

rich peptides. The method remove the assumption of non-concerted motions in the native state and replaces the reference structure with a closure constraint such as cyclization [SKC08b]. The method is based on an adaptive search that (i) uses multiple levels of detail to represent a protein chain to efficiently extract low-energy conformations and (ii) switches from exploring the entire conformational space to a targeted iterative exploration of emerging energy minima that continues until no lower-energy minima emerge. Applications on naturally-occurring and engineered cyclic peptides $20 - 30$ aas long reveal detailed conformational ensembles populated under native conditions. Figure 1.1(b)-(c) shows that adding a linker in a cyclic peptide changes the native state by giving rise to a novel conformational ensemble.



(a) (b) (c)

Fig. 1.2: The three conformational ensembles obtained by MuSE are shown in transparent, with the lowest-energy conformation in each ensemble shown in opaque. The three ensembles capture well the known functional states of calmodulin [SKC08a].

1.4.3 No Structural Information, only Sequence to Capture Motions

Finally, the Multiscale Space Exploration (MuSE) method proposed in chapter 6 extracts from amino-acid sequence the conformational ensembles comprising the na-

tive state [SKC08a] in proteins that undergo large-scale concerted motions. MuSE implements a two-stage multiscale exploration, first exploring an entire coarse-grained conformational space, then focusing the exploration to free-energy minima in the free energy landscape associated with the explored space. Both stages make use of multiple levels of detail in order to efficiently generate low-energy conformations. Applications on three sequences up to 214 aas long show that the method captures conformational ensembles which correspond well to known functional states. Figure 1.2 shows three such ensembles obtained from the calmodulin sequence.

1.5 Thesis Overview

Chapter 2 provides a brief background on protein structure and flexibility. Work related to the methods presented in this thesis precedes the description of each method in the following chapters. Chapter 3 presents FEM, its applications on loop fragments, and its coverage of conformational space. Chapter 4 shows how PEM employs FEM to characterize the native flexibility of an entire protein. Chapter 4 presents applications of PEM on proteins of various folds and lengths and compares PEM-obtained measurements with experimental data. Chapter 5 presents NcCYP and its applications on three cyclic chains. Chapter 6 presents MuSE and its applications on three different proteins that populate diverse functional states under native conditions through large-scale concerted motions. Finally, chapter 7 discusses the work presented in this thesis and directions of extending this work in future research.

Chapter 2

Protein Modeling

This chapter summarizes protein modeling in terms of structure representation and energy calculations. The issues of what detail is chosen to represent a protein conformation and what amount of information is stored internally directly affect the efficiency of a computational method that searches for protein conformations. Moreover, deciding what structural detail is needed to represent a conformation impacts the accuracy of predicted results and hence their relevance. The issue of how and when energetic calculations are carried out during a search for protein conformations is also crucial both for the efficiency of a search method and the confidence with which computed conformations can be employed to make predictions in silico.

2.1 Proteins as Linear Chains

A protein molecule consists of repeated blocks of atoms known as amino acids (interchangeably referred to as residues). An amino acid has an alpha carbon (C_α) atom connected to a hydrogen atom, an amino group, a carboxylic group, and a group of atoms known as a side chain. Consecutive peptide bonds between the amino

nitrogen and the carboxylic carbon link amino acids in a polypeptide chain, as shown in Fig. 2.1. Amino acids are numbered from the N- to the C-terminus which refer to the amino and carboxyl not involved in peptide bonds. The backbone is what remains after stripping the side chains off a polypeptide chain (removing C_β in Fig. 2.1).

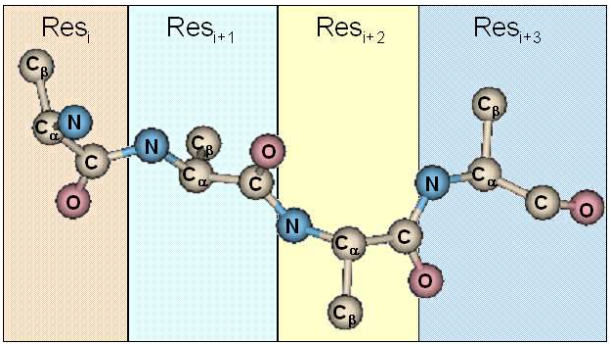


Fig. 2.1: (a) Polypeptide chain with four amino acids. C_β is the only side-chain atom shown. Figure is generated with MOLMOL [KBW96].

2.2 From Secondary to Tertiary Structure

The sequence of amino acids, referred to as the primary structure, determines the 3D structure [Anf73] of proteins, referred to as the tertiary structure. Though possessing different tertiary structures, all proteins share common repeated 3D blocks [oBN70] such as α-helices, β-sheets, β-turns, and others that form the secondary structures of a protein. An illustration of some of these building blocks is given in Figure 2.2. While secondary structure elements are generally well-conserved among proteins, other fragments known as random coils or loops exhibit large structural variability. It is the compact arrangement of secondary structure elements and loops that give rise to the tertiary structure of a protein.

(a) β-turn connecting β-sheets (b) loop connecting α-helix to β-sheet

Fig. 2.2: (a) A β-turn, depicted in black, connects two β-sheets, drawn in grey as arrows pointed along the protein backbone. The α-helix is drawn in silver. (b) Another representation of an α-helix as a cylinder is given. A long loop connects the α-helix to the β-sheet. Figure is generated with VMD [HDS96].

2.3 Internal Representation of a Protein Conformation

Deciding how to internally represent a conformation is critical for efficiently computing protein conformations.

2.3.1 Cartesian Coordinates

A conformation C that uniquely describes the 3D structure of a protein with N atoms may be represented as a vector $\langle A_{1x}, A_{1y}, A_{1z}, \ldots, A_{Nx}, A_{Ny}, A_{Nz} \rangle$, where A_{ix}, A_{iy}, A_{iz} are atom A_i coordinates. The parameters needed to represent C, in this case the atom coordinates, are often referred to as degrees of freedom (DOFs).

Even though this representation uniquely defines the location of every atom A_i, note that most of the $3N$ DOFs in this representation are redundant. Atom positions can be measured using only bond lengths, bond angles, and dihedral angles in a more succinct representation referred to as *internal coordinates*.

2.3.2 Internal Coordinates

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, labeled θ in Figure 2.3(b), is defined as the angle between the plane π_1 defined by the first and second bond and the plane π_2 defined by the second and third bond. Rotation by θ changes positions of atom A_{i+1} and others down the 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 bong 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 chain.

2.3.3 Idealized Geometry Model

Not all bonds are rotatable in a protein conformation. 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 C_α atom and between the

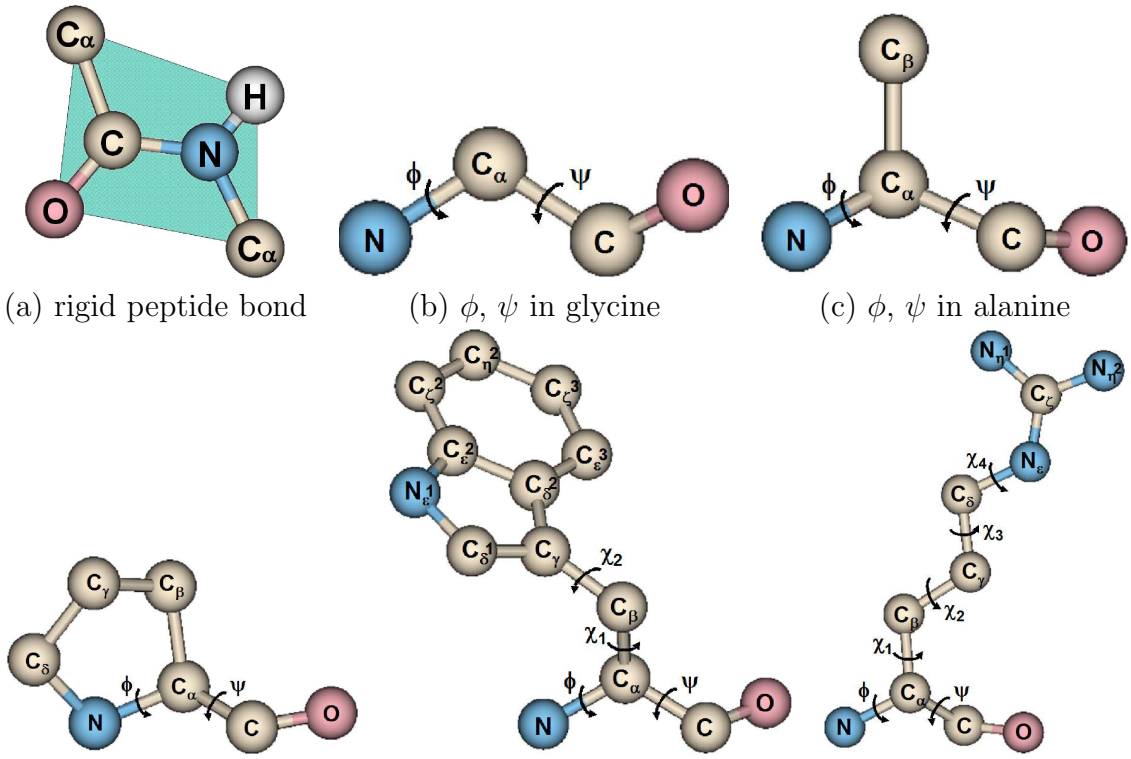
(a) internal coordinates (b) dihedral angle

Fig. 2.3: (a) Illustration of internal coordinates, where b refers to bond length, α to bond angle, and θ 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. Figures are generated with MOLMOL [KBW96].

nitrogen of the amino group and the C_α allow rotations by angles labelled ϕ and ψ, respectively, as shown in Figure 2.4(b) and Figure 2.4(c).

Amino acids on the other hand can contribute at most 4 dihedral angles, χ_1-χ_4. Amino acids with short or no side chains contribute no dihedrals, as Figure 2.4(b) and Figure 2.4(c) illustrate. Other side chains, being bulky, are constrained in their motion and contribute no side-chain dihedrals, as Figure 2.4(d) shows. Amino acids with long side chains such as tryptophan or arginine contribute a maximum of 4 side-chain dihedrals, as shown in Figure 2.4(e) and Figure 2.4(f).

Analysis of PDB protein structures reveals that bond lengths and angles do not vary much among structures [GK97]. Such evidence is used to assume an idealized or rigid geometry model which allows to fix bond lengths and bond angles to idealized equilibrium values and employ only dihedral angles as DOFs. Considering only dihedrals focuses the exploration of conformational space to large and more interesting atomic rearrangements and reduces the dimensionality of this space to $3N/7$ because



(a) rigid peptide bond (b) ϕ, ψ in glycine (c) ϕ, ψ in alanine

(d) rigid ring in proline (e) χ_1, χ_2 in tryptophan (f) χ_3, χ_4 in arginine

Fig. 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 ϕ, ψ on two amino acids, glycine, and alanine. (d) No side-chain dihedrals for the rigid ring in Proline (e) Tryptophan contains 2 side-chain dihedrals (f) Arginine has the highest number of side-chain dihedrals, 4. Figures are generated with MOLMOL [KBW96].

at most 6 DOFs - (ϕ, ψ, χ_1, χ_2, χ_3, χ_4) - are needed per residue [ATK94].

2.3.4 Modeling Proteins as Robotic Manipulators

The methods in this thesis model a protein chain as a kinematic chain with revolute joints [MZ94, SLB99, HA01, ADS02, ABG+03, XA03, CSRST04, LvdBDL04, CSdA+05, vdBLLD05] by employing the idealized geometry model illustrated in Figure 2.5. 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 a polypeptide chain requires a set of kinematic chains: one for the backbone, and others for the side chains connected to the backbone. As in forward kinematics [Cra89], where a joint rotation changes positions of following links, rotation about a dihedral bond changes positions of following atoms. Methods in this thesis propagate rotations down the chain as in [ZK02] to compute atom positions.

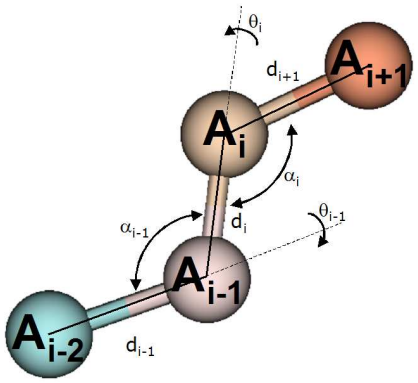


Fig. 2.5: A protein can be modeled as an articulated mechanism, where d_i is the length of the bond between atoms A_{i-1} and A_i, α_{i-1} the bond angle between A_{i-2}, A_{i-1}, and A_i, θ_{i-1} the ψ dihedral, and θ_i the ϕ dihedral. Figure is generated with MOLMOL [KBW96].

2.4 The Modeling Question: What is the Necessary Amount of Detail?

It is unclear what detail is necessary to expose nature's secrets in telling a protein chain what set of conformations to occupy under native conditions. Is atomic detail always needed? That is, does accuracy entail having to follow each atom in 3D space as the linear chain of amino acids explores conformational space? Or is it that coarse-grained representations that retain only a few atoms from each amino acid or

discretize conformational space in other ways will perform just as well as fine-grained representations. The question of what representational detail is needed to capture physical properties of proteins remains open in biophysics [Cle08].

2.4.1 Representing a Protein Conformation in All-atom Detail

Representing conformations in all-atom detail is desirable, as it allows to obtain a very detailed picture of how each single atom moves in space [DB01]. Because a protein molecule can have at least a few thousand atoms, no matter whether the idealized geometry or the cartesian representation are chosen, the number of DOFs is prohibitive for current computational resource. Indeed, the few all-atom studies of proteins are massively distributed among thousands of CPUs [SP00] and carried out on relatively small systems [SSP04, JVP06]. On bigger systems, employing all-atom detail limits the kinds of motions that can be computed to timescales no longer than a few nanoseconds [Dag00, PB02, HOvG02, Hes02, Tai04, vGBB$^+$06].

The statistical formulation of the free energy landscape, however, makes it possible to capture essential features of the free energy surface of a protein with only a limited set of parameters [CJO00, CNO00, CJO01, CGO03, COC04]. Studies suggest that coarse representations can accurately describe large-scale protein motions [DMS$^+$06].

2.4.2 Coarse-grained Representation of a Protein Conformation

Very early simulations of protein chains showed that important physical properties could be obtained with considerable less than atomic detail. For example, the first Go

models were based on lattices, explicitly representing only C_α atoms of a protein chain and restriciting atoms to lie on a lattice [TUG75]. Lattice modeling not only opened up the possibility of modeling in silico very large protein chains, but incidentally exposed an interesting complexity result: finding the minimum-energy conformation on a 3D cubic lattice is NP-hard [UM93].

In spite of their simplicity, only off-latice versions of the Go model are able to model real protein structures [CJO00,CNO00,CJO01,CGO03,MC04,DMC05,MC06]. Other successful coarse-grained representations include united-residue models, where an additional coordinate is added to indicate the position of the side chain on each amino acid. Backbone-resolution models, where only the backbone atoms of a protein chain are explicitly represented, have also been employed to predict native structures of proteins [OCL$^+$05, HLSW99, CJS$^+$06, GFR05].

2.4.3 Combining Multiple Resolutions in this Thesis

Recent work focuses on combining coarse- and fine-grained representations of proteins. For example, methods that focus on predicting a lowest-energy native conformation from a protein sequence conduct most of the search in a coarse-grained space. Atomic detail is added only at the end to refine candidate conformations [BMB05].

Going from a coarse- to a fine-grained representation of a protein conformation is possible without much loss of accuracy, as demonstrated by the work in [HKC07]. The methods in this thesis actually adapt components of the algorithm proposed in [HKC07] to add atomic detail to computed coarse-grained conformations.

The focus of today's computational studies on proteins is on actually combining multiple resolutions in order to exploit the efficiency from searching in a coarser-grained space and yet retain the biochemical detail and specificity provided by all-atom representations. Different levels of detail are sometimes maintained even among the set of conformations computed during the same iteration of an iterative search method [LYZ06, CGR06, KH05, PMDS+07, MCP+07].

The methods developed in this thesis belong to the class of multi-resolution or multiscale methods. They initially employ coarse-grained representations of a protein chain to efficiently provide an initial picture of the conformational space that characterizes protein molecules. The space is then analyzed to detect low-energy minima that are relevant for further exploration in all-atom detail [SCK06, SCK07, SKC07, SKC08a].

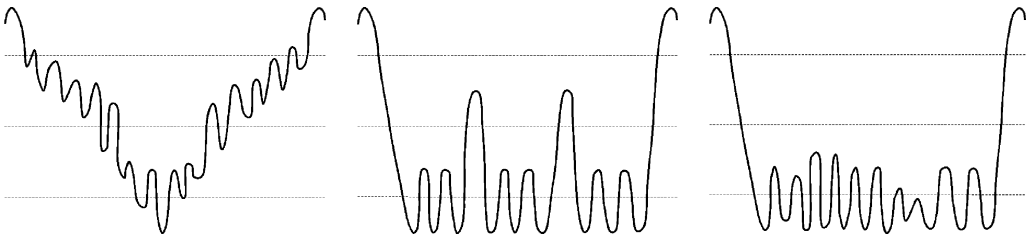
2.5 Protein Energy Landscapes

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 [OLSW97].

The potential surface of a protein is high-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 theoretical free energy landscape formulation, which advocates the use of statistical mechanics to organize the multitude of protein conformational states in terms of a minimal number of collective parameters [OLSW97, Gru02, Cle08].

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 high-dimensionality of conformational space, 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



(a) single minimum (b) multiple minima (c) shallow basin

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

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 native 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 these three main cases.

2.5.1 Empirical All-atom Force Fields

Although the atomic interactions in proteins are well understood, only empirical models exist to model a protein's potential energy. When using all-atom detail for computed conformations, the methods in this thesis employ empirical force fields such as CHARMM [BBO$^+$83] or AMBER [WCB$^+$94, DWC$^+$03]. These force fields, of similar functional form, sum over all favorable and unfavorable interactions to calculate the potential energy of a conformation.

As illustrated below for CHARMM [MBB$^+$98], all-atom force fields consider interactions due to all atoms. Parameters to the energy function are empirically de-

termined from experimental data and supplemented with ab initio results, as shown below:

$$\sum_{\text{bonds}} K_b(b - b_0)^2 \quad + \quad \text{(Bond Term)}$$

$$\sum_{\text{UB}} K_{\text{UB}}(S - S_0)^2 \quad + \quad \text{(Urey-Bradley Term)}$$

$$\sum_{\text{angle}} K_\theta(\theta - \theta_0)^2 \quad + \quad \text{(Bond Angle Term)}$$

$$\sum_{\text{dihedrals}} K_\chi(1 + cos(n\chi - \delta)) \quad + \quad \text{(Dihedral Term)}$$

$$\sum_{\text{impropers}} K_{\text{imp}}(\phi - \phi_0)^2 \quad + \quad \text{(Improper Term)}$$

$$\sum_{\text{nonbonded}} \epsilon[(\frac{r_{\min_{ij}}}{r_{ij}})^{12} - (\frac{r_{\min_{ij}}}{r_{ij}})^6] \quad + \quad \text{(Van der Waals Term)}$$

$$\sum_{atoms} \frac{q_i q_j}{\epsilon_1} r_{ij} \quad \text{(Electrostatic Term)}$$

where K_b, K_{UB}, K_θ, K_χ, and K_{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, θ and θ_0 to the current and equilibrium bond angle, respectively, χ to the dihedral angle value, n to the periodicity, ϕ to the improper angle value, ϵ to the Lennard-Jones well depth, $r_{\min_{ij}}$ 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 ϵ_1 to the effective dielectric constant.

2.5.2 Combining Coarse-grained and All-atom Force Fields

The methods developed in this thesis make use of various levels of detail during the search for native conformations. In particular, the MuSE method described in chapter 6 explores a coarse-grained space, where conformations are energetically

evaluated with the use of a coarse-grained energy function. This function is designed to associate low energies with native-like conformations and to compensate for the lack of all-atom detail. The inherent approximations in a coarse-grained energy function necessitate systematic comparisons of coarse-grained energies with all-atom energies. Multiple energy functions and representations of different resolutions are employed in this thesis to improve the accuracy of offered predictions.

2.5.3 Dealing with Errors in Force Fields and Solvation Models

Methods proposed in this thesis calculate discrepancies in measurements when using different state-of-the-art force fields. In addition, the methods developed in this thesis rely on the implicit Generalized Born (GB) solvation model [STHH90] to mimick interactions of the atoms in a protein conformation with the surrounding solvent. Explicit solvent calculations, due to their high demand on computational, are only carried out for few computed conformations to ensure that predictions of the methods proposed in this thesis are not solvent-dependent.

Applications of the proposed methods have shown that AMBER ff03 [DWC$^+$03] and the GB model [STHH90] yield reliable results, as already supported by available literature [HAO$^+$06, BB00]. In addition, work in [BB00] shows that the overall shape and location of the native basin are found to be less sensitive to whether implicit GB or explicit solvation models are used. Currently, development of accurate yet computationally feasible force fields and solvation models remains an area of active research [PC03, Mac04, JT05, HAO$^+$06, WB06, ROW$^+$07].

Chapter 3

Characterizing the Native Flexibility of a Fragment

This chapter presents the Fragment Ensemble Method (FEM) to characterize the native flexibility of a protein fragment such as a missing loop. The underlying approach combines a geometric exploration of the fragment's conformational space with a statistical mechanics formulation. This approach yields an ensemble of low-energy conformations on which thermodynamic quantities are measured as ensemble averages for direct comparison with experimental data. FEM is validated by applying it to characterize native flexibility in both instances of strongly stable and flexible loop fragments. In each instance, fluctuations measured over generated ensembles are consistent with available data from both experiment and simulation studies.

3.1 Introduction on the Native Flexibility of a Fragment

The native state of a protein may consist of a large ensemble of different conformations. In particular, fragments such as loops are often highly flexible even in generally stable proteins. Flexible loops are not easily characterized by X-ray crystallography

as they introduce significant disorder in a protein crystal. In fact, partially resolved protein structures are often reported in these cases, with the loop fragment missing.

Finding a physically-relevant conformation for a missing loop fragment in a given protein structure is an important problem (known as "loop modeling"[1]) in automated crystallographic protein structure determination, homology modeling, and ab initio structure prediction [DJ99, RDCB97, BB01, MFZH03]. The problem involves generating a loop conformation whose N- and C- terminal residues attach to the fixed anchor residues of the two protein segments at either end of the loop. However, proposing a single conformation fails to address the flexibility of the missing loop. In light of the high variation of loop structures in proteins, one or few conformations may not adequately represent the diversity in the ensemble of conformations assumed by a flexible missing loop.

This chapter proposes FEM to address native flexibility in the loop modeling problem. Given an incomplete protein structure and the amino-acid sequence of the missing loop, FEM generates an ensemble of low-energy loop conformations that complete the given protein structure. The method combines a statistical mechanics formulation with an efficient probabilistic exploration of conformational space. The exploration exploits analogies between proteins and robots [MZ94, SLB99] to model a loop fragment as a kinematic chain. FEM employs a multi-resolution approach. Backbone-resolution loop conformations are first generated to satisfy the

[1] Alternative names include loop/fragment completion, gap completion, loop closure, or fragment fitting.

geometric and energetic constraints imposed by the given protein structure. These conformations are then structurally and energetically refined to obtain an ensemble of low-energy all-atom loop conformations.

FEM is not limited to applications on missing loops but can generate an ensemble of physical conformations for any protein fragment. This capability is exploited by employing FEM as a core computational unit in a method developed to characterize flexibility of an entire protein chain, described in chapter 4. FEM is validated by using it to characterize loop structure and flexibility in both instances of strongly stable and completely disordered loops. In each instance, fluctuations measured over a generated ensemble fully agree with experimental and simulation data.

This chapter is organized as follows. Context is provided for the loop modeling problem through a review of related work in section 3.2. FEM is described in section 3.3. Section 3.4 analyzes FEM-obtained ensembles for non-internal loops in chymotrypsin inhibitor 2 (CI2), the variable surface antigen (VlsE), and α-lactalbumin (α-Lac) (loops of length 12, 20, and 26 residue, respectively). Good agreement is obtained between computed ensembles and data available from experiment and simulation. Section 3.5 presents additional ensembles obtained from applications of FEM on internal loops in protein structures. An analysis in section 3.6 interprets the good agreement in the context of the FEM coverage of the conformational space of a fragment. Additional analysis of the energetic refinement in FEM is presented in section 3.7. The chapter concludes with a summary and discussion in section 3.8.

3.2 Related Work on Addressing Kinematic Constraints

FEM explores the native flexibility of a missing protein fragment by dealing with the core problem of fitting a generated fragment conformation with a given reference protein structure. Driven by applications in X-ray crystallography, homology modeling, and ab initio structure prediction, existing work [BK90, SK90, FR92, MJ93, ZRVD93, ZRDK94, vVK97, DB00, FDS00, TBHM02, DAL03, CD03, LvdBDL04, vdBLLD05, CSdA$^+$05, KGLK05] focuses on fitting a generated loop conformation to model an unknown loop. The following summary organizes related work into inverse kinematics methods, (ab initio) search-based methods, and database methods.

3.2.1 Inverse Kinematics Methods

Methods that solve an Inverse Kinematics (IK) problem to model a missing loop exploit the fact that steering a terminal residue of the loop so that it assumes the pose of the corresponding fixed anchor is very similar to controlling motions of a robot arm so that the robot hand/gripper assumes a specified target position and orientation. By modeling the polypeptide chain of a missing loop as an open kinematic chain [MZ94], the problem of attaching the terminal residues of a loop to their corresponding fixed anchors can be posed as an IK problem: solve for the DOFs of the kinematic chain so that a terminal anchor of the loop assumes its target pose. This problem was first introduced in the context of robot manipulators [Cra89].

Exact Inverse Kinematics Methods

Robotics-inspired techniques [HA01, CSRST04] that employ exact IK solvers to enumerate all solutions [Chi93, MC94, MZ94, WS99, CSJD04] can do so on chains with no more than 6 DOFs. 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] to address the IK problem in hyper-redundant robots, where the number of regularly distributed joints is very large.

Specialized Exact Inverse Kinematics Methods in Biology

Specialized IK solutions in biology appeared as early as 1970 [GS70], where fragments with ≤ 6 DOFs were predicted by solving a set of polynomial equations representing geometric transformations. These equations were applied to build tripeptide loops [GS70]. Later work offered efficient analytical solutions for three consecutive residues through spherical geometry and polynomial equations [BK85, PS91, MZW95, WS99]. Bounding IK solutions for chains with ≤ 6 DOFs within small intervals was applied in the context of drug design [ZK02]. A new formulation extending the domain of solutions to any three residues, not necessarily consecutive, was recently proposed [CSJD04]. An efficient subdivision of the solution space pushes the dimen-

sionality limit from 6 to 9 DOFs [ZWW$^+$04].

Optimization-based Inverse Kinematics Methods

Currently, only optimization-based IK solvers [FWS$^+$86, WC91] can deal with an arbitrary number of DOFs. Two such methods, Random Tweak [FWS$^+$86, SYF$^+$87] and Cyclic Coordinate Descent (CCD) [Lue84,WC91,CD03], iteratively solve a system of equations until the kinematic constraints on the loop termini are satisfied. Both methods are based on iteratively changing dihedral DOFs of a fragment/kinematic chain until the terminal residue reaches its target pose.

Random Tweak Random Tweak relies on computing the Jacobian of atom distances from their target positions with respect to the dihedrals. This computation is demanding and even numerically unstable as the Jacobian may loose rank. In addition to not being free of 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 proposed dihedral change on all the others. Additional constraints on the dihedrals may result in the unpredictable motion of a residue away from rather than toward its target pose.

Cyclic Coordinate Descent Avoiding the use of a Jacobian, CCD is computationally inexpensive, numerically stable, and free of singularities. This method avoids the inter-dependence of dihedrals by adjusting one DOF at a time. CCD allows for additional constraints on dihedrals with a predictable motion of atoms towards target

positions. First introduced in the context of non-linear programming [Lue84], this method was applied to robotics [WC91], and later to the loop closure problem for proteins [CD03, Lot04, LvdBDL04, vdBLLD05]. In particular, [Lot04, LvdBDL04, vdBLLD05] combined this method with motions in the self-motion manifold, where local motions do not influence the kinematic constraints and can be used to move towards a local minimum of an objective function.

3.2.2 Ab Initio Search-based Methods

Search-based methods rely on a generate-and-test paradigm to first generates many fragment conformations and then filter out the ones that do not satisfy kinematic constraints. This class of methods can be categorized into two subclasses, loop construction methods and motion-planning based approaches.

Loop Construction Methods

The dimensionality limitation of loop closure algorithms was addressed in [BK87] by analytically solving for short fragments and enumerating solutions for longer fragments. Combinatorial approaches [MJ86, BHN88, DS90, BVSD93, Bru93, FSBM94, PM95, DB00] limit a combinatorial explosion by restricting the set of (ϕ, ψ) dihedral angles to distributions that are biased towards more populated regions of the (ϕ, ψ) Ramachandran map [RRS63].

Other algorithms propose local moves [EGE95], importance sampling by local minimization of randomly generated conformations [LS89a, LS89b, LS89c], or glob-

ally minimize energy by mapping a trajectory of local minima [DRP98]. Other optimization-based search algorithms consist of molecular dynamics simulations [BK90, TNM92,RT93,NHK00], Monte Carlo (MC) combined with Molecular Dynamics (MD) [RF99], biased probability MC [AT94, EMCG95, TZM+97], and MC with Simulated Annealing [HCG92, CHG93, CE93, CE96, VCD94, FDS00]. Even dynamic programming algorithms [FR92] and genetic algorithms [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 methods suffer from practical limitations on fragment length and low success rate in fragment closure.

Motion-Planning based Approaches

Robotics-inspired methods employ a probabilistic sampling framework [KSLO96a]. Central to all these algorithms is the use of the probabilistic roadmap [KSLO96b] for sampling conformational space. Loop conformations can first be sampled ignoring the kinematic constraints and later enforcing them through gradient descent [YLK01]. The satisfaction of constraints can alternatively be integrated in the sampling process [HA01]. In the latter case, the loop is broken into an *active* part, for which conformations are generated disregarding the constraints, and a *passive* part that is closed through exact IK methods [HA01]. 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 is a natural extension of the above idea [CSL02]. The authors have applied this algorithm to the closure of long protein loops [CSRST04]. Other approaches subject loop conformations to attractive forces that pull the end effector of the chain, the robot hand or gripper, to its target position and orientation [LSB05].

3.2.3 Database Methods

Database methods [SK90, vVK97, TBHM02, DAL03] search for candidate loops that satisfy constraints on length and geometry in homologous proteins available in structural databases such as the PDB [BWF$^+$00]. These methods were first proposed in the context of electron density fitting [JT86], and then in comparative modeling [CLP$^+$87, CCLW89, SK90, TL92, Lev92, TME$^+$93, LS94, MT96, LLL99]. Database approaches rely on the assumption that loops in proteins submitted in the PDB provide natural templates to model missing loops. For antibody hyper-variable loops, database methods yield satisfactorily results [CLP$^+$87, CLT$^+$89, MT96, MTR$^+$98] due to these loops forming very specific folds based on key residues.

Database methods suffer from limited loop diversity in the PDB [TL92] and so can model loops of ≤ 15 aas [DAL03]. This issue can be addressed by constructing loops from shorter fragments sampled from structural libraries [KGLK05]. In particular, divide-and-conquer approaches [TBHM02] recursively break loops into equal-sized fragments that are analytically solved and combined to yield loops of arbitrary length. While limited in the length of loops they can model, by employing the PDB, database methods have the advantage of producing physically-realistic conformations [MJ86].

3.3 FEM: Computing Low-energy Conformations of a Protein Fragment

Since the proposed FEM is generally applicable to any protein fragment and not just a loop, the method is described hereafter in terms of generating an ensemble of physical conformations for a protein fragment.Given an incomplete protein structure and the amino-acid sequence of the missing fragment, FEM generates an ensemble of physical fragment conformations that fit with the given protein structure through essentially a three-step multi-resolution approach:

(i) *Backbone Geometric Exploration:* The conformational space available to the backbone of a missing fragment is explored to generate fragment conformations that fit with a given protein structure without introducing collisions (details are found in section 3.3.2). Obtained conformations are passed on to step (ii).

(ii) *Side-chain Exploration for a Fixed Backbone:* The configurational space available to the side chains of a fragment is explored to add all-atom detail to each fitted fragment conformation (details are found in section 3.3.3). Obtained conformations are passed on to step (iii) for energetic refinement.

(iii) *All-atom Energy Refinement:* Obtained conformations are subjected to extensive energy minimization that seeks stabilizing interactions between atoms of the fitted fragment and the rest of the protein (details are found in section 3.3.4). Each fragment conformation is retained in the ensemble if the corresponding

completed protein conformation has energy lower than a cutoff value.

Steps (i)-(iii) of FEM allow efficiently generating a large ensemble of fragment conformations whose corresponding completed protein conformations are physically relevant. Each of these steps is detailed below after preliminary definitions.

3.3.1 Modeling a Protein Fragment

Let protein residues from the N- to C- terminus be numbered 1 to n. A protein fragment $[n_1, n_2]$ is said to be missing if atomic coordinates are available only for residues 1 to n_1 and n_2 to n. The missing fragment $[n_1, n_2]$ is defined as the polypeptide chain consisting of residues n_1 to n_2, including n_1 and n_2. Finding a conformation for the missing fragment involves generating coordinates for all atoms of its polypeptide chain. Doing so in a way that fits the fragment with the given protein structure requires that the coordinates of residues n_1 and n_2 in the fragment conformation be as those of n_1 and n_2 in the given protein structure. In this sense, residues n_1 and n_2 are "duplicated": those in the given protein structure are fixed and referred to as stationary/fixed anchors; those in the fragment move as one tries to find coordinates for the fragment's atoms and are referred to as mobile anchors.

Hence, modeling an unknown fragment involves finding coordinates for its residues so that its mobile anchors attach to the stationary anchors in the given protein structure. Attaching a mobile anchor to its stationary counterpart means translating the mobile anchor so that one of its backbone atoms assumes its target position in the stationary anchor and orienting the anchor so that its N, C_α, and C backbone atoms

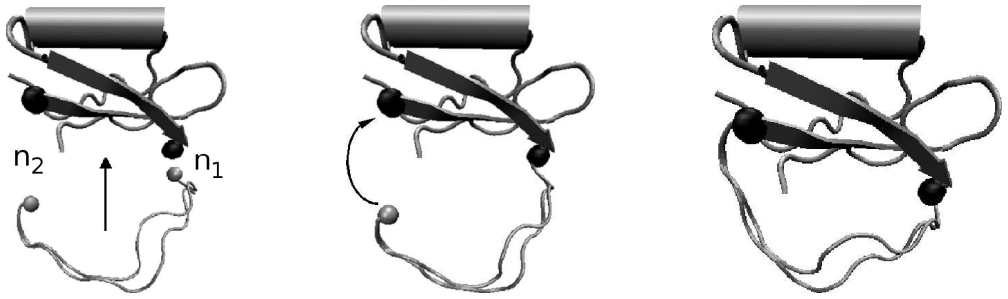
align with their counterparts in the stationary anchor residue. A mobile anchor reaches its target pose when it assumes its target position and orientation in space.

The problem of modeling an unknown fragment consists of two steps: (i) obtain initial atomic coordinates for the fragment; and (ii) modify the fragment conformation so the mobile anchors assume the target poses in the stationary anchors.

Obtaining Initial Coordinates for the Unknown Fragment The first step can be addressed in different ways. In the implementation of FEM in this thesis, a biologically relevant polypeptide chain for an unknown fragment such as a missing loop is initially obtained from a sequence-homologous protein structure selected from the PDB [BWF+00]. Any missing atom information[2] is completed through the PSFGEN [GP02] package. A large set of different conformations of the polypeptide chain of the fragment are then obtained by modifying the chain's dihedral angles, as described in section 3.3.2. Conformations obtained in this way do not generally fit with the given protein structure, as illustrated in Figure 3.1(a) for a loop, since the mobile anchors n_1 and n_2 may not be attached to their stationary counterparts. Indeed, fragment conformations depend in a non-trivial way on the amino-acid sequence of the fragment and the environment provided by the rest of the protein.

Closing the Fragment One mobile anchor of the fragment, n_1 in Figure 3.1(b), is easily attached to its stationary counterpart through rigid body transformations,

[2]Structures reported in the PDB commonly miss hydrogen or side-chain atoms.



(a) (b) (c)

Fig. 3.1: (a) Mobile anchors in two conformations of the CI2 VAL53-ASP64 fragment, drawn in grey, are not attached to the stationary anchors in black. (b) n_1 is attached to its corresponding stationary anchor through rigid body transformations. (c) Rotations of the dihedral bonds of the fragment steer n_2 towards its target pose in the stationary anchor.

a translation and two rotations to align the backbone atoms of the mobile anchor to their stationary counterparts in the fixed anchor. As illustrated in Figure 3.1(c), the resulting fragment conformation needs to be modified so as to attach the remaining mobile anchor n_2 to its stationary anchor. This problem is often referred to as "closing the fragment" or "closing the loop" in the context of loop modeling. It is addressed in the *Backbone Geometric Exploration* step, which takes as inputs the given protein structure and the polypeptide chain of the missing fragment already attached to one fixed anchor and outputs fragment conformations that fit with the given protein structure by solving the kinematic constraint on n_2.

3.3.2 Step (i): Backbone Geometric Exploration

FEM samples kinematically-constrained conformations of fragment $[n_1, n_2]$ as shown in pseudocode in Algorithm 1. The exploration starts by stripping away all but backbone atoms off the polypeptide chain obtained for the missing fragment. Working with a coarse resolution allows making direct use of analogies between proteins and

robots [MZ94, SLB99] that are often exploited to adapt powerful robotic space exploration methods to the study of protein systems [CSdA+05, ADS02, ABG+03, LFKL00]. As shown in line 1 of Algorithm 1 and in keeping with these analogies, FEM models the backbone chain of the fragment as an open kinematic chain, where a protein's atoms are equivalent to robotic links and rotatable bonds connecting atoms to joints connecting links. Fragment $[n_1, n_2]$ is modeled as a kinematic chain whose base is at n_1 and end-effector is at n_2. The method employs the idealized geometry, where bond lengths and bond angles are kept fixed in their equilibrium (native) values. As shown in line 2 of Algorithm 1, the only DOFs used at this stage are the ϕ, ψ backbone dihedral angles starting at residue $n_1 + 1$ and ending at residue $n_2 - 1$.

Algorithm 1 ExploreWithConstraints (C_{ref}, $[n_1, n_2]$, $n_{\max}$, ϵ, σ)

Input:
C_{ref}: conformation corresponding to reference protein structure
$[n_1, n_2]$: fragment $[n_1, n_2]$ for which to explore conformations
$n_{\max}$: maximum number of CCD cycles
ϵ: criterion for evaluating satisfaction of kinematic constraint on n_2
σ: permutation of DOFs of $[n_1, n_2]$
Output: Conformation C that satisfies kinematic constraint on n_2

1: $K \leftarrow$ kinematic chain modeling $[n_1, n_2]$
2: $B \leftarrow$ DOFs of K corresponding to backbone dihedral angles of $[n_1, n_2]$
3: $\theta|_B \leftarrow$ DOF values sampled uniformly at random in $[-\pi, \pi]^{|B|}$
4: $C \leftarrow$ apply rotations by $\theta|_B$ dihedral angles
5: **for** $n \leftarrow 1$ to $n_{\max}$ **do**
6: $\quad B_\sigma \leftarrow$ permutation of DOFs B
7: $\quad \bar{C} \leftarrow$ CCD(B_σ, pose of n_2 in C, target pose of n_2 in C_{ref})
8: $\quad d \leftarrow$ Euclidean distance between pose of n_2 in $\bar{C}$ and target pose of n_2 in C_{ref}
9: $\quad$ **if** $d \leq \epsilon$ **then**
10: $\quad\quad$ exit **for** loop

Line 3 shows that conformations for the backbone of the fragment are first sampled by ignoring the kinematic constraint on n_2. Values for the DOFs are sampled uniformly at random in $[-\pi, \pi]$. Considering only the backbone reduces the dimensionality of the sampled conformational space and allows for an efficient exploration. Rotations by the sampled angles do not change the atom positions of n_1 but violate the constraint on n_2. As shown in line 7, each sampled conformation is subjected to CCD. The implementation of CCD in Algorithm 1 follows closely that in [CD03].

CCD closes each generated fragment conformation by solving the following IK problem: given the positions of the backbone atoms of the stationary anchor n_2, assign values to the DOFs of the kinematic chain modeling the fragment so that the backbone atoms of the mobile anchor n_2 assume their target positions in the stationary anchor. CCD recasts this problem as a minimization problem. Given one particular DOF (i.e., backbone dihedral angle) of the kinematic chain, the algorithm analytically finds the value yielding the minimum distance between residue n_2 of the fragment and its target pose in the given protein structure.

CCD proceeds in cycles. Each cycle iterates over all DOFs according to a prespecified order (shown by permutation σ of DOFs in Algorithm 1), updating each DOF one at a time, until n_2 reaches a pose within an ϵ-neighborhood of the target pose in C_{ref}. As shown in line 5 of Algorithm 1, the number of cycles is limited to $n_{\max}$.

Each conformation closed with CCD depends on the initial fragment conformation sampled. This dependence is a useful feature exploited to generate many fragment

conformations to complete a given protein structure. The completed structure is deemed collision-free if its energy is below a maximum energy value $E_{\max}$ [3].

3.3.3 Step (ii): Side-chain Exploration for a Fixed Backbone

The *Backbone Geometric Exploration* step generates many different backbone-resolution conformations for a missing fragment. Since it only modifies the backbone of a fragment, the side chains of the polypeptide chain of the fragment are not in their optimal configurations in each generated backbone conformation. Therefore, values for the dihedral angles of these side chains are sampled uniformly at random in $[-\pi, \pi]$. For each backbone conformation, the side-chain dihedral space is explored until an all-atom fragment conformation is found whose corresponding completed protein conformation C is collision-free.

3.3.4 Step (iii): Energetic Refinement of a Modeled Fragment

To render interactions between atoms of a fragment conformation and the rest of the protein favorable, each completed protein conformation C is subjected to extensive energy minimization. Energy is measured by physical force fields such as CHARMM [MBB+98] or AMBER [WCB+94]. The energetic refinement of C is designed to attribute unfavorable interactions mainly to a fragment's atoms, since the conformation corresponding to the given protein structure is considered feasible.

[3] Parameters are introduced to keep the description of the method general. Values to these parameters are empirically determined and listed in section 4.4.5.

To achieve this goal, the refinement interleaves two strategies that mainly explore fluctuations of a closed fragment to minimize the energy of C while maintaining the given protein structure. The first, closure-constrained backbone refinement, inspired by work in [LvdBDL04, vdBLLD05], modifies the backbone dihedrals of a fragment during minimization. The second, closure-constrained conjugate gradient descent, relaxes the idealized geometry model and allows all atoms' coordinates to change as dictated by the force field for crucial interactions of the fragment with the rest of the protein. While exploring small fluctuations of the given protein structure, this strategy attributes most of the mobility to the fragment's atoms.

Since both minimization strategies are local searches that may converge to local minima, they serve as relaxation steps for each other. If after N steps of the closure-constrained conjugate gradient descent, the improvement in energy is less than a cutoff value η, this indicates failure to escape from a local energy minimum. Therefore, the minimization switches to the closure-constrained backbone refinement which can further minimize energy. The two strategies interleave with each other for a maximum of $N_{\max}$ minimization steps, testing after every N steps whether to terminate the minimization (if the improvement in energy is less than a convergence value μ).

Closure-constrained Backbone Refinement - Exploring the Self-motion Manifold Due to the 6 constraints on n_2, the remaining $m = 2\,(n_2 - n_1 - 1) - 6$ of the q DOFs of the fragment are redundant. They define a sub-space, the self-motion manifold [Bur89], which is explored for fluctuations of backbone atoms to

minimize the energy of a conformation and yet maintain the pose of n_2. The manifold is approximated with its tangent space as in [vdBLLD05]. An instantaneous change in q is then obtained with $\dot{q} = J^{\ddagger}(q)\dot{x} + N(q)N^T(q)g(q)$ as in [Bur89], where $J^{\ddagger}(q)$ is the pseudo-inverse of a $6\times m$ Jacobian matrix relating the linear and angular velocities of a frame x attached to n_2, $N(q)$ is an orthonormal basis for the null-space, and $g(q)$ is the gradient of the energy function. The constraints on n_2 force $\dot{x} = 0$. Projecting $g(q)$ on the null space yields a motion $\dot{q}$ in dihedral space that minimizes the energy function while keeping n_2 in its pose. The manifold is explored with a steepest descent that at each step updates $J(q)$ to compute $\dot{q}$ as in [CK00]. A singular value decomposition [ABB+99] yields $J(q) = U\Sigma V^T$, where the vectors of V corresponding to zero-valued singular values provide $N(q)$. Due to these computational requirements, it is necessary to limit the number of steps of the descent.

Closure-constrained Conjugate Gradient Descent: A conjugate gradient descent is performed on the energy landscape defined by the pseudo-energy function $E = E_{\text{forcefield}} + \sum_{atom\, i \notin \text{fragment}} K_{d_i} \cdot |\vec{x_i}(C) - \vec{x_i}(C_{\text{rest}})|^2$, where $\vec{x_i}$ indicates the 3D position of atom i during minimization and C_{rest} refers to the conformation corresponding to the rest of the reference protein structure. Minimizing the second term as well as the energy (first term) ensures that more mobility is asked of the fragment's atoms for stable interactions with the given protein structure. The extent to which atom i outside the fragment moves away from its position in C_{rest} depends on the strength of interactions between atoms of the fragment and C_{rest} and is modeled

through the damping constant K_{d_i}, empirically determined for each protein.

3.3.5 Obtaining an Ensemble of Low-energy Fragment Conformations

Steps (i) through (iii) of FEM yield many low-energy all-atom closed fragment conformations. Closed fragment conformations whose corresponding completed protein conformations are of energy no higher than a cutoff value of 20 kcal/mol from a reference energy[4] are deemed physically relevant and are added to an ensemble $\Omega_{[n_1,n_2]}$ of physical fragment conformations. Fragment conformations are generated independently from one another; hence, their computation is easily distributed. The issue of ensemble convergence, i.e., how many fragment conformations need to be generated to obtain a reliable native conformational ensemble, is resolved by measuring converge in ensemble averages of after the addition of every $1,000$ conformations.

Probability of a Local Fluctuation: A statistical mechanics formulation is employed to weight each conformation $C \in \Omega_{[n_1,n_2]}$ with energy E_C according to its Boltzmann probability $P(C) = P_{\text{ref}} e^{-\frac{E(C)-E_{\text{ref}}}{RT_0}}$, where P_{ref} and E_{ref} are the probability and the energy of C_{ref}, T_0 is the room temperature (300K), and R is the gas constant. The reference probability P_{ref} can be arbitrarily set equal to 1 as the calculation of average quantities is independent of the actual value of P_{ref}. A cutoff value of 20 kcal/mol for $E(C) - E_{\text{ref}}$ allows to discard generated conformations that do not contribute to thermodynamic averages measured over the ensemble of completed

[4]When a reference energy is not available, as in the case of a partially-resolved protein structure, the minimum-energy completed conformation is used instead.

conformations (conformations where this cutoff is higher than 20 kcal/mol have an extremely low Boltzmann probability ($\lesssim 10^{-15}$) at room temperature T_0). The Boltzmann average $\langle X_i \rangle_{[n_1,n_2]}$ of a measurable quantity X_i at a given position i (such as, for instance, the value of the least root-mean-square deviation - lRMSD - for a given residue) is computed over all conformations $\{C\}$ of the ensemble $\Omega_{[n_1,n_2]}$ associated with fragment $[n_1, n_2]$:

$$\langle X_i \rangle_{[n_1,n_2]} = \frac{\sum_{C \in \Omega_{[n_1,n_2]}} e^{-\frac{E(C)-E_{\text{ref}}}{RT_0}} X_i(C)}{Z},$$

where $Z = \sum_{C \in \Omega_{[n_1,n_2]}} e^{-\frac{E(C)-E_{\text{ref}}}{RT_0}}$ is the partition function associated with $\Omega_{[n_1,n_2]}$.

3.3.6 Implementation Details

Backbone Geometric Exploration: In the implementation of the CCD algorithm in this thesis, the maximum number $n_{\max}$ of CCD cycles is 500. The closure criterion $\epsilon = 0.001$Å. The $E_{\max}$ employed is empirically valued at $5,000$kcal/mol.

All-atom Energy Refinement: The maximum number of minimization steps, the frequency of testing whether the convergence criterion has been met, and the actual definition of convergence are all empirically determined quantities that work well for all the proteins presented here: $N_{\max} = 1,000$, $N = 300$, $\eta = 2$ kcal/mol, and $\mu = 20$ kcal/mol. Due to the complexity of approximating the self-motion manifold and the numerical computation of the force field gradient, the steepest descent employed in the closure-constrained backbone refinement to explore motions on the self-motion manifold is limited to 50 steps. In the closure-constrained conjugate gra-

dient descent, in CI2, α-Lac, ubiquitin, and protein G, where interactions between atoms of a fragment and of C_{ref} are strong, $K_{d_i} = 10$. In other systems, such as VlsE, $K_{d_i} = 100$. In the exploration of the self-motion manifold $\dot{q}$ is numerically computed through the implementation of finite differences in the OPT++ nonlinear optimization package [Mez94] modeling the energy function as an FDNLF1 object. The conjugate gradient descent is implemented through the OPTCG procedure in the same package modeling the pseudo-energy function as an NLF1 object since its gradient can be computed analytically.

Employed Packages: Missing atoms are filled in with the PSFGEN [GP02] package. Given a file that specifies types and charges of atoms in amino acids and a PDB file with the coordinates of the existing atoms of the polypeptide chain, PSFGEN creates a new PDB file where coordinates of the missing atoms are guessed and incorporated in the respective amino acids of the polypeptide chain. The OPT++ [Mez94] package is employed for the efficient implementation of the conjugate gradient descent algorithm. The algorithms implemented in OPT++ provide robust and efficient solutions to nonlinear optimization problems that require expensive function evaluations.

Hardware and Software Setup: The implementation was carried out in `ANSI C/C++` using the `Intel`8.0 compilers and libraries. The experiments were run on the Rice Terascale Cluster, a 1 TeraFLOP Linux cluster of `Intel Itanium`2 processors. Each node has two 64-bit processors running at 900MHz with 1.5MB of L2 data cache

and 2GB memory per processor. On such architecture, it takes on average 67 minutes to obtain $1,000$ conformations for a surface fragment of 30 residues.

3.4 Applications of FEM on Non-internal Loops

To first test the accuracy of FEM, native non-internal loops are first reproduced in stable proteins, such as CI2, PDB code 1COA [JMe+93], and α-Lac, PDB code 1HML [RSA93], respectively[5]. The loops considered include the 12-residue loop between VAL53 and ASP64 in CI2 and the 26-residue loop between LYS51 and THR76 in α-Lac. FEM is used to generate an ensemble of conformations for the considered loop in each protein.

Figure 3.2(a1) shows the ensemble of generated conformations for the VAL53-ASP64 loop in CI2. Qualitatively, the obtained loop conformations are clustered around the native loop as found in the equilibrated crystal structure of CI2. The native flexibility of the loop is quantified by plotting in Figure 3.2(a2) the energy profile of the generated ensemble versus the lRMSD of the generated loop conformations from the equilibrated native loop conformation. The obtained energy profile is clearly funnel-like, in full agreement with the known role and stability of this loop for the activity of CI2 [LD94, JF94].

Residue fluctuations obtained on the generated ensemble are validated against B factors [JMe+93] available for CI2. Fluctuations of each residue are obtained by

[5]The PDB structures are refined through the above conjugate gradient descent, with $K_{d_i} = 0$.

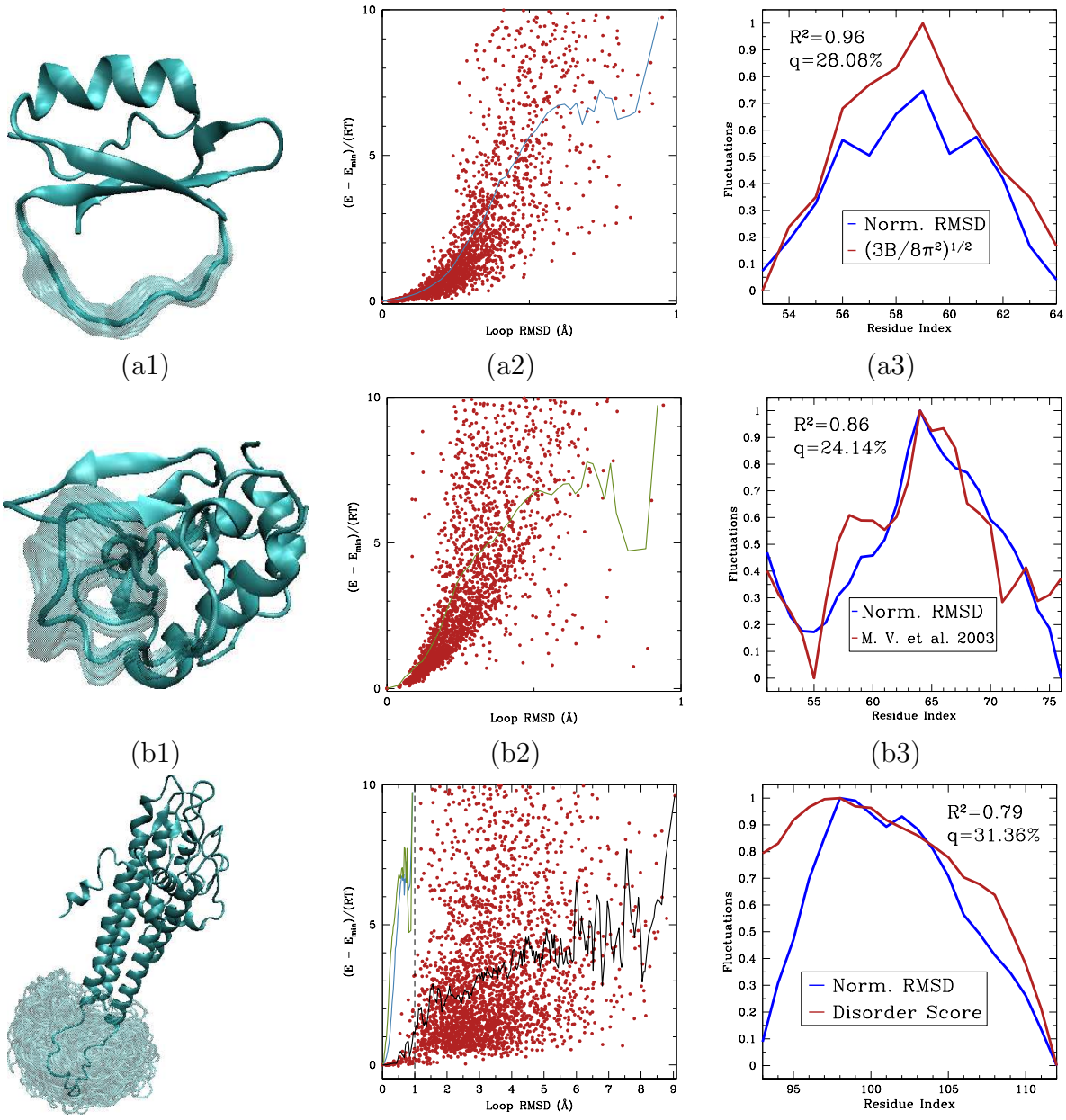


Fig. 3.2: (a1)-(c1) 5,000 transparent loop conformations vs. opaque reference are rendered with VMD [HDS96]. (a2)-(c2) Associated energy landscapes are shown as energetic difference vs. lRMSD of each conformation from reference. Only conformations with energy $\leq$ 10 RT units from reference are shown (2499, 2022, and 2755, respectively). An average profile is computed by binning conformations every 0.001 Å away from reference and averaging energies of a bin. CI2 and α-Lac profiles are steep, whereas VlsE profile is flat. (a3)-(c3) Obtained fluctuations vs. B factor-derived ones for CI2, fluctuations in [VPDK03] for α-Lac, and disorder scores for VlsE.

averaging through the Boltzmann statistics the residue lRMSD measured in each loop conformation relative to the native loop conformation as found in the equilibrated crystal structure of CI2. Since fluctuations derived from B factors are different in magnitude from fluctuations obtained over the ensemble generated by FEM, both sets of fluctuations are normalized. As shown in Figure 3.2(a3), the obtained fluctuations are consistent with those derived from the available B factors; the data agree with a Pearson correlation of 96% and q-factor of 28%. This agreement indicates that fluctuations of this loop are mainly local and can be obtained in isolation, even when immobilizing the rest of the protein structure.

The generated ensemble for α-Lac is shown in Figure 3.2(b1). As expected, the obtained loop conformations are clustered around the native loop of the equilibrated crystal structure. Figure 3.2(b2) reveals a funneled energy landscape with a global minimum around the native conformation found in the equilibrated crystal structure of α-Lac, similarly to the energy landscape associated with the ensemble of loop conformations generated for CI2.

The fluctuations observed over the generated ensemble for α-Lac are fully consistent with what is obtained from an MC simulation guided to agree with hydrogen exchange protection factors [VPDK03]. Figure 3.2(a3) compares residue fluctuations measured over the ensemble generated by FEM with the fluctuations reported in [VPDK03] (data courtesy of M. Vendruscolo). Due to their different magnitudes, fluctuations are normalized in the comparison. A Pearson correlation of 86% and a

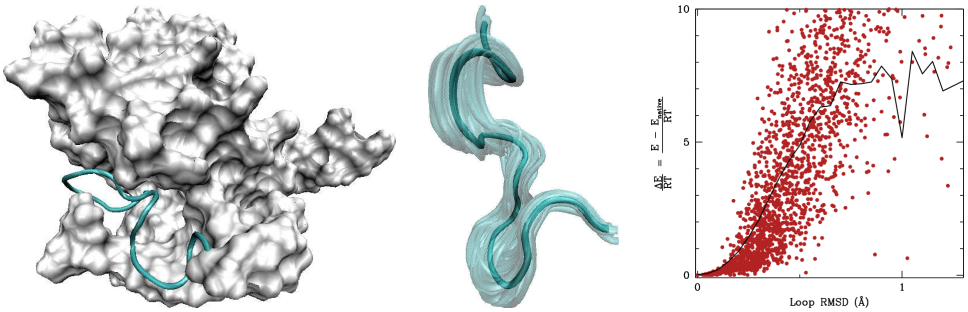
q-factor of 24% are obtained. Interestingly, the Pearson correlation between fluctuations measured over the FEM-obtained ensemble with fluctuations derived from B factor data for α-Lac [RSA93] is 63%, comparable to the 61% Pearson correlation obtained when comparing fluctuations derived from the B factor data to fluctuations reported in [VPDK03].

The examples described above provide a good testbed for the accuracy of FEM in producing ensembles of native-like loop conformations with associated steep funnel-like energy landscapes for strongly stable proteins. The most interesting application of FEM, however, is the generation of a large ensemble of loop conformations for proteins with highly flexible loops. Results are presented here when when applying this method to generate an ensemble of conformations for the LYS93-GLY112 loop in the crystal structure of VlsE, PDB code 1L8W [ESK$^+$02]. This 20-residue loop is missing in the crystal structure due to its high flexibility [ESK$^+$02]. The analysis reveals that there are many geometrically variable conformations relevant for this loop at room temperature. The high conformational heterogeneity of the closed loop conformations can be seen in Figure 3.2(c1). The heterogeneity of these loop conformations is quantified in Figure 3.2(c2), where the energy landscape associated with the generated ensemble is plotted as a function of the lRMSD from the most stable complete protein conformation obtained through FEM. Figure 3.2(c2) shows a plateau-like energy landscape, which is very different from the funnel-like landscapes obtained for the loops in CI2 and α-Lac.

To validate the ensemble generated for the missing loop of VlsE, the magnitudes of the structural fluctuations per residue (measured relative to the lowest energy structure obtained) are compared with disorder scores computed from the amino-acid sequence of VlsE through the PONDR package [LRR$^+$99, LRR$^+$01]. Disorder scores predicted by the PONDR package [LRR$^+$99, LRR$^+$01] for the loop are all well above 0.5 (the boundary between disorder and order), consistent with the fact that the LYS93-GLY112 loop in VlsE is highly disordered. Since the comparison between fluctuations and disorder scores is between two different quantities of different magnitudes, both quantities are normalized. As shown in Figure 3.2(c3), the agreement between the residue fluctuations and the PONDR-predicted disorder scores is with a Pearson correlation of 79% and q-factor of 31%. It is worth noting that this comparison is qualitative since residue fluctuations and the disorder scores represent different quantities. Interestingly, both quantities, as shown in Figure 3.2(c3), indicate that ILE98 is the most flexible and disordered residue in the missing loop of VlsE.

3.5 Applications of FEM on Internal Loops

An additional consideration when characterizing the native flexibility of protein loops is the space available to fluctuations. Internal loops in proteins represent an extreme case where the available space is limited by the rest of the protein structure. An instance of an internal loop is the L1 loop, residues $37 - 51$, of the human 120-residue SWI1 AT-rich interaction domain (ARID) protein (SWI1 ARID), PDB code



(a) (b) (c)

Fig. 3.3: (a) The equilibrated representative NMR structure of SWI1 ARID and its L1 loop are drawn with VMD [HDS96]. The loop, in cyan, is surrounded by the rest of the protein structure, whose solvent accessible surface, in white, is computed by sliding a 1.4 Å radius sphere approximation of a water molecule. (b) Using VMD [HDS96], the FEM-obtained ensemble of $5,000$ loop conformations is shown transparent vs. the opaque reference structure. (c) The energy landscape associated with the ensemble is shown in red. The black line represents the average energy profile. The energy landscape is funnel-like and the average energy profile is steep, indicating that FEM recovers the native structure of the loop.

1RYU [KZU+04]. As shown in Figure 3.3(a), the equilibrated representative NMR structure of SWI1 ARID surrounds the L1 loop.

FEM is applied to obtain an equilibrium conformational ensemble for this loop. Because of the reduced conformational space available, it is computationally more costly to obtain physical conformations of an internal loop. The number of conformations generated during the exploration of FEM before obtaining a collision-free conformation where atoms of the loop are not in collision with the surrounding structure of the protein is on average 4 times higher than in the case of the non-internal loops presented above. This computational cost can be lowered by, for instance, taking into consideration the surrounding environment when sampling loop conformations that fit with the rest of the protein structure.

Despite the higher computational cost, the obtained ensemble, shown in Fig-

ure 3.3(b), accurately models the fact that the mobility of the L1 loop is highly limited due to the surrounding environment. Obtained conformations are clustered around the native loop of the equilibrated reference structure. This is consistent with Figure 3.3(c), which shows a funneled energy profile. These results indicate that the flexibility of internal loops is highly constrained from the surrounding environment.

3.6 A Closer Look at the FEM Exploration

Results in sections 3.4-3.5 can be interpreted in light of the FEM exploration. Modeling a fragment as a kinematic chain and using CCD allows FEM to map the uniformly sampled space $\mathcal{C}$ of chain conformations to the space $\bar{\mathcal{C}}$ of IK solutions, conformations that satisfy the end-effector kinematic constraints. To sufficiently explore the sub-space of low-energy conformations to which the energy minimization procedure maps $\bar{\mathcal{C}}$, FEM needs to provide a good coverage of $\bar{\mathcal{C}}$.

The solution space $\bar{\mathcal{C}}$ can be described by a system of multi-variable non-linear polynomial equations that relate the chain DOFs to the end-effector constraints [Cra89]. $\bar{\mathcal{C}}$ may contain components of different dimensions such as isolated solutions, solution curves, and solution surfaces [SVW04]. A notion of coverage of $\bar{\mathcal{C}}$ can be given through that of dispersion [CLH$^+$05], which measures the largest portion of $\bar{\mathcal{C}}$ where FEM samples no conformations. A good coverage of $\bar{\mathcal{C}}$ involves minimizing dispersion in each component of $\bar{\mathcal{C}}$. Covering each component uniformly, as provided through the notion of discrepancy [CLH$^+$05], might be desirable as well.

The question whether applying CCD to $\mathcal{C}$ provides a good coverage of each component of $\bar{\mathcal{C}}$ remains challenging and open to theoretical analysis. Answering this question is further complicated by the not yet understood dependence of the $\bar{\mathcal{C}}$ exploration on the σ permutation of the chain DOFs employed by CCD. Demonstration of an inadequate coverage of $\bar{\mathcal{C}}$ does not necessarily mean that the exploration of the equilibrium conformational space of a protein fragment is insufficient. The reason is that not all components of $\bar{\mathcal{C}}$ may be accessible to a protein. It has been shown that certain equilibrium conformations may be kinematically inaccessible, i.e., unreachable within biological timescales [BA94].

In light of these open questions, to provide insight into the coverage of $\bar{\mathcal{C}}$, an experimental analysis is provided on the FEM exploration of $\bar{\mathcal{C}}$ for kinematic chains with increasing number of DOFs. The analysis first start with 6R chains, where the upper bound of 16 IK solutions [Pri86] allows directly comparing these solutions to those obtained by FEM when mapping $\mathcal{C}$ with CCD. On kinematic chains with more than 6 DOFs, the dimensionality of $\bar{\mathcal{C}}$ does not allow for a direct comparison. In this case, the analysis focuses on solutions obtained when applying CCD to neighborhoods of conformations in $\mathcal{C}$. To investigate how the σ permutation of DOFs employed by CCD affects the exploration of $\bar{\mathcal{C}}$, each experiment is repeated with three obvious choices for σ; counting from the base to the end-effector (N- to C-terminus) of the chain one can define: (i) the random permutation, where the order of DOFs changes randomly in each CCD cycle; (ii) the identity permutation, where the value for DOF

i is found before the one for DOF $i+1$; and (iii) the reverse permutation, which refers to the reverse of the identity permutation.

3.6.1 On the Space of Kinematically-constrained 6-DOF Chains

It is interesting to determine first whether for 6R chains, mapping $\mathcal{C}$ with CCD allows to sample all $\bar{\mathcal{C}}$. This is done on a comprehensive list of 20 IK problems for which all IK solutions, obtained with a polynomial continuation method, are documented [WM89]. For each problem, IK solutions are compared to the solutions sampled by FEM. Two conformations are deemed close if their geodesic distance in $SO(2)^n$ normalized by the number n of DOFs is no more than 0.1 rad. Sampled solutions of a problem with i IK solutions are discretized into i bins, each bin corresponding to an IK solution. A sampled solution goes into a particular bin if it is closest to the IK solution associated with that bin. Solutions are sampled until no bin is empty. For each problem, for each choice of σ, all bins are filled after sampling a maximum of 100 solutions. The maximum distance between a sampled solution and its closest IK solution is no more than 0.02 radians. For 6R chains, for each choice of σ, CCD allows to obtain all isolated IK solutions of $\bar{\mathcal{C}}$.

3.6.2 On the Space of Kinematically-constrained Redundant Chains

For redundant kinematic chains (more DOFs than constraints) the solution space $\bar{\mathcal{C}}$ is not discrete but may contain components of different dimensions [SVW04]. The analysis now assesses the ability of FEM to sample different regions of $\bar{\mathcal{C}}$ for redundant

	[30 DOFs]			[50 DOFs]			[100 DOFs]		
	(i)	(ii)	(iii)	(i)	(ii)	(iii)	(i)	(ii)	(iii)
1 °	0.6115	7.0175	5.8640	0.2241	7.7586	5.9319	0.2279	6.2142	5.0963
5 °	0.0045	0.0988	0.0395	0.0010	0.0036	0.0018	0.0010	0.0010	0.0010
10 °	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010

Table 3.1: A conformation A sampled uniformly at random from $\mathcal{C}$ maps with CCD to an IK solution $B \in \bar{\mathcal{C}}$. Table shows how many (in %) of 1000000 neighbor conformations of A sampled uniformly at random map with CCD to B. Rows show results obtained when neighbor conformations of A are sampled from neighborhoods of radii $\{1°, 5°, 10°\}$. Columns show results obtained for chains of 30, 50, and 100 DOFs when CCD employs three different choices of the σ permutation of DOFs. (i)-(iii) refer to the random, identity, and reverse permutations, respectively. Results are averaged over 100 instances of A.

chains by analyzing the solutions obtained when applying CCD to neighborhoods of a conformation A sampled uniformly at random in $\mathcal{C}$. 1000000 neighbor conformations of A are sampled uniformly at random from neighborhoods of radii $\{1°, 5°, 10°\}$ deviation per DOF. CCD is applied to A and its neighbor conformations so the end-effector reaches a target pose that is randomly sampled in $SE(3)$. Differences are then measured between the solutions to which CCD maps a neighborhood of A and the conformation B to which A maps under CCD. This is done for 100 instances of conformations A and different choices of σ on chains of 30, 50, and 100 DOFs.

Measurements first focus on the probability that CCD maps a neighbor conformation of A to B. Table 3.1 shows that when employing CCD to increasing perturbations of a random conformation A, the probability of obtaining the IK solution B to which A maps decreases rapidly. In general the probability gets respectively smaller when employing CCD with the identity, reverse, and random permutation. In fact, when employing CCD with the random permutation on a chain of 50 DOFs, the probability

Fig. 3.4: (a) A distribution of conformations in $\mathcal{C}$ is obtained by sampling uniformly at random 1000000 conformations of kinematic chains of $30, 50, 100$ DOFs from neighborhoods of radii $\{1°, 5°, 10°\}$ around a conformation A sampled uniformly at random in $\mathcal{C}$. Mapping this distribution with CCD yields a distribution of conformations in $\bar{\mathcal{C}}$. For each distribution, the distance between the mean and median conformations is measured through ρ, the geodesic distance in $SO(2)^n$ normalized by the number n of DOFs. The ratio of the distance corresponding to the distribution in $\bar{\mathcal{C}}$ over that corresponding to the distribution in $\mathcal{C}$ averaged over 100 instances of A is plotted here. (b) Conformations of a chain with 12 DOFs are sampled uniformly at random from a 10° radius neighborhood that maps with CCD to the conformations shown in (c). (a)-(c) Results are obtained with the random permutation of DOFs.

of obtaining the same IK solution B drops quickly to 0.0001% when increasing the perturbation to 10°. The decrease in the probability of obtaining the same solution in $\bar{\mathcal{C}}$ upon increasing neighborhood radii in $\mathcal{C}$ indicates that the sampling of solutions is not limited to a particular region of $\bar{\mathcal{C}}$.

The obtained distribution of solutions in $\bar{\mathcal{C}}$ is now compared with the distribution of the sampled neighbor conformations. For each distribution, the distance between the mean and median conformations is measured. The median conformation of the neighborhood of A corresponds to the median distance between A and its neighbor conformations. The median conformation of the obtained distribution in $\bar{\mathcal{C}}$ corresponds to the median distance between B and obtained solutions in $\bar{\mathcal{C}}$. Fig. 3.4(a)

shows that for each of the chains the distance between the mean and median conformations in the distribution of sampled solutions is persistently larger than in the distribution of neighbor conformations of A. The difference between the two distributions gets smaller as neighbor conformations of A get more diverse with the increase of neighborhood radius and number of DOFs. Fig. 3.4(b)-(c) illustrates how small perturbations around a conformation A can map to a diverse set of solutions. Similar results are obtained for permutations other than the random. While the observed diversity is not desirable when kinematically-constrained chains need to follow a particular trajectory [BK02], this very feature of CCD allows in this work to obtain different solutions and so explore different regions of $\bar{\mathcal{C}}$.

This analysis shows that CCD allows FEM to explore different conformations of $\bar{\mathcal{C}}$ for redundant chains. While the question whether applying CCD to $\mathcal{C}$ allows to cover all components of $\bar{\mathcal{C}}$ remains, in practice, the FEM exploration of $\bar{\mathcal{C}}$ is sufficient to model equilibrium fluctuations. As shown in applications of FEM, good agreement is obtained between $\langle X_i \rangle$ measurements computed over the sub-space of low-energy conformations to which $\bar{\mathcal{C}}$ maps under the energy minimization procedure and measurements provided from experiments or guided simulations.

3.7 Analysis of the Interleaving Minimization in FEM

Exploration of the self-motion manifold has been previously employed to refine the backbone of a closed loop in partially resolved X-ray structures [LvdBDL04, vd-

BLLD05] according to a coarse-grained objective function. In FEM, the self-motion manifold is explored to minimize physical energy functions such as the CHARMM [MBB+98] all-atom force field. Since the computation of the self-motion manifold and its exploration are computationally expensive, an analysis is conducted to determined whether combining this exploration with the closure-constrained conjugate gradient descent as opposed to only employing the closure-constrained conjugate gradient descent for the minimization offers significant benefits.

The analysis compares the improvement in energy of conformations minimized by the interleaving procedure as opposed to using closure-constrained conjugate gradient descent minimization only. Two opposite scenarios are plotted in Figure 3.5 on the energetic refinement of ubiquitin conformations for fragments $[15, 45]$ and $[35, 65]$. For brevity, the closure-constrained conjugate gradient descent minimization is referred to as CG and to the interleaving procedure as CGSMM since it combines the closure-constrained conjugate gradient descent minimization with the exploration of the self-motion manifold. Figures 3.5(a) and (b) plot the final energies of conformations refined through the interleaving procedure versus the final energies obtained when refining conformations with the closure-constrained conjugate gradient descent minimization only, for fragments $[15, 45]$ and $[35, 65]$, respectively.

Figures 3.5(a) and (b) show a positive outcome, where lower energy completed protein conformations are obtained in the case of fragment $[14, 45]$ by interleaving than by using the closure-constrained conjugate gradient descent minimization alone.

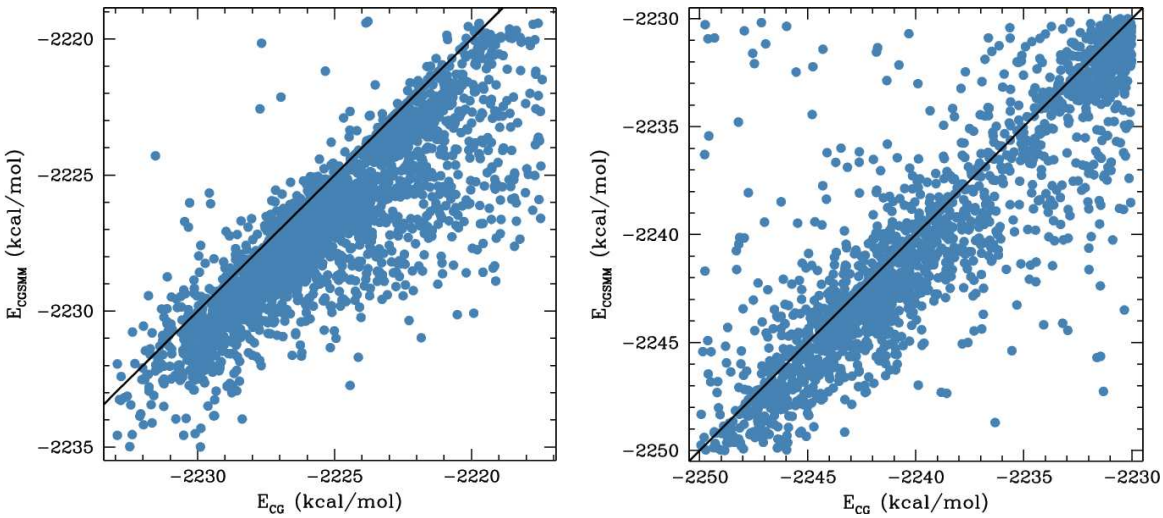


Fig. 3.5: (a) Conformations generated for fragment $[15, 45]$ in ubiquitin, refined with CGSMM, have lower energies than those refined with CG alone. (b) Almost 50% of the conformations generated for fragment $[35, 65]$ in ubiquitin, refined with CGSMM, have higher energies than would be obtained if refining them with CG alone.

Figure 3.5(c) shows an opposite scenario, where interleaving changes the energy landscape and results in conformations with higher energy than would be obtained by the closure-constrained conjugate gradient descent minimization alone. Whether using the interleaving procedure or the closure-constrained conjugate gradient descent alone, the ensemble of generated conformations does not fundamentally change. The conclusion from this analysis is that the computational time devoted to the exploration of the self-motion manifold is not crucial to the minimization procedure. The closure-constrained conjugate gradient descent minimization alone is sufficient to lower the energy of completed protein conformations.

3.8 FEM: Discussion and Conclusion

The FEM proposed to model native loop flexibility makes use of an efficient robotics-inspired exploration to sample the conformational space available to a missing fragment that fits with a given protein structure. This exploration allows FEM to explore the space of arbitrarily-long fragments, an advantage over database and ab initio search-based methods [SK90, vVK97, TBHM02, DAL03, KGLK05, DB00, BK90, FDS00, MJ93, FR92, ZRVD93, ZRDK94].

The multi-resolution approach employed in FEM allows efficiently modeling protein fragments as kinematically-constrained chains. In addition, the use of all-atom force fields allows estimation of conformational energies. A statistical mechanics formulation then provides a natural way to associate a weight to each obtained conformation and as a result obtain a native conformational ensemble. This is an obvious advantage over existing exploration methods applied to protein fragments [CSRST04, LvdBDL04, vdBLLD05].

When applied to stable proteins such as CI2 and α-Lac, the proposed FEM recovers the native loops of these proteins. The generated ensembles are clustered around the native loops, and the associated energy landscapes are funnel-like. Fluctuations measured over each ensemble are fully consistent with experimental data and existing simulations. A novel application of FEM on VlsE with a missing loop of 20 amino acids generates an ensemble whose conformational heterogeneity is consistent with the high disorder of the missing loop. These results point to a potential application

of FEM where consideration of the crystal environment as in [JPR+04] may allow to even model the effects of crystal packing on loop flexibility.

The results obtained by FEM motivate the employment of the method as a core computational unit in a larger scheme aimed at characterizing the native flexibility not just of one protein fragment but of an entire protein chain. Usage of FEM for this purpose is detailed in chapter 4.

Chapter 4

Characterizing the Native Flexibility of a Protein

This chapter proposes the Protein Ensemble Method (PEM) to characterize native flexibility in proteins where fragments of the protein chain move independently of one another. PEM combines local fluctuations of consecutive overlapping fragments. Local fluctuations are obtained with FEM as described in chapter 3. Using the theory of statistical mechanics, the Boltzmann-weighted fluctuations corresponding to each fragment are combined to obtain fluctuations for the entire protein. The agreement obtained between PEM-modeled fluctuations, wet-lab experiment, and guided simulation measurements, indicates that PEM is able to reproduce with high accuracy protein native fluctuations that occur over a broad range of timescales.

4.1 Introduction on the Native Flexibility of a Protein

PEM complements experimental and simulation techniques by characterizing the native flexibility of an entire protein. Unlike existing simulation techniques [Dag00, PB02, HOvG02, Tai04, RVS04, vGBB+06], PEM does not follow trajectories in conformational space but samples conformations independently of one another.

PEM is based on the premise that, in proteins where fragments of the protein chain do not move in concert with one-another, global native fluctuations of the protein chain can be obtained by combining local native fluctuations of fragments. Fragments are defined consecutively and with overlap by sliding a window over the chain. PEM measures amino-acid fluctuations of each fragment as Boltzmann-weighted averages over the sampled space of low-energy conformations of each fragment. Low-energy conformations of each fragment are obtained with FEM, detailed in chapter 3.

The results obtained by applications of PEM to various proteins indicate that one computationally effective strategy to model global native fluctuations of a protein is to combine local native fluctuations of consecutive overlapping protein fragments. This strategy is appealing because fluctuations of different fragments can be obtained in parallel. The strategy is well suited for proteins with non-concerted fluctuations, that is, where native fluctuations of a fragment can be obtained while the rest of the polypeptide chain is unperturbed. Extensions to capture concerted fluctuations, briefly discussed in section 4.6, are motivation of further work presented in chapter 6.

Focusing on proteins with non-concerted motions is of broad interest. There is no evidence of correlation between global physico-chemical properties such as stability or contact order [PSB98] and the nature, local or correlated, of protein fluctuations. Moreover, despite the limited information on protein structures and motions available in current databases [BWF$^+$00] and literature, proteins with non-concerted motions represent a significant portion of proteins with known structure [Fer99, BSM$^+$06]. For

the proteins presented, PEM-modeled fluctuations are fully consistent with multiple timescale measurements obtained from NMR wet-lab experiments and guided simulation techniques. Thus, for the examples considered, PEM can be employed to provide a microscopic level of understanding of protein function.

The rest of this chapter is organized as follows. Section 4.2 summarizes related work to place PEM in context of other methods aimed at capturing native flexibility in proteins. Section 4.3 is devoted to a thorough comparison and discussion of the advantages of PEM over existing simulation techniques. This section also provides biophysical background and rationale behind the design of PEM. Details and analysis of PEM are related in section 4.4. Section 4.5 analyzes PEM-obtained fluctuations for ubiquitin and protein G. This section shows that measured thermodynamic quantities correlate remarkably well NMR data such as order parameters, residual dipolar couplings, and 3-bond scalar couplings. The chapter concludes in section 4.6.

4.2 Related Work on Characterizing Native Flexibility

What follows is a survey of existing simulation techniques. The survey is by no means comprehensive. Rather, it focuses on related work that places PEM in context.

4.2.1 Survey of Simulation Techniques

Current simulation techniques to sample conformational space are either systematic or random searches [vGBB$^+$06]. MD simulations [NN03, KK05, AM06, vGBB$^+$06]

systematically update atom coordinates of a conformation to obtain a new one by numerically solving Newton's equations of motions. The occurrence of a conformation obtained with a constant temperature MD simulation is proportional to the Boltzmann probability. Since the solution accuracy demands a small timestep in the order of femtoseconds, obtaining a physical trajectory of conformations is computationally demanding [vGBB$^+$06, Elb05]. Moreover, thoroughly sampling conformational space may require many trajectories. The sampling of rare events such as crossing local maxima of the energy landscape adds to the computational cost of sampling conformational space in a sequential fashion. Thus, in a reasonable amount of time, MD simulations sample a small sub-space of the conformational space available to a protein and are often limited to exploring events that occur within nanoseconds [Dag00, PB02, HOvG02, Hes02, Tai04, vGBB$^+$06].

Rather than solving Newton's equations of motions, random search techniques such as MC simulations [RVS04, vGBB$^+$06] conduct a biased probabilistic walk in conformational space to obtain a sequence of conformations. The biased probabilistic walk ensures through the Metropolis criterion [MRR$^+$53] that a conformation is obtained with frequency proportional to its Boltzmann probability. While sometimes computationally more efficient than MD simulations, MC simulations also obtain conformations sequentially. Hence, they also spend considerable time sampling rare events such as crossing maxima in the energy landscape. Extensions to enhance sampling include methods such as importance [KLV74] and umbrella sampling [TV77],

replica MC [SW86], jump walking [FFD90], multicanonical ensemble [BN92], entropic sampling [Lee93], weighted histograms [KRB$^+$93], local elevation [HTvG94], parallel tempering/replica exchange [Han97], smart walking [ZB97], multicanonical jump walking [XJ99], conformational flooding [SGE00], local energy flattening [ZKS02], activation relaxation [MM00], Markov state models [SSP04], and guided simulation techniques [VPDK03, BV04, LLBD$^+$05] that use experimental data to guide trajectories to relevant regions of conformational space.

The proposed PEM classifies as a random search that transforms a non-Boltzmann collection of randomly sampled conformations into a Boltzmann ensemble by weighting each conformation with its Boltzmann probability. Rather than obtaining conformations sequentially, PEM probes the energy landscape through a probabilistic exploration that samples conformations independently of one another.

4.3 PEM, Related Methods, and Biophysical Rationale

Conformations of a polypeptide chain are often kinematically constrained, e.g., by the bond network of a protein [JRKT01]. PEM employs a probabilistic space exploration with kinematic constraints (implemented in FEM as described in chapter 3). In FEM, conformations are obtained independently of one another, which allows modeling native fluctuations with no inherent timescale limitations. This is an advantage over existing simulation techniques which, due to their exploration of protein conformational space one trajectory at a time, are limited to modeling native fluctua-

tions up to the nanosecond timescale [Dag00,PB02,HOvG02,Hes02,Tai04,vGBB$^+$06]. A comparison with existing simulation techniques, presented in the following for both accuracy and running time, highlights the advantages of the proposed PEM. The purpose of comparing running times is mainly to illustrate the orders of magnitude difference between PEM and existing simulation techniques since running times of simulation studies are reported in different machines and by different authors.

Applications of PEM to ubiquitin, protein G, and other proteins show that PEM-obtained fluctuations agree very well with available experimental data over multiple timescales. Pearson correlations no lower than 0.80 are achieved between native fluctuations obtained by PEM after no more than 164 CPU hours on a current processor and experimental and guided simulation measurements (details on the accuracy of the results obtained by PEM can be found in section 4.5). On the other hand, achieving these same high correlations with existing simulation techniques is either possible through simulations that require orders of magnitudes longer CPU time (year-long) [HAO$^+$06] or through simulations that shorten running time to a few months or a few weeks by incorporating experimental data to guide MD or MC trajectories to relevant regions of conformational space [VPDK03, BV04, LLBD$^+$05].

The high computational time demand of existing simulation techniques limits a direct comparison between these techniques and PEM to a few well-studied proteins. On α-Lac, a protein presented in [SCK07], obtaining native fluctuations beyond the ns timescale remains challenging for existing simulation techniques. To the best of

my knowledge, the only simulation study that overcomes the timescale limitations on α-Lac is an MC simulation that employs a coarse-grained representation of this protein and guides trajectories by incorporating available experimental data [VPDK03]. With no a priori knowledge of experimental data and at the same time employing an all-atom representation of protein conformations, in 164 CPU hours PEM obtains native fluctuations of α-Lac that occur over a wide range of timescales. As detailed in [SCK07], PEM-obtained fluctuations for α-Lac agree very well with the ensemble of conformations obtained in [VPDK03].

An important reference protein system for comparisons is ubiquitin, where 6ns of an MD simulation in explicit solvent are needed to obtain a correlation of 0.62 with NMR data [LLBD$^+$05]. While running times are not reported, our experience estimates that 2ns of simulation time on an AMD Athlon 1900MP machine require one week of CPU time. Longer CPU times are needed to achieve higher accuracy: 80ns, estimated to about one year of CPU time, are needed to obtain a 0.96 correlation with NMR order parameters [HAO$^+$06]. The only successful simulation study to my knowledge to obtain good agreement with experimental data (correlation of 0.96) in a few months (22.5ns) guides MD trajectories to relevant regions of conformational space with NMR data [LLBD$^+$05]. While this result is very significant [Bor05], the required a priori knowledge of high-quality experimental data limits the predictive power of guided simulation techniques. With no additional knowledge of experimental data, native fluctuations obtained by PEM after 120 hours of CPU time on ubiquitin

agree with available NMR data with correlations no lower than 0.95.

PEM focuses on obtaining native fluctuations in proteins where fluctuations of fragments of the polypeptide chain are uncorrelated (overcoming this assumption is the subject of chapter 6). To obtain such fluctuations, PEM employs a first-order approximation that is a powerful algorithmic approach well-rooted in biophysics, particularly in the context of protein folding [Sch58,MnTHE97,HF96,WLH01]. In protein folding, the enumeration of all configurations of a protein (where each amino acid is considered either in an ordered or a disordered state) is often addressed through a first-order approximation which groups all ordered amino acids on one single continuous stretch of the protein sequence. Considering one single continuous stretch of the protein sequence at a time (or one fragment at a time) is known as the "single sequence approximation." The single sequence approximation was first proposed in the context of the helix-coil theory [Sch58, MnTHE97] and lately has been shown sufficient in enumerating folding propensities of amino acids of many different proteins [HF96, WLH01].

PEM uses the single sequence approximation in a novel context; the method samples conformations of a fragment while the rest of the polypeptide chain is unperturbed in order to obtain detailed all-atom information about protein conformations under native conditions. The applicability of the single sequence approximation in this context is justified in proteins where there are no correlated motions between fragments of the polypeptide chain that are far in sequence and where, as a consequence,

fluctuations of one fragment can be obtained independently of another. As discussed in detail in section 4.4, in the absence of correlated fluctuations, PEM constructs fluctuations of the entire polypeptide chain in a multiscale fashion by combining together the fluctuations obtained for fragments covering the chain. Treatment of native fluctuations in the context of even correlated fluctuations is presented in section 4.6, which motivates the work presented in chapter 6.

4.4 PEM: A Local-to-Global Approach

Section 4.4.1 shows how PEM defines fragments on a protein polypeptide chain and then combines FEM-obtained fluctuations measured over the native conformational space of each fragment to model global native fluctuations of an entire protein chain.

As shown in pseudocode in Algorithm 2, PEM takes as input an experimentally determined conformation C_{PDB} from the PDB [BWF$^+$00]. Since C_{PDB} is an average over protein conformations populated under native conditions, PEM initially minimizes the energy of C_{PDB} with a conjugate gradient descent on the energy landscape, as detailed in chapter 3, to obtain a conformation C_{ref} whose energy E_{ref} is assumed to correspond to the global minimum of the energy landscape. PEM employs C_{ref} as a reference conformation to sample low-energy conformations near the global minimum.

Algorithm 2 PEM (C_{PDB}, l, δl, dl, w)

Input:
C_{PDB}: protein conformation obtained from the PDB
l: length of window sliding over polypeptide chain of protein
δl: overlap between consecutive windows
dl: size of increment to l and δl
w: function to weight fluctuation of an amino acid of a fragment
Output: Native fluctuations $\langle X_i \rangle$ of each amino acid i

1: $C_{\text{ref}} \leftarrow$ energetically refine C_{PDB}
2: $P \leftarrow$ protein polypeptide chain comprising amino acids 1 to N
3: Slide over P a window of length l amino acids with overlap of δl amino acids between consecutive windows to define fragments $[n_1, n_2]$
4: **for** each fragment $[n_1, n_2]$ **do**
5: $\Omega_{[n_1,n_2]} \leftarrow$ ensemble of FEM-obtained low-energy conformations of fragment $[n_1, n_2]$
6: associate $e^{-(E(C)-E_{\text{ref}})/(RT_0)}$ to each $C \in \Omega_{[n_1,n_2]}$ to obtain Boltzmann ensemble
7: $Z \leftarrow \sum_{C \in \Omega_{[n_1,n_2]}} e^{-(E(C)-E_{\text{ref}})/(RT_0)}$ ▷partition function-normalization factor
8: $\langle X_i \rangle_{[n_1,n_2]} \leftarrow \frac{1}{Z} \sum_{C \in \Omega_{[n_1,n_2]}} e^{-(E(C)-E_{\text{ref}})/(RT_0)} X_i(C)$ for amino acid $i \in [n_1, n_2]$
9: **for** each amino acid $i \in P$ **do**
10: $\mathcal{N}_i \leftarrow \sum_{\{\,[n_1,n_2]\,:\, i \in [n_1,n_2]\}} w(i, [n_1, n_2])$ ▷normalization factor
11: $\langle X_i \rangle \leftarrow \frac{1}{\mathcal{N}_i} \sum_{\{\,[n_1,n_2]\,:\, i \in [n_1,n_2]\}} \langle X_i \rangle_{[n_1,n_2]}\, w(i, [n_1, n_2])$
12: $\{\langle X_i \rangle_{\min}, \langle X_i \rangle_{\max}\} \leftarrow \{\min, \max\}_{\{\,[n_1,n_2]\,:\, i \in [n_1,n_2]\}}\ \langle X_i \rangle_{[n_1,n_2]}$
13: **if** $\langle X_i \rangle_{\max} - \langle X_i \rangle_{\min} \geq \langle X_i \rangle_{\min}$ **then**
14: $l \leftarrow l + dl$ and $\delta l \leftarrow \delta l + dl$
15: **goto** line 3

4.4.1 Consecutive Overlapping Fragments over a Protein Chain

As shown in line 3 of Algorithm 2, PEM slides a window l residues long over the protein chain P to split P into consecutive fragments. The window is slid so that neighboring fragments overlap significantly with one another in $\delta l \approx l$ residues (by definition, $\delta l < l$). As illustrated in Fig. 4.1(a), sliding a window of length 30 with overlap of 25 residues defines 19 fragments on the 123-residue chain of α-Lac.

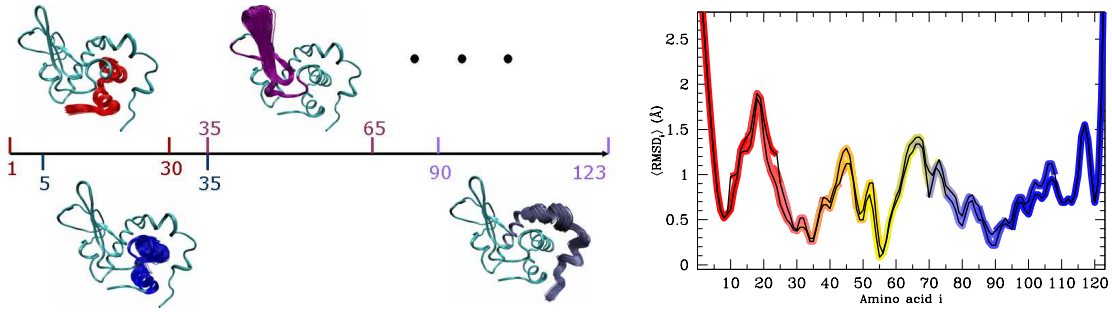
4.4.2 Obtaining Local Native Fluctuations of a Fragment

PEM employs FEM to sample an ensemble $\Omega_{[n_1,n_2]}$ of low-energy conformations of a fragment $[n_1, n_2]$ while keeping the rest of the polypeptide chain as in C_{ref}. Fig. 4.1(a) shows such ensembles for fragments defined on the polypeptide chain of α-Lac. As shown in line 6 of Algorithm 2, the theory of statistical mechanics [Cha87] is employed to transform the sampled ensemble $\Omega_{[n_1,n_2]}$ into a Boltzmann ensemble of conformations by weighting each conformation C of $\Omega_{[n_1,n_2]}$ with its Boltzmann probability $e^{-(E(C)-E_{\text{ref}})/RT_0}$ (as described in chapter 3).

Let $X_i(C)$ measure the fluctuation of amino acid i around C_{ref} as witnessed by a conformation C. An example of $X_i(C)$ is the lRMSD of amino acid i from C_{ref}:

$$\text{lRMSD}_i(C) = \sqrt{\frac{1}{\#\text{ atoms in i}} \sum_{\text{atom j} \in \text{ i}} \| \vec{p_j}(C) - \vec{p_j}(C_{\text{ref}}) \|^2},$$

where $\vec{p_j}$ is the position of atom j, and $\|\cdot\|$ is the L_2 norm. Other choices for $X_i(C)$ include the deviation from C_{ref} of the orientation of a particular bond vector in amino acid i (order parameters [Kay05], presented in detail in section 4.5, constitute another choice for $X_i(C)$). As described in chapter 3, the transformation of $\Omega_{[n_1,n_2]}$ into a Boltzmann ensemble allows to measure a statistical average of $X_i(C)$ over conformations $C \in \Omega_{[n_1,n_2]}$. As shown in line 8 of Algorithm 2, PEM sums over all $X_i(C)$, weighting each by the Boltzmann probability of the corresponding conformation $C \in \Omega_{[n_1,n_2]}$, to obtain a Boltzmann-weighted average $\langle X_i \rangle_{[n_1,n_2]}$. This average quantifies the FEM-obtained fluctuation of amino acid i as witnessed by the $\Omega_{[n_1,n_2]}$ ensemble of conformations available to fragment $[n_1, n_2]$ under native conditions.



(a) (b)

Fig. 4.1: (a) Sliding a window of length 30 and overlap of 25 amino acids on the 123-aa chain of α-Lac defines 19 fragments, starting with $[1, 30]$ and ending with $[90, 123]$. An ensemble of low-energy conformations is sampled for each fragment through the FEM exploration detailed in chapter 3. Each ensemble is shown in different colors while the rest of C_{ref} is in cyan. Conformations are drawn with VMD [HDS96]. (b) $\langle \text{lRMSD}_i \rangle_{[n_1,n_2]}$ values, measured as in line 8 of Algorithm 2, are drawn in different colors for different fragments $[n_1, n_2]$. Values for the first and last 5 amino acids of each fragment are discarded. $\langle \text{lRMSD}_i \rangle_{\text{min}}$ and $\langle \text{lRMSD}_i \rangle_{\text{max}}$, measured as in line 12 of Algorithm 2, are drawn in black.

The average $\langle X_i \rangle_{[n_1,n_2]}$ measured over ensemble $\Omega_{[n_1,n_2]}$ can change with the addition of sampled conformations to the ensemble. To determine a termination condition for sampling, one can measure whether, after adding conformations to ensemble $\Omega_{[n_1,n_2]}$, there are any changes in ensemble-averaged measurements such as $\langle \text{lRMSD}_i \rangle_{[n_1,n2]}$. When such measurements converge, the sampling of low-energy conformations of fragment $[n_1, n_2]$ terminates as no new information is obtained about native fluctuations of the fragment.

4.4.3 Global Native Flexibility: Combining Fragment Fluctuations

Modeling global native fluctuations of a polypeptide chain involves quantifying the fluctuation of any amino acid i of the chain as witnessed by the available native conformational space. As the fluctuation of an amino acid i is a statistical average

over the available conformations of the chain, it is denoted here by $\langle X_i \rangle$. PEM estimates the global fluctuation $\langle X_i \rangle$ of an amino acid i by combining local fluctuations $\langle X_i \rangle_{[n_1,n_2]}$. As shown in line 11 of Algorithm 2, PEM estimates $\langle X_i \rangle$ as a weighted average over all fluctuations $\langle X_i \rangle_{[n_1,n_2]}$ measured over the sampled ensembles of the fragments $[n_1, n_2]$ overlapping in i.

As illustrated for α-Lac in Fig. 4.1, $\langle X_{19} \rangle$ is averaged over $\langle X_{19} \rangle_{[1,30]}$, $\langle X_{19} \rangle_{[5,35]}$, $\langle X_{19} \rangle_{[10,40]}$, and $\langle X_{19} \rangle_{[15,45]}$. Due to the method employed by PEM to satisfy the kinematic constraints on amino acids n_1 and n_2 of a fragment $[n_1, n_2]$, amino acids i close to n_1 or n_2 do not deviate significantly from their configurations in C_{ref} in the sampled ensemble $\Omega_{[n_1,n_2]}$. Their fluctuations are consequently low and not representative of native conditions. To take this into account, their contribution to the global native fluctuation $\langle X_i \rangle$ is downplayed through a weighting function $w(i, [n_1, n_2])$.

An example of a weighting function that downplays fluctuations of the first and last 5 amino acids of each fragment is $w(i, [n_1, n_2]) = 0$ if $\min\{|i-n_1|, |i-n_2|\} < 5$ and $w(i, [n_1, n_2]) = 1$ otherwise. Figure 4.1(b) shows measured $\langle \text{lRMSD}_i \rangle$ for each amino acid i in α-Lac using this weighting function. Fig. 4.1(b) shows that $\langle \text{lRMSD}_i \rangle_{[n_1,n_2]}$ values measured over ensembles of fragments that encompass amino acid i are similar, as indicated by the small difference between $\langle \text{lRMSD}_i \rangle_{\max}$ and $\langle \text{lRMSD}_i \rangle_{\min}$, where $\langle \text{lRMSD}_i \rangle_{\max}$ and $\langle \text{lRMSD}_i \rangle_{\min}$ are measured as shown in line 12 of Algorithm 1. A large difference between $\langle \text{lRMSD}_i \rangle_{\min}$ and $\langle \text{lRMSD}_i \rangle_{\max}$ would indicate that the length of the window limits fluctuations, in which case, as shown in lines 13-15 of

Algorithm 2, window length and overlap are incremented by dl amino acids.

An alternative weighting scheme involves a Gaussian distribution that progressively decreases the contribution of residues closer to the fragments ends, that is $w(i,[n_1,n_2]) = e^{-\frac{1}{2}\left(\frac{\Delta i}{\sigma}\right)^2}$, where $\Delta i = |i - (n_1 + n_2)/2|$ measures the distance of residue $i \in [n_1, n_2]$ from the central residue $(n_1+n_2)/2$ in fragment $[n_1, n_2]$ ($\sigma = l/2$).

4.4.4 Measuring Robustness to Different Approximations

The weighting scheme is one approximation made by PEM. Here is a comprehensive list of all identified approximations:

(i) The order in which the DOFs are progressively updated in the CCD routine. The associated error is estimated by computing differences between averages obtained from two independently generated ensembles: one where the DOFs are ordered sequentially from the N- to the C- terminus, the other by selecting the DOFs in random order.

(ii) The inaccuracy of the energy force field employed. The associated error is estimated by repeating the ensemble generation with two different force fields, CHARMM [MBB$^+$98] and AMBER [WCB$^+$94], and measuring the differences between corresponding thermodynamic averages measured over each ensemble.

(iii) Finite-size effects introduced by the definition of fragments and the nature of the CCD algorithm. Differences between averages obtained from the two different weighting schemes described above provide an estimate for the associated error.

(iv) The interleaving minimization used on the obtained conformations. The associated error is estimated by computing differences between averages obtained from two generated ensembles: one employing the interleaving minimization and the other employing the closure-constrained conjugate gradient descent only.

The errors associated with these approximations are incorporated in the error bars for ensemble averages of NMR data such as order parameters, residual dipolar couplings, and 3-bond scalar couplings. The small error bars (as shown in Figures 3 and 5 in section 4.5) allow concluding that these approximations do not significantly affect the native flexibility captured for the proteins presented here. In particular, the small size of the error bars indicates that the developed PEM is robust against these approximations. Thus, the results obtained in different fragments can be combined to produce a global picture of fluctuations over an entire protein.

4.4.5 Implementation Details

Initial values for window length l and overlap δl are 20 and 15 amino acids, respectively. If the obtained fluctuations appear biased by these values, both l and δl are incremented by 5 amino acids. Though theoretical maximum values for l and δl can reach the entire chain length N, PEM applications suggest $20 \leq l \leq 40$ to maintain accuracy and efficiency. Convergence of PEM-obtained fluctuations for the proteins presented here is attained on $l = 30$ and $\delta l = 25$ amino acids. For each of the proteins here, around $13,000$ conformations with energy within 20 kcal/mol from the reference structure are generated for each 30-aa fragment. Of these, around $5,000$

conformations per fragment have energies no higher than 5 kcal/mol from the energy of the equilibrated solution structure used as reference.

PEM is implemented in `ANSI C/C++` using `Intel`8.0 compilers and libraries. All results presented here were obtained on the Rice University Terascale cluster of 900 MHz `Intel Itanium2` processors and on the Rice University ADA cluster of 2.2 GHz AMD Opteron processors. The computation for each protein required less than 100 CPU hours.

4.5 Applications of PEM on Small- to Medium-size Proteins

This section presents applications of PEM to characterize the native flexibility of proteins of various sizes and folds. Results are presented for streptococcal protein G, human ubiquitin, eglin c, the SH3 domain of Fyn tyrosine kinase (FynSH3), the tenth type III domain of fibronectin (FNfn10), and the P. magnus albumin-binding second GA module of PAB (ALB8-GA). These proteins are 61, 76, 70, 58, 90, and 53 aa long, respectively. Because of their small- to medium-size, these proteins represent ideal applications of PEM.

Moreover, the selected proteins encompass different folds, from all α, α+β, mainly β, to all β. For all these proteins, the PEM-obtained native fluctuations are fully consistent with NMR data such as order parameters (S^2), 3-bond scalar couplings (^3J), and residual dipolar couplings (RDCs). Additionally, for ALB8-GA, where side-chain S^2 data are not available, predictions are made on side-chain fluctuations.

For the proteins presented here, a reference conformation is obtained by equilibrating PDB-obtained structures to remove unfavorable atomic interactions. X-ray structures of protein G [DW94] and ubiquitin [VKBC87] are available in the PDB under entries 1igd and 1ubq, respectively. NMR ensembles of solution structures of eglin c [HGW92], FynSH3 [MPBR96], FNfn10 [MHBC92], and ALB8-GA [JdCW$^+$97] are available under PDB entries 1egl, 1nyg, 1ttf, and 1gab. The X-ray structure or, in the case of NMR ensembles, the solution structure that is reported as the best, representative, or the average of the NMR ensemble for each protein is subjected to a conjugate gradient descent detailed in chapter 3 ($K_d = 0$ in this case). The average structures of the NMR ensembles of FynSH3, FNfn10, and ALB8-GA are reported under PDB entries 1nyf, 1ttg, and 1prb. When a best, representative, or average structure is not reported in the PDB, which is the case for eglin c, the first structure of the NMR ensemble is chosen to be refined.

Equilibrated X-ray structures of protein G and ubiquitin differ from their corresponding X-ray structures by an all-atom lRMSD of no more than 1.2Å. Equilibrated structures of eglin c, FynSH3, FNfn10, and ALB8-GA differ from their corresponding solution structures by all-atom lRMSDs of 1.8, 1.7, 2.0, and 2.5 Å, respectively. The effect of the equilibration on the obtained results is discussed in detail in section 4.6.1.

On all proteins presented here, windows of length 30 residues with 25-residue overlap suffice to reveal consistent fluctuations measured over ensembles of neighboring overlapping fragments. In particular, Figures 4.2(a1) and (b1) qualitatively show the

variability of generated conformations for protein G and ubiquitin. The consistency of fluctuations measured over ensembles of neighboring overlapping fragments can be seen in Figures 4.2(a2) and (b2), where the average lRMSD are plotted for each residue as measured over ensembles of the fragments that encompass that residue.

4.5.1 Thermodynamic Quantities Measured for Validation

Ensembles obtained for each protein are validated by comparing thermodynamic quantities measured over them with NMR data that probe the dynamics of each protein. These quantities are compared to S^2, ${}^3J$, and RDC data available from NMR. Additional measurements by PEM consist of probabilities of contacts and hydrogen bonds. Two amino acids are considered in contact with one another if the Euclidean distance between two of their atoms is no more than 4.5 Å. A hydrogen bond is considered formed if the OH distance is less than 2.4 Å and the maximum NHO angle for the hydrogen bond alignment is 2.44 rad.

Amide order parameters S^2_{NH} provide information on the reorientational averaging of the NH bond, whereas methyl order parameters do so for the methyl bond. S^2 data for a bond are measured by averaging over the distribution of vectors assumed by the bond in a generated ensemble [BV04]. The calculation of S^2 data is based on the Lipari-Szabo model-free formalism [LS82] that does not assume a particular model of internal motions. The model-free formalism allows for a direct comparison of calculated S^2 values with experimental order parameters under the assumption that motions of the methyl symmetry axis and of the protons about this axis are

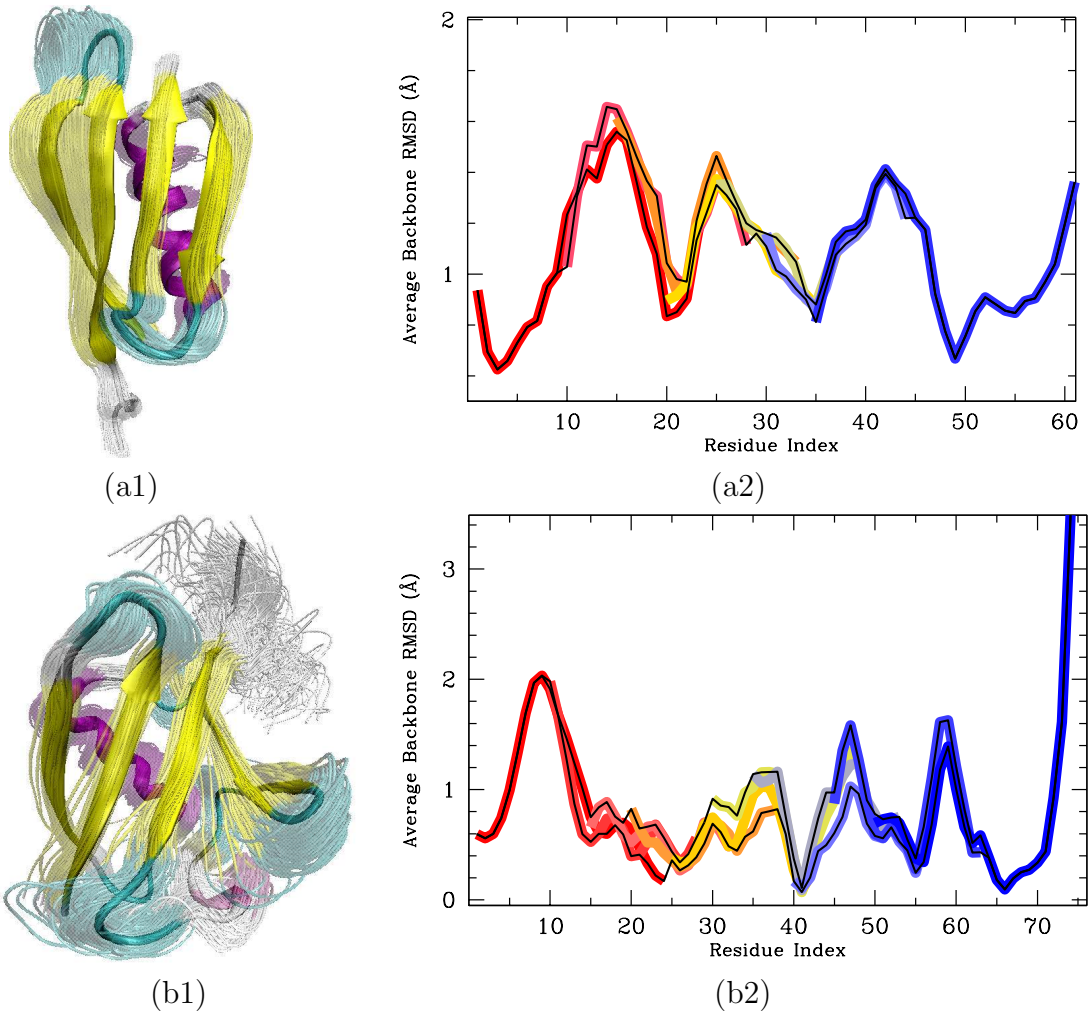


Fig. 4.2: (a1)-(b1) Obtained ensembles for protein G and ubiquitin, respectively. (a2)-(b2) Average lRMSD per residue obtained by combining fluctuations of all fragments regions. Results for different regions are shown in different colors, from red to blue as a window of 30 residues slides from the N- to the C- terminus of the protein. The black lines mark the highest and lowest lRMSD values recorded from all the different windows embracing each given residue, and provide an estimate for the uncertainty of the procedure. Two consecutive 30-residue windows have an overlap of 25 residues. The results corresponding to the first and last 5 residues of each fragment are discarded as they are biased by the finite size of the window.

decoupled [LSL82]. A thorough discussion on the model-free formalism can be found in [LS82, LSL82].

Based on the Lipari-Szabo model-free formalism [LS82], the order parameter $S^2_{i,j}$ for a bond between atoms i and j is calculated through the formula $S^2_{i,j} = \frac{3}{2}(\langle\hat{x}^2_{i,j}\rangle + \langle\hat{y}^2_{i,j}\rangle + \langle\hat{z}^2_{i,j}\rangle + 2\langle\hat{x}_{i,j}\hat{y}_{i,j}\rangle^2 + 2\langle\hat{x}_{i,j}\hat{z}_{i,j}\rangle^2 + 2\langle\hat{y}_{i,j}\hat{z}_{i,j}\rangle^2 - \frac{1}{2})$, where $\hat{x}, \hat{y}, \hat{z}$ denote the components of the unit vector along the bond. Since bond lengths remain essentially unchanged from their native (equilibrium) values during PEM's execution, the above formula can be simplified as in [BV04] to $S^2_{i,j} = \frac{3}{2(r^{\min}_{i,j})^4}(\langle x^2_{i,j}\rangle + \langle y^2_{i,j}\rangle + \langle z^2_{i,j}\rangle + 2\langle x_{i,j}y_{i,j}\rangle^2 + 2\langle x_{i,j}z_{i,j}\rangle^2 + 2\langle y_{i,j}z_{i,j}\rangle^2 - \frac{1}{2})$, where $r^{\min}_{i,j}$ refers to the equilibrium length of the bond connecting atoms i and j. The ensemble-averaged S^2 for a particular bond is thus obtained by Boltzmann-averaging over the distribution of x, y, z components of vectors assumed by the bond. $S^2 = 1$ indicates no heterogeneity in the distribution of these vectors, whereas $S^2 = 0$ is indicative of a uniform distribution.

The ${}^3J$ parameters quantify the side-chain population of rotameric states and are related through the Karplus equation to the probability of occupation of different rotamer states for torsion angles of specific side chains [CCB03]. As outlined in [CCB03], the ensemble of rotameric states can be used to parameterize the Karplus equation. Optimal values to the Karplus parameters A, B, C, δ can be defined to improve the agreement between observed and calculated scalar coupling data. Rather than optimize such parameters, we choose to perform a golden test and use the Karplus equation empirically parameterized for the ubiquitin X-ray structure in [CCB03].

Finally, RDCs quantify fluctuations on the direction of different bond vectors. RDCs are measured over fragment ensembles as in [TB97], normalized with respect to RDCs measured for an amide NH in the same orientation by scaling according to bond lengths and gyromagnetic ratios [TB97].

The difficulty of classic MD simulations in reproducing these NMR data is related to the timescales captured by these data: S^2 data extracted from ^{15}N relaxation experiments capture from the ps to the ns timescale [Kay05]. RDCs report on averages over longer timescales of up to millisecond range and so can reveal slower protein motions over a very broad timescale [Kay05]. Characterizing S^2 and ^3J can also be highly nontrivial since the timescale for the slowest side-chain rotations may be in the millisecond range [MB04]. In particular, ^3J data can report on rotameric averaging on timescales from few hundredths of a second to picoseconds [BVG$^+$94].

The error bars associated with the PEM-calculated thermodynamic averages measure the inherent error originating from the various approximations in PEM (as detailed in chapter 4). The Pearson correlation R^2, the q factor, and the reduced χ^2 factor that are used to quantify the agreement between PEM-obtained and NMR quantities are measured as defined in Bevington & Robinson [BR02].

4.5.2 Validation of protein G Fluctuations with NMR Measurements

Protein G, a cell surface streptococcal protein, binds immunoglobulin with high affinity and potentially enhances microbial virulence. It is important in labeling and purification of antibodies and the study of protein-protein interactions [SBK91]. The

availability of NMR data for protein G [HF03, BBM$^+$05, URDB03] makes it possible to quantitatively validate PEM-obtained fluctuations.

The experimentally available S^2 data for protein G are derived from ^{15}N NMR relaxation experiments [HF03] and capture the fast dynamics of this protein in the picosecond to nanosecond timescale. For brevity, let us refer to them as fast S^2. Figure 4.3(a) shows the agreement between the PEM-obtained backbone (amide) S^2 and the fast S^2 data of protein G. The Pearson correlation between the two quantities is 73%. It is worth noting that no scaling has been applied to the measured S^2 order parameters to match to the fast S^2 data (no scaling is applied in the comparisons with the experimental data for protein G and ubiquitin). The agreement is better on α-helix and the β_2- and β_3-helix loops (residues $22-48$), indicating that most of the flexibility captured by PEM for these residues happens on the picosecond to nanosecond timescale. However, the agreement drops on the N- and C- terminal chains and on residues $14-22$ due to a higher heterogeneity reported for these regions from the corresponding FEM-obtained ensembles. The region between residues $14-22$ incidentally includes the "melting hot spot " [DLG04] loop of residues $14-17$ and the beginning of the β_2-strand, residues $18-22$. The order parameters obtained by PEM for residues $14-17$ point to slower timescale motions for this region.

To validate the high heterogeneity in this region, ensemble-averaged S^2 data are compared with order parameters for the NH bond derived in [BBM$^+$05] as they provide information on reorientational averaging of the NH bond up to the millisecond

timescale. For brevity, let us refer to these as slow S^2. Figure 4.3(b) shows a better agreement between the S^2 data obtained by PEM and the slow S^2 data derived in [BBM$^+$05] as the Pearson correlation improves up to 83%. As Figure 4.3(b) shows, the agreement between the S^2 data for the residues on the N- and C- terminal chains improves, indicating that the motions in these residues happen in a slower timescale. In addition, while the magnitudes of the calculated S^2 data for residues $14-22$ are higher than those derived in [BBM$^+$05], the two profiles for this region of the protein are comparable. This further confirms that motions of this important region in protein G happens in a slower timescale.

To further validate slower timescale fluctuations captured by PEM for protein G, PEM-obtained RDCs are compared with five sets of experimental RDC data used in refining the X-ray structure [DW94] to obtain the NMR structure [URDB03] of protein G. Figure 4.3(c) shows that RDCs obtained by PEM and those obtained from NMR in bicelle medium [URDB03] agree with a Pearson correlation of 97% and q-factor of 21%. Naturally, a lower q-factor of 6% is obtained when comparing this NMR RDC data to the RDC-refined NMR structure [URDB03] itself. Comparison of PEM-obtained RDC data with experimental RDCs measured over the other four media [URDB03] reveals agreement with Pearson correlation varying from 94% to 98% and q-factor varying from 18% to 24% (data not shown). A complete comparison of the RDC-refined NMR structure in [URDB03] with each of the five experimentally measured RDCs reveals a q-factor varying from 5% to 7%, with an average of 6%.

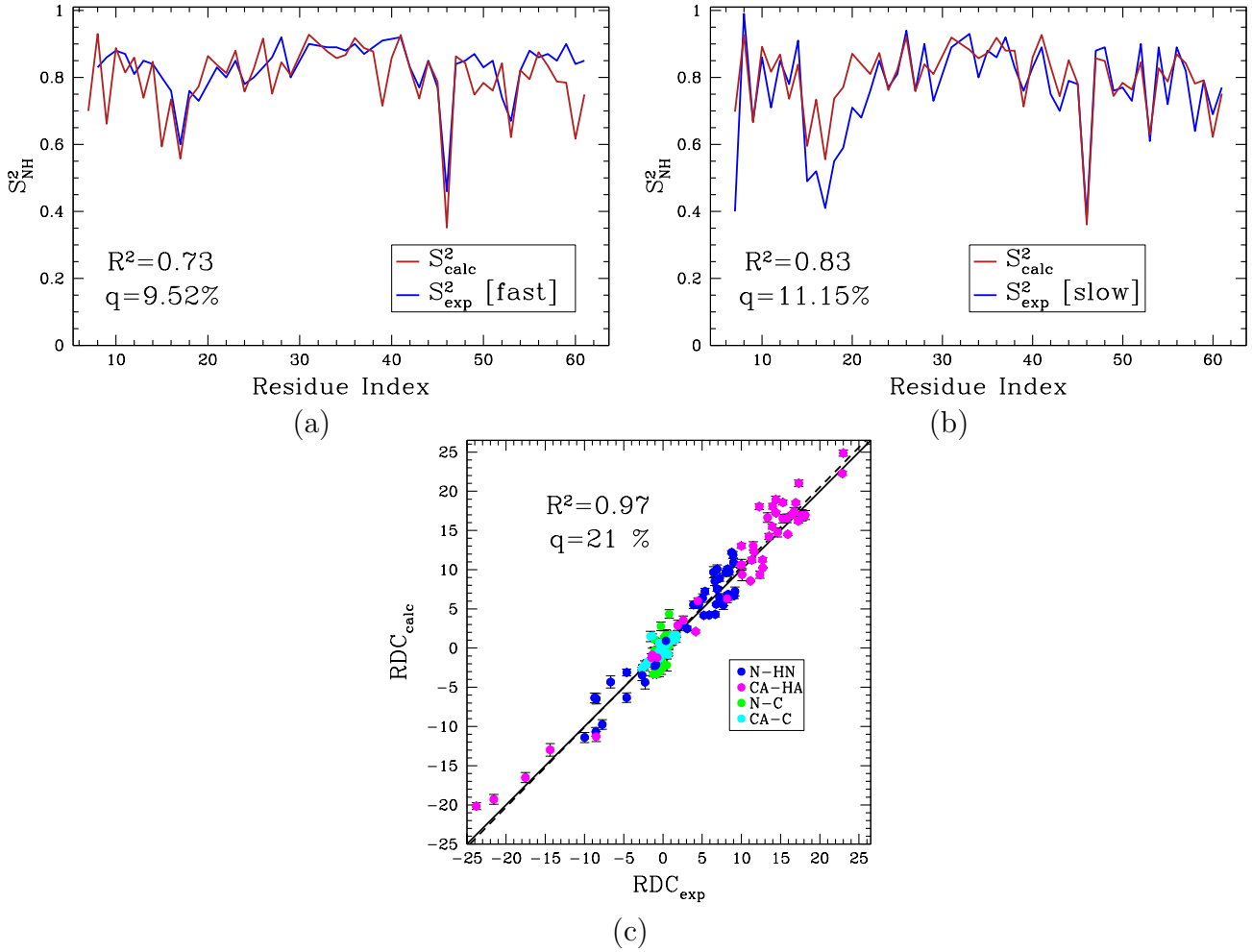


Fig. 4.3: Comparison of NMR data with thermodynamics data obtained by PEM for protein G. (a) Comparison of PEM-obtained S^2 backbone (amide) order parameters (S^2_{calc}) with fast S^2_{NH} data obtained from NMR relaxation measurements (S^2_{exp}). (b) Comparison of PEM-obtained S^2 backbone (amide) order parameters (S^2_{calc}) with slow S^2_{NH} data obtained from NMR relaxation measurements (S^2_{exp}). (c) Comparison of residual dipolar coupling (RDC) parameters obtained by PEM (RDC_{calc}, on the y-axis), and obtained from NMR relaxation experiments (RDC_{exp}, on the x-axis). Results for different bond types are shown in different colors. (a)-(c) The dashed black line indicates the linear least squares regression fit on the two sets of data, while the continuous line represents the identity line.

4.5.3 Validation of Ubiquitin Fluctuations with NMR Measurements

Ubiquitin regulates multiple intracellular pathways in eukaryotic cells [Pic04] and is involved in labeling proteins for proteolysis. Its involvement in protein degradation makes it important for anticancer drug discovery. The availability of abundant NMR data for ubiquitin [CCB03,CMOB98,TFPB95] allows a detailed comparison of PEM-obtained fluctuations on this protein.

Figure 4.4(a) shows the agreement between PEM-obtained and the experimentally available backbone (amide S^2) and side-chain (methyl S^2) data [CCB03, TFPB95]. Figure 4.4(a) shows a Pearson correlation of 96% and indicates that low S^2 order parameters are found not only for residues in the carboxy-terminal region of ubiquitin, residues from $72-76$, but also in residues that form the protein core. Fluctuations of each residue can also be seen as residue lRMSDs obtained by PEM in Figure 4.2(b2).

Figure 4.4(b) shows the agreement between PEM-obtained RDCs and those available from NMR [CMOB98]. The RDCs obtained by PEM agree with NMR RDCs with a Pearson correlation of 97% and q-factor of 23%. The only better agreement with the NMR RDCs comes from the NMR ensemble itself, a Pearson correlation of 99% and q-factor 14%, which is not a surprise since the NMR ensemble in [CMOB98] is derived from the NMR RDCs [CMOB98].

Due to the availability of NMR ${}^3J$ data for ubiquitin [CMOB98], comparisons are also provided between PEM-obtained ${}^3J$ and NMR data [CMOB98]. Figure 4.4(c) shows the agreement between the NMR and the PEM-obtained ${}^3J_{NC_\gamma}$ and ${}^3J_{CC_\gamma}$,

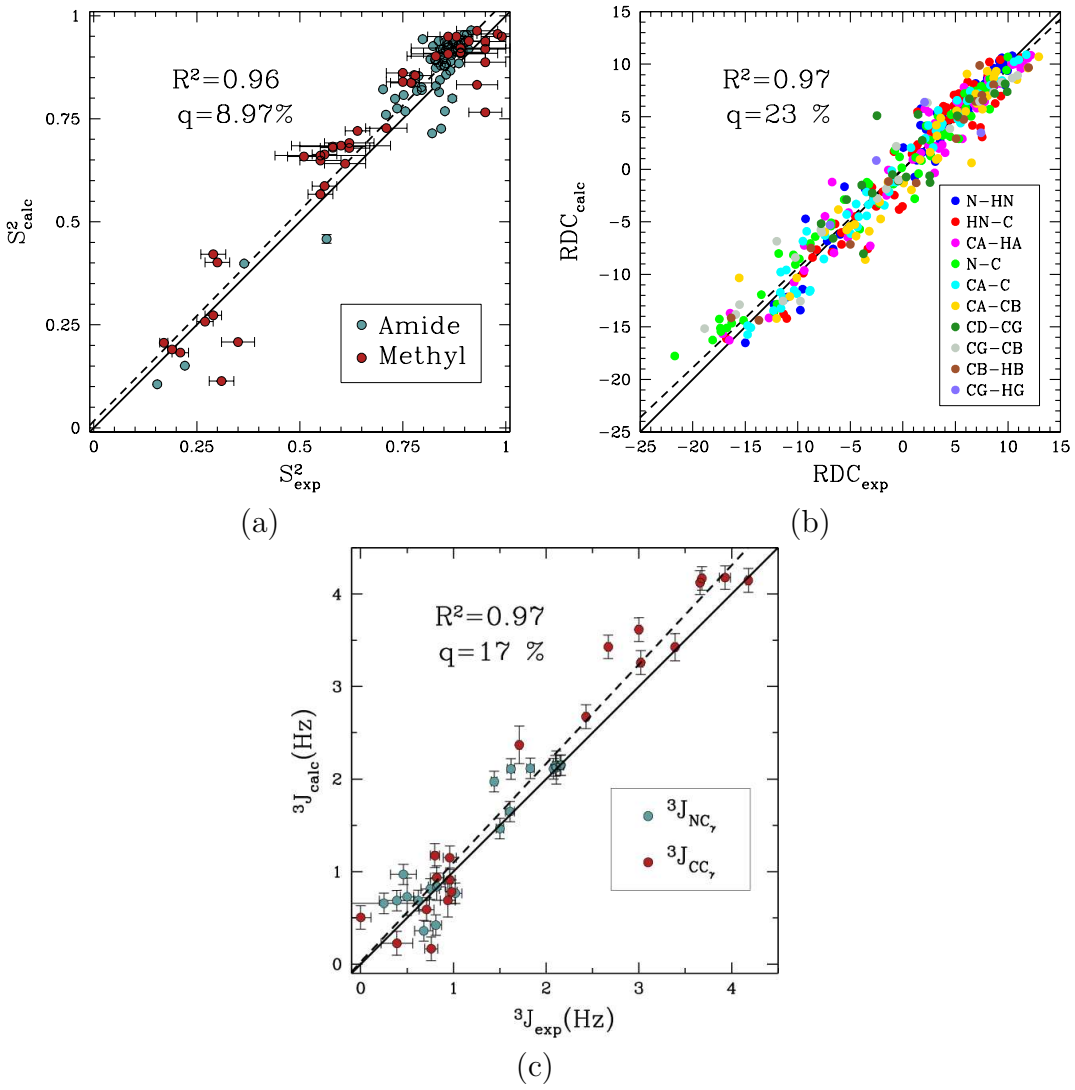


Fig. 4.4: (a) Comparison of PEM-obtained S^2 order parameters for backbone (amide S^2) and side chains (methyl S^2), (S^2_{calc}, on the y-axis), with NMR relaxation measurements (S^2_{exp}, on the x-axis). (b) Comparison of PEM-obtained residual dipolar coupling (RDC) parameters (RDC_{calc}, on the y-axis), with NMR relaxation measurements (RDC_{exp}, on the x-axis). Different bond types are shown in different colors. (c) Comparison of PEM-obtained 3-bond scalar coupling parameters ${}^3J_{NC_\gamma}$ and ${}^3J_{CC_\gamma}$ (${}^3J_{calc}$, on the y-axis) with NMR relaxation experiments (${}^3J_{exp}$, on the x-axis). (a)-(c) Dashed black line indicates linear least squares regression fit on the two sets of data, while continuous line represents the identity.

which are the 3-bond scalar couplings between the side-chain gamma carbon and the backbone amide nitrogen and carbonyl carbon, respectively.

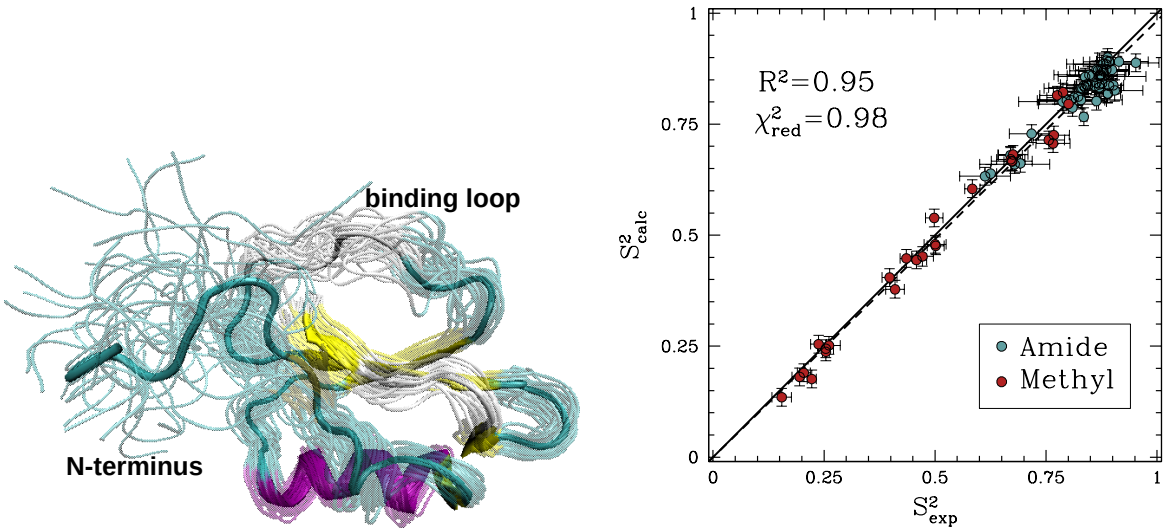
Comparing the ^3J data obtained by PEM with the ones available from NMR reveals a Pearson correlation of 97%, which indicates that the side chains in the conformations generated by PEM populate the right rotameric states. Such a correlation is higher than the 84% and 89% Pearson correlation obtained when comparing the scalar couplings measured on the ubiquitin crystal structure [VKBC87] and NMR ensemble [CMOB98], respectively, with experimental scalar coupling data. This result indicates that the ensemble-averaging of the side-chain dihedrals improves the agreement with experimental scalar coupling data.

4.5.4 Validation of Eglin C Fluctuations with NMR Measurements

Figure 4.5(a) superimposes native conformations obtained by PEM for all consecutive overlapping fragments defined over eglin c. Figure 4.5(a) clearly shows the structural heterogeneity among these conformations. The largest native fluctuations obtained for this protein are located in the Thr1-Gly15 N-terminus, which is practically disordered. Interestingly, the protease-binding loop, encompassing amino acids Ser41-Arg48, is also very mobile. Of all the amino acids of the loop, Val43-Leu47 are the most mobile. The mobility of the loop is also reflected in the low average of 0.7 of the S^2_{calc} data corresponding to the amide bonds of the loop's amino acids.

The entire amide and methyl S^2_{calc} data obtained by PEM for eglin c are shown in Figure 4.5(b). Figure 4.5(b) shows that S^2_{calc} agree with S^2_{exp} data [CL04] with a

Pearson correlation of 95% and reduced χ^2 of 0.98. Methyl S^2_{calc} data, measured by PEM on computed fragment ensembles of eglin c (as detailed in chapter 4), are on average as low as 0.49. This is mostly due to the disordered Thr1-Gly15 N-terminus.



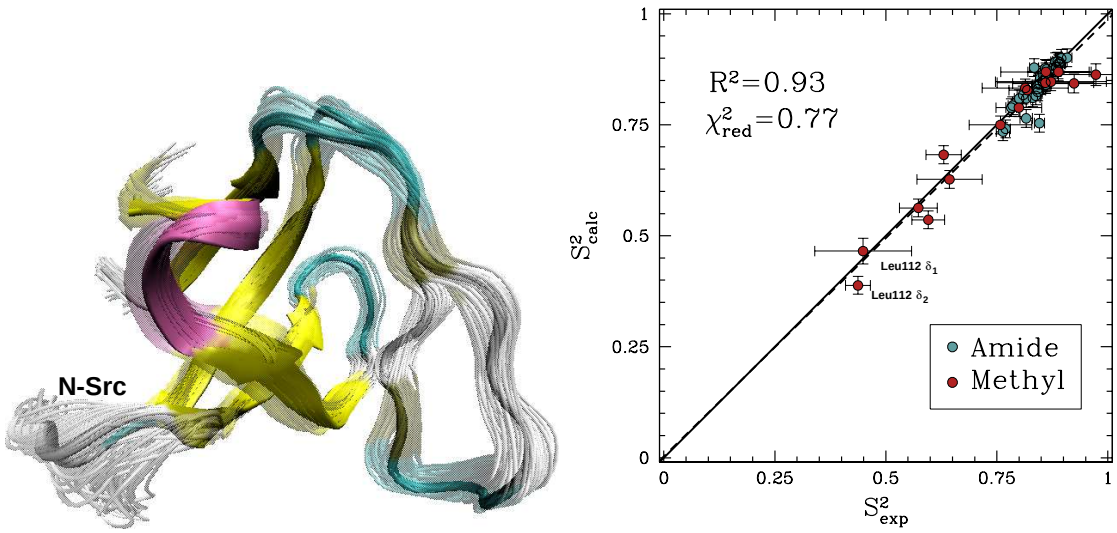
(a) (b)

Fig. 4.5: (a) Eglin c conformations with energy no higher than 5 kcal/mol from the equilibrated solution structure, shown in opaque, are drawn in transparent. (b) Calculated amide and methyl S^2 data (S^2_{calc} on the y-axis) are compared to NMR S^2 data (S^2_{exp} on the x-axis). (c) Calculated $^3J_{NC_\gamma}$ and $^3J_{CC_\gamma}$ ($^3J_{calc}$ on the y-axis) are compared to NMR ^3J data ($^3J_{exp}$ on the x-axis). (b)-(c) The dashed black line indicates the linear least squares regression fit on the data sets. The continuous line is the identity line.

4.5.5 Validation of Fyn SH3 Fluctuations with NMR Measurements

PEM-obtained conformations are shown in Figure 4.6(a). Unlike the results obtained for eglin c, Figure 4.6(a) shows that the PEM-obtained native fluctuations for FynSH3 are prevalently small-scale. The largest fluctuations are located in the N-Src loop, which encompasses amino acids Asn113-Trp119. Interestingly, the N-Src loop discriminates between class I and class II ligands binding to FynSH3 [MPBR96]. Of all this loop's amino acids, its central amino acid, Glu116 is the most mobile.

The obtained native fluctuations for FynSH3 are validated by comparing S^2_{calc} data to the corresponding S^2_{exp} NMR data [MK04]. Figure 4.6(b) shows that S^2_{calc} and S^2_{exp} data [MK04] for FynSH3 agree with a Pearson correlation of 93% and reduced χ^2 of 0.77. The small-scale fluctuations qualitatively shown in Figure 4.6(a) are reflected in the S^2_{calc} data: amide and methyl S^2_{calc} data have high averages of 0.84 and 0.72. This result agrees with experimental findings that large amplitude microsecond-millisecond motions are unlikely in the FynSH3 native state [MK04].



(a) (b)

Fig. 4.6: (a) Fyn SH3 conformations with energy no higher than 5 kcal/mol from the equilibrated solution structure, shown in opaque, are drawn in transparent. (b) Calculated amide and methyl S^2 data (S^2_{calc} on the y-axis) are compared to NMR S^2 data (S^2_{exp} on the x-axis). (c) Calculated ${}^3J_{NC_\gamma}$ and ${}^3J_{CC_\gamma}$ (${}^3J_{calc}$ on the y-axis) are compared to NMR ${}^3J$ data (${}^3J_{exp}$ on the x-axis). (b)-(c) The dashed black line indicates the linear least squares regression fit on the data sets. The continuous line is the identity line.

An interesting instance is represented by amino acid Leu112, located at the border between a β-sheet and the beginning of the N-Src loop. The methyl S^2_{calc} values associated with the χ_1 and χ_2 angles of Leu112 (highlighted in Figure 4.6(b)) are the

lowest in the whole protein, even though the backbone fluctuations at this position are limited. Figure 4.7 shows the distribution of the side-chain χ_1 and χ_2 angles in Leu112 and reveals that the low methyl S^2_{calc} data result from averaging over multiple rotameric states populated by the side chain of Leu112 in the ensemble.

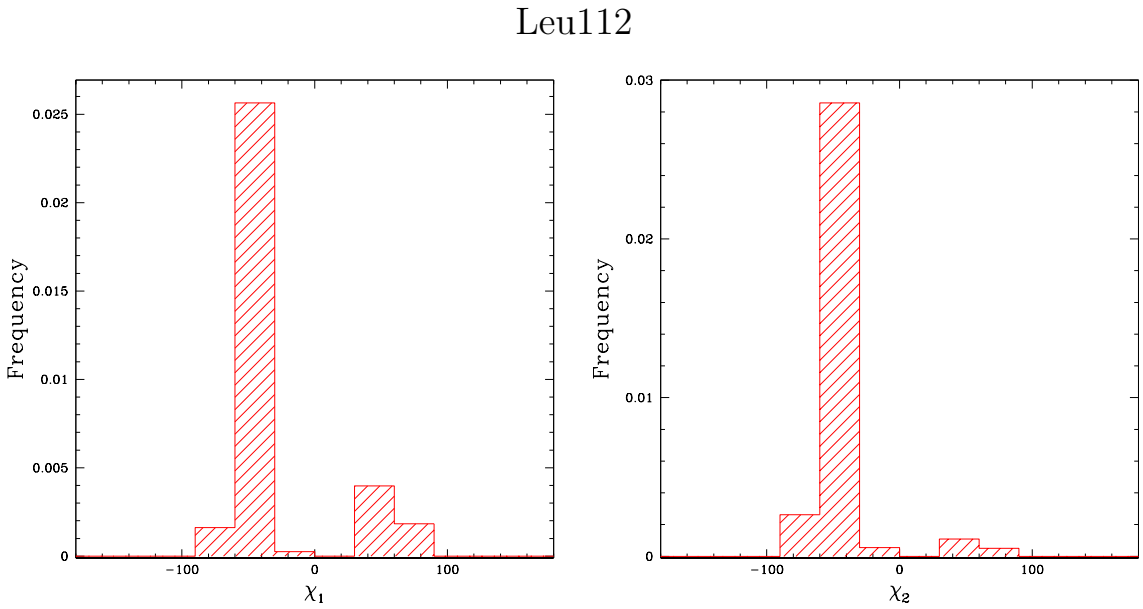


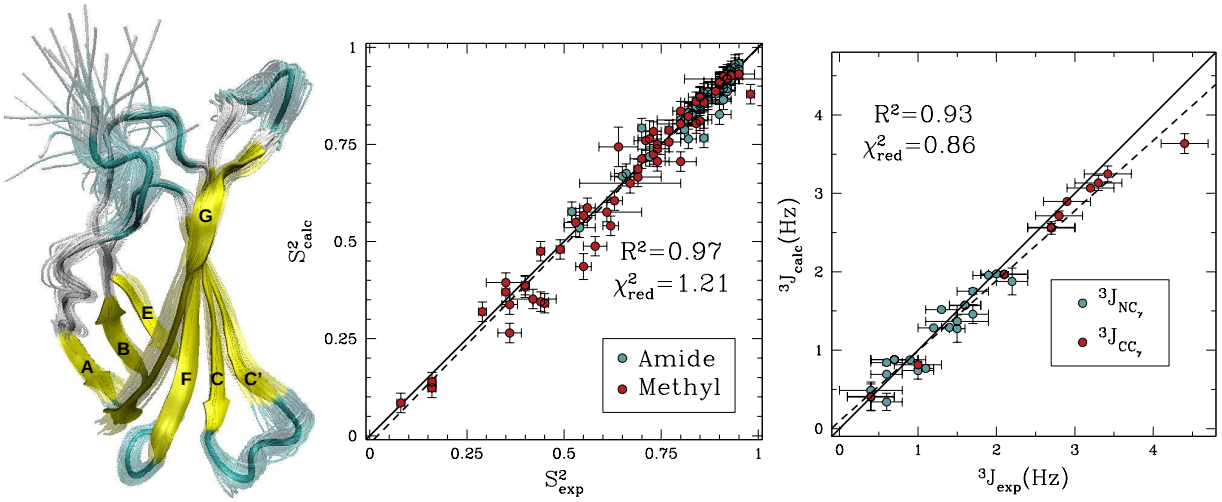
Fig. 4.7: Distributions of χ_1 and χ_2 angles (χ_1 and χ_2 correspond to the dihedral angles associated with the $C_\gamma - C_{\delta_1}$ and the $C_\gamma - C_{\delta_2}$ bonds, respectively) for Leu112 in FynSH3 reveal that Leu112 prefers more than one rotameric state.

4.5.6 Validation of FNfn10 Fluctuations with NMR Measurements

Fragment conformational ensembles obtained for FNfn10 are shown in Figure 4.8(a). The N-terminal amino acids appear disordered, while the 7 β-strands of FNfn10, A, B, C, C', E, F, and G, are well-defined and practically rigid. The surface loops connecting the β-sheets (AB, BC, CC', C'E, EF, and FG), however, are shown to be mobile. The PEM-obtained mobility for these loops agrees with the hypothesis that motions of these loops play a role in the induced-fit recognition of

FNfn10 by multiple receptors [CEP97]. In particular, the most mobile amino acids, Val27, Ser43, and Arg78, are located in the BC, CC', and FG loops. Interestingly, the FG loop, which includes the RGD cell-adhesion motif, encompassing amino acids Arg78-Asp80 [CEP97], is the most flexible of all the surface loops in FNfn10.

Figure 4.8(b)-(c) show that S^2_{calc} and ${}^3J_{\text{calc}}$ for FNfn10 agree with S^2_{exp} and ${}^3J_{\text{exp}}$ data [BRFC04] with Pearson correlations of 97%, 93% and reduced χ^2s of 1.21, 0.86, respectively. Amide S^2 data with a high average of 0.86 indicate small-scale fluctuations and a practically rigid hydrophobic core. This result agrees with the findings in [CEP97] where microsecond-millisecond motions in FNfn10 are not observed.



(a) (b) (c)

Fig. 4.8: (a) FNfn10 conformations with energy no higher than 5 kcal/mol from the equilibrated solution structure, shown in opaque, are drawn in transparent. (b) Calculated amide and methyl S^2 data (S^2_{calc} on the y-axis) are compared to NMR S^2 data (S^2_{exp} on the x-axis). (c) Calculated ${}^3J_{NC_\gamma}$ and ${}^3J_{CC_\gamma}$ (${}^3J_{calc}$ on the y-axis) are compared to NMR ${}^3J$ data (${}^3J_{exp}$ on the x-axis). (b)-(c) Dashed black line is the linear least squares fit on the data sets. Continuous line is the identity line.

While most side chains have a single staggered rotamer, Val4, Val11, and Val50 have unusually low ${}^3J$ values, indicative of rotamer averaging. Distributions of side-

chain γ_1 and γ_2 angles in these amino acids are measured over native conformations of FNfn10 and shown in Figure 4.9. Figure 4.9 confirms that Val4, Val11, and Val50, while preferring one rotamer, are found on average in 4-5 rotamers.

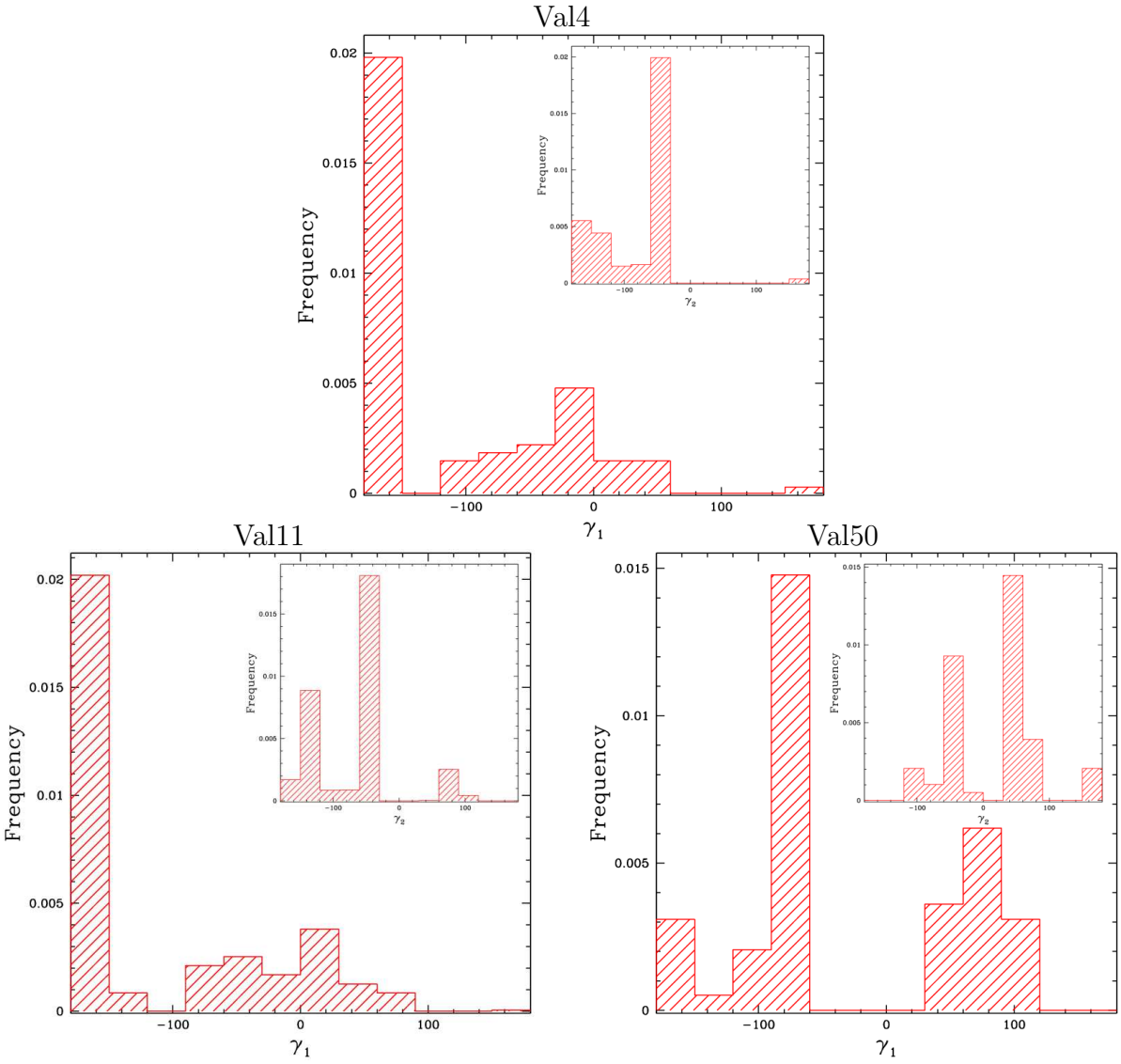


Fig. 4.9: Distributions of γ_1 and γ_2 angles for Val4, Val11, and Val50 in FNfn10 reveal that these amino acids visit on average 4-5 other rotamers. Distributions of γ_2 angles are shown inside. Averaging over rotameric states explains these amino acids' unusually low ^3J data, even though small-scale backbone fluctuations are detected.

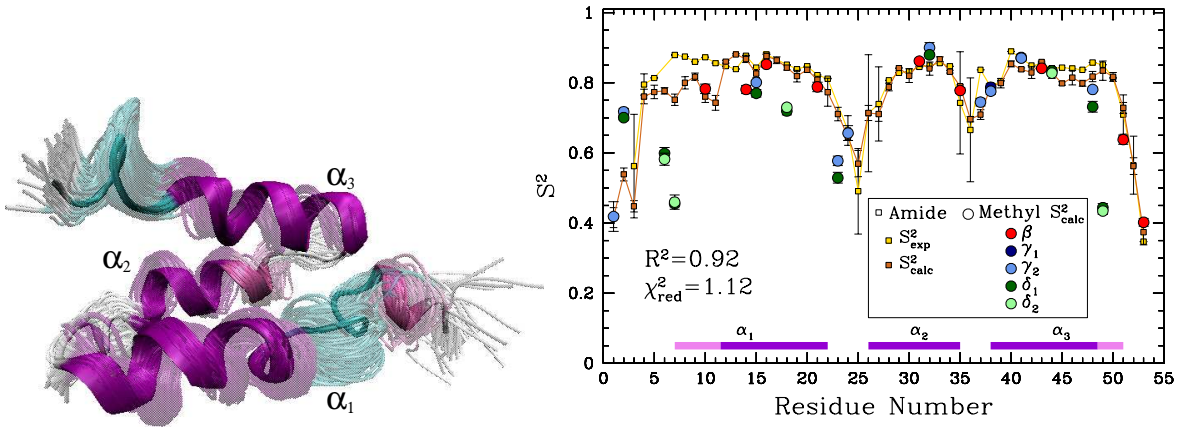
4.5.7 Validation of ALB8-GA Fluctuations with NMR Measurements

The fragment conformational ensembles obtained by PEM for ALB8-GA are shown in Figure 4.10(a). Figure 4.10(b) plots the amide and methyl S^2_{calc} data measured by PEM. Amide S^2_{calc} and S^2_{exp} data [JNE$^+$02] for ALB8-GA agree with a Pearson correlation of 92% and reduced χ^2 of 1.12. Since NMR methyl S^2 data are currently not available for comparison, in Figure 4.10(b) we show predictions of methyl S^2 data as obtained by PEM.

The ensemble drawn in Figure 4.10(a) shows that the second α-helix of ALB8-GA, α_2, is tightly packed between the other two helices, α_1 and α_3. Figure 4.10(b) shows that obtained backbone fluctuations of α_2 are small (amide S^2 data > 0.8). This result supports the loss of conformational flexibility resulting from selective pressure on α_2, which has evolved to bind human serum albumin with high affinity [JNE$^+$02].

In contrast, disorder is observed in the N-terminus of α_1. Amino acids Leu7-Lys11, located at the beginning of the α_1 helix of the solution structure of ALB8-GA [JdCW$^+$97], are found to be highly mobile. These amino acids' high fluctuations can be seen in Figure 4.10(b). Moreover, Leu7-Lys11 can populate both helical and coil configurations. Indeed, while occasionally populating helical configurations in the PEM-obtained ensemble, these amino acids have a high probability to visit unfolded coil-like configurations.

The low helical content of these amino acids predicted by PEM can be seen in Figure 4.11(a). Figure 4.11(a) shows a square symmetric matrix where a blue square

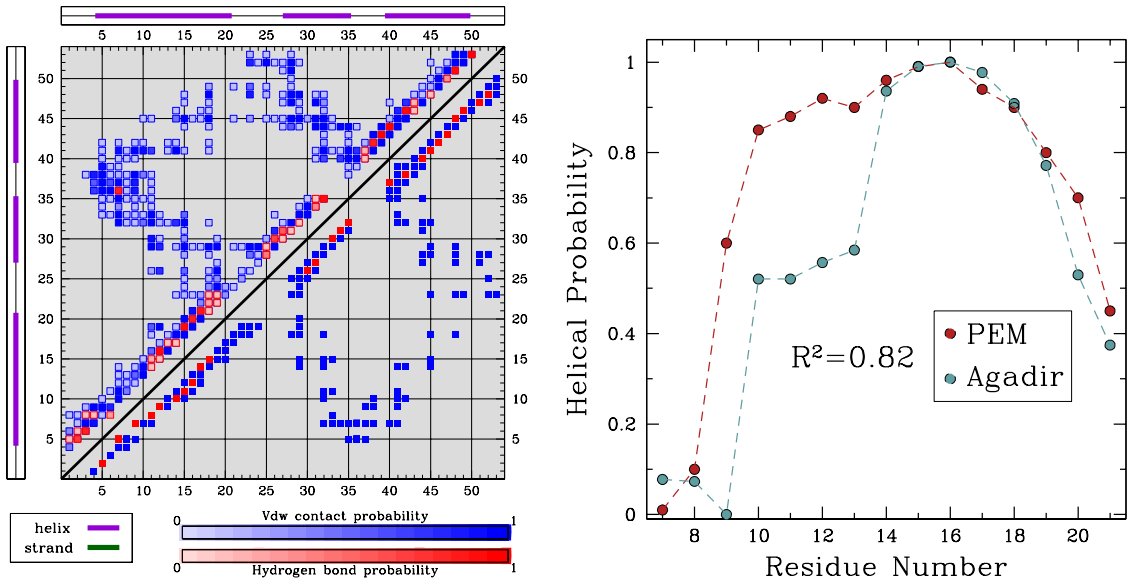


(a) (b)

Fig. 4.10: (a) Conformations with energy no higher than 5 kcal/mol from equilibrated solution structure, shown in opaque, are superimposed in transparent. (b) Calculated amide S^2_{calc} data (orange squares), are compared to NMR S^2_{exp} data (yellow squares). PEM-obtained methyl S^2_{calc} data are shown in colored circles (no NMR data are available for comparison). Horizontal bars on the $x-$axis show the position of the three α-helices in ALB8-GA. The parts of these bars drawn in lighter colors indicate amino acids that are found in unfolded configurations as well.

at position (i, j) indicates the presence of a contact between amino acid i and amino acid j, and a red square indicates the formation of a hydrogen bond between amino acids i and j. Figure 4.11(a) contrasts the contacts and hydrogen bond network as present in the PEM-obtained conformations, shown top left, with the network present in the representative NMR structure of ALB8-GA, shown bottom right. The bottom right half of the map reveals that in the NMR structure hydrogen bonds are present for amino acids Leu7-Lys11 to be in helical configurations. On the other hand, the top left half of the map shows both the scarcity and the low probabilities for hydrogen bonds in this region, indicating that amino acids Leu7-Lys11 visit coil-like configurations in the PEM-obtained conformations with high probability.

The relative populations of helical and coil configurations visited by amino acids



(a) (b)

Fig. 4.11: (a) Formation of a contact between amino acids i, j is indicated with a blue square at position (i, j). Formation of a hydrogen bond is indicated with a red square. Darker shades denote higher formation probabilities. Top left half shows probabilities measured over PEM-obtained conformations. For reference, bottom right shows contacts and hydrogen bonds in representative NMR structure. The hydrogen bonds in the NMR structure indicate that Leu7-Lys11 are in helical configurations. The PEM-obtained map shows either missing or less probable hydrogen bonds in this region, indicating that Leu7-Lys11 visit unfolded configurations. (b) Probabilities for Leu7-Ala21 to be in α_1, measured over PEM-obtained conformations, are in red. Secondary structure is assigned with STRIDE [FA95]. Normalized helicity scores for each amino acid obtained with Agadir [MnS97] are in blue.

Leu7-Lys11 can be quantified by measuring the probabilities of the N-terminus amino acids Leu7-Ala21 to be in helical configurations in the ALB8-GA conformations obtained by PEM. Secondary structure assignment for these amino acids on every generated conformation is computed with STRIDE [FA95]. The measured probabilities are shown in Table 4.1(b). These probabilities have been compared with the helicity scores produced by Agadir [MnS97], a program that predicts the helical behavior of polypeptide chains given only amino acid sequence information. The complete amino

acid sequence of Leu7-Ala21 is shown in Table 4.1(a). The helicity scores predicted by Agadir are shown in Table 4.1(c).

(a)	L	K	N	A	K	E	D	A	I	A	E	L	K	K	A
(b)	0.01	0.10	0.60	0.85	0.88	0.92	0.90	0.96	0.99	1.00	0.94	0.90	0.80	0.70	0.45
(c)	4.7	4.6	3.0	14.4	14.4	15.2	15.8	23.5	24.7	24.9	24.4	22.9	19.9	14.6	11.2

Table 4.1: The ALB8-GA sequence of amino acids 7-21 is shown in (a). The probability of each amino acid to be part of the first α-helix in ALB8-GA as obtained by PEM is shown in (b). Helicity scores predicted for each amino acid by Agadir [MnS97] are shown in (c).

The helicity scores predicted by Agadir agree with the PEM prediction that amino acids Leu7-Lys11 of α_1 have lower probabilities of being found in helical configurations in the native state of ALB8-GA compared to amino acids Lys12-Lys19. This can be seen in Figure 4.11(b), which plots and correlates the probabilities measured over the PEM-obtained conformations with the Agadir-predicted scores. Although the comparison with the Agadir-predicted scores can only be interpreted at a qualitative level (the two data sets measure different quantities), the Pearson correlation with these scores is interestingly high, 82%. This agreement further supports the PEM prediction that these five amino acids (Leu7-Lys11) at the beginning of the α_1 helix have a high probability to visit unfolded configurations under native conditions.

Since helix-to-coil transitions happen on timescales longer than nanoseconds [DMn04], the unfolding observed for amino acids Leu7-Lys11 cannot be detected by the NMR amide S^2_{exp} data [JNE$^+$02]. The conformations obtained by PEM for ALB8-GA may contain additional information to what is present in the available NMR data. It would be interesting to devise wet-lab experiments that can observe native fluctuations of

α_1 over longer timescales. By capturing helix-coil transitions, such experiments could allow to test the PEM prediction of low helical content for Leu7-Lys11.

4.6 PEM: Discussion and Conclusion

The above applications of PEM to proteins of various lengths and folds show that PEM fully characterizes native local fluctuations of small- to medium-size proteins in all-atom detail in good agreement with available NMR data. These results give weight to the conclusion that, as a sampling-based approach with no inherent timescale limitations, PEM complements current simulation techniques in highlighting structural and thermodynamic properties of the native state in proteins with non-concerted motions. In particular, as demonstrated for ALB8-GA, PEM can also complement experimental techniques and formulate hypotheses that can be tested in wet labs.

More than 90% of PEM's computation time is spent in energy minimization for two main reasons. First, the all-atom energy function employed to compute the energy of a conformation is of quadratic complexity in the number of atoms. Second, the 20 kcal/mol cutoff employed for the energetic difference of a computed conformation from the reference energy and the ruggedness of the energy landscape require a high number of minimization steps. This computation cost underscores the need for more efficient energy functions and minimization techniques that still maintain the physicochemical details needed to relate computation and theory with wet-lab experiments.

4.6.1 Effect of Equilibration on Obtained Results

One persistent issue with computational techniques that rely on the equilibration of PDB-obtained structures is the effect of the equilibration on the results. The results presented below make the case that such effect is minimal; for example, the scalar coupling data measured on the X-ray structure of human ubiquitin are virtually the same as the scalar coupling data measured on the equilibrated structure; both quantities correlate with the same Pearson correlation of 84% with the experimentally available scalar coupling data [CCB03].

In addition, the equilibration of the NMR ensemble of 10 human ubiquitin structures [CMOB98] also does not change the 89% Pearson correlation between the experimental scalar coupling data [CCB03] and the calculated scalar coupling data over the minimized ensemble. The correlation between the experimental calculated S^2 order parameters [CCB03, TFPB95] and the calculated S^2 order parameters slightly improves, increasing from a Pearson correlation of 62% for the NMR ensemble to a Pearson correlation of 67% for the equilibrated NMR ensemble. No significant changes are observed upon the equilibration of NMR structures in the correlation with experimental RDCs [CMOB98]. The Pearson correlation remains 99% (the RDCs [CMOB98] are used to refine the NMR ensemble reported in ref. [CMOB98]). Similar minimal effects of the equilibration are observed for the other proteins.

4.6.2 Significance of Agreement with NMR Data

All results presented here have been obtained by using two different force fields: CHARMM22 [MBB$^+$98] and AMBER94 [WCB$^+$94]. These force fields have similar functional form but different parameterization strategies. It has been recently shown that MD simulations with these force fields allow to obtain similar structural and dynamical properties of proteins [PB02]. The results obtained are also found to be essentially independent of the choice of CHARMM22 or AMBER94. The small differences observed in the results obtained with the two force fields are incorporated in the error bars in Figures 4.3(a)-(c) and Figures 4.4(a)-(c). The effect of other approximations used by PEM besides the choice of the force field is also measured as outlined in section 4.4 and incorporated in the error bars.

It is worth stressing the importance of the recovery of RDCs by PEM with both force fields, (shown in Figure 4.3(c) for protein G and Figure 4.4(b) for ubiquitin). While NMR [Kay05] and molecular dynamics [KK05] simulations can characterize local backbone fluctuations in the ps-ns timescale, slower motions in the millisecond-second range, of crucial interest to many functionally important biological processes [TP04, KZ03], are not well understood. Recovering RDC data that report on slow timescale motions, up to the millisecond range, is an important result and confirms the validity of PEM in capturing native flexibility in proteins.

In addition, the correct prediction of NMR data related to side-chain motion, such as the methyl S^2 order parameters (Figure 4.4(a)), and 3-bond scalar cou-

plings ^3J (Figure 4.4(c)), is a significant result. The NMR ensemble available for ubiquitin [CMOB98] correlates with a Pearson correlation of 62% with the experimentally available S^2 order parameters, significantly lower than the Pearson correlation of 96% obtained with PEM. In addition, it has been previously reported that a 6ns MD simulations on ubiquitin performed in explicit solvent and reported in [LLBD$^+$05] cannot capture the heterogeneity of the native state of the protein as given in the experimental S^2 order parameters [CCB03] (the Pearson correlation with the experimental S^2 order parameters is 62%). Another successful effort in recovering the NMR data for human ubiquitin, presented in [LLBD$^+$05] guides replica exchange MD simulations to generate ubiquitin conformations that correlate well with NOE derived distances [CMOB98] and S^2 order parameters [CCB03] and reports Pearson correlations of no lower than 96% with experimental S^2, RDCs, and scalar couplings.

Finally, the recovery of NMR data related to side-chain dynamics, scalar couplings and S^2 order parameters, is an important result since the timescale for the slowest side-chain rotations may be about milliseconds [MB04]. As a consequence, the equilibrium distribution of side-chain conformers cannot be observed directly in MD simulations [PB02]. Since different conformations are generated independently in the ensembles, different low-energy conformers for a given side chain can be sampled even if they are separated by a large barrier, which would hinder the transition from one to the other in MD simulations. Indeed, a closer look at the ensemble of ubiquitin conformations reveals that 88% of the allowed side-chain rotamers are populated,

although some are found with much smaller frequency than others (as expected in the human ubiquitin native ensemble - see [LLBD$^+$05, CCB03]). The successful recovery of these side-chain NMR data by PEM (Figures 4.4(a)-(c)) further corroborates the validity of PEM in characterizing local native fluctuations.

4.6.3 Accuracy of Fluctuations and Higher-order Approximations

PEM is a first-order method that samples conformations of a fragment while the rest of the polypeptide chain is unperturbed. The results presented in section 4.5 indicate that PEM can be used as a framework to test whether local fluctuations are sufficient to explain experimental data. In the absence of experimental data, higher-order approximations are needed to capture concerted motions. In most proteins, such motions involve no more than two fragments of the polypeptide chain at a time [Fer99]. Employing a second-order approximation may therefore be sufficient in detecting the presence of correlated fluctuations in the native state of a protein.

The second-order approximation would involve two windows sliding over a protein chain. All ordered pairs of non-intersecting windows can be easily enumerated, as shown in Figure 4.12 for windows drawn in black and gray. For each ordered pair of non-intersecting windows, conformations would first be obtained for the fragment defined by the black window. These conformations would serve as C_{ref} conformations to obtain conformations for the fragment defined by the gray window. Figure 4.12 considers all ordered pairs of non-intersecting windows since deciding which window to use to obtain C_{ref} conformations may affect the ensemble of final conformations.

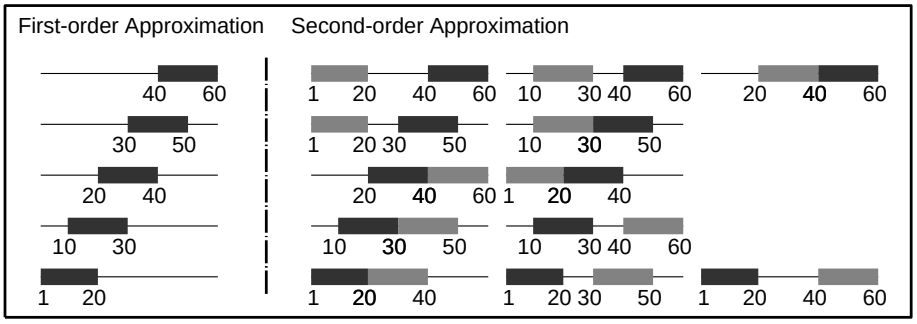


Fig. 4.12: Under the first-order approximation employed by PEM, shown in the left panel, a window slides over a polypeptide chain. This is illustrated by black windows of length $l = 20$ and overlap $\delta l = 10$ on a polypeptide chain of $N = 60$ amino acids. The second-order approximation is shown on the right panel. All possible ordered pairs of non-intersecting windows with length l and overlap δl are considered. In this case, conformations are first obtained by PEM for the fragments defined by the windows drawn in black. With each so-obtained conformation as initial reference structures, final conformations are then obtained by applying PEM to the fragments defined by the windows drawn in gray.

Algorithm 3 Second-order Model(C_{ref}, l, δl)

Input:
C_{ref}: reference protein conformation
l: length of window sliding over polypeptide chain of protein
δl: overlap between consecutive windows
Output: Ensemble of low-energy protein conformations Ω

1: $\Omega \leftarrow \emptyset$
2: $P \leftarrow$ protein polypeptide chain comprising amino acids 1 to N
3: Slide over P a window A of length l with overlap of δl to define fragments $[n_1, n_2]$
4: **for** each fragment $[n_1, n_2]$ defined by A **do**
5: slide over P a window B of of length l with overlap of δl to define fragments $[m_1, m_2]$
6: **if** $[n_1, n_2] \cap [m_1, m_2] == \emptyset$ **then**
7: $C_{[n_1,n_2]} \leftarrow$ low-energy conformation of $[n_1, n_2]$ with rest of P fixed as in C_{ref}
8: **for** each fragment $[m_1, m_2]$ defined by B **do**
9: $C_{[n_1,n_2],[m_1,m_2]} \leftarrow$ low-energy conformation of $[m_1, m_2]$ with rest of P fixed as in $C_{[n_1,n_2]}$
10: $\Omega \leftarrow \Omega \cup C_{[n_1,n_2],[m_1,m_2]}$

Algorithm 3 provides a glimpse on how to obtain native conformations when considering all pairs of fragments over a chain. Average quantities $\langle X_i \rangle$ for each amino acid i can be measured over conformations as in section 4.4.3. However, rather than consider all pairs of fragments, a more efficient and general approach is presented in chapters 5 and 6. The approach captures even concerted motions and employs PEM for a finer sampling of low-energy regions in the conformational space.

Chapter 5

Capturing Native State in Cysteine-rich Cyclic Peptides

This chapter describes a method to characterize the conformational diversity of the native state of cysteine-rich cyclic peptides. The method uses minimal information, namely amino acid sequence and cyclization as a geometric constraint that characterizes the native state. The method does not assume a specific disulfide bond pairing for cysteines and allows the possibility of unpaired cysteines. A detailed view of the conformational space relevant for the native state is obtained through a hierarchic multiscale exploration. Application to three long cyclic peptides of different folds shows that the conformational ensembles and cysteine arrangements associated with free energy minima are fully consistent with experimental data.

5.1 Introduction on Cysteine-rich Cyclic Peptides

Chemical and physical studies on cysteine-rich enzymes led Anfinsen to postulate that the amino-acid sequence encodes for the correct tertiary structure and arrangement of cysteine residues in disulfide bonds in the protein native state [Anf73]. Since

then, experiment, computation, and theory have shown functional relevance both in excursions of a protein from an average experimentally-determined native structure [EML+05,KK05,SDW04] and in rearrangements of cysteines in different disulfide bonds under native conditions (Hogg, 2003). Experimental and computational characterization of the conformational diversity of the native state and the diversity of cysteine arrangements remain active areas of research [CSLS04, HODF98, LLBD+05, PKL01, SCK06].

This chapter proposes a method to characterize the native state of cysteine-rich cyclic peptides. In the following, the method is referred to as NcCYP for Native state characterization of cysteine-rich CYclic Peptides. NcCYP uses minimal information, more specifically (i) amino-acid sequence and (ii) backbone cyclization, to generate low-energy conformations comprising the native state. No a priori assumptions are made about the native disulfide bond pairing of cysteines. Proximity and energetic criteria determine how to feasibly arrange cysteines in each conformation generated, also allowing the possibility of unpaired cysteines.

Related work on existing methods that capture both conformational diversity in the native state and diversity in native disulfide bonds is presented in section 5.2. The NcCYP method is described in detail in section 5.3. Applications of NcCYP are then presented in section 5.4. The chapter concludes with a discussion in section 5.5.

5.2 Related Work on Cysteine-rich Cyclic Peptides

Many methods (mostly based on MD or MC) have been proposed to target cyclic peptides [KPPR00, LGS+03, RSG04] as these peptides' enhanced stability and diverse biological activities are appealing for peptidomimetics and pharmaceutical purposes [Cra06, CCD06, SYTK07]. NcCYP presents two improvements over current computational methods: (i) an efficient exploration of the conformational space allows generating a very large number (hundreds of thousands) of low-energy cyclic conformations in reasonable time; (ii) the diversity of cysteine arrangements is considered. In practice, this is achieved first by obtaining conformations satisfying the geometric constraints imposed by cyclization, then subjecting these conformations to energetic refinement. This two-step procedure is necessary, since geometrical considerations alone do not guarantee low-energy conformations. Such treatment has been shown both general and efficient in generating large ensembles of native conformations for proteins [SCK06, SCK07, SKC07].

As summarized in chapter 2, the high-dimensionality of protein conformational space poses significant demands on computational methods searching for the global minimum corresponding to the native state on the free energy landscape. Current computational methods perform this search using additional information about the native state, in the form of experimental data [LLBD+05] or average native structures [SCK06, SCK07]. NcCYP uses a hierarchic multiscale exploration to efficiently explore the high-dimensional space of conformations relevant for the native state in

cyclic peptides without using additional information of experimental data. While conformational space remains vast (peptides considered here are up to 31 aas long), compared to proteins, the space is more tractable to exploration.

Methods that search for native conformations traditionally do not allow for cysteine rearrangements in different disulfide bonds, even though experiments show that rearrangements may drive function, misfolding, or disease [BM04, Hog03]. The prediction of cysteine arrangements has been addressed with statistical mechanics [FCTS92], optimized threading potentials [DC00], neural networks [FC05, MFC04], or sequence information [MGHT02]. Usually, methods focused on generating conformations in the context of protein structure prediction or protein folding prefer to treat disulfide bonds as fixed constraints during the course of the simulation to reduce the dimensionality of the search space [AS00, SKO97]. Recent attempts to allow cysteine rearrangements during the search for native conformations often result in low-energy conformations with non-native arrangements [CSLS04].

5.3 NcCYP: Hierarchical Multiscale Search for Cyclic Conformations

The NcCYP method proposed in this chapter does not assume a specific disulfide bond pairing between cysteines; neither all cysteines need to be paired for a resulting conformation to be energetically feasible a priori. The method obtains a large ensemble of low-energy conformations populating the native state of a cysteine-rich

cyclic peptide using only (i) the amino-acid sequence and (ii) backbone cyclization as a geometric constraint that characterizes the native state.

Conformations are generated through a multiscale approach. Cyclic backbones are first obtained employing a backbone representation that models only backbone atoms. Each cyclic backbone is converted to an all-atom conformation, which is then energetically refined. A physically realistic all-atom force field in implicit solvent is used to associate an energetically-favorable cysteine arrangement to each conformation (detailed below).

This multiscale approach allows generating a large number of all-atom conformations with distinct cyclic structures and feasible cysteine arrangements. Conformations are clustered according to their cysteine arrangements to reveal low-energy minima associated with different arrangements. Conformations representative of energy minima are used as starting points in the search for new lower-energy conformations. This iterative exploration continues until no lower-energy minima are obtained.

Generated conformations are analyzed through a spatial and energetic analysis. Non-linear dimensionality reduction is employed to reveal global reaction coordinates that structurally distinguish among conformations. These coordinates allow visualizing conformational clusters and associate a free energy landscape to the explored space. Comparing free energies of emerging clusters yields a probability distribution for the possible cysteine arrangements.

Like the FEM and PEM methods presented in chapters 3 and 4, NcCYP initially

uses a backbone representation. This coarse-grained level of detail allows for a fast treatment of the geometric constraint imposed by cyclization on the termini. NcCYP employs CCD to efficiently obtain conformations that satisfy the geometric constraint on the termini. The cyclic backbone conformations are obtained independently of one another. A parallel computation framework, detailed below, is employed to efficiently generate a large number (hundreds of thousands) of cyclic peptide conformations.

After generating a large number of backbone conformations, NcCYP switches to a high-level representation. Each cyclic backbone is converted to an all-atom conformation. A search for optimal side chains is conducted for each backbone as in [HKC07]. The resulting all-atom conformation is refined with a physically realistic all-atom energy function. The AMBER9 ff03 force field [DWC$^+$03] and the implicit Generalized Born (GB) solvation model [STHH90] are used for this purpose. A discussion on the choice of the force field is presented below.

In addition to the all-atom refinement, a feasible cysteine arrangement is assigned to each conformation. Proximity and energetic criteria are used to associate an energetically-favorable cysteine arrangement to each generated conformation. It is worth stressing that a particular disulfide bond pattern is not enforced a priori. In the end, a large number of low-energy all-atom conformations with distinct cyclic backbone structures and feasible cysteine arrangements are obtained.

This initial ensemble provides a broad view of the conformational space. In the following, this initial stage of NcCYP is referred to as Sampling the Equilibrium

Ensemble with Dynamic Disulfide Bond Formation (SEEDD). Because of the large number and diversity of independently generated conformations, SEEDD can overcome the problem of getting trapped in false energy minima. Clustering SEEDD-obtained conformations according to their cysteine arrangements is a natural way to reveal populated conformational states associated with different arrangements.

After this broad view of energy minima, the NcCYP exploration proceeds iteratively. This second stage is referred to as POPulate MINima (POPMIN). POPMIN uses conformations representative of energy minima as starting points from which to structurally guide the search towards new lower-energy conformations, which are in turn clustered as above to reveal even more minima. This continues until convergence, that is, until no new lower-energy minima appear in successive iterations.

Finally, obtained conformations are subjected to a spatial and energetic analysis. By means of non-linear dimensionality reduction, the high-dimensional conformational space of generated conformations is reduced to a low-dimensional space spanned by few coordinates. These coordinates reveal conformational clusters and allow defining a low-dimensional free energy landscape. This analysis, together with SEEDD and POPMIN, are now described in detail.

5.3.1 SEEDD - Obtaining a Broad View of Conformational Space

An all-atom cyclic conformation is generated as follows:

(i) An initial random conformation for the backbone chain is generated first.

(ii) The chain is cyclized by bringing its termini close enough by means of CCD. A peptide bond is then imposed between the termini.

(iii) Energetically feasible side-chain configurations are then added onto the cyclic backbone by following the side-chain reconstruction proposed in [HKC07].

(iv) The resulting all-atom conformation undergoes an energetic refinement and is passed on to step (v) if its potential energy is no higher than 20 kcal/mol of the minimum energy E_{min_ndis} obtained thus far. Else, the search resumes from step (i). E_{min_ndis} is updated if the retained conformation's energy is lower.

(v) Cysteines are arranged in disulfide bonds as described below. If the resulting conformation's potential energy is higher than 20 kcal/mol of a minimum energy E_{min_dis} obtained at this stage, the search resumes from step (i). Otherwise, the conformation is retained and E_{min_dis} is updated accordingly.

Steps (i)-(v) of SEEDD are detailed below:

(i) Generating a Random Backbone Chain

SEEDD generates a random backbone chain by sampling values for the backbone ϕ, ψ angles, keeping bond lengths and angles fixed at equilibrium values. The implementation of SEEDD in this thesis allows sampling values for these angles from different distributions. Two distributions have been thoroughly tested and compared: (i) angle values are sampled uniformly at random in the $[-\pi, \pi]$ interval; (ii) values for the $[\phi, \psi]$ pair of each amino acid are sampled employing Ramachandran

maps [RRS63]. Since scheme (i) does not take into account the amino acid sequence identity and hence does not exploit the stereo-chemical constraints of amino acids as observed in protein structures in the PDB [BWF$^+$00], it is less probable to obtain self-avoiding backbone chains. Hence, the random backbone chains in this work have been obtained through scheme (ii). Ramachandran probability maps are constructed from a non-redundant subset of protein structures([DJC97]) in the PDB. The $[-\pi, \pi] \times [-\pi, \pi]$ map of possible angle values is discretized into $2^\circ \times 2^\circ$ bins. The population of each bin is then normalized to yield a probability distribution for each amino acid, which is then used to sample angle values for the amino acid's ϕ, ψ angles.

(ii) Cyclizing a Random Backbone Chain

The C-terminus is steered towards the N-terminus through CCD, an optimization-based inverse kinematics method described in chapter 3. When the C-termini steered with CCD is within 2 Å of the N-termini, a peptide bond is imposed. A short conjugate gradient minimization of $n = 150$ steps employing a 12-6 Lennard-Jones potential is then used to remove steric clashes and obtain a cyclic self-avoiding backbone chain.

(iii) Adding Energetically Feasible Side-chain Configurations

Given a cyclic self-avoiding backbone chain, side-chain configurations are obtained from backbone-dependent rotamer libraries [XH01] as in [HKC07]. The most energetically feasible side-chain configuration for each amino acid is chosen from entries in

the rotamer library that align best with the amino acid's backbone. Side-chain configurations of each amino acid are iteratively replaced with lower-energy ones until no lower-energy configurations can be found. A short energy minimization of $n = 150$ conjugate gradient steps of all the placed side chains with the backbone kept fixed helps remove steric clashes.

Side-chain Adding Schemes In addition to obtaining side-chain configurations from backbone-dependent rotamer libraries, a second scheme has been tested for comparison: keeping side-chain bond lengths and angles at equilibrium values, values for dihedral angles are sampled uniformly at random in $[-\pi, \pi]$. Comparing previous experience with side-chain addition schemes under FEM and work in [HKC07] experience shows that fewer minimization steps are needed to remove steric clashes and improve side-chain packing when using rotamer libraries as in [HKC07]. Hence, results presented in this chapter have been obtained with the second scheme.

(iv) Energetic Refinement of an All-atom Cyclic Conformation

The resulting all-atom cyclic conformation undergoes a conjugate gradient descent of a maximum of $N = 5000$ steps, employing the AMBER9 ff03 force field [DWC$^+$03] and the GB implicit solvation model [STHH90]. The minimization terminates early (i.e., converges) if there is no more than an $\eta = 2.0$ kcal/mol improvement over $k = 300$ consecutive steps. The resulting conformation is passed on to the next step if its potential energy is no higher than 20 kcal/mol of a current minimum energy

recorded in E_{min_ndis}. The conjugate gradient descent algorithm employed in this work is available as part of the AMBER9 package [CDC+06].

Role of Force Fields and Solvation Models Applications of NcCYP when different force fields are employed show that the AMBER9 ff03 force field and the implicit GB solvation model allow correctly locating the native basin for the peptides presented here. Available literature confirms that ff03 and the GB model yield reliable results [HAO+06]. This finding is also supported by work in [BB00] where the overall shape and location of the native basin are found to be less sensitive to whether implicit GB or explicit solvation models are used.

(v) Formation of Disulfide Bonds

NcCYP does not enforce a specific disulfide bond pairing between cysteines. Both proximity and energetic criteria are used to find a feasible cysteine arrangement for each conformation. First, cysteines closer than a certain threshold are identified. Disulfide bonds are then formed between cysteine pairs and optimized through a short energy minimization. These bonds are allowed to break and form iteratively until convergence. Even if specific cysteine pairings were enforced a priori, this would not significantly reduce the dimensionality of the conformational space. On the other hand, considering all possible disulfide bond patterns is not practical.

The number of ways to pair all given 2k cysteines in k disulfide bonds is $\frac{\prod_{i=0}^{k-1}\binom{2k-2i}{2}}{k!}$ (e.g., 15 pairings for 6 cysteines). Requiring that all cysteines be paired imposes a

priori information on the actual native state. On the other hand, enumerating all possible cysteine arrangements, allowing for unpaired cysteines, is also not practical, as the number of arrangements is $1 + \sum_{j=1}^{k} \frac{\prod_{i=0} j-1 \binom{2k-2i}{2}}{j!}$ (e.g., 76 arrangements for 6 cysteines). For these reasons, NcCYP avoids explicit enumeration.

One way to limit the number of possible cysteine arrangements is to focus only on those that maximize sequence separation between the cysteines involved in disulfide bonds. Based on a statistical observation [ZLT+05] over the PDB that cysteines involved in disulfide bonds maximize sequence separation, a protocol was designed to enumerate only arrangements that maximize sequence separation. Experimental evidence, however, shows that cysteines close in sequence can indeed form disulfide bonds [AACV06]. Hence, the final protocol decided upon in the implementation of NcCYP in this thesis relies only on proximity and energetic considerations.

The protocol used in NcCYP proceeds iteratively, at each iteration determining the closest cysteine pairs, forming disulfide bonds between the pairs according to a proximity threshold that gets increasingly stricter over the iterations, and finally optimizing the length of imposed disulfide bonds through a short conjugate gradient minimization of $t = 1000$ steps (the convergence criterion in step (iv) can terminate the minimization earlier). All disulfide bonds are broken between iterations and the resulting conformation is subjected to the same short minimization.

The number of iterations is limited to two for computational efficiency. The proximity threshold gets increasingly stricter over the iterations: initially a disulfide

bond is imposed if either the C_α-C_α or the S-S distance between a closest cysteine pair is less than 6.0 Å . The final iteration imposes a disulfide bond only if the S-S distance between a closest cysteine pair is no more than 4.0 Å (twice the equilibrium 2.0 Å length). In this way, disulfide bonds are maintained only if paired-up cysteines remain spatially close even when broken and subjected to minimization between iterations.

It is important to point out that the protocol allows for unpaired cysteines if these do not meet the stricter proximity thresholds. The resulting conformation is retained only if its potential energy is no higher than 20 kcal/mol of a minimum potential energy recorded in E_{min_dis}. If a conformation meets this energetic criterion, it is added to the collection of conformations obtained by SEEDD and $E_{\text{min\_dis}}$ is updated accordingly. The same force field and solvation model are used to determine the feasibility of a cysteine arrangement. In particular, no special terms in the AMBER ff03 force field promote formation of disulfide bonds.

A Parallel Computation Framework The lack of dependence between the generation of one conformation from another makes the computation trivial to spread across multiple processors. Parallelization allows sampling a very large number of conformations (the entire computation for the results presented in this paper takes about a week when distributed among 50 CPUS). The minimum energy values E_{min_ndis} and E_{min_dis} are not synchronized among processors so as to obtain a large number of conformations without biasing the search for them to particular regions.

5.3.2 POPMIN - An Iterative Exploration of Energy Minima

SEEDD-obtained conformations are clustered (color-coded) according to cysteine arrangements to reveal those arrangements associated with energy minima. Few (1-2) lowest-energy (with energies $\leq E_{min\_dis} + 5$ kcal/mol) conformations are selected to represent an energy minimum associated with a particular cysteine arrangement. These conformations, deemed *seeds*, are used as reference structures from which to structurally guide POPMIN to generate lower-energy conformations.

POPMIN extends the PEM method presented in chapter 4 in using a seed conformation as a reference structure from which to generate more low-energy all-atom conformations. Like PEM, POPMIN defines consecutive fragments of length $l = 10$ and overlap of $\delta l = 3$ amino acids over a seed's backbone. (e.g., on a cyclic chain of length 18, POPMIN defines fragments [1-10], ..., [11-2], ..., [17-8]). $l = 10$, roughly half the length of the peptides in this work, ensures that large structural fluctuations will be explored around the seed. $\delta l = 3$ ensures consistent fluctuations between neighboring fragments. An ensemble of low-energy conformations is obtained for each fragment, maintaining the rest of the chain fixed as in the reference structure.

Adding onto PEM, obtained conformations are now refined in the state-of-the-art AMBER9 ff03 force field and GB implicit solvation model. Moreover, feasible cysteine arrangements are computed on each generated conformation with the new proximity-based protocol. Generated conformations are retained only if their potential energy is no higher than 20 kcal/mol from the seed's energy value.

When a large number (1000-3000) of conformations are obtained starting from a particular cysteine arrangement (color-code), the obtained conformations are clustered as above to reveal potentially lower-energy minima from which to start the next iteration. If a newly generated conformation has an energy value ≤ 1.0 kcal/mol from that of the seed used to generate it and is further than 1.0 Å lRMSD from the seed, then the conformation is considered as the new representative, and it replaces the seed. When seeds do not change between iterations, i.e., no lower-energy minima emerge, POPMIN is considered converged.

5.3.3 Spatial Analysis of Generated Conformations

The high-dimensional conformational space populated by NcCYP-generated conformations is projected onto a lower-dimensional space through ScIMAP [DMS+06, PSCK07]. ScIMAP is a non-linear dimensionality reduction method for the analysis of non-linear surfaces associated with protein simulation data. ScIMAP uses proximity relations and dimensionality reduction to project a high-dimensional space into a lower-dimensional one that preserves distances between conformations.

ScIMAP computes the nearest-neighbors graph by connecting each conformation to its nearest neighbors. The shortest distance between two conformations is defined as the length of the shortest path that connects them in the nearest-neighbor graph. The shortest-path distances between L conformations selected as landmarks and the remaining conformations are computed and stored in a matrix M. The top eigenvectors of M are used as an orthogonal base set for the low-dimensional projection.

NcCYP applies ScIMAP as in [DMS+06] to obtain a few coordinates that span the space of generated conformations. Different numbers of landmarks (1000-2000) and nearest neighbors (20-30) have been tested to ensure accuracy and robustness of the obtained projections. The so-obtained coordinates naturally reveal conformational clusters in lower-dimensional spaces.

5.3.4 Free Energy Analysis of Conformational Landscape

Free energy values are calculated on the low-dimensional space through a modified version of the weighted histogram method (WHAM) [FS88, FS89]. The modification takes into account that NcCYP-generated conformations are not obtained with a constant temperature constraint; therefore, they do not define a canonical ensemble. Assuming that conformational space is sampled uniformly (as the analysis in chapter 3 seems to indicate), and that the sampling is dense, the low-dimensional space is divided in cells and a density is associated to each cell. The potential energy associated with conformations whose projections fall on a particular cell is averaged to smooth out any inherent noise in the force field or solvation model used. Different cell sizes have been tested to ensure the robustness of the results obtained by NcCYP.

5.3.5 Equilibration in Explicit Solvent

Conformations representative of lowest free energy minima are equilibrated in explicit solvent. Retainment of structural integrity post-equilibration (lRMSD from pre-equilibrated structure is found to be less than 2.0 Å) leads to the conclusion that

conformations predicted by NcCYP as representative of lowest free energy minima are not overly sensitive to whether implicit GB or explicit solvation models are used.

The equilibration, based on [Wal04], is carried out in polyhedral boxes of pre-equilibrated TIP3 water over four stages. In the first three, constant volume periodic boundaries are maintained. Each of the first two stages consist of 1000 steepest descent steps followed by 4000 conjugate gradient descent steps. In the first stage, positional restraints keep the solute atoms fixed. Restraints are removed in the second stage. The third stage consists of 20 ps of an NVE MD simulation that heats the system from 0 to 300 K using Langevin dynamics for temperature regulation. Weak positional restraints are imposed on the solute atoms. The final stage consists of 100 ps of an NPT MD simulation. Constant pressure and density are maintained on average. Other system properties such as temperature, volume, potential, kinetic, and total energy also monitored show full system relaxation.

5.4 Applications on RTD-1, cMII-6, and Kalata B8

Rhesus θ-defensin-1 The first system selected for application is rhesus θ-defensin-1 (RTD-1), a cyclic peptide found in Rhesus macaque leukocytes. As part of the immune system [TYG$^+$99], RTD-1 is microbicidal for bacteria and fungi and three times more potent than its open-chain human analogue [TYG$^+$99]. RTD-1 consists of 18 amino acids and assumes a β-hairpin fold under native conditions [DCR$^+$05, TSC01]. The NMR ensemble (PDB code 2atg) [DCR$^+$05, TSC01] in Figure 5.1(a) shows flexi-

ble turns connecting the β-sheets. The three disulfide bonds in this ensemble are in a ladder arrangement, between cysteines 4-17, 6-15, and 8-13. Figure 5.1(a) also shows the RTD-1 amino acid sequence.

cMII-6 The second system selected for the application of the NcCYP method is cyclized by adding a linker of 6 amino acids to the naturally-occurring α-conotoxin (α-CTX) MII [CFD$^+$05]. Found in the venom of Conus Magus, MII is a potent inhibitor of nicotinic acetylcholine receptors and a potential lead in the design of drugs against Parkinson's disease [QPKM01]. The MII NMR ensemble (PDB code 1MII) and sequence of 16 amino acids sequence are shown in Figure 5.1(b). Cyclization of MII is possible because the 11.2 ± 0.3 Å distance between the N- and C- termini can be easily spanned by a few amino acids [CFD$^+$05]. The sequence of the linker and the resulting cMII-6 sequence of 22 amino acids are shown in Figure 5.1(c), with the cMII-6 NMR ensemble (PDB code 2AJW) [CFD$^+$05] shown in Figure 5.1(d). Figure 5.1(d) shows that, while the linker is highly flexible (as expected from its richness in GLY and ALA), cMII-6 retains both MII's central α-helix and the two disulfide bonds between cysteines 8-14 and 9-22.

Kalata B8 The final system selected is kalata B8, a cyclic peptide found in Oldenlandia affinis. Kalata B8 is a hybrid of the two major Möbius and bracelet subfamilies of the cyclotide family [CCD06]. Like other cyclotides, kalata B8 displays anti-HIV activity. Unlike other cyclotides, the peptide exhibits significant conformational flex-

ibility in the native state while maintaining its cysteine knot motif [CCD06]. Figure 5.1(e) shows the 20 structures of the NMR ensemble of kalata B8 (PDB code 2b38), with the sequence of 31 amino acids superimposed over the ensemble. Figure 5.1(e) also shows the six cysteines paired up in the 1-10, 5-17, and 10-23 arrangement.

5.4.1 Generation of Conformational Ensembles with NcCYP

NcCYP is applied to the RTD-1, cMII-6, and kalata B8 sequences to generate a large number of low-energy all-atom conformations, $534,299$, $518,100$, and 539205, respectively. This number of conformations provides a broad initial view of the conformational space relevant for the native state. Analysis of the associated energy landscape is performed by considering generated conformations with potential energies no higher than 20 kcal/mol from the global minimum potential energies obtained for each peptide. This choice is justified by the fact that conformations with higher energy values have a a negligible Boltzmann probability $\lesssim 10^{-15}$ at room temperature.

The $8,034$, $5,380$, and 4420 RTD-1, cMII-6, and kalata B8 conformations, respectively, that satisfy this energetic criterion are projected on a lower-dimensional space through ScIMAP. The application of ScIMAP to the generated conformations reveals that 2 coordinates are sufficient to capture more than 90% of the structural variability among the conformations obtained for each peptide. The low-dimensional landscapes presented here are obtained with 2000 landmarks, 20 nearest neighbors, using lRMSD for nearest neighbor calculations. Free energy calculations on the low-dimensional landscapes highlight the cysteine arrangements that are the most prob-

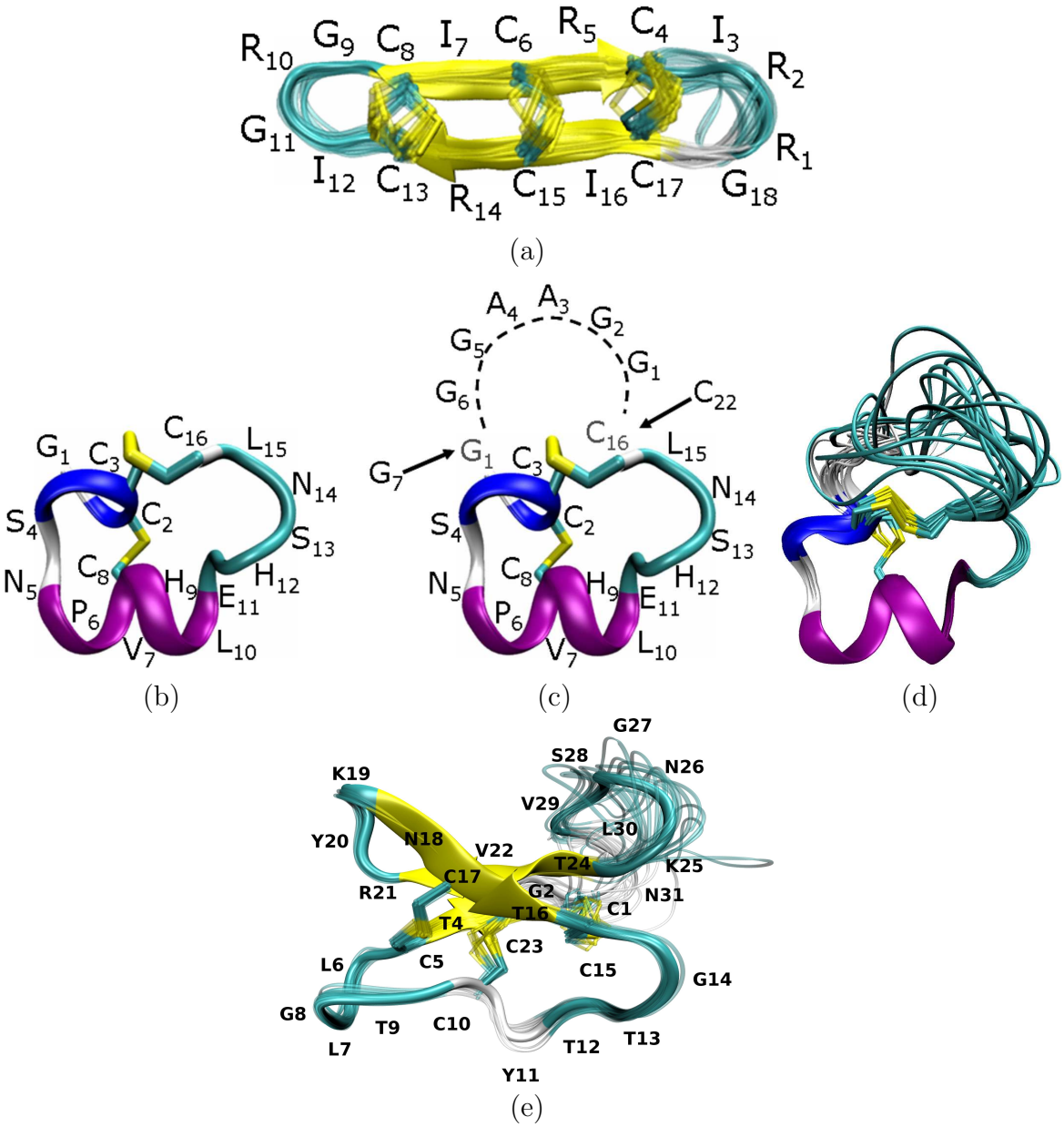


Fig. 5.1: (a)-(e) 20 conformations of NMR ensembles of RTD-1, MII, cMII-6, and kalata B8 are superimposed over one another, with the first of each ensemble drawn thicker for reference. Sequences are shown for each peptide. (c) Dashed line delineates the linker sequence, which shifts amino-acid positions in cMII-6 by 6.

able at room temperature. Finally, conformations associated with lowest free energy minima are subjected to explicit solvent equilibrations to ensure that structural stability is maintained across different solvation models.

Two main results emerge from the analysis detailed below: (i) on both naturally-occurring RTD-1 and kalata B8 peptides, the native cysteine arrangement present in the NMR ensemble is correctly recovered as the lowest free energy minimum by NcCYP; (ii) on the engineered cMII-6 peptide, NcCYP predicts that two distinct cysteine arrangements are populated under native conditions. Interestingly, the conformational ensembles associated with each cysteine arrangement are both consistent with the available NMR ensemble. While one arrangement stabilizes a central α-helix and lowers the flexibility of the linker, the other arrangement offers an entropical compensation by destabilizing the helix and increasing the flexibility of the linker. This result is an example of the capability of the proposed method to complement NMR ensembles by providing additional all-atom detail for the native state.

5.4.2 Analysis of Generated Ensembles of RTD-1

Figure 5.2(a) shows the projections of generated RTD-1 low-energy (no higher than 20 kcal/mol of the global minimum energy obtained) conformations onto the first two coordinates obtained with ScIMAP. Each point in this 2D landscape is color-coded according to the cysteine arrangement in the corresponding conformation as shown in Figure 5.2(a). Cysteine arrangements where not all three disulfide bonds are formed are practically all filtered out when considering only conformations with potential energies no higher than 20 kcal/mol of the global minimum energy obtained for RTD-1. Including conformations with higher potential energies in this 2D landscape reveals abundant cases where not all three disulfide bonds are formed (data not shown).

Figure 5.2(a) reveals two well-separated clusters in blue and yellow. Generated conformations with cysteines arranged in native disulfide bonds as in the NMR ensemble are clustered together in the blue cluster. The yellow cluster, albeit smaller, is associated with conformations with cysteines in the 4-6 8-13 15-17 arrangement.

Free energy values calculated over this landscape are used to color-code the 2D landscape in Figure 5.2(b) according to a red-to-blue color spectrum denoting high-to-low free energy values. The comparison of Figures 5.2(a) and (b) shows that the lowest free energy minimum (labeled A) corresponds to the blue cluster of projections of conformations with the native 4-17 6-15 8-13 cysteine arrangement. The second-lowest free energy minimum (labeled B) corresponds to the yellow cluster of projections of conformations with the non-native 4-6 8-13 15-17 cysteine arrangement.

Figure 5.2(c), plotting free energy values as a function of the first ScIMAP-obtained coordinate, allows to directly compare free energies of the regions labeled A and B in Figure 5.2(b). Figure 5.2(c) shows that the A-labeled region is a global minimum, and the B-labeled region is a local minimum. The free energy difference between these two minima is about 6 kcal/mol ($\sim$ 10 RT units at room temperature). Therefore, NcCYP correctly predicts the 4-17 6-15 8-13 cysteine arrangement present in the NMR ensemble as the native one populated by RTD-1 at equilibrium.

The conformational ensembles corresponding to the two free energy minima are shown in Figure 5.3. The ensemble corresponding to the native minimum is shown in Figure 5.3(a), whereas that corresponding to the alternative, higher energy, minimum

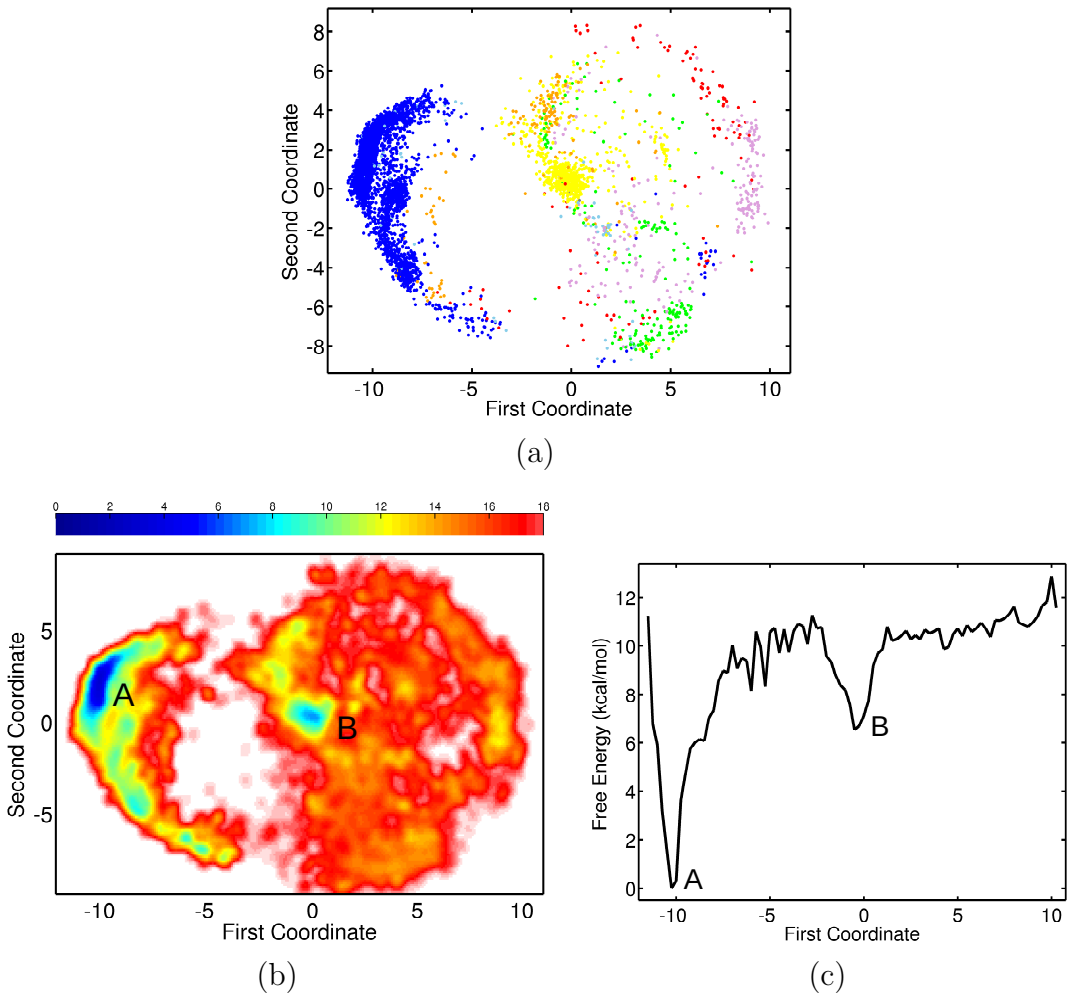


Fig. 5.2: RTD-1 landscape associated with conformations with energies $\leq$ 20 kcal/mol from global minimum. Each point is color-coded with cysteine arrangement in corresponding conformation: blue for the 4-17 6-15 8-13 arrangement observed in NMR ensemble; sky blue for at least one native disulfide bond and rest of the cysteines unpaired; red for 4-17 formed as under native conditions, with rest scrambled; yellow for 4-6 8-13 15-17; green, plum, and orange for remaining all-scrambled arrangements. (b) Red-to-blue spectrum shows high-to-low free energies. Lowest free energy minimum labeled A corresponds to blue (4-17 6-15 8-13) cluster in (a). Second-lowest free energy minimum labeled B corresponds to yellow (4-6 8-13 15-17) cluster in (a). (c) Free energies measured over first coordinate reveal that A is $\leq$ 10 RT units than B.

is shown in Figure 5.3(b). Figure 5.3(a) shows that the conformations associated with the native state have $\beta-$hairpin folds and highly flexible turns connecting well-formed β-sheets. This result confirms the hypothesis in [DCR$^+$05] that the turns connecting

the sheets are highly flexible in this peptide. A comparison between these conformations and the NMR ensemble in Figure 5.1(a) reveals that the native state predicted by NcCYP, though very similar in fold to the NMR ensemble, has higher structural variability. Figure 5.3(b) shows that the non-native conformations corresponding to the higher-energy minimum are structurally similar to one-another, lacking any particular secondary structure, with few of them showing partially formed β-sheets. The structural similarity, considering that these conformations are generated completely independently of one another in NcCYP, suggests that these conformations are not artifacts of the force field or the method but belong to an actual higher free energy minimum. This is supported by the fact that conformations representative of the ensembles associated with the two lowest free energy minima retain their structural integrity when subjected to equilibration in explicit solvent.

The average secondary structure in the two ensembles (a and b in Figure 5.3) can be quantified. A probability value is calculated as a Boltzmann average over amino acid secondary structure assignments (obtained with STRIDE [FA95]) over each conformation of an ensemble. Figure 5.3(c) compares probabilities measured over conformations in Figure 5.3(a), blue line, to those measured over conformations in Figure 5.3(b), yellow line. Figure 5.3(c) clearly shows the presence of β-sheets among conformations with the native 4-17 8-13 15-17 cysteine arrangement and the negligible presence of any secondary structure among the conformations with the non-native 4-6 8-13 15-17 arrangement. These results show that NcCYP accurately

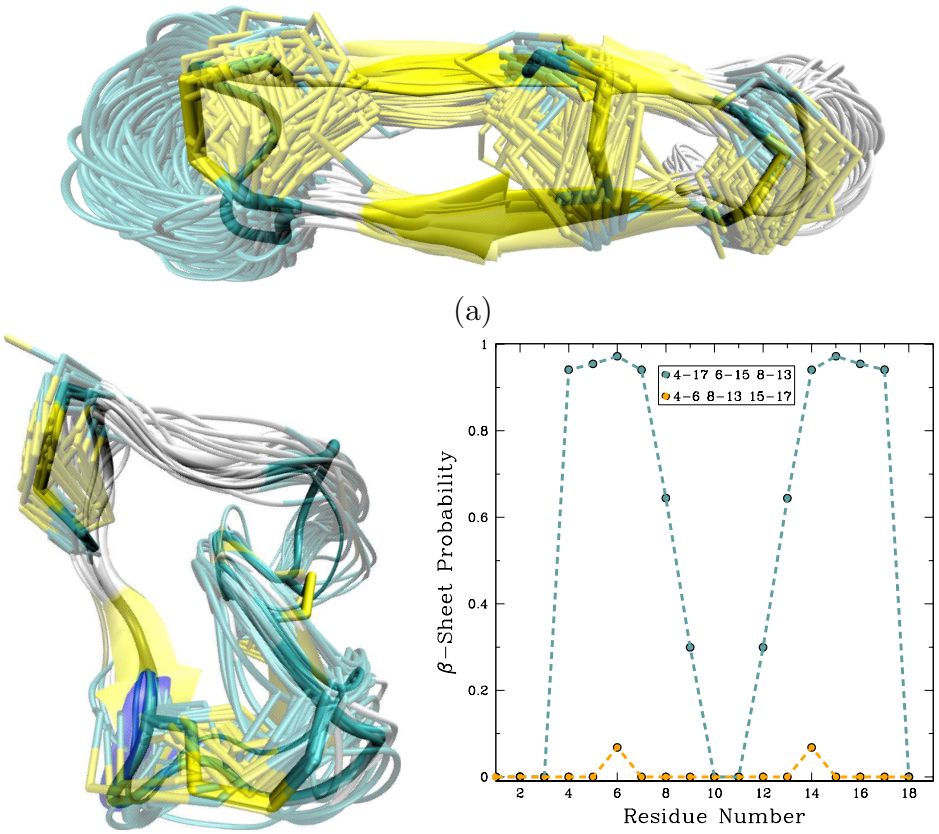


Fig. 5.3: Conformations associated with free energy minima A and B (with free energies ≤ 10 kcal/mol) are respectively shown in (a) and (b), superimposed in transparent over the minimum energy conformation. (c) Blue line, showing secondary structure probabilities for each amino acid over ensemble in (a), reveals well-formed β-sheets. Yellow line, showing probabilities over ensemble in (b), reveals negligible secondary structure.

predicts the exclusive preference of RTD-1 for the 4-17 6-15 8-13 cysteine arrangement under native conditions and the β-sheet fold, consistent with the NMR ensemble.

5.4.3 Analysis of Generated Ensembles of cMII6

As for RTD-1, Figure 5.4(a) shows the 2D landscape obtained with ScIMAP for generated cMII-6 conformations with potential energies no higher than 20 kcal/mol from the global minimum energy obtained. Each point in the landscape is color-coded

according to the cysteine arrangement in the corresponding conformation as shown in Figure 5.4(a). Cysteine arrangements where not all two disulfide bonds are formed are all filtered out when considering only conformations with energies no higher than 20 kcal/mol of the global minimum energy obtained. Cases where not all disulfide bonds are formed become more abundant when considering conformations with higher energies (data not shown). Figure 5.4(a) reveals an abundance of blue- and green-colored projections and a smaller cluster of yellow-colored ones. An interesting separation of conformations emerges around the $x = 0$ line, where x denotes the first coordinate.

Calculated free energy values color-code the 2D landscape in Figure 5.4(b). The red-to-blue color spectrum denotes high-to-low free energies. Comparing Figures 5.4(a) and (b) reveals two free energy minima separated by the $x = 0$ line. These A- and B-labeled minima correspond to green- and blue-colored projections in Figure 5.4(a), respectively. These minima, corresponding to the 8-14 9-22 and the 8-22 9-14 cysteine arrangements, appear equally probable at room temperature.

Free energy values calculated over the first coordinate, shown in Figure 5.4(c), reveal two minima labeled A and B for direct comparison with Figure 5.4(b). The B-labeled minimum associated with the 8-14 9-22 arrangement has a lower free energy value than the A-labeled minimum associated with the 8-22 9-14 arrangement. The free energy difference between these minima, however, is relatively small ($\sim$ 1 RT unit), indicating that both arrangements can be populated at room temperature. An inspection of the NMR ensemble in Figure 5.1(c) shows that, from a purely geomet-

rical consideration, it is not difficult for cMII-6 to accommodate both arrangements.

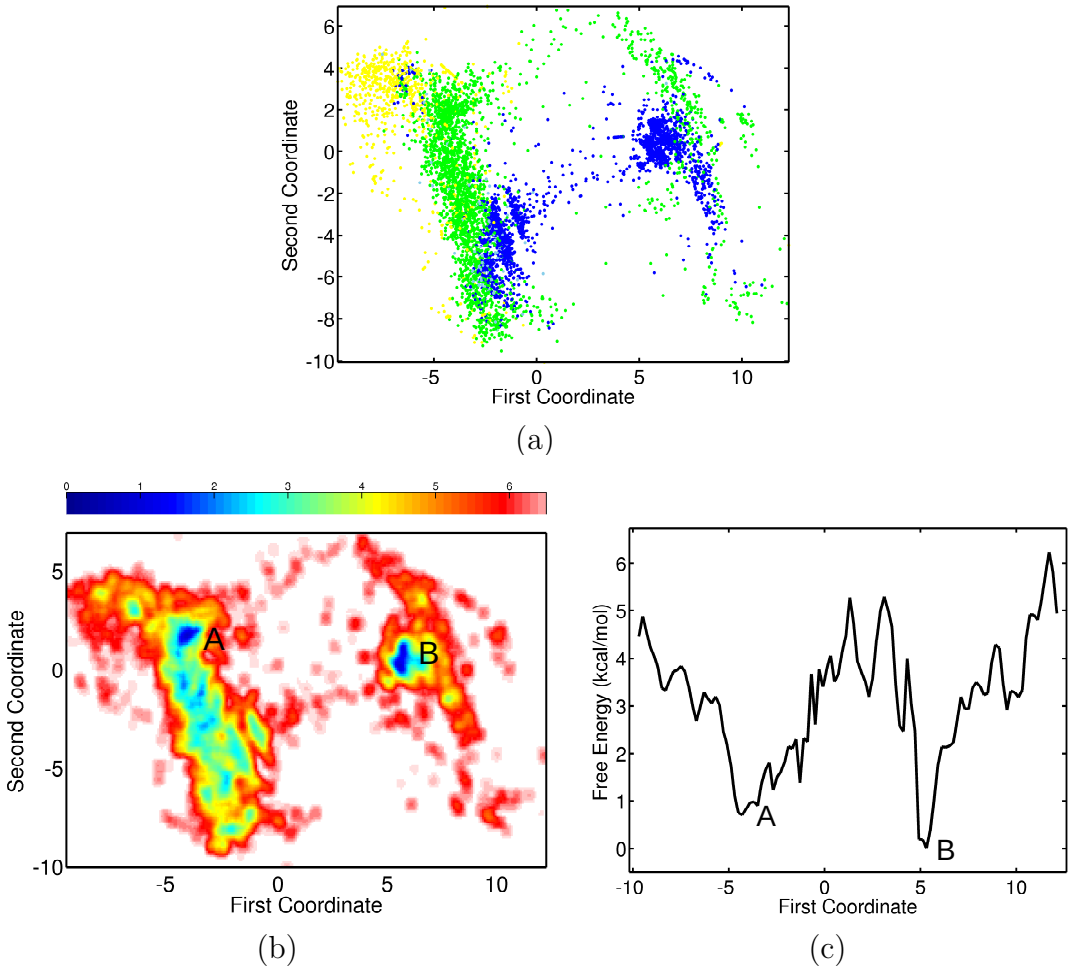


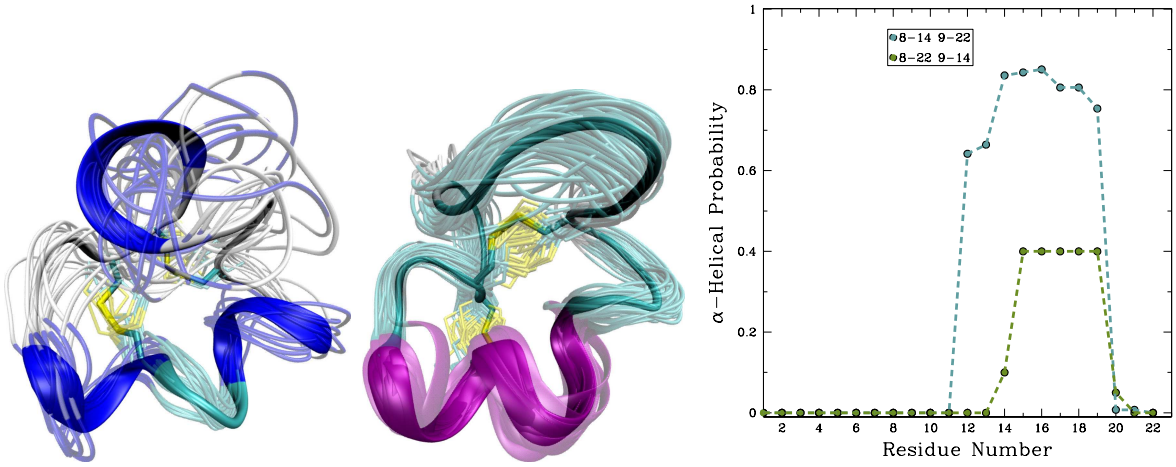
Fig. 5.4: cMII-6 landscape associated with conformations with energies no higher than 20 kcal/mol from the global minimum. Each point is color-coded with the cysteine arrangement in corresponding conformation: blue for the 8-14 9-22 native one in the NMR ensemble; green and yellow for the remaining 8-22 9-14 and 8-9 14-22 arrangements, respectively. (b) Red-to-blue spectrum shows high-to-low free energies. The lowest free energy minima A and B correspond to green (8-22 9-14) and blue (8-14 9-22) projections. (c) Free energies measured over first coordinate reveal that the difference between A and B is about 1 RT.

The separation of the two minima around the $x = 0$ line is structurally meaningful. Inspection of conformations corresponding to the two minima, shown in Figure 5.5(a)-(b), reveal that the separation denotes the formation of an α-helix. Conformations with the 8-22 9-14 cysteine arrangement, shown in Figure 5.5(a), lack well-formed sec-

ondary structure, whereas those associated with the 8-14 9-22 arrangement, shown in Figure 5.5(b), have a central α-helix, similarly to the published NMR ensemble [CFD$^+$05] in Figure 5.1(d). This is quantified in Figure 5.5(c), which compares amino-acid secondary structure probabilities calculated over conformations in Figure 5.5(a), green line, to those calculated over conformations in Figure 5.5(b), blue line. Figure 5.5(c) clearly shows the lack of helical structure in the conformations in Figure 5.5(a). In contrast, Figure 5.5(c) shows high helical probabilities for the central amino acids in the conformations shown in Figure 5.5(b). The structural similarity among conformations in Figure 5.5(a) suggests that these conformations are not an artifact but belong to another low free energy minimum. As for the first peptide, this is also confirmed by the fact that, when subjected to explicit solvent equilibrations, conformations representative of the ensembles associated with the two lowest free energy minima retain their structural integrity.

The ensembles corresponding to the 8-14 9-22 and 8-22 9-14 cysteine arrangements shown respectively in Figure 5.5(a) and (b) are both consistent with the cMII-6 NMR ensemble in cysteine arrangement, shown in Figure 5.1(d). The 8-14 9-22 arrangement stabilizes the central α-helix also present in the NMR ensemble and so lowers the potential energy of associated conformations. On the other hand, the 8-22 9-14 arrangement, while destabilizing to the helix, displays higher linker flexibility as in the NMR ensemble, thus offering an entropic compensation under native conditions.

The small free energy difference between these two cysteine arrangements leads



(a) (b) (c)

Fig. 5.5: Conformations associated with minima A and B (free energies ≤ 7 kcal/mol) are respectively shown in (a) and (b) superimposed in transparent over the minimum energy conformation. (a) There is no distinguishable helical structure among conformations associated with minimum A (8-22 9-14 arrangement) (b) A well-formed central α-helix can be seen in conformations associated with minimum B (8-14 9-22 arrangement). (c) Green line shows secondary structure probabilities for amino acids over ensemble in (a). Blue line shows probabilities over ensemble in (b).

to the hypothesis that both arrangements can be populated under native conditions. This is not surprising, considering that cMII-6 is engineered from a naturally occurring peptide, hence not necessarily highly optimized to uniquely fold to a particular structure. Application on this peptide indicates that, by ranking the feasibility of different native-like conformational substates, the NcCYP method can direct experimental procedures toward further refinement of the native state ensemble.

5.4.4 Analysis of Generated Ensembles of Kalata B8

Figure 5.6(a) shows the 2D landscape obtained for generated kalata B8 conformations with potential energies no higher than 20 kcal/mol from the global minimum energy obtained. Each point in the landscape is color-coded according to the cysteine

arrangement in the corresponding conformation. Figure 5.6(a) reveals two highest populated clusters in blue and red: blue color-codes the native 1-15 5-17 10-23 cysteine arrangement also present in the NMR ensemble; red color-codes arrangements with one native disulfide bond 5-17 and the rest of the cysteines unpaired or scrambled. Free energy values calculated over the landscape are shown in Figure 5.6(b).

Comparing Figures 5.6(a) and (b) shows that the lowest free energy minimum (labeled A) corresponds to the blue cluster of projections. The second-lowest free energy minimum (labeled B) corresponds to the red cluster. The projection of the free energy landscape on the first coordinate is shown in Figure 5.6(c), where the two minima are labeled accordingly. The free energy difference between the two minima is about 4 kcal/mol ($\simeq$ 7 RT). The 1-15 5-17 10-23 cysteine arrangement (the one consistent with the NMR ensemble) is correctly recovered as the native one.

The conformational ensembles corresponding to the two free energy minima A and B are shown in Figures 5.7(a) and (b), respectively. Comparing Figures 5.7(a) and Figure 5.1(e) reveals that the ensemble corresponding to A is very similar to the NMR ensemble but has overall higher structural variability in the loop regions. Interestingly, an α-helix is partially populated in one of the loops in this ensemble. Figure 5.7(b) shows that the non-native conformations corresponding to the higher-energy minimum are also structurally similar to one-another, overall lacking secondary structure, with few showing partially formed β-sheets. Conformations representative of the ensembles associated with the two lowest free energy minima retain

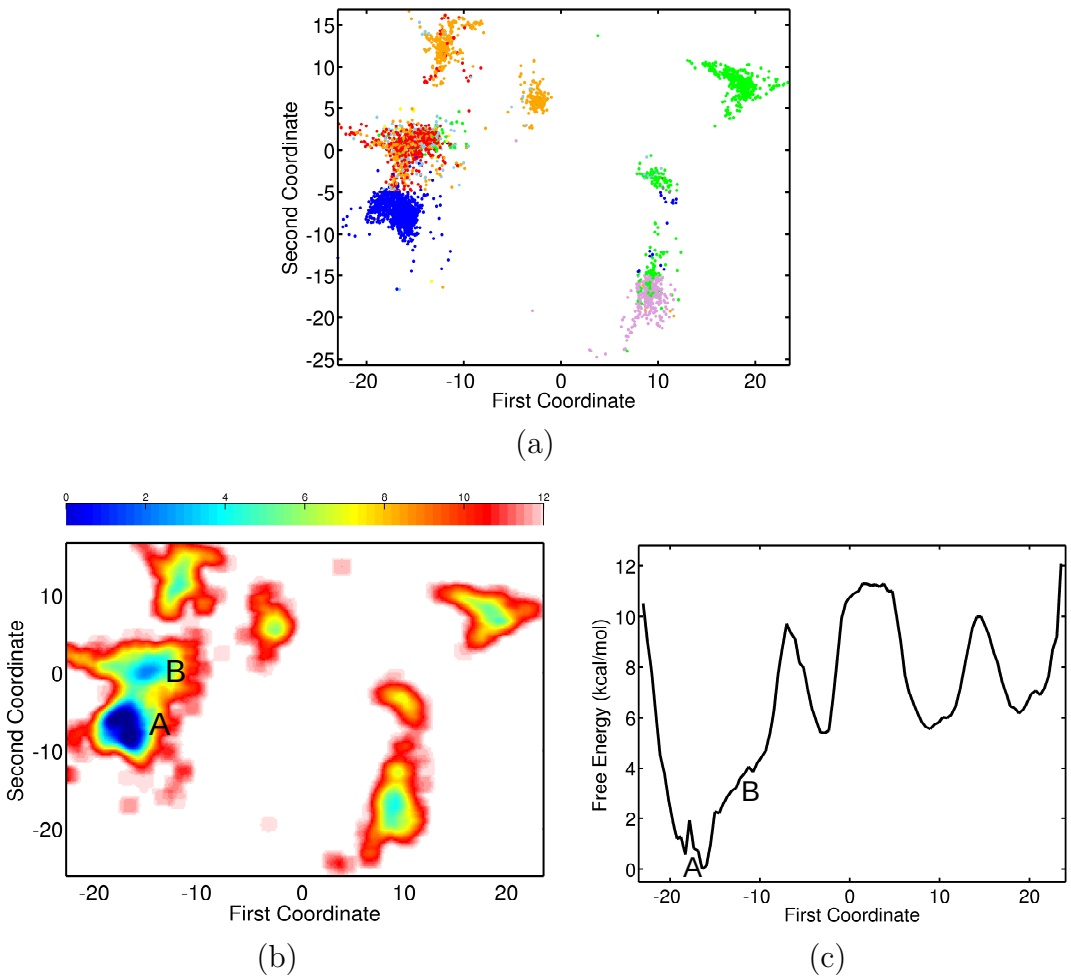


Fig. 5.6: Kalata B8 landscape associated with conformations with energies $\leq$ 20 kcal/mol from global minimum. Each point is color-coded with cysteine arrangement in corresponding conformation: blue for 1-15 5-17 10-23 native arrangement in NMR ensemble; sky blue for at least one native disulfide bond, with the rest unpaired; yellow for 1-15 native bond and the rest scrambled; red for 5-17 native bond and the rest scrambled; plum and orange for remaining all-scrambled arrangements. (b) Red-to-blue spectrum shows high-to-low free energies. Minimum labeled A corresponds to blue (1-10 5-17 10-23) cluster in (a). (c) Free energies measured over first coordinate show that A is 4 RT units lower than B.

their structural integrity when subjected to equilibration in explicit solvent.

Figure 5.7(c) compares the probability of β-sheet formation between the two different minima. Figure 5.7(c) shows the presence of β-sheets among conformations with the native 1-15 5-17 10-23 cysteine arrangement and negligible secondary struc-

ture among conformations with the non-native arrangement. Figure 5.7(d) similarly compares the probability of helix formation in each of the ensembles, showing a partial helix formed with low probability in the lowest free energy minima. These results suggest that, in agreement with the NMR ensemble, NcCYP captures the exclusive preference of kalata B8 for the 1-15 5-17 10-23 cysteine knot motif, the β-sheet fold, and the high flexibility of the loops connecting the β-strands under native conditions.

5.4.5 Additional Application of NcCYP to Enhance NMR Ensembles

An additional application for the NcCYP method is the refinement and enhancement of NMR conformational ensembles. NcCYP can further pursue the exploration of populated low-energy basins. In particular, POPMIN can be applied to refine and enhance a conformational ensemble obtained through NMR. By using the NMR structures as representatives of low-energy basins, i.e., seeds, POPMIN can structurally guide the exploration towards new lower-energy conformations. Using the RTD-1, cMII-6, and 2b38 NMR structures as starting points from where to launch the exploration, POPMIN has been used to enhance the NMR conformational ensembles for each of these peptides. The enhanced ensembles are found to be consistent with the results presented above that were obtained without structural information: the native cysteine arrangement is recovered for RTD-1 and kalata B8; for cMII-6, both arrangements predicted by NcCYP are recovered.

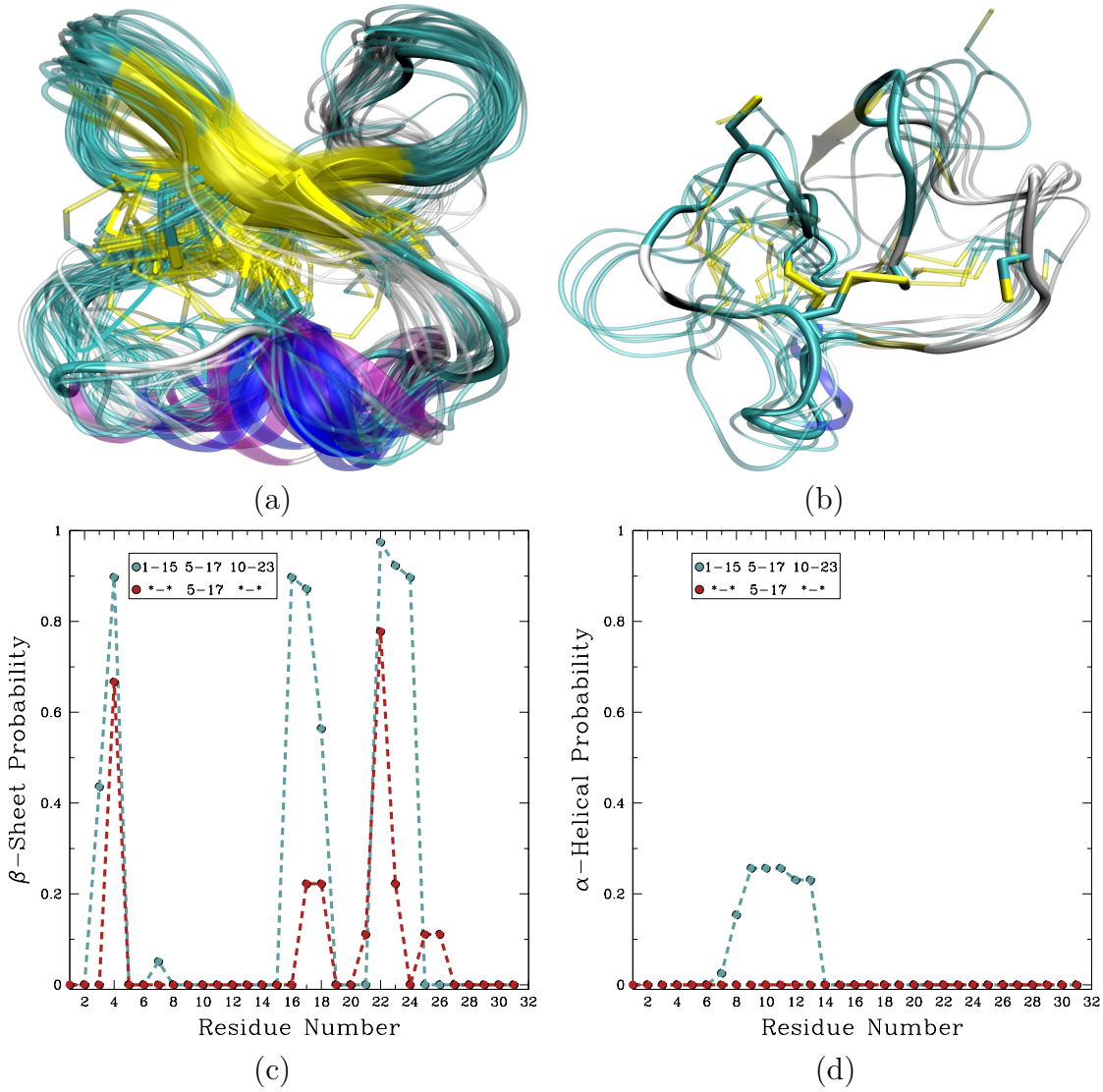


Fig. 5.7: Conformations associated with minima A and B (with free energies ≤ 8 kcal/mol) are respectively shown in (a) and (b) superimposed in transparent over minimum energy conformation. (c) Blue line shows secondary structure probabilities of amino acids over ensemble in (a). Red line shows probabilities obtained over ensemble in (b). (d) Some helicity is observed with low probability in ensemble in (a).

5.5 NcCYP: Discussion and Conclusion

The described NcCYP method predicts native conformational diversity of cysteine-rich cyclic peptides using minimal information, namely, amino-acid sequence and backbone cyclization. The native conformational ensembles obtained by NcCYP for

two naturally-occurring peptides, 18 and 31 aas long are fully consistent with the respective NMR ensembles of these peptides. The native cysteine arrangement is recovered as the global free energy minimum in each case. Application of the method on an engineered sequence of 22 amino acids reveals two equally probable cysteine arrangements. The conformational ensembles associated with each arrangement are consistent with the published NMR ensemble of the peptide.

NcCYP's multiscale hierarchic exploration obtains a detailed view of the large conformational space relevant for the native state. The ability to generate a large number of distinct cyclic conformations, model diversity in cysteine arrangements, and focus the exploration towards increasingly relevant energy minima are key ingredients to the success of the method. The applications presented here show that it is possible to predict a few relevant features of the native state from minimal a priori information in reasonable time.

The main limitation of NcCYP is that it becomes computationally expensive for longer sequences. To overcome this limitation, a coarser level of detail and exploration scheme is proposed in the next chapter to extend application to small and medium-size proteins that exhibit concerted motions under native conditions.

Chapter 6

Extracting Native State from Protein Sequence

This chapter describes a method to characterize the conformational space populated by a protein at equilibrium in two stages: first exploring the entire space at a coarse-grained level of detail, then narrowing a refined exploration to selected low-energy regions. The coarse-grained exploration periodically adds all-atom detail to selected conformations so that the search leads to regions which maintain low energies in all-atom detail. The second stage reconstructs selected low-energy coarse-grained conformations in all-atom detail. A low-dimensional free energy landscape associated with all-atom conformations then focuses the exploration to free energy minima and their conformational ensembles. The lowest free energy ensembles obtained from the application of the method to three different proteins up to 214 aas long correctly capture known functional states of the considered proteins.

6.1 Introduction And Related Work

Work presented in chapters 4-5 builds up to the method presented here. Chapter 4 presented, PEM, a method that was developed to characterize the native state of a

protein where fragments of the chain do not move in concert with one another. This assumption, relevant for a large class of proteins, does not allow for a direct extension to proteins that may exhibit concerted motions under native conditions. NcCYP, described in chapter 5, removes the assumption of non-concerted motions but is able to obtain a detailed view of the native state for short cyclic peptides.

Chapters 4-5 reveal two points: (i) a broad view of conformational space combined with a PEM-like exploration of energy minima can capture diverse equilibrium (native) conformational ensembles that make up the native state; (ii) computational demands ($\simeq$ 1 week on 50 CPUs for cyclic peptides) underscore the need for efficiently computing low-energy conformations when exploring protein conformational space.

The method described in this chapter removes the assumption of non-concerted motions in the native state; reduces the need for a priori information from experimental data to just amino-acid sequence; and extends the hierarchical exploration proposed in NcCYP to longer sequences (up to 214 amino acids). The method is referred to as MuSE for Multiscale Space Exploration.

To efficiently obtain a large set of equilibrium conformations, MuSE proceeds in two stages: first obtaining a broad view of the entire conformational space at a coarse-grained level of detail, then narrowing a refined exploration to selected low-energy regions in the space. In the end, this two-stage exploration yields conformational ensembles associated with different minima in the free energy landscape corresponding to a protein. The lowest free energy ensembles provide good candidates for a protein's

functional states. Indeed, ensembles obtained when applying MuSE to three different proteins correctly capture populated functional states of these proteins at equilibrium.

The coarse-grained exploration in the first stage employs a Monte Carlo Simulated Annealing (MC-SA) scheme to generate conformations. The coarse graining is based on the Associative Memory Hamiltonian with Water (AMW) model [PUE+04]. Within the MC-SA scheme, conformations are generated by assembling backbone fragments from a database of protein structures.

Unlike PEM and NcCYP, MuSE does not sample values for the backbone dihedral angles because of the overwhelming number of such DOFs in protein chains. This issue was side-stepped in PEM by focusing on a subset of these angles in the fragment-based approach, and in NcCYP, which focused on cyclic sequences up to 31 amino-acids long. Instead, MuSE employs the fragment assembly approach, popular in structure prediction methods [BB01, CFT03, GFR05, BMB05, CJS+06], to address the high-dimensionality of the conformational space.

The MC-SA exploration in MuSE is multiscale. All-atom detail is periodically added to few generated coarse-grained conformations. This ensures that the coarse-grained exploration leads to regions which maintain low energies in all-atom detail.

In the second stage, low-energy coarse-grained conformations are systematically reconstructed in all-atom detail. The ScIMAP nonlinear dimensionality reduction technique [DMS+06] is employed to obtain a few global coordinates that span the space of all-atom conformations. The global coordinates help to reveal a low-

dimensional free energy landscape that underlies the all-atom conformations. The conformational ensembles corresponding to the free energy minima in this landscape are enriched with additional low-energy all-atom conformations generated with PEM [SCK06]. Since PEM switches between coarse-grained and all-atom detail to efficiently explore the equilibrium all-atom conformational space around a given reference conformation, the second stage of MuSE is again multiscale.

The MuSE two-stage multiscale exploration produces several all-atom conformational ensembles with associated free energies that allow comparing the relevance of the ensembles at equilibrium. Applications of MuSE to three different proteins show that the obtained ensembles faithfully capture the well-known functional states of the considered proteins. The obtained ensembles provide robust starting points to characterize functional motions, that can be tested by means of further experimental or simulation techniques [OKT$^+$06, ZJZ07].

The rest of this chapter is organized as follows: The MuSE method is described in detail in section 6.2. Applications on three different protein sequences are then presented in section 6.3. A discussion follows in section 6.4.

6.2 MuSE: A Two-stage Multiscale Exploration

The goal of MuSE is to obtain all-atom conformational ensembles accessible to a protein at equilibrium. Exploring the conformational space in all-atom detail is a daunting task even for relatively short proteins ($\sim$ 100 aas). MuSE proceeds in two

stages to efficiently explore such a space.

In the first stage, the exploration gradually focuses from the entire conformational space to low-energy regions. An MC-SA scheme explores the space through many MC simulations at increasingly lower temperatures. Conformations are generated by putting together 3-aa fragments (trimers) of a database compiled over a non-redundant set of protein structures from the PDB. Generated conformations are accepted or rejected according to the Metropolis criterion. A coarse-grained level of detail is maintained, modeling only backbone heavy atoms and side-chain C_β atoms of a protein. Bond lengths and angles are kept fixed in equilibrium values. The MC simulations are launched from "seed" conformations that are carefully selected to guide the exploration to regions of the coarse-grained space that are also low-energy in all-atom detail. The selection involves switching between coarse-grained and all-atom detail on a few conformations. This multiscale exploration yields in the end a large number of low-energy coarse-grained conformations that are suitable starting points for further refined exploration of emerging low-energy regions.

In the second stage, all-atom detail is introduced to low-energy coarse-grained conformations obtained from the first stage. Energies are minimized with an all-atom energy function and implicit solvent model. A low-dimensional free energy landscape is associated with the all-atom space, revealing conformational ensembles associated with free energy minima in the landscape. Regions associated with the minima are then explored in detail by employing PEM, a multiscale method that

switches between different levels of detail to efficiently generate low-energy all-atom conformations. The end result is a large ensemble of all-atom conformations corresponding to the free energy minima. The two stages of MuSE are described next. Implementation details follow.

6.2.1 Stage 1: Exploration of a Coarse-grained Conformational Space

The coarse-grained conformational space of a protein is explored by iterating over the following steps: (i) select coarse-grained conformations (seeds) from which to start the exploration; (ii) from each seed, initiate several MC simulations to generate more coarse-grained conformations; (iii) analyze generated conformations to select seeds for the next iteration. This iterative scheme is the MC-SA introduced above, where the effective MC temperature is lowered after every iteration. At each temperature, several MC simulations are launched from each selected seed.

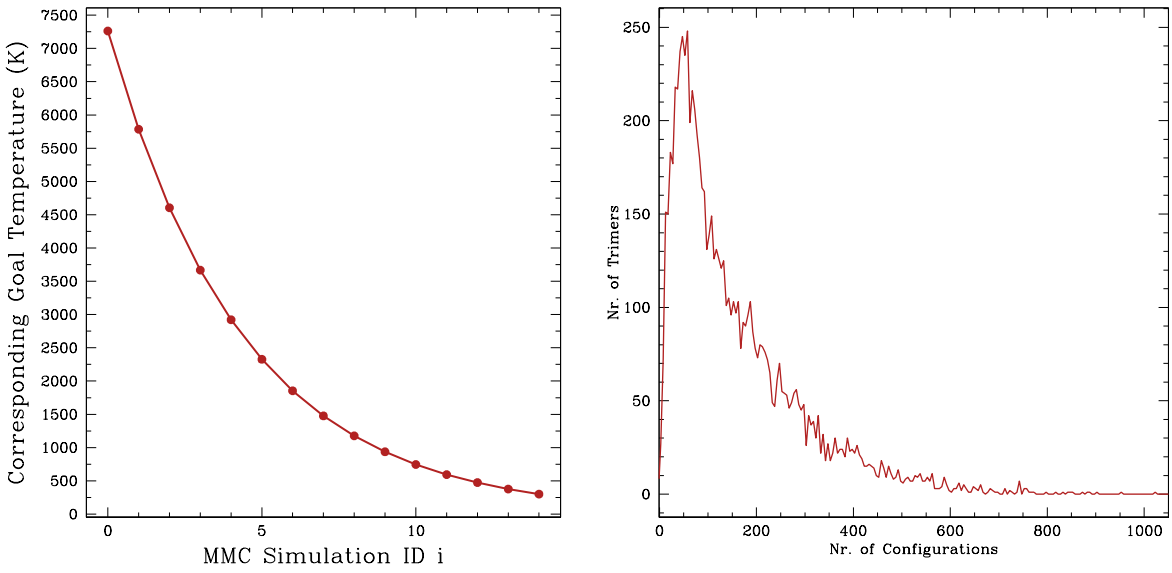
The choice of seeds is critical to the success of the method. At the beginning of the exploration, as no information is yet available on the coarse-grained conformational space, initial seeds are generated by slightly perturbing fully extended chains of a given amino-acid sequence. During the MC-SA, low-energy coarse-grained conformations generated at the previous (higher) temperature are selected as seeds for the next (lower) temperature. The selection involves adding all-atom detail to a few coarse-grained conformations to pick as seeds among them those conformations that are also low-energy in all-atom detail. Each step of the MC-SA scheme is described next.

Monte Carlo Simulated Annealing With a Coarse-grained Model

The MC-SA in MuSE, based on the original work in [KGV83], gradually lowers the MC temperature from a high value T_0 down to room temperature $T_k = 300$K (equilibrium conditions). At higher temperatures, uphill movements in the underlying energy landscape are accepted with high probability, allowing crossing energy barriers and obtaining a broad view of conformational space. As temperature is lowered, downhill movements become prevalent and focus the search in local minima. The MC-SA implemented here lowers temperature k times according to the following cooling schedule.

The initial temperature T_0 for the MC-SA scheme is determined from a desired acceptance probability associated with generated conformations. Initially, a coarse-grained conformation whose energy is $\delta E = 10$ kcal/mol higher than that of the previously generated conformation is accepted with a probability of 0.5. This acceptance probability corresponds to $e^{-\delta E/(k_{\mathrm{B}} T_0)} = 0.5$, which gives an initial temperature T_0 of ~ 7261 K (k_{B} denotes the Boltzmann constant). The final temperature T_f is set to 300 K. The temperature T_0 is progressively lowered k times according to a proportional cooling schedule that updates the MC temperature as in $T_{i+1} = T_i \cdot \frac{T_f}{T_0}^{\frac{1}{k+1}}$ until $T_k = T_f$. Temperatures for each $0 \leq i \leq k$ are shown in Figure 6.1(a).

At each temperature T_i ($0 \leq i \leq k$), n_s seed conformations are chosen. Then, several MC simulations are launched from each seed conformation. The MC simulations launched from a seed conformation differ from one another in how much they confine



(a) (b)

Fig. 6.1: (a) shows the temperatures during the MC-SA. (b) shows through a histogram the population of configurations for trimers in the local database.

conformations they generate around a given radius of gyration (Rg), which is the average distance of atoms from the center of mass. The confinement is enforced through an energetic penalty in the coarse-grained energy function described below. Confining different MC simulations to search inside spheres of a goal radius $\mathrm{Rg}_{\mathrm{goal}}$ allows discretizing the conformational space available to a protein and effectively expediting the exploration.

Low-energy conformations that capture a protein in different functional states may have different radii of gyration. This is certainly the case in the proteins considered here, which assume different functional states through large-scale motions. Therefore, it is not reasonable to bias the exploration to a single and prefixed $\mathrm{Rg}_{\mathrm{goal}}$ value. To allow for large-scale motions, MuSE considers various $\mathrm{Rg}_{\mathrm{goal}}$ values which are

determined before the MC-SA begins. n_s long MC simulations are carried out from slightly perturbed extended conformations at the highest temperature T_0 without any confinement. The distribution of Rg values of the generated conformations is then discretized to determine m $\mathrm{Rg}_{\mathrm{goal}}$ values. These m values are then used in the MC-SA as follows: at each temperature T_i, from each of the n_s seeds, m MC simulations are launched, each one confining generated conformations by one of the m $\mathrm{Rg}_{\mathrm{goal}}$ values.

Traditionally, structure prediction methods bias towards native-like conformations with ideal radii of gyration $R_0 = 2.83 \times N^{0.34}$ [PUE+04, GFR05, PHE+06]. This value, close to that predicted by theory [FR05], biases assembly towards collapsed conformations. MuSE instead aims to capture diverse functional states of proteins: non-collapsed conformations (assumed by proteins such as CaM) should not be discarded if they are energetically feasible. For this reason, each temperature in the MC-SA launches many MC simulations that employ different goal radii of gyration to allow for the possibility of non-collapsed conformations.

Choosing Seed Conformations From Which to Explore

The choice of seeds is crucial to the success of the MC-SA exploration. Seeds for simulations at the highest temperature T_0 are obtained by randomly applying $\leq 2^\circ$ perturbations to the $\phi = -120^\circ, \psi = 120^\circ$ backbone dihedral angles of an extended chain. For each lower temperature T_{i+1} during the MC-SA, seeds are selected as follows.

Conformations generated at a previous temperature T_i are collected in an ensem-

ble Ω_{T_i}. A structural analysis is first performed over Ω_{T_i} to select n_c low-energy conformations that are either obtained very often during the MC-SA or are geometrically distinct. These n_c conformations represent n_c "basins" in the coarse-grained space. These basins are mapped to an all-atom space by adding all-atom detail to the n_c conformations and energetically minimizing them in an all-atom energy function and implicit solvation model. In the end, n_s out of the n_c conformations are chosen that are low in energy in all-atom detail. The all-atom detail is then stripped off the n_s chosen seeds so the MC-SA exploration can continue in the coarse-grained space. Implementation details on the seed selection strategy are related in section 6.2.1.

Metropolis Monte Carlo Simulations To Generate Coarse-grained Conformations

Each MC simulation starts from a seed conformation and lasts for a total of N_{MC} cycles. A cycle consists of N-4 moves, where N is the number of amino acids in a protein chain (there are at most N-4 trimers on such a chain). Each move involves choosing a trimer randomly over the chain. The local database of trimer configurations is then queried with the amino-acid sequence of the trimer. A configuration (6 backbone dihedral angles) is selected randomly over the ones available for the trimer in the database. The selected configuration that is proposed to replace that of the trimer in the current conformation is accepted or rejected according to the Metropolis criterion.

In particular, MuSE compiles and maintains a local fragment database during

the coarse-grained exploration in the first stage. A PDB subset of nonredundant protein structures (as of July 2007) is obtained through the PISCES server [WD03]. Chosen proteins have $\leq 40\%$ sequence similarity, ≤ 2.5 Å resolution if the structure is obtained through X-ray crystallography, or R-factor ≤ 0.2 if obtained through NMR. The $6,056$ protein chains in this subset are split into all possible overlapping fragments of three consecutive amino acids. For each trimer, the local database maintains the list of configurations (6 backbone dihedral angles) populated by the trimer over all protein chains (a total of $10,072,004$ trimer configurations).

Figure 6.1(b) shows that the populations of different trimers are very heterogeneous. However, all possible 20^3 trimers are populated, and only $70/8,000$ have less than 10 configurations. Low populations for this small percentage of trimers are associated with bulky amino acids that are energetically penalized for being neighbors in a protein chain. It is worth pointing out that for all the protein sequences in this study, no trimers have less than 21 configurations in the local database.

A coarse-grained energy function is used to evaluate the energy of the conformation resulting after each move. Since trimer configurations are compiled over PDB structures, local terms are not included in the energy function. The energy is a linear combination of the non-local terms $E_{Lennard-Jones}$, E_{H-Bond}, $E_{contact}$, E_{water}, E_{burial}, E_{Rg}. In particular, $E_{Lennard-Jones}$ is an adaption of the 12-6 Lennard-Jones potential employed in AMBER9 [CDC$^+$06]. The adaptation allows for a soft penetration of vdw spheres. E_{H-Bond} is a hydrogen-bonding term implemented as in [GFR05].

The E_{contact}, E_{water}, and E_{burial} terms, implemented as in the AMW energy function [PUE+04], allow considering water-mediated interactions in coarse-grained conformations. The C_β positions that are needed to evaluate these three terms are computed from the backbone of a conformation as in [MKS97]. The E_{Rg} term implements the energetic penalty $(\text{Rg} - \text{Rg}_{\text{goal}})^2$ if a conformation's radius of gyration Rg is above Rg_{goal}.

An important consideration during each move in an MC simulation is how to choose trimers. In the proposed implementation of the method, each of the $N - 4$ moves in a cycle of an MC simulation in MuSE chooses a trimer randomly over the sequence of N amino acids. If instead, each move proceeds in order down the sequence, it is easy to get stuck trying to find acceptable configurations for the chosen trimer. Randomly picking trimers over the sequence allows getting out of such "local minima."

Determining the length of an MC simulation is also crucial to ensure a broad exploration. Averages in each of the terms of coarse-grained energy values obtained during an MC simulation are observed to determine the duration, number of N_{MC} cycles, of an MC simulation. If the averages converge between windows of w cycles, the simulation terminates. For the proteins here, convergence in averages of energy terms is checked every $w = 500$ cycles and is achieved at $\sim$1000 cycles. Thus, an MC simulation is carried for $N_{\text{MC}} = 2000$ cycles.

Moreover, the acceptance probability (conformations can be accepted or rejected

according to the Metropolis criterion) during an MC simulation may diverge from the goal acceptance probability associated with the goal temperature T_i of the simulation. The effective MC temperature of the simulation is updated every $l = 100$ cycles to improve agreement with the goal temperature. Frequent updates are avoided and no updates are performed after $\text{update}_{\text{stop}} = 500$ cycles.

In particular, every l cycles, the current acceptance ratio $\text{Racc}_{\text{curr}}$ is measured over the number of accepted conformations by the MC simulation. This ratio is compared to the goal acceptance ratio $\text{Racc}_{\text{goal}}$ set out for the simulation. If $(1-\text{tol}) \times \text{Racc}_{\text{goal}} \leq \text{Racc}_{\text{curr}} \leq (1+\text{tol}) \times \text{Racc}_{\text{goal}}$ or the MC simulation has reached its $\text{update}_{\text{stop}}$ cycle, the temperature of the simulation is not updated ($\text{tol} = 0.1$). The reason for avoiding frequent updates is so that the temperature will not be overly sensitive to local findings of the exploration. Moreover, the temperature is actually locked after $\text{update}_{\text{stop}}$ cycles to maintain the detailed balance within an MC simulation. If none of these criteria are met and $\text{Racc}_{\text{curr}}$ falls below (above) the allowed interval, the temperature of the simulation is increased (decreased) by a factor of $\alpha = 0.1$, small enough for a gradual effect.

Seeding Monte Carlo Simulations

After all MC simulations at temperature T_i terminate, their conformations are collected in the ensemble Ω_{T_i}. Conformations with energies no higher than $\langle E_{T_i} \rangle$ averaged over Ω_{T_i} are retained in the ensemble $\Omega^*_{T_i}$. This energetic criterion ensures that conformations selected as seeds for the next MC simulations will come from

low-energy regions in the coarse-grained energy landscape.

A structural analysis is then conducted on $\Omega^*_{T_i}$ to identify "basins" in the coarse-grained energy landscape. Conformations are binned by radii of gyration to yield $t = 11$ sub-ensembles $\Omega^*_{T_i,R_{\text{goal}}}$. A conformation C with Rg(C) is placed in a specific bin Rg_{goal} if $|\text{Rg(C)} - \text{Rg}_{\text{goal}}| \leq \delta\text{Rg}$, where $\delta\text{Rg} = \min\{1.4\text{Å}, \sqrt{\text{Rg}_{\text{goal}}\text{T}_{\text{i+1}}}\}$. This binning limits the expected increase in energy by the confinement penalty at the next temperature T_{i+1} to 1.0 kcal/mol.

Conformations to seed MC simulation at the next temperature T_{i+1} are now chosen from ensemble $\Omega^*_{T_i,R_{\text{goal}}}$. Different strategies can be implemented to choose n_s conformations from the ensemble $\Omega^*_{T_i,R_{\text{goal}}}$: (i) conformations can be chosen randomly; (ii) they can be cluster centroids; (iii) an energetic criterion can be employed; or (iv) a combination of the above.

MuSE uses the following strategy. Conformations in $\Omega^*_{T_i,R_{\text{goal}}}$ are clustered according to lRMSD with the Leader algorithm [JDC87]. Obtained clusters (each limited to a radius of $c_{\text{rad}} = 2.0$ Å) are ordered according to their populations to reveal those with at least $n_{\text{pop}} = 5$ conformations. At most $n_c = 100$ centroid conformations are chosen[1], each one from clusters that meet the population cutoff.

If less than n_c clusters meet this cutoff, the rest of the conformations are chosen from $\Omega^*_{T_i,R_{\text{goal}}}$ in such a way that each selected conformation maximizes its lRMSD from those already selected. This strategy identifies geometrically distinct conformations in the absence of highly-populated ones.

[1]The centroid of a cluster is the lowest-energy conformation populating the cluster.

The resulting n_c conformations now capture distinct "basins" in the coarse-grained landscape. However, due to inherent approximations in the coarse-grained energy function, it is not clear that these basins are also low-energy minima in an all-atom space. Hence, an energetic analysis follows. The analysis first adds atomic detail to the n_c conformations by adding side chains onto each of their backbones as in [HKC07]. The resulting all-atom n_c conformations are then refined with the AMBER ff03 force field [DWC$^+$03] using the GB implicit solvation model [STHH90]. The refinement is a conjugate gradient descent that checks for convergence in energy. The correlation between coarse-grained energies and all-atom energies of conformations after the refinement is computed to associate a score to each conformation. The score ensures that, when correlation is high, $n_s = 5$ conformations whose all-atom energies best match their coarse-grained energies are chosen as seeds. Otherwise, n_s conformations with lowest all-atom energy are selected.

This strategy deals with the issue of error and uncertainty in the coarse-grained energy function that may affect the determination of basins. Since this function uses a coarse-grained representation that integrates out all DOFs besides backbone heavy atoms and side-chain C_β atoms, it is important to regularly estimate whether basins in the coarse-grained landscape remain low-energy regions when adding atomic detail. After n_s conformations are finally selected, the extra DOFs (side chains) are removed to guide the MC simulations at the next temperature back in the coarse-grained space.

6.2.2 Stage 2: Exploration in an All-atom Conformational Space

The first stage allows efficiently sampling a large number of coarse-grained conformations. Conformations obtained during the lowest three temperatures in the MC-SA are considered, and their energy distribution is evaluated. Among them, conformations with energy no higher than one standard deviation from the average energy are selected and used as starting points in the second stage of MuSE to explore the all-atom conformational space.

All-atom detail is added to the lowest-energy coarse-grained conformations as in [HKC07]. Each all-atom conformation is then energetically minimized with the AMBER ff03 energy function [DWC$^+$03] and the Generalized Born (GB) implicit solvation model [STHH90]. Out of the resulting all-atom conformations, only those with energies no higher than 100 kcal/mol from the global minimum energy are retained. This cutoff discards conformations with negligible Boltzmann probabilities at equilibrium.

The all-atom conformations are projected on a low-dimensional landscape through ScIMAP, a nonlinear dimensionality reduction technique proposed [DMS$^+$06] and tested in [PSCK07, SKC08b]. Free energy values are calculated over this landscape to yield a low-dimensional free energy landscape. Free energy minima emerging in the landscape highlight conformational ensembles that are possibly relevant at equilibrium. The lowest-energy conformations associated with the free energy minima are chosen as reference to further explore the conformational space around the minima.

This focused exploration of the minima is implemented through PEM, a multiscale method proposed and tested in [SCK06, SKC07, SCK07] to explore the all-atom conformational space around a given conformation.

6.2.3 Implementation Details

All simulations have been performed on 2.2 GHz AMD64 Opteron CPUs. The MC-SA lowers temperature $k = 14$ times. An MC simulation of 2000 cycles takes between 1-4 hours on a single CPU for the protein sequences considered here. At each temperature and from each of the $n_s = 5$ seed conformations, m MC simulations are launched. Implementations with $n_s > 5$ seeds have been considered for the broader exploration that would be obtained. The number of generated conformations, however, becomes too large for storage and time demands.

The distribution of Rg values among conformations generated from $10,000$ cycles long MC simulations is analyzed to determine $\langle \text{Rg} \rangle$. $\langle \text{Rg} \rangle = 23.4$ Å for calbindin $\text{D}_{9\text{k}}$, and $\langle \text{Rg} \rangle = 33.4$ Å for the CaM and ADK proteins here. Then, $m = 11$ Rg_{goal} values are specified as follows: $\text{Rg}_{\text{goal}} = \infty$ (meaning no confinement is imposed), $\text{Rg}_{\text{goal}} = \langle \text{Rg} \rangle$, and 9 more at consecutive 1.4 Å decrements from $\langle \text{Rg} \rangle$. For example, in the most confined MC simulation for calbindin $\text{D}_{9\text{k}}$, $\text{Rg}_{\text{goal}} = 10.8$Å, which matches that predicted from theory for a chain of 76 amino acids [FR05].

In total $5 \times 11 = 55$ MC simulations are run (in parallel) at each temperature on different CPUs. A total of $14 \times 55 \times 2000 = 1,540,000$ coarse-grained conformations are generated in 14-56 hours on 55 CPUs. For the calbindin $\text{D}_{9\text{k}}$, CaM, and ADK

proteins considered here, 45363, 54820, and 48394 conformations, respectively, are selected from the first stage and handed off to the second stage. From this point on, the second stage focuses on 29290, 33166, and 29424 all-atom conformations, respectively, that meet the 100 kcal/mol cutoff described above. The PEM exploration around each free energy minimum yields on average $2,000$ low-energy all-atom conformations. The all-atom energy minimizations are the most computational demanding in PEM, bringing the total time of the second stage to 2-4 weeks on 50 CPUs for the results presented here.

6.3 Applications to Various Protein Sequences

Results presented below are obtained from the application of MuSE to three increasingly long proteins that are known to undergo large-scale functional motions.

6.3.1 Calbindin D_{9k}

The first protein selected is the 76-aa sequence of calbindin D_{9k}, a protein that transports Ca^{2+}, Mg^{2+}, and Mn^{2+} ions [CGR89, AML+97]. Calbindin D_{9k} is an EF-hand protein, a four-helix bundle with two helix-loop-helix EF-hand motifs. The N-terminus EF-hand contains helix H1, metal-binding loop L1, and helix H2. The C-terminus EF-hand contains helix H3, metal-binding loop L2, and helix H4. A linker region links the EF-hands.

Figure 6.2(a2) superimposes 160 experimental structures available for calbindin

D_{9k} in the PDB. There are differences among these structures in the network of van-der-waals contacts and hydrogen bonds. In particular, contacts between helices H1 and H4 and between H2 and H3 are formed with widely varying probabilities, indicating these helices can fluctuate away from each other. A wide range of contact probabilities also indicate that the linker and loops L1 and L2 are highly mobile.

These 160 structures capture calbindin D_{9k} in different functional states. Comparison of the metal-free (apo) [SKC95], Ca^{2+}- [STF92], Mg^{2+}-, and Mn^{2+}-binding [AML+97] states reveals that the internal structures of the helices remain largely unperturbed. On the other hand, L1, L2, and the linker act as hinges to pack the helices more tightly in the Mg^{2+}- and the Mn^{2+}-binding states. It is reasonable to assume that these functional states coexist at equilibrium with different probabilities depending on ion concentrations [LJC95].

6.3.2 Calmodulin

The second protein considered is the 144-aa sequence of calmodulin (CaM), an EF-hand protein that binds calcium and regulates > 100 proteins, such as kinases, phosphodiesterases, calcium pumps, and motility proteins [MK84, Mea88, OD90]. CaM resembles a dumbbell structure where the terminal domains, linked through a flexible α-helix, are in a trans orientation from each other on either side of the linker.

Three main functional states are observed in experiment. Differences among them are mostly due to a partial unfolding around residue 77 in the central α-helix linker. Further bending of the linker around this point brings the terminal domains in contact

with each other. Figure 6.2(b) superimposes three X-ray structures capturing CaM in its main functional states. The apo state, PDB code 1cfd [KTG$^+$95], is shown in magenta; the calcium-binding state, PDB code 1cll [BBC88, CMMQ92], is in blue; the collapsed peptide-binding state, PDB code 2f3y [FHHQ05], is in green.

The terminal domains are similar in structure (lRMSD between them is 1.0 Å). The central helix is fully formed in the calcium-binding state, partially unfolds in its middle in the apo state, and bends in the collapsed state, bringing the domains in contact. Transitions between the apo and collapsed states are observed in experiment and simulation [FED$^+$95, ZJZ07]. Novel collapsed states are reported in a few MD studies [SV04, PFNG06].

6.3.3 Adenylate Kinase

The third protein considered is the 214-aa sequence of adenylate kinase (ADK), a phosphotransferase enzyme that maintains energy balance in cells by catalyzing the reversible reaction $Mg^{2+} \cdot ATP + AMP \rightleftharpoons Mg^{2+} \cdot ADP + ADP$ [RL68]. The enzyme has a CORE domain and AMP- and ATP-binding domains. The substrate-binding domains undergo large-scale motions to independently bind substrates, giving rise to four functional states: the apo state, where both substrate-binding domains are open, the collapsed state, where both are closed, and two intermediate states, where one of the domains is open and the other closed.

Figure 6.2(c) superimposes four X-ray structures. The structure shown in magenta, PDB code 4ake [MSRS96], shows ADK in its apo state. The collapsed state,

PDB code 2aky [AS95], is shown in blue. The two intermediate states, PDB codes 1dvr [SPS96] and 2ak3 [DS91], are respectively shown in orange and green. The AMP-binding domain is open and the ATP-binding domain is closed in 1dvr, while the ATP-binding domain is open and the AMP-binding domain is closed in 2ak3.

Transitions of ADK between the apo and collapsed states are observed both in experiment and simulation [ÅW07, SQH07, LW08]. While the free energy difference between the apo and collapsed states has been reported to be negligible at room temperature, the two intermediate states of ADK have been associated with higher energies [LW08]. As such, ADK is an ideal system to test the performance of MuSE.

6.3.4 Generation of Conformational Ensembles

The analysis presented here focuses on 29290, 33166, and 29424 lowest-energy all-atom conformations obtained from applying MuSE on calbindin D_{9k}, CaM, and ADK, respectively. These conformations are projected onto a lower-dimensional space by using the ScIMAP nonlinear dimensionality reduction technique [DMS+06] to extract a few global coordinates that best distinguish among conformations.

The ScIMAP analysis reveals that 2 global coordinates capture more than 70% of the structural variability in the ensemble of conformations of each protein. The low-dimensional landscapes presented below are obtained with 3000 landmarks, 50 nearest neighbors, and using lRMSD for nearest neighbor calculations. Free energy calculations on the low-dimensional landscapes highlight free energy minima for each

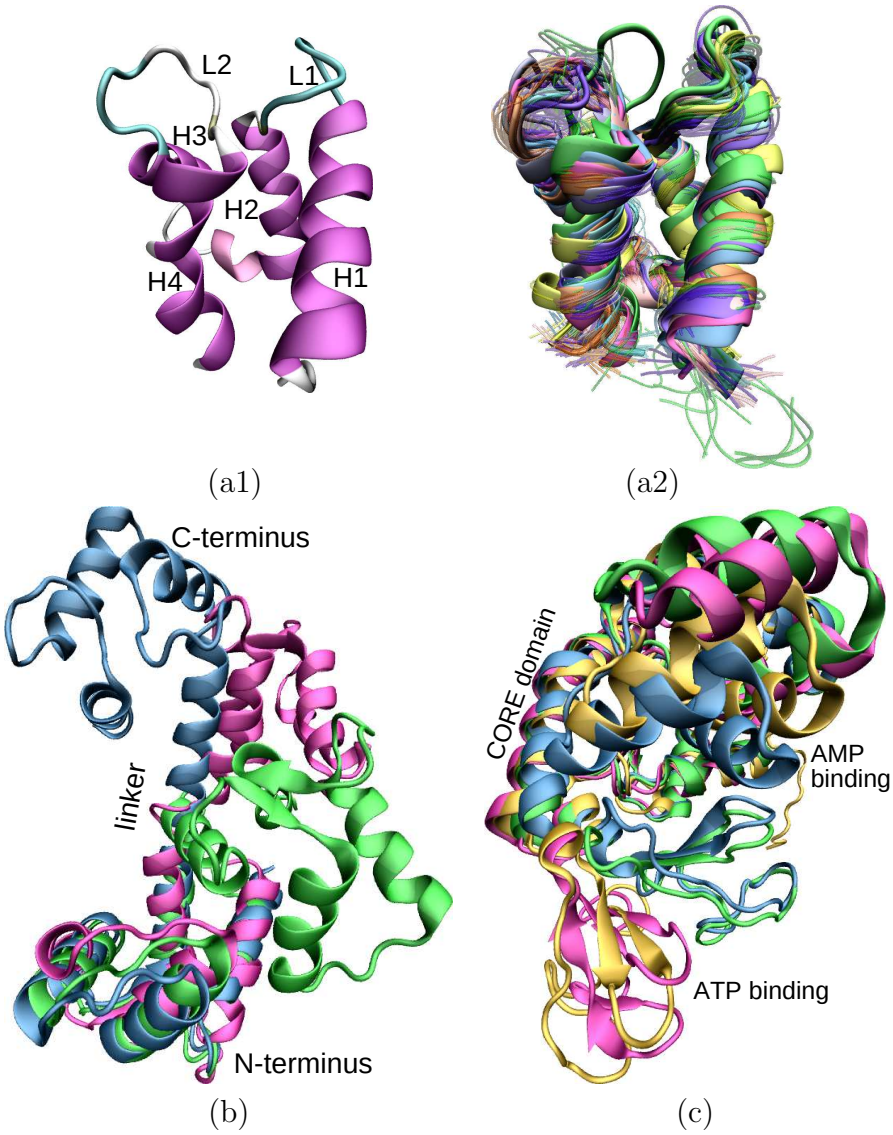


Fig. 6.2: (a1) Helices H1-H4 and loops L1 and L2 are labeled over the PDB structure 4icb of calbindin D_{9k}. (a2) 160 PDB structures are superimposed over one another. X-ray structures and first structures of NMR ensembles are in opaque. Additional NMR structures are in transparent. (b) CaM PDB structures are superimposed over one another: 1cfd is in magenta, 1cll in blue, and 2f3y is in green. (c) ADK PDB structures are superimposed over one another: 4ake is in magenta, 2ak3 in orange, 1dvr in green, and 2aky in blue.

protein. The conformational space around the minima is further explored with PEM.

Three main results emerge from the analysis of the free energy landscapes obtained for the three proteins: (i) on calbindin D_{9k}, the two free energy minima obtained

capture well the difference in packing of EF-hand helices in the various functional states; (ii) on both CaM and ADK, the free energy minima are in correspondence to known functional states; (iii) higher-energy ensembles are also observed. Interestingly, these ensembles for CaM have been reported on a few MD studies and may correspond to (yet) unobserved metastable collapsed functional states.

6.3.5 Analysis of Generated Ensembles of Calbindin D_{9k}

Figure 6.3(a) shows the free energy landscape of calbindin D_{9k} as a function of the first two global coordinates revealed by ScIMAP for the ensemble of conformations obtained by MuSE. Free energy values are color-coded in a red-to-blue spectrum that denotes high-to-low values. The experimental structures of calbindin D_{9k} are projected on this 2D landscape and shown as black circles.

The two lowest free energy minima are labeled A and B in Figure 6.3(a). The experimental structures, when projected on the 2D landscape, cluster around minimum A. The energetic separation between A and B is ~ 3 kcal/mol. Given understandable approximations in empirical energy functions and the approximations used in MuSE, the two minima can be considered energetically equivalent.

The conformational ensembles corresponding to minima A and B are shown in Figures 6.3(a1) and 6.3(b1), respectively. The lowest-energy conformations within each ensemble, shown in opaque in Figures 6.3(a1)-(b1), have an lRMSD of 2.36 and 5.14 Å from PDB structure 4icb (Ca^{2+}-binding state), respectively. The lRMSD lowers down to 1.89 and 2.59 Å, respectively, when the Local-Global Alignment (LGA) tool

is used [Zem03]. The lower values result from the fact that LGA localizes structural differences between two conformations[2].

Both ensembles in Figures 6.3(a1)-(b1) capture the overall fold of calbindin D_{9k}. The helices are well-formed, whereas L1, L2, and the linker are very mobile. The main difference between the two ensembles is in the packing of the EF-hand helices: tighter packing is observed in the ensemble in Figure 6.3(a1). In particular, the distance between central residues in loops L1 and L2 has an average of 12.69 Å in the ensemble in Figure 6.3(a1) and an average of 14.33 Å in the ensemble in Figure 6.3(b1). For comparison, this average is 11.33 Å in PDB structure 4icb. The tighter packing in the ensemble associated with minimum A is in good agreement with what is observed in the Mg^{2+}- and Mn^{2+}-binding states [AML+97]. On the other hand, looser packing is observed in the apo and the Ca^{2+}-binding states [SKC95, STF92], consistent with what is observed in the ensemble associated with minimum B.

The lowest-energy conformations of the ensembles corresponding to minima A and B are used as references to further explore the conformational space through PEM. The PEM-generated ensembles, shown in Figures 6.3(a2)-(b2), reproduce the structural features that distinguish minima A and B and corroborate the results obtained from the coarse-grained exploration.

The conformational ensembles associated with minima A and B can be further characterized by considering the network of van-der-waals contacts and hydrogen

[2]LGA is used to assess similarity between predictions and targets in structure prediction [MFK+07].

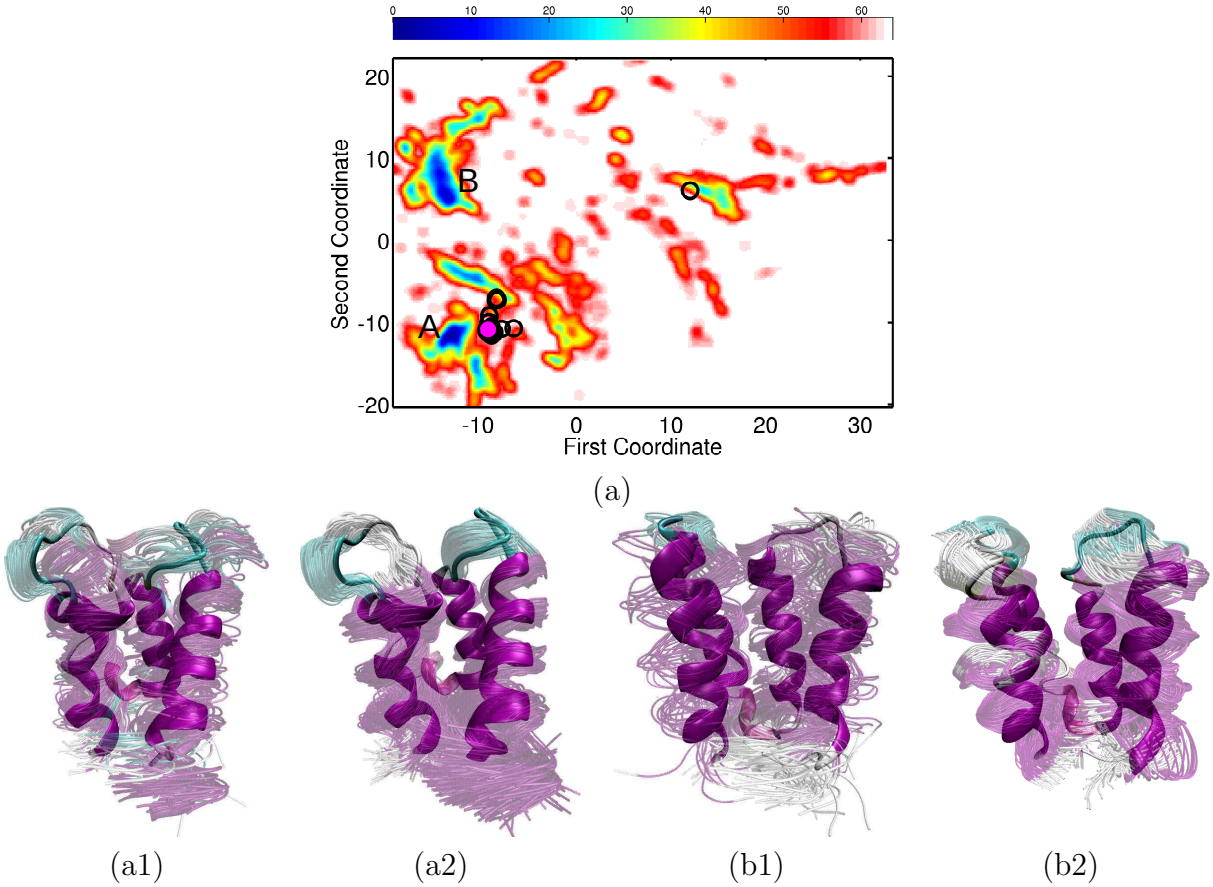
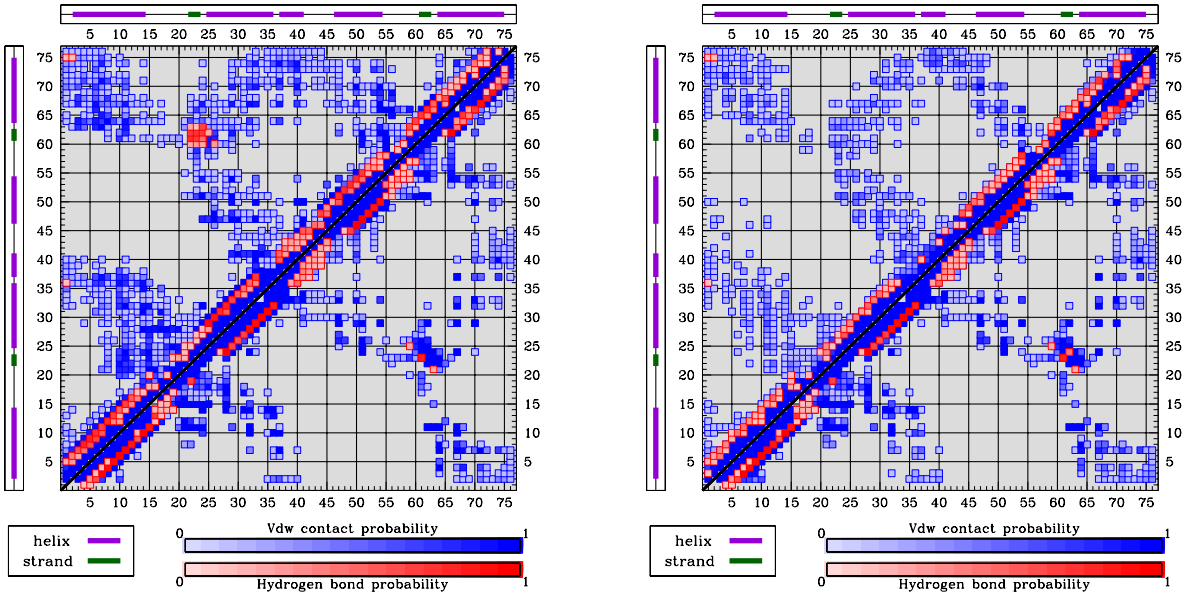


Fig. 6.3: (a) Red-to-blue color spectrum in 2D landscape obtained for calbindin D_{9k} denotes high-to-low free energy values. Black circles show projections of PDB structures over the landscape. The projection of PDB structure 4icb is drawn in magenta. The lowest free energy minima are labeled A and B. Conformational ensembles corresponding to A and B are shown in (a1) and (b1), respectively. Conformations are superimposed in transparent over lowest-energy one drawn in opaque. (a2) and (b2) show ensembles obtained with PEM from each lowest-energy conformation.

bonds. Probabilities of contacts and hydrogen bonds formation are measured as Boltzmann averages over each ensemble. These probabilities are shown by the color-coded maps in Figures 6.4(a3)-(b3). Darker colors denote higher probabilities. For comparison, the bottom halves of these maps show the formation probabilities measured by averaging over the 160 experimental structures of calbindin D_{9k}. Two amino acids are defined in contact if the Euclidean distance between two of their heavy atoms

is ≤ 4.5 Å. A hydrogen bond is considered formed if the OH distance is less than 2.4 Å and the maximum NHO angle for the hydrogen bond alignment is 2.44 rad [NB08].



(a3) (b3)

Fig. 6.4: (a3) and (b3) compares contacts and hydrogen bonds measured over enriched conformational ensembles corresponding to A and B (top half) to contacts and hydrogen bonds averaged over the PDB structures. Darker shades denote higher probabilities.

The high similarity between the top and bottom halves of the map in Figure 6.4(a3) indicates that the conformational ensemble associated with minimum A captures the main interactions present in the experimental structures. In particular, interactions between loops L1-L2, H1-H2, and H3-H4 occur with high probability. This result confirms the tight packing of the helices that characterizes conformations associated with minimum A. On the other hand, Figure 6.4(b3) shows that some of these interactions occur rarely in the ensemble associated with minimum B, as a result of the looser packing that characterizes conformations in minimum B.

6.3.6 Analysis of Generated Ensembles of Calmodulin

The free energy landscape of CaM lowest-energy conformations generated by MuSE is shown as a function of the first two global coordinates in Figure 6.5(a). Free energy values are color-coded in a red-to-blue spectrum to denote high-to-low values. Three low-energy minima emerge, labeled A, B, and C in Figure 6.5(a). The first global coordinate separates A from B and C, whereas the second coordinate separates C from A and B. The projection of PDB structure 1cll (drawn in blue) on this landscape falls near minimum A, that of 1cfd (in magenta) falls near B, and that of 2f3y (in green) falls near C. Energy differences among the minima are < 1 kcal/mol.

The conformational ensembles corresponding to minima A, B, and C are shown in Figures 6.5(a1), 6.5(b1), and 6.5(c1), respectively. The main feature in the ensemble corresponding to minimum A is a well-formed α-helix in the linker, as in PDB structure 1cll. The helix is partially unfolded in the ensemble corresponding to minimum B, as in PDB structure 1cfd. The linker bends further in the ensemble corresponding to minimum C, as in PDB structure 2f3y. The three ensembles show that the terminal domains exhibit some mobility while largely preserving their secondary structures.

The lowest-energy conformations in ensembles A, B, and C, shown in opaque in Figures 6.5(a1)-(b1), have LGA lRMSDs of 2.572, 2.201, and 2.792 Å from PDB structures 1cll, 1cfd, and 2f3y, respectively. These conformations are used as references to further search in all-atom detail the conformational space around the minima. The PEM-generated ensembles, shown in Figures 6.5(a3)-(b3), reproduce well the struc-

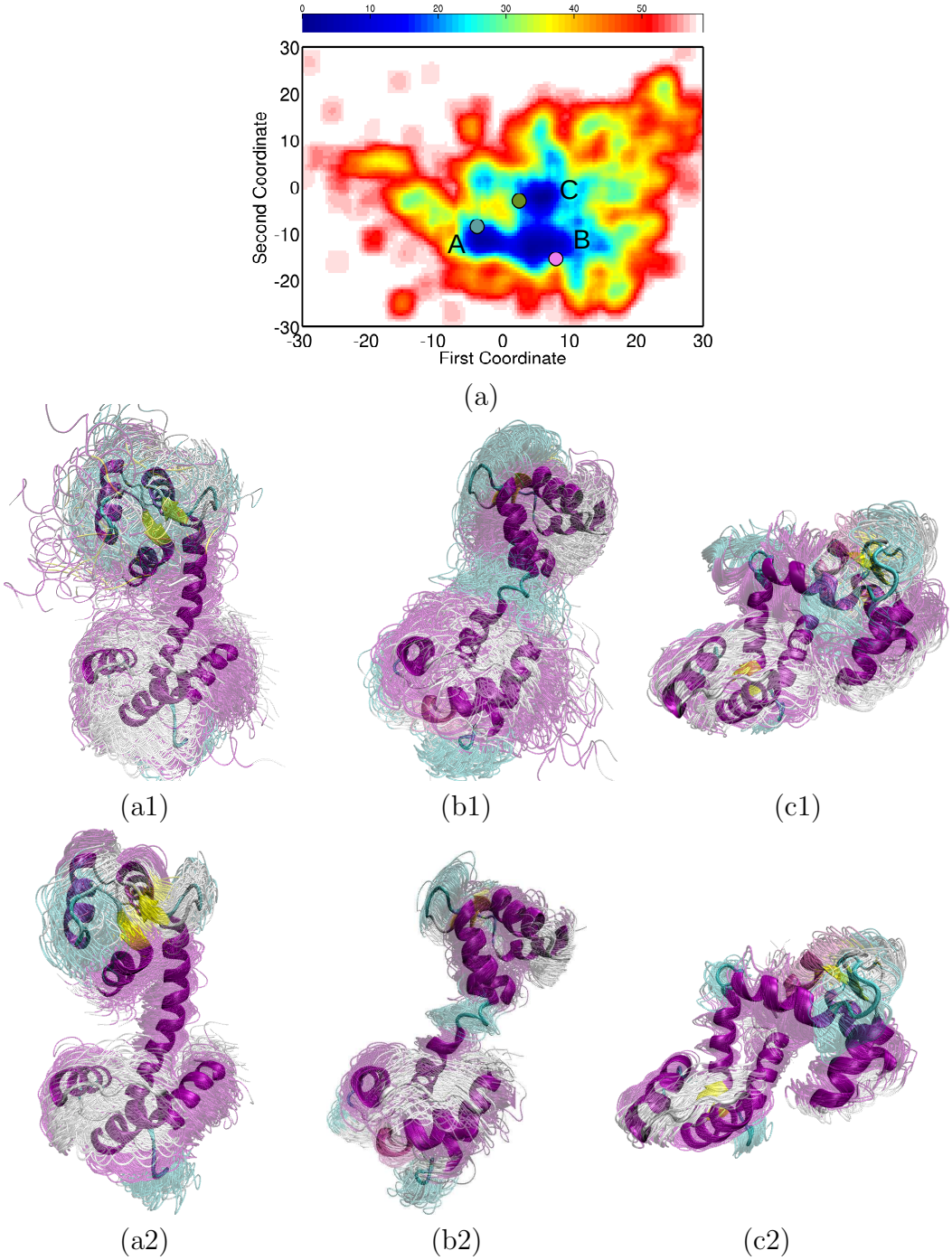


Fig. 6.5: (a) Red-to-blue color spectrum denotes high-to-low free energies. Free energy minima are labeled A, B, and C. PDB structures projected on landscape are 1cfd in magenta, 1cll in blue, and 2f3y in green. (a2)-(c2) show ensembles corresponding to A, B, and C. Conformations are superimposed in transparent over lowest-energy ones in opaque. (a3)-(c3) show conformational ensembles obtained with PEM from each lowest-energy conformation.

tural differences among the minima and further support the conformational diversity captured from the coarse-grained exploration.

Figure 6.5(a) shows that minimum B is broader than A and C. The corresponding conformational ensemble in Figure 6.5(b1) provides an explanation: the partial unfolding of the helix linker in this ensemble allows access to a large configurational space. Figure 6.5(a) also shows that minima A, B, and C are not isolated from one another. Conformations bridging A and B exhibit the helix linker gradually unfolding in its middle, whereas conformations bridging B and C further bend the linker. The conformations bridging the minima may provide transitions between the three main functional states of CaM. Conformations mediating between the calcium-binding and collapsed states have been observed in 20-ns MD simulations [SV04].

Inspection of MuSE-obtained conformations reveals higher-energy collapsed ensembles not (yet) observed in experiment. Similar collapsed structures have been observed in MD simulations when CaM is depleted of a calcium ion [PFNG06].

Formation probabilities of van-der-waals contacts and hydrogen bonds are measured over each ensemble associated with the minima and are shown by color-coded maps in Figures 6.6(a3)-(c3). The bottom halves of the maps show contacts and hydrogen bonds measured over PDB structures 1cll, 1cfd, and 2f3y, respectively. Darker colors denote higher probabilities. The maps associated with the ensembles corresponding to minima A, B, and C largely reproduce those of PDB structures 1cll, 1cfd, and 2f3y, respectively. Figure 6.6(a3) shows additional rare interactions

between the terminal domains that arise as the domains move closer to a linker that bends slightly without unfolding in minimum A. Similar interactions are present in Figure 6.6(b3), as the linker unfolds in minimum B. Figure 6.6(c3) shows that this interdomain coupling becomes more prevalent as the linker bends further in minimum C. Such coupling has been observed in MD studies [SV04].

6.3.7 Analysis of Generated Ensembles of Adenylate Kinase

Figure 6.7(a) shows the free energy landscape associated with MuSE-generated ADK conformations as a function of the two global coordinates obtained from ScIMAP. Color-coding free energy values in a red-to-blue spectrum that denotes high-to-low values reveals two free energy minima, labeled A and B in Figure 6.7(a). The energetic difference between the minima is $\simeq 1.3$ kcal/mol.

The four X-ray structures of ADK that capture this protein's functional states are projected and drawn on the landscape in Figure 6.7(a) in different colors: the projection of PDB structure 4ake is shown in magenta, 1dvr is in orange, 2ak3 is in green, and 2aky is in blue. Figure 6.7(a) shows that, when projected on the free energy landscape, 4ake, which captures ADK in its apo state, falls near minimum A, and 2aky, which captures the enzyme in its collapsed state, falls near minimum B.

The conformational ensembles corresponding to minima A and B are shown in Figures 6.7(a1) and 6.7(b1), respectively. The main features in the ensemble in Figure 6.7(a1) are open AMP- and ATP-binding domains, as in 4ake. Both domains are closed in the ensemble in Figure 6.7(b1), as in 2aky. The lowest-energy conformations

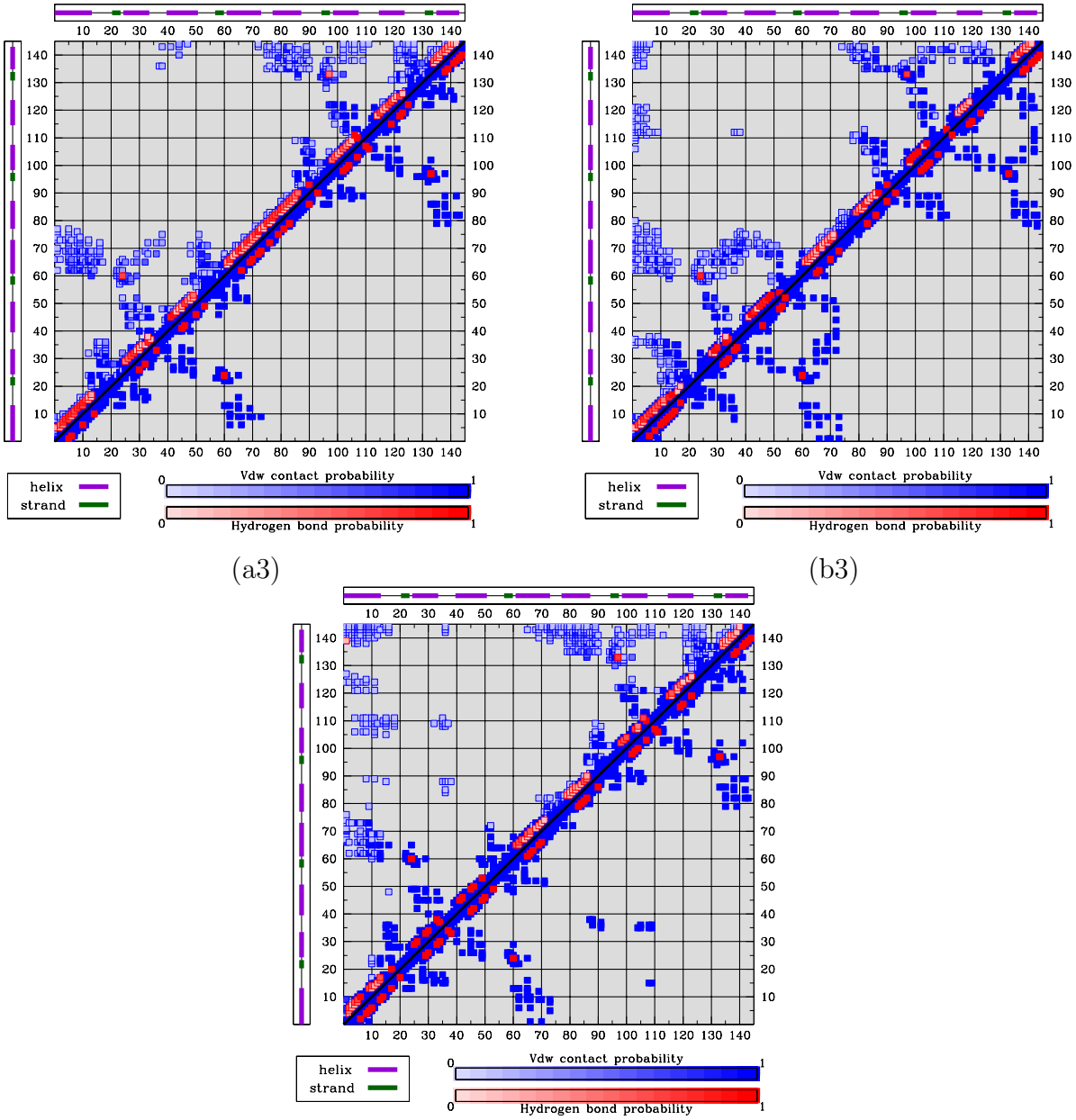


Fig. 6.6: Top halves of maps in (a3), (b3), and (c3) show contact and hydrogen-bond formation probabilities measured over ensembles associated with minima A, B, and C. Bottom halves show contacts and hydrogen bonds in PDB structure 1cll in (a3), 1cfd in (b3), and 2f3y in (c3). Darker shades denote higher probabilities.

in each ensemble, shown in opaque in Figures 6.7(a1)-(b1), are within within 2.951 and 3.265 Å LGA lRMSD from PDB structures 4ake and 2aky, respectively.

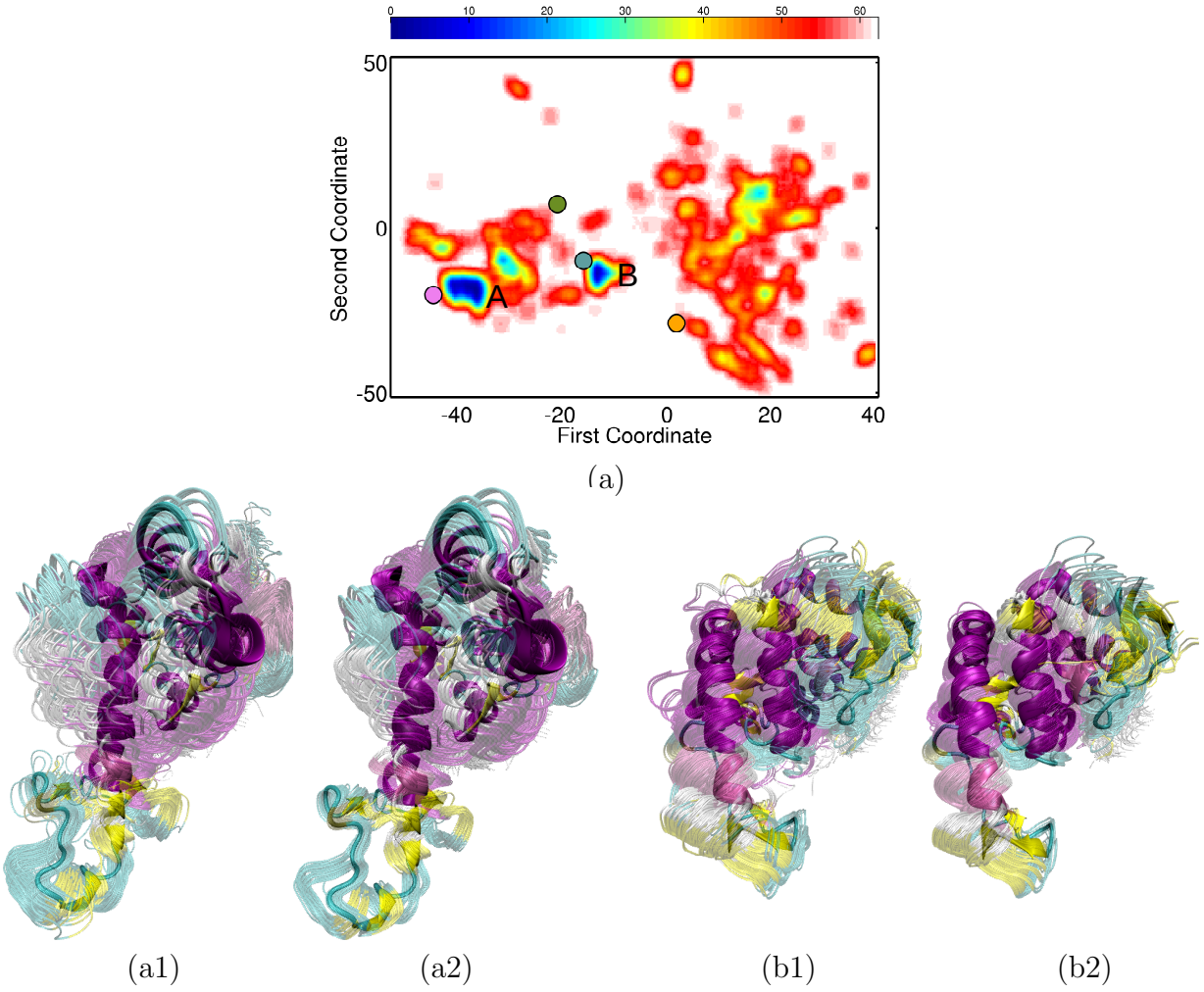


Fig. 6.7: (a) Red-to-blue color spectrum in 2D landscape obtained for ADK denotes high-to-low free energy values. Free energy minima are labeled A and B. PDB structures are projected on the landscape: 4ake in magenta, 2ak3 in orange, 1dvr in green, and 2aky in blue. (a1) and (b1) show ensembles corresponding to A and B. Conformations are superimposed in transparent over lowest-energy ones drawn in opaque. (a2) and (b2) show conformational ensembles obtained with PEM from each lowest-energy conformation.

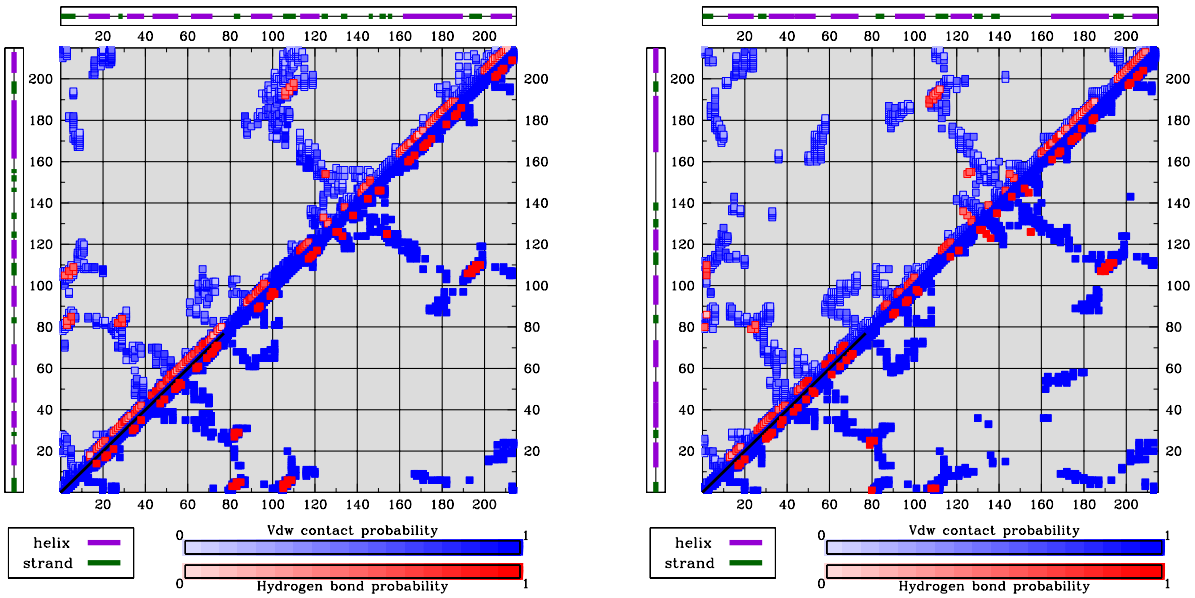
The lowest-energy conformations in each minimum are used as references for PEM to further explore the space around A and B. The obtained ensembles, shown in

Figures 6.7(a2)-(b2), reproduce the apo and collapsed states of ADK, further supporting the structural features associated with the predicted minima. Figures 6.8(a3) and 6.8(b3) juxtapose formation probabilities of contacts and hydrogen bonds measured over the ensembles associated with minima A and B to contacts and hydrogen bonds in PDB structures 4ake and 2aky. The maps associated with the conformational ensembles respectively reproduce the maps associated with the PDB structures.

The projection of PDB structures on the landscape in Figure 6.7(a) reveals that the intermediate functional states, where one terminal domain is open and the other closed, are not obtained as free energy minima. These states are associated with energy barriers in the transition between the apo and collapsed states [LW08]. The higher energies associated with these intermediate states disqualify conformations representative of these states from being selected for further exploration in MuSE. Instead, the lower-energy apo and collapsed states of ADK prevail in the landscape offered by MuSE as relevant at equilibrium. Folding simulations like the one in [LW08] could be employed to launch MD trajectories and capture the intermediate states as ADK transitions between the apo and collapsed states obtained by MuSE.

6.4 MuSE: Discussion and Conclusion

The application of a multiscale strategy to explore the conformational space reproduces well known functional states for the three proteins considered. In particular, on calbindin D_{9k}, the obtained free energy minima capture the variation in the packing



(a3) (b3)

Fig. 6.8: (a3) and (b3) show contacts and hydrogen bonds measured over enriched conformational ensembles corresponding to A and B (top half). The bottom halves show contacts and hydrogen bonds in PDB structure 4ake in (a3) and 2aky in (b3). Darker shades denote higher probabilities.

of EF-hand helices. One minimum is associated with a tighter packing, as observed in the Mg^{2+}- and the Mn^{2+}-binding states. The other minimum shows a looser packing, as in the apo and Ca^{2+}-binding states. On CaM, the three obtained free energy minima reproduce the three documented functional states. The two free energy minima obtained for ADK capture well the apo and collapsed states of this protein.

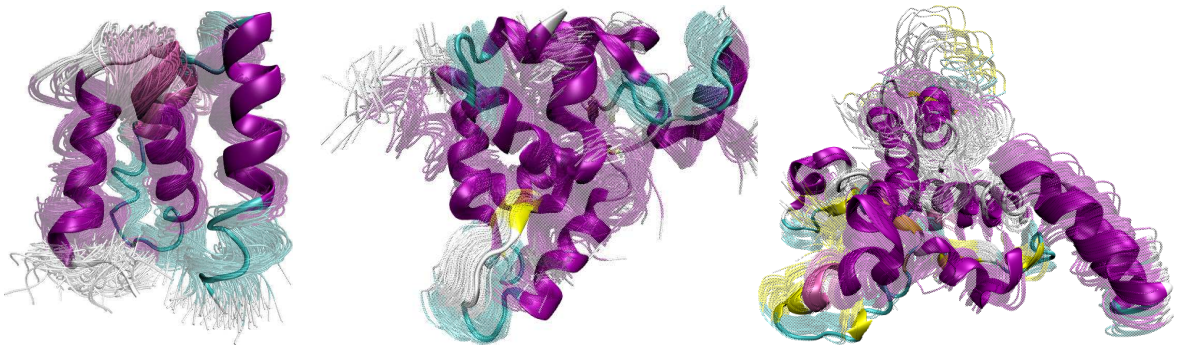
Intermediate functional states are not directly observed for ADK. The higher energy of these states probably discards their conformations in the first stage of the exploration. However, by capturing the apo and collapsed states, MuSE could be further enhanced by simulation studies as in [LW08]. MD trajectories could provide details on states mediating the transition of a protein between the main functional

states associated with the free energy minima obtained with MuSE.

The selection of seed conformations, the multiscale search for these seeds, the parallel simulations launched from selected seeds in the first stage, and the coarse-grained energy function employed during this stage are all critical components that together guide the exploration to relevant regions in the all-atom conformational space. High-dimensional conformational spaces associated with longer ($\geq$ 300 aas) proteins than the ones considered here may require employing different seed selection strategies, loosening some of the energetic cutoffs, and even employing coarser representations than the one used in the first stage. Realistic coarse-grained energy functions need to be devised for even coarser representations. Alternatively, multiple levels of coarse-graining could be employed to explorate a larger high-dimensional space.

Analysis of the conformational ensembles obtained here reveals higher-energy ensembles of collapsed conformations, some of which are shown in Figure 6.9. Similar collapsed conformations have been reported, for instance, for CaM in MD studies but not (yet) in experiment. One cannot rule out that the presence of these conformations may be due to approximations in empirical energy functions or in the method. Multiple energy functions and representations can be employed in the future to improve predictions made by the method.

A different explanation for the presence of additional conformational ensembles may be offered by considering the difference between thermodynamics and kinetics. MuSE is thermodynamic in nature. Association of timescales to access different



(a) (b) (c)

Fig. 6.9: Higher-energy conformational ensemble obtained for calbindin D_{9k} in (a), CaM in (b) and ADK in (c). The lowest-energy conformations within each ensemble are drawn in opaque, superimposing the remaining conformations in an ensemble in transparent. The ensemble in (a) corresponds to the region $\{10 \leq x \leq 20, 2 \leq y \leq 8\}$ in the 2D free energy landscape obtained for calbindin D_{9k}. The ensemble in (b) corresponds to the region $\{5 \leq x \leq 10, 5 \leq y \leq 10\}$ in the 2D free energy landscape obtained for CaM. The ensemble in (c) corresponds to the region $\{-35 \leq x \leq -25, -15 \leq y \leq 0\}$ in the 2D free energy landscape obtained for ADK.

conformational ensembles is an obvious direction for future work, that can improve accuracy and strengthen the connection with experiment.

The proposed MuSE marks a first step towards obtaining a picture of the conformational diversity of proteins at equilibrium. The results obtained by MuSE can serve as a robust starting point to characterize functional motions in proteins, either in combination with more refined computational methods, or with experiments.

Chapter 7

Discussion

Obtaining the ensemble of conformations populated by a protein under native conditions is important to understand how the protein uses these conformations to modulate its biological function. The methods described here have made several contributions to the in silico treatment of the protein native state. They are leading to a comprehensive view of the native conformational ensemble in proteins that employ large-scale concerted motions to assume diverse functional states.

The presented methods search for native conformations using systematically less information from experimental techniques: (i) first employing an experimental structure to guide the search in proteins $\sim$ 100 aas long ; (ii) then replacing the experimental structure with a closure constraint in cyclic peptides 20$-$30 aas long; (iii) finally, employing only the amino-acid sequence of small- to medium-size proteins.

A probabilistic exploration conducted hierarchically and at multiple levels of detail is proposed to efficiently probe the vast high-dimensional protein conformational space and detect energy minima emerging in the free energy surface associated with the space. As less experimental information is employed, combining a broad view of a

coarse-grained conformational space with a detailed all-atom view of emerging energy minima becomes crucial to the success of the methods presented here.

Applications of the PEM method on proteins of various folds reproduce with high accuracy wet-lab data that span a broad range of timescales. Applications of the NcCYP method on cysteine-rich cyclic peptides capture well both the conformational diversity of the native state in these peptides and the diversity of their disulfide bonds under native conditions. Finally, applications of the MuSE method on protein sequences up to 214 aas long show the ability of this method to extract from amino-acid sequence the conformational ensembles populated in different functional states.

The problem of extracting the native conformational ensemble from just amino-acid sequence has proven challenging in computational biology for four decades. The results obtained in this thesis are promising enough to motivate further work in this direction. Adding timescale information to computed conformational ensembles is one potential direction of future research to improve in silico predictions.

The methods presented in this thesis bridge between computer science, biophysical theory, and wet-lab experiments. By obtaining the conformational ensemble(s) associated with the protein native state, these methods have the potential to complement experiments in the study of function and mechanism in proteins.

Bibliography

[AACV06] Arolas JL, Aviles FX, Chang JY, and Ventura S. Folding of small disulfide-rich proteins: clarifying the puzzle. Trends Biochem Sci 31(5):292–301 (2006)

[ABB+99] Anderson E, Bai Z, Bischof C, Blackford S, Demmel J, Dongarra J, Du Croz J, Greenbaum A, Hammarling S, McKenney A, and Sorensen D. LAPACK Users' Guide. Society for Industrial and Applied Mathematics, Philadelphia, PA, 3rd edition (1999)

[ABG+03] Apaydin MS, Brutlag DL, Guestrin C, Hsu D, and Latombe JC. Stochastic roadmap simulation: an efficient representation and algorithm for analyzing molecular motion. J Comp Biol 10(3-4):257–281 (2003)

[ADS02] Amato NM, Dill KA, and Song G. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. J Comp Biol 10(3-4):239–255 (2002)

[AM06] Adcock SA and McCammon JA. Molecular dynamics: Survey of methods for simulating the activity of proteins. Chem Rev 106(5):1589–1615 (2006)

[AML+97] Andersson M, Malmendal SL, Linse S, Ivrasson I, Forsén S, and Svensson A. Structural basis for the negative allostery between Ca^{2+}- and Mg^{2+}-binding in the intracellular Ca^{2+}-receptor calbindin D_{9k}. Protein Sci 6(6):1139–1147 (1997)

[Anf73] Anfinsen CB. Principles that govern the folding of protein chains. Science 181(4096):223–230 (1973)

[AS95] Abele U and Schulz GE. High-resolution structures of adenylate kinase from yeast ligated with inhibitor Ap5A, showing the pathway of phosphoryl transfer. Protein Sci 4(7):1262–1271 (1995)

[AS00] Abkevich VI and Shakhnovich EI. What can disulfide bonds tell us about protein energetics, function, and folding: Simulations and bioinformatics analysis. J Mol Biol 300(4):975–985 (2000)

[AT94] Abagyan R and Totrov M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 235(3):983–1002 (1994)

[ATK94] Abayagan R, Totrov M, and Kuznetsov D. ICM - a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comput Chem 15(5):488–506 (1994)

[ÅW07] Ådén J and Wolf-Watz M. NMR identification of transient complexes critical to adenylate kinase catalysis. J Am Chem Soc 129(45):14003 –14012 (2007)

[BA94] Baker D and Agard DA. Kinetics versus thermodynamics in protein folding. Biochemistry 33(24):7505–7509 (1994)

[BB00] Bursulaya BD and Brooks CLI. Comparative study of the folding free energy landscape of a three-stranded β-sheet protein with explicit and implicit solvent models. J Phys Chem 104(51):12378–12383 (2000)

[BB01] Bonneau R and Baker D. Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys and Biomolec Struct 30(1):173–189 (2001)

[BBC88] Babu YS, Bugg CE, and Cook WJ. Structure of calmodulin refined at 2.2 Å resolution. J Mol Biol 204(1):191–204 (1988)

[BBM$^+$05] Bouvignes G, Bernadó P, Meier S, Cho K, S G, and Brueschweiler R. Identification of slow correlated motions in proteins using residual dipolar couplings and hydrogen-bond scalar couplings. Proc Natl Acad Sci USA 102(39):13885–13890 (2005)

[BBO$^+$83] Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, and Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 4(2):187–217 (1983)

[BHN88] Bruccoleri RE, Haber E, and Novotny J. Structure of antibody hypervariable loops reproduced by a conformational search algorithm. Nature 335(6190):564–568 (1988)

[BK85] Bruccoleri RE and Karplus M. Chain closure with bond angle variations. Macromolecules 18(12):2676–2773 (1985)

[BK87] Bruccoleri RE and Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26(1):137–168 (1987)

[BK90] Bruccoleri RE and Karplus M. Conformational sampling using high temperature molecular dynamics. Biopolymers 29(14):1847–1862 (1990)

[BK97] Becker OM and Karplus M. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. J Chem Phys 106(4):1495–1517 (1997)

[BK02] Brock O and Khatib O. Elastic strips: A framework for motion generation in human environments. Int J Robot Res 21(12):1031–1052 (2002)

[BKP88] Brooks CLI, Karplus M, and Pettit BM. Proteins: A theoretical perspective of dynamics, structure and function. John Wiley and Sons, 1st edition (1988)

[BKW+77] Bernstein FC, Koetzle TF, Williams G, Meyer DJ, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, and Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 112(3):535–542 (1977)

[BM04] Baneyx F and Mujacic M. Recombinant protein folding and misfolding in Escherichia Coli. Nat Biotechnol 22(11):1399–1408 (2004)

[BMB05] Bradley P, Misura KMS, and Baker D. Toward high-resolution de novo structure prediction for small proteins. Science 309(5742):1868–1871 (2005)

[BN92] Berg BA and Neuhaus T. Multicanonical ensemble: A new approach to simulate first-order phase transitions. Phys Rev Lett 68(1):9–12 (1992)

[Bor05] Borman S. Protein structure wed to dynamics: Technique determines structure and motions of native proteins simultaneously. Chemical And Engineering News 83(3):12 (2005)

[BR02] Bevington PR and Robinson DK. Data reduction and error analysis for the physical sciences. McGraw-Hill, New York, NY, 3rd edition (2002)

[BRFC04] Best RB, Rutherford TJ, Freund SMV, and Clarke J. Hydrophobic core fluidity of homologous protein domains: Relation of side-chain dynamics to core composition and packing. Biochemistry 43(5):1145–1155 (2004)

[Bru93] Bruccoleri RE. Application of systematic conformational search to protein modeling. Mol Simulat 10(2-6):151–174 (1993)

[BSM+06] Bhalla J, Storchan GB, MacCarthy CM, Unversky VN, and Tcherkasskaya O. Local flexibility in molecular function paradigm. Mol Cell Proteomics 5(7):1212–1223 (2006)

[Bur89] Burdick JW. On the inverse kinematics of redundant manipulators: characterization of the self-motion manifold. In Proc. IEEE Int. Conf. Robot. Autom., vol. 1, 264–270. IEEE, Scottsdale, AZ (1989)

[BV04] Best RB and Vendruscolo M. Determination of ensembles of structures consistent with NMR order parameters. J Am Chem Soc 126(26):8090–8091 (2004)

[BVG+94] Bax A, Vuister GW, Grzesiek S, Delaglio F, Wang AC, Tschudin R, and Zhu G. Measurement of homo- and heteronuclear J couplings from quantitative J correlation. Methods Enzymol 239:79–105 (1994)

[BVSD93] Brower RC, Vasmatzis G, Silverman M, and DeLisi C. Exhaustive conformational search and simulated annealing for models of lattice peptides. Biopolymers 33(3):320–334 (1993)

[BWF+00] Berman HM, Westbrook J, Feng Z, Gilliand G, Bhat TN, Weissig H, N SI, and Bourne PE. The Protein Data Bank. Nucl Acids Res 28(1):235–242 (2000)

[CB05] Ciu Q and Bahar I. Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems. CRC Press, 1st edition (2005)

[CCB03] Chou JJ, Case DA, and Bax A. Insights into the mobility of methyl-bearing side chains in proteins from $^3J_{CC}$ and $^3J_{CN}$ couplings. J Am Chem Soc 125(29):8959–8966 (2003)

[CCD06] Craik DJ, Cemazar M, and Daly NL. The cyclotides and related macrocyclic peptides as scaffolds in drug design. Curr Opin Drug Discov Devel 9(2):251–260 (2006)

[CCLW89] Claessens M, Cutsem E, Lasters I, and Wodak S. Modeling the polypeptide backbone with 'spare parts' from known protein structures. Protein Eng 4(5):335–345 (1989)

[CD03] Canutescu AA and Dunbrack RL. Cyclic coordinate descent: A robotics algorithm for protein loop closure. Protein Sci 12(5):963–972 (2003)

[CDC+06] Case DA, Darden TA, Cheatham TEI, Simmerling CL, Wang J, Duke RE, Luo R, Merz KM, Pearlman DA, Crowley M, Walker RC, Zhang W, Wang B, Hayik S, Roitberg A, Seabra G, Wong KF, Paesani F, Wu X, Brozell S, Tsui V, Gohlke H, Yang L, Tan C, Mongan J, Hornak V, Cui G, Beroza P, Mathews DH, Schafmeister C, Ross WS, and Kollman PA. Amber 9 (2006)

[CE93] Carlacci L and Englander L. The loop problem in proteins: a Monte Carlo simulated annealing approach. Biopolymers 33(8):1271–1286 (1993)

[CE96] Carlacci L and Englander SW. Loop problem in proteins: developments on the Monte Carlo simulated annealing approach. J Comput Chem 17(8):1002–1012 (1996)

[Cen94] Center PSP. Critical Assessment of Techniques for Protein Structure Prediction (1994). http://predictioncenter.org/

[CEP97] Carr PA, Erickson HP, and Palmer AGI. Backbone dynamics of homologous fibronectin type III cell adhesion domains from fibronectin and tenascin. Struct Fold Des 5(7):949–959 (1997)

[CFD+05] Clark RJ, Fischer H, Dempster L, Daly N, Rosengren KJ, Nevin ST, Meunier FA, Adams DJ, and Craik DJ. Engineering stable peptide

toxins by means of backbone cyclization: Stability of the α-conotoxin MII. Proc Natl Acad Sci USA 102(39):13767–13772 (2005)

[CFT03] Chikenji G, Fujitsuka Y, and Takada S. A reversible fragment assembly method for de novo protein structure prediction. J Chem Phys 119(13):6895–6903 (2003)

[CGO03] Clementi C, Garcia A, and Onuchic JN. Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: an all-atom representation study. J Mol Biol 326(3):933–954 (2003)

[CGR89] Christakos S, Gabrielides C, and Rhoten WB. Vitamin D-dependent calcium binding proteins: Chemistry, distribution, functional considerations, and molecular biology. Endocr Rev 10(1):3–26 (1989)

[CGR06] Christakos S, Gabrielides C, and Rhoten WB. Multigraining: an algorithm for simultaneous fine-grained and coarse-grained simulation of molecular systems. J Chem Phys 125(15):154106 (2006)

[Cha87] Chandler D. Introduction to modern statistical mechanics. Oxford University Press, New York, NY, 1st edition (1987)

[CHG93] Collura V, Higo J, and Garnier J. Modelling of protein loops by simulated annealing. Protein Science 2:1502–1510 (1993)

[Chi93] Chirikjian GS. General methods for computing hyper-redundant manipulator inverse kinematics. In Proc IEEE/RSJ Int Conf Intell Robot Sys (IROS), vol. 2, 1067–1073. IEEE, Yokohama, Japan (1993)

[CJO00] Clementi C, Jennings PA, and Onuchic JN. How native-state topology affects the folding of dihydrofolate reductase and interleukin-1β. Proc Natl Acad Sci USA 97(11):5871–5876 (2000)

[CJO01] Clementi C, Jennings PA, and Onuchic JN. Prediction of folding mechanism for circular-permuted proteins. J Mol Biol 311(4):879–890 (2001)

[CJS+06] Colubri A, Jha AK, Shen MY, Sali A, Berry RS, Sosnick TR, and Freed KF. Minimalist representations and the importance of nearest neighbor effects in protein folding simulations. J Mol Biol 363(4):835–857 (2006)

[CK00] Chang K and Khatib O. Operational space dynamics: efficient algorithms for modeling and control of branching mechanisms. In Proc. IEEE Int. Conf. Robot. Autom., 850–856. IEEE, San Francisco, CA (2000)

[CL04] Clarkson MW and Lee AL. Long-range dynamic effects of point mutations propagate through side chains in the serine protease inhibitor eglin c. Biochemistry 43(39):12448–12458 (2004)

[Cle08] Clementi C. Coarse-grained models of protein folding: Toy-models or predictive tools? Curr Opinion Struct Biol 18(1):10–15 (2008)

[CLH+05] Choset H, Lynch KM, Hutchinson S, Kantor G, Burgard W, Kavraki LE, and Thrun S. Principles of robot motion. MIT Press, Cambridge, MA, 1st edition (2005)

[CLP+87] Chothia C, Lesk A, Prilusky J, , and Manning N. Canonical structures for the hypervariable loops of immunoglobulins. J Mol Biol 196(4):901–917 (1987)

[CLT+89] Chothia C, Lesk A, Tramontano A, M Levitt S, Gill SJ, Air G, Sheriff S, Padla E, Davies D, Tulip WR, Colman PM, Spinelli S, Alzari PM, and Poljak RJ. Conformation of immunoglobulin hypervariable regions. Nature 342(6252):877–883 (1989)

[CMMQ92] Chattopadhyaya R, Meador WE, Means AR, and Quiocho FA. Calmodulin structure refined at 1.7 Å resolution. J Mol Biol 228(4):1177–1192 (1992)

[CMOB98] Cornilescu G, Marquardt JL, Ottiger M, and Bax A. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J Am Chem Soc 120(27):6836–6837 (1998)

[CNO00] Clementi C, Nymeyer H, and Onuchic JN. Topological and energetic factors: what determines the structural details of the transition state ensemble and "on-route" intermediates for protein folding? An investigation for small globular proteins. J Mol Biol 298:937–953 (2000)

[COC04] Chavez LL, Onuchic JN, and Clementi C. Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates. J Am Chem Soc 126(27):8426–8432 (2004)

[Cra89] Craig J. Introduction to robotics: mechanics and control. Addison-Wesley, Boston, MA, 2nd edition (1989)

[Cra06] Craik DJ. Seamless proteins tie up their loose ends. Science 311(5767):1563–1564 (2006)

[CSdA+05] Cortes J, Simeon T, de Angulo R, Guieysse D, Remaud-Simeon M, and Tran V. A path planning approach for computing large-amplitude motions of flexible molecules. Bioinformatics 21(S1):116–125 (2005)

[CSJD04] Coutsias E, Seok C, Jacobson M, and Dill K. A kinematic view of loop closure. J Comput Chem 25(4):510–528 (2004)

[CSL02] Cortes J, Simeon T, and Laumond JP. A random loop generator for planning the motions of closed kinematic chains using PRM methods. In IEEE International Conference on Robotics and Automation, vol. 2, 2141–2146 (2002)

[CSLS04] Czaplewski C, Stanislaw O, Liwo A, and Scheraga HA. Prediction of the structures of proteins with the UNRES force field, including dynamic formation and breaking of disulfide bonds. Protein Eng Des Sel 17(1):29–36 (2004)

[CSRST04] Cortes J, Simeon T, Remauld-Simeon M, and Tran V. Geometric algorithms for the conformational analysis of long protein loops. J Comput Chem 25(7):956–967 (2004)

[Dag00] Dagget V. Long timescale simulations. Curr Opinion Struct Biol 10(2):160–164 (2000)

[DAL03] Du PC, Andrec M, and Levy RM. Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. Protein Eng 16(6):407–414 (2003)

[DB00] Deane CM and Blundell TL. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. Proteins 40(1):135–144 (2000)

[DB01] Dominy BN and Brooks CLI. Identifying native-like protein structures using physics-based potentials. J Comput Chem 23(1):147–160 (2001)

[DC00] Dombkowski AA and Crippen GM. Disulfide recognition in an optimized threading potential. Protein Eng Des Sel 13(10):679–689 (2000)

[DCR+05] Daly NL, Chen YK, Rosengren KJ, Marx UC, Phillips ML, Waring AJ, Wang W, Lehrer RI, and Craik DJ. Built by association: oligomerization of the antiviral theta-defensin, retrocyclin-2. to be published (2005)

[DJ99] Dunbrack Jr RL. Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL. Proteins: Struct Funct Bioinf 37(S3):81–87 (1999)

[DJC97] Dunbrack Jr RL and Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 6(8):1661–1681 (1997)

[DLG04] Ding K, Louis JM, and Gronenborg AM. Insights into conformation and dynamics of protein GB1 during folding and unfolding by NMR. J Mol Biol 335(5):1299–1307 (2004)

[DMC05] Das P, Matysiak S, and Clementi C. Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes. Proc Natl Acad Sci USA 102(29):10141–10146 (2005)

[DMn04] Doshi UR and Muñoz V. The principles of α-helix formation: Explaining complex kinetics with nucleation-elongation theory. J Phys Chem B 108(24):8497–8506 (2004)

[DMS+06] Das P, Moll M, Stamati H, Kavraki LE, and Clementi C. Low-dimensional free energy landscapes of protein folding reactions by nonlinear dimensionality reduction. Proc Natl Acad Sci USA 103(26):9885–9890 (2006)

[Dod07] Dodson EJ. Computational biology: Protein predictions. Nature 450(7167):176–177 (2007)

[DRP98] Dudek MJ, Ramnarayan K, and Ponder JW. Protein structure prediction using a combination of sequence homology and global energy minimization ii. energy functions. J Comput Chem 19(5):548–573 (1998)

[DS90] Dudek M and Scheraga HJ. Protein structure prediction using a combination of sequence homology and global energy minimization. i. global energy minimization of surface loops. J Comput Chem 11(1):121–151 (1990)

[DS91] Diederichs K and Schulz GE. The refined structure of the complex between adenylate kinase from beef heart mitochondrial matrix and its substrate AMP at 1.85 Å resolution. J Mol Biol 217(3):541–549 (1991)

[DW94] Derrick JP and Wigley DB. The third IgG-binding domain from streptococcal protein G. An analysis by X-ray crystallography of the structure alone and in a complex with Fab. J Mol Biol 243(5):906–918 (1994)

[DWC+03] Duan Y, Wu C, Chowdhury S, Lee MC, Xiong GM, Zhang W, Yang R, Cieplak P, Luo R, Lee T, Caldwell J, Wang JM, and Kollman P. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J Comput Chem 24(16):1999–2012 (2003)

[EGE95] Elofsson A, Grand SL, and Eisenberg D. Local moves: an efficient algorithm for simulation of protein folding. Proteins: Struct Funct Genet 23(1):73–82 (1995)

[Elb05] Elber R. Long-timescale simulation methods. Curr Opinion Struct Biol 15(2):151–156 (2005)

[EMCG95] Evans JS, Mathiowetz AM, Chan SI, and Goddard WA. De novo prediction of polypeptide conformations using dihedral probability grid Monte Carlo methodology. Protein Sci 4(6):1203–1216 (1995)

[EML+05] Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, Skalicky JJ, Kay LE, and Kern D. Intrinsic dynamics of an enzyme underlies catalysis. Nature 438(7064):117–121 (2005)

[ESK+02] Eicken C, Sharma V, Klabunde T, Lawrenz MB, Hardham JM, Norris SJ, and Sacchettini JC. Crystal structure of lyme disease variable surface

antigen VlsE of borrelia burgdorferi. J Biol Chem 277(24):21691–21696 (2002)

[FA95] Frishman D and Argos P. Knowledge-based protein secondary structure assignment. Proteins: Struct Funct Genet 23(4):566–579 (1995)

[FC05] Ferre F and Clote P. DIANNA: a web server for disulfide connectivity prediction. Nucl Acids Res 33:W230–W232 (2005)

[FCTS92] Fiser A, Cserzo M, Tudos E, and Simon I. Different sequence environments of cysteines and half cystines in proteins. Application to predict disulfide forming residues. FEBS Lett 302(2):117–120 (1992)

[FDS00] Fiser A, Do RK, and Sali A. Modeling of loops in protein structures. Protein Sci 9(9):1753–1773 (2000)

[FED+95] Finn BE, Evenäs J, Drakenberg T, Waltho JP, Thulin E, and Forsén S. Calcium-induced structural changes and domain autonomy in calmodulin. Nat Struct Biol 2(9):777–783 (1995)

[Fer99] Fersht AR. Structure and Mechanism in Protein Science. A Guide to Enzyme Catalysis and Protein Folding. W. H. Freeman and Co., New York, NY, 3 edition (1999)

[FFD90] Frantz DD, Freeman DL, and D DJ. Reducing quasi-ergodic behavior in Monte Carlo simulations by j-walking. J Chem Phys 93(4):2769–2784 (1990)

[FHHQ05] Fallon JL, Halling DB, Hamilton SL, and Quiocho FA. Structure of calmodulin bound to the hydrophobic IQ domain of the cardiac Ca(v)1.2 calcium channel. Structure 13(12):1881–1886 (2005)

[FR92] Finkelstein AV and Reva BA. Search for the stable state of a short chain in a molecular field. Protein Eng 5(7):617–624 (1992)

[FR05] Fleming PJ and Rose GD. Protein Folding Handbook. Wiley-VCH, Weinheim, Germany, 1st edition (2005)

[FS88] Ferrenberg AM and Swendsen RH. New Monte Carlo technique for studying phase transitions. Phys Rev Lett 61(23):2635–2638 (1988)

[FS89] Ferrenberg AM and Swendsen RH. Optimized Monte Carlo data analysis. Phys Rev Lett 63(12):1185–1198 (1989)

[FSBM94] Fidelis K, Stern PS, Bacon D, and Moult J. Comparison of systematic search and database methods for constructing segments of protein structure. Protein Eng 7(8):953–960 (1994)

[FW94] Frauenfelder H and Wolynes PG. Biomolecules: Where the physics of complexity and simplicity meet. Physics Today 47(2):58–64 (1994)

[FWS+86] Fine RM, Wang HJ, Shenkin P, Yarmush D, and Levinthal C. Predicting antibody hypervariable loop conformations. ii: Minimization and molecular dynamics studies of mcpc603 from many randomly generated loop conformations. Proteins 1(4):342–362 (1986)

[GFR05] Gong H, Fleming PJ, and Rose GD. Building native protein confrmation from highly approximate backbone torsion angles. Proc Natl Acad Sci USA 102(45):16227–16232 (2005)

[GK97] Grosberg A and Khokhlov A. Giant molecules: here, there, and everywhere. Academic Presss, 1st edition (1997)

[GP02] Gullingsrud J and Phillips J. PSFGEN User's Guide. Technical report, University of Illinois at Urbana-Champaign (2002). http://www.ks.uiuc.edu/Research/vmd/plugins/psfgen/

[Gru02] Gruebele M. Protein folding: the free energy surface. Curr Opinion Struct Biol 12(2):161–168 (2002)

[GS70] Go N and Scheraga HJ. Ring closure and local conformational deformations of chain molecules. Macromolecules 3(2):178–187 (1970)

[HA01] Han L and Amato NM. A kinematics-based probabilistic roadmap method for closed chain systems. In Donald BR, Lynch KM, and Rus D, eds., Algorithmic and Computational Robotics: New Directions, 233–246. AK Peters, MA (2001)

[Han97] Hansmann UHE. Parallel tempering algorithm for conformational studies of biological molecules. Chem Phys Lett 281(1-3):140–150 (1997)

[HAO+06] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, and Simmerling C. Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins: Struct Funct Bioinf 65(3):712–725 (2006)

[HCG92] Higo J, Collura V, and Garnier J. Development of an extended simulated annealing method : application to the modeling of complementary determining regions of immunoglobulins. Biopolymers 32(1):33–43 (1992)

[HDS96] Humphrey W, Dalke A, and Schulten K. VMD - Visual Molecular Dynamics. J Mol Graph Model 14(1):33–38 (1996). http://www.ks.uiuc.edu/Research/vmd/

[Hes02] Hess B. Convergence of sampling in protein simulations. Phys Rev E Stat Nonlin Soft Matter Phys 65(3.1):1–10 (2002)

[HF96] Hilser VJ and Freire E. Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors. J Mol Biol 262(5):756–772 (1996)

[HF03] Hall JB and Fushman D. Characterization of the overall and local dynamics of a protein with intermediate rotational anisotropy: Differentiating between conformational exchange and anisotropic diffusion in the B3 domain of protein G. J Biomol NMR 27(3):261–275 (2003)

[HGW92] Hyberts SG, Goldberg MS, and Wagner G. The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with X-ray structures. Protein Sci 1(6):736–751 (1992)

[HKC07] Heath AP, Kavraki LE, and Clementi C. From coarse-grain to all-atom: Towards multiscale analysis of protein landscapes. Proteins: Struct Funct Bioinf 68(3):646–661 (2007)

[HLSW99] Hardin C, Luthey-Schulten Z, and Wolynes PG. Backbone dynamics, fast folding, and secondary structure formation in helical proteins and peptides. Proteins: Struct Funct Genet 34(3):281–294 (1999)

[HM05] Huang YJ and Montellione GT. Structural biology: Proteins flex to function. Nature 438(7064):36–37 (2005)

[HODF98] Hilser VJ, Oas T, Dowdy D, and Freire E. The structural distribution of cooperative interactions in proteins: Analysis of the native state ensemble. Proc Natl Acad Sci USA 95(17):9903–9908 (1998)

[Hog03] Hogg PJ. Disulfide bonds as switches for protein function. Trends Biochem Sci 28(4):210–214 (2003)

[HOvG02] Hansson T, Oostenbrink C, and van Gunsteren WF. Molecular dynamics simulations. Curr Opinion Struct Biol 12(2):190–196 (2002)

[HTvG94] Huber T, Torda AE, and van Gunsteren WF. Local elevation: a method for improving the searching properties of molecular dynamics simulation. J Comput Aided Mol Design 8(6):695–708 (1994)

[JDC87] Jain AK, Dubes RC, and Chen CC. Bootstrap techniques for error estimation. IEEE Trans Pattern Analysis and Machine Intelligence 9(55):628–633 (1987)

[JdCW+97] Johansson MU, de Cháteau M, Wikström M, Forsén S, Drakenberg T, and Björck L. Solution structure of the albumin-binding GA module: a versatile bacterial protein domain. J Mol Biol 266(5):859–865 (1997)

[JF94] Jackson SE and Fersht AR. Contribution of residues in the reactive site loop of chymotrypsin inhibitor 2 to protein stability and activity. Biochemistry 33(46):13880–13887 (1994)

[JMe+93] Jackson SE, Moracci M, elMasry N, Johnson CM, and Fersht AR. Effect of cavity-creating mutations in the hydrophobic core of chymotrypsin inhibitor 2. Biochemistry 32(42):11259–11269 (1993)

[JNE+02] Johansson MU, Nilson H, Evenäs J, Forsén S, Drakenberg T, Björck L, and Wikström M. Differences in backbone dynamics of two homologous bacterial albumin-binding modules: implications for binding specificity and bacterial adaptation. J Mol Biol 316(5):1036–1099 (2002)

[JPR+04] Jacobson PJ, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, and Friesner RA. A hierarchical approach to all-atom protein loop prediction. Proteins: Struct Funct Bioinf 55(2):351–367 (2004)

[JRKT01] Jacobs DJ, Rader AJ, Kuhn LA, and Thorpe MF. Protein flexibility predictions using graph theory. Proteins: Struct Funct Bioinf 44(2):150–165 (2001)

[JT86] Jones TA and Thirup S. Using known substructures in protein model building and crystallography. EMBO 5(4):819–822 (1986)

[JT05] Jorgensen WL and Tirado-Rives J. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. Proc Natl Acad Sci USA 102(19):6665–6670 (2005)

[JVP06] Jayachandran G, Vishal V, and Pande VS. Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece. J Chem Phys 124(16):164902–164914 (2006)

[Kay05] Kay LE. NMR studies of protein structure and dynamics. J Magn Reson 173(2):193–207 (2005)

[KBW96] Koradi R, Billeter M, and Wuthrich K. MOLMOL - a program for display and analysis of macromolecular structures. J Mol Graph Model 14(1):51–55 (1996). http://www.bruker-biospin.de/NMR/nmrsoftw/prodinfo/molmol/

[KD95] Koehl P and Delarue M. New mean field self-consistent formalism providing simultaneously both gap closure and side chain positioning in protein homology modelling. Nat Struct Biol 2(2):163–170 (1995)

[KGLK05] Kolodny R, Guibas L, Levitt M, and Koehl P. Inverse kinematics in biology: the protein loop closure problem. Int J Robot Res 24(2-3):151–163 (2005)

[KGV83] Kirkpatrick S, Gelatt CD, and Vecchi MP. Optimization by simulated annealing. Science 220(4598):671–680 (1983)

[KH05] Kwak W and Hansmann UH. Efficient sampling of protein structures by model hopping. Phys Rev Lett 95(13):138102 (2005)

[KK05] Karplus M and Kuriyan J. Molecular dynamics and protein function. Proc Natl Acad Sci USA 102(19):6679–6685 (2005)

[KLV74] Kalos MH, Levesque D, and Valleau L. Helium at zero temperature with hard-sphere and other forces. Phys Rev A 9(5):2178–2195 (1974)

[KPPR00] Krause K, Pineda LF, Peteranderl R, and Reissman S. Conformational properties of a cyclic peptide bradykinin B-2 receptor antagonist using experimental and theoretical methods. J Pept Res 55(1):63–71 (2000)

[KRB+93] Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, and Kollman PA. The weighted histogram analysis method for free-energy calculations on biomolecules: I. The method. J Comput Chem 13(8):1011–1021 (1993)

[KSLO96a] Kavraki LE, Svetska P, Latombe JC, and Overmars M. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. IEEE T Robotic Autom 12(4):566–580 (1996)

[KSLO96b] Kavraki LE, Svetska P, Latombe JC, and Overmars M. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. IEEE Transaction on Robotics and Automation 12:566–580 (1996)

[KTG+95] Kuboniwa H, Tjandra N, Grzesiek S, Ren H, Klee CB, and Bax A. Solution structure of calcium-free calmodulin. Nat Struct Biol 2(9):768–776 (1995)

[KZ03] Kern D and Zuiderweg ER. The role of dynamics in allosteric regulation. Curr Opinion Struct Biol 13(6):748–757 (2003)

[KZU+04] Kim S, Zhang Z, Upchurch S, Isern N, and Chen Y. Structure and DNA-binding sites of the SWI1 AT-rich interaction domain (ARID) suggest determinants for sequence-specific DNA recognition. J Biol Chem 279(16):16670–16676 (2004)

[LD94] Li A and Daggett V. Characterization of the transition state of protein unfolding by use of molecular dynamics: chymotrypsin inhibitor 2. Proc Natl Acad Sci USA 91(22):10430–10434 (1994)

[Lee93] Lee Y. New Monte Carlo algorithm: entropic sampling. Phys Rev Lett 71(2):211–214 (1993)

[Lev92] Levitt M. Accurate modeling of protein conformation by automatic segment matching. J Mol Biol 226(2):507–533 (1992)

[LFKL00] LaValle SM, Finn PW, Kavraki LE, and Latombe JC. A randomized kinematics-based approach to pharmacophore-constrained conformational search and database screening. J Comput Chem 21(9):731–747 (2000)

[LGS+03] Loiseau N, Gomis JM, Santolini J, Delaforge M, and Andre F. Predicting the conformational states of cyclic tetrapeptides. Biopolymers 69(3):363–385 (2003)

[LJC95] Linse S, Jonsson B, and Chazin WJ. The effect of protein concentration on ion binding. Proc Natl Acad Sci USA 92(11):4748–4752 (1995)

[LLBD+05] Lindorff-Larsen K, Best RB, DePristo MA, Dobson CM, and Vendruscolo M. Simultaneous determination of protein structure and dynamics. Nature 433(7022):128–132 (2005)

[LLL99] Li W, Liu Z, and Lai L. Protein loops on structurally similar scaffolds: Database and conformational analysis. Biopolymers 49(6):481–495 (1999)

[Lot04] Lotan I. Algorithms exploiting the chain structure of proteins. Ph.D. thesis, Stanford University (2004)

[LRO07] Lee D, Redfern O, and Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8(12):995–1005 (2007)

[LRR+99] Li X, Romero P, Rani M, Dunker AK, and Obradovic Z. Predicting protein disorder for N-, C-, and internal regions. Genome Informatics 10:30–40 (1999)

[LRR+01] Li X, Romero P, Rani M, Dunker AK, and Obradovic Z. Sequence complexity of disordered protein. Proteins: Struct Funct Bioinf 42(1):38–48 (2001)

[LS82] Lipari G and Szabo A. Model-free approach to the interpretation of nuclear magnetic resonance relaxation in macromolecules. 1. Theory and range of validity. J Am Chem Soc 104(17):4546–4559 (1982)

[LS89a] Lambert MH and Scheraga HA. Pattern recognition in the prediction of protein structure. i. tripeptide conformational probabilities calculated from the amino acid sequence. J Comput Chem 10(6):770–797 (1989)

[LS89b] Lambert MH and Scheraga HA. Pattern recognition in the prediction of protein structure. ii. chain conformation from a probability-directed search procedure. J Comput Chem 10(6):798–816 (1989)

[LS89c] Lambert MH and Scheraga HA. Pattern recognition in the prediction of protein structure. iii. an importance-sampling minimization procedure. J Comput Chem 10(6):817–831 (1989)

[LS94] Lessel U and Schomburg D. Similarities between protein structures. Protein Eng 7(10):1175–1187 (1994)

[LSB05] Lee A, Streinu I, and Brock O. A methodology for efficiently sampling the conformation space of molecular structures. J Phys Biol 2(4):108–S115 (2005)

[LSL82] Lipari G, Szabo A, and Levy RM. Protein dynamics and NMR relaxation: comparison of simulations with experiment. Nature 300(5888):197–198 (1982)

[Lue84] Luenberger DG. Linear and Nonlinear Programming. Addison-Wesley, 2nd edition (1984)

[LvdBDL04] Lotan I, van den Bedem H, Deacon AM, and Latombe JC. Computing protein structures from electron density maps: The mising loop problem. In Erdman M, Hsu D, Overmars M, and van der Stappen F, eds., Algorithmic Foundations of Robotics VI, 153–168. Springer STAR Series (2004)

[LW08] Lu Q and Wang J. Single molecule conformational dynamics of adenylate kinase: energy landscape, structural correlations, and transition state ensembles. J Am Chem Soc 130(14):4772–4783 (2008)

[LYZ06] Lyman E, Ytreberg FM, and Zuckermann DM. Resolution exchange simulations. Phys Rev Lett 96(2):028105 (2006)

[Mac04] Mackerell AD. Empirical force fields for biological macromolecules: Overview and issues. J Comput Chem 25(13):1584–1604 (2004)

[MB04] Ming D and Brueschweiler R. Prediction of methyl-side chain dynamics in proteins. J Biomol NMR 29(3):363–368 (2004)

[MBB+98] Mackerell JAD, Bashford D, Bellot M, L DJR, Evanseck JD, J FM, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, K LFT, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher III WE, Roux B, Schlenkrich B, Smith JC, Stote RH, Straub J, Wiórkiewicz-Kuczera J, Yin D, and Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102(18):3586–3616 (1998)

[MC94] Manocha D and Canny J. Efficient inverse kinematics for general 6r manipulator. IEEE T Robotic Autom 10(5):648–657 (1994)

[MC04] Matysiak S and Clementi C. Optimal combination of theory and experiment for the characterization of the protein folding landscape of S6: How far can a minimalist model go? J Mol Biol 343(8):235–248 (2004)

[MC06] Matysiak S and Clementi C. Minimalist protein model as a diagnostic tool for misfolding and aggregation. J Mol Biol 363(1):297–308 (2006)

[MCP+07] Matysiak S, Clementi C, Praprotnik M, Kremer K, and Delle Site L. Modeling diffusive dynamics in adaptive resolution simulation of liquid water. J Chem Phys 128(2):024503 (2007)

[Mea88] Means AR. Molecular mechanisms of action of calmodulin. Recent Prog Horm Res 44:223–262 (1988)

[Mez94] Meza JC. OPT++: An object-oriented class library for nonlinear optimization. Technical Report SAND94-8225, Sandia National Laboratories (1994). http://csmr.ca.sandia.gov/opt++/

[MFC04] Martelli PL, Fariselli P, and Casadio R. Prediction of disulfide-bonded cysteines in proteomes with a hidden neural network. Proteomics 4(6):1665–1671 (2004)

[MFK+07] Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, and Tramontano A. Critical assessment of methods of protein structure prediction (CASP) round VII. Proteins: Struct Funct Bioinf 69(S8):3–9 (2007)

[MFZH03] Moult J, Fidelis K, Zemla A, and Hubbard T. Critical assessment of methods of protein structure prediction (CASP) round V. Proteins: Struct Funct Bioinf 53(S6):334–339 (2003)

[MGHT02] Mucchielli-Giorgi MH, Hazout S, and Tuffery P. Predicting the disulfide bonding state of cysteines using protein descriptors. Proteins: Struct Funct Bioinf 46(3):243–249 (2002)

[MHBC92] Main AL, Harvey TS, Baron MJ, and Campbell ID. The three-dimensional structure of the tenth type III module of fibronectin: an insight into RGD-mediated interactions. Cell 71(4):671–678 (1992)

[MJ86] Moult J and James MNG. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1(2):146–163 (1986)

[MJ93] McGarrah DB and Judson RS. Analysis of the genetic algorithm method of molecular conformation determination. J Comput Chem 14(11):1385–1395 (1993)

[MK84] Manalan AS and Klee C. Calmodulin. Advan Cyclic Nucleot Protein Phosphoryl Res 18:227–278 (1984)

[MK04] Mittermaier A and Kay LE. The response of internal dynamics to hydrophobic core mutations in the SH3 domain from the Fyn tyrosine kinase. Protein Sci 13(4):1088–1099 (2004)

[MKS97] Milik M, Kolinski A, and Skolnick J. Algorithm for rapid reconstruction of protein backbone from alpha carbon coordinates. J Comput Chem 18(1):80–85 (1997)

[MM00] Malek R and Mousseau N. Dynamics of Lennard-Jones clusters: A characterization of the activation-relaxation technique. Phys Rev E 62(6):7723–7728 (2000)

[MnS97] Muñoz V and Serrano L. Development of the multiple sequence approximation within the Agadir model of α-helix formation. Comparison with Zimm-Bragg and Lifson-Roig formalisms. Biopolymers 41(5):495–509 (1997)

[MnTHE97] Muñoz V, Thompson PA, Hofrichter J, and Eaton WA. Folding dynamics and mechanism of beta-hairpin formation. Nature 390(6656):196–199 (1997)

[MPBR96] Morton CJ, Pugh DJ, Brown EL, and Renzoni DA. Solution structure and peptide binding of the SH3 domain from human Fyn. Structure 4(6):705–714 (1996)

[MRR+53] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, and Teller E. Equation of state calculations by fast computing machines. J Chem Phys 21(6):1087–1092 (1953)

[MSRS96] Müller CW, Schlauderer GJ, Reinstein J, and Schulz GE. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. Structure 4(2):147–156 (1996)

[MT96] Martin ACR and Thornton JM. Structural families in loops of homologous proteins – automatic classification, modeling and application to antibodies. J Mol Biol 263(5):800–815 (1996)

[MTR+98] Morea V, Tramontano A, Rustici M, Chothia C, and Lesk AM. Conformations of the third hypervariable region in the VH domain of immunoglobulins. J Mol Biol 275:269–294 (1998)

[MZ94] Manocha D and Zhu Y. Kinematic manipulation of molecular chains subject to rigid constraints. In Altman RB, Brutlag DL, Karp PD, Lathrop RH, and Searls DB, eds., Proc Int Conf Intell Sys Mol Biol (ISMB), vol. 2, 285–293. AAAI, Stanford, CA (1994)

[MZW95] Manocha D, Zhu Y, and Wright W. Conformational analysis of molecular chains using nano-kinematics. Comput Appl Biosci 11(1):71–86 (1995)

[NB08] Niimura N and Bau R. Neutron protein crystallography: beyond the folding structure of biological macromolecules. Acta Crystallogr A 64(1):12–22 (2008)

[NH07] Nigham A and Hsu D. Protein conformational flexibility analysis with noisy data. In Research in Computational Molecular Biology, vol. 4453, 396–411. Singer-Verlag Berlin/Heidelberg, Berlin, Heidelberg (2007)

[NHK00] Nakajima N, Higo J, and Kidera A. Free energy landscapes of short peptides by enhanced conformational sampling. J Mol Biol 296(1):197–216 (2000)

[NN03] Norberg J and Nilsson L. Advances in biomolecular simulations: methodology and recent applications. Quart Rev Biophys 36(3):257–306 (2003)

[oBN70] on Biochemical Nomenclature IIC. Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules. Biochemistry 245(24):6489–6497 (1970)

[OCL+05] Oldziej S, Czaplewski C, Liwo A, Chinchio M, Nanias M, Vila JA, Khalili M, Arnautova YA, Jagielska A, Makowski M, Schafroth HD, Kazmierkiewicz R, Ripoll DR, Pillardy J, Saunders JA, Kang YK, Gibson KD, and Scheraga H. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. Proc Natl Acad Sci USA 102(21):7547–7552 (2005)

[OD90] O'Neal KT and DeGrado WF. How calmodulin binds its targets : sequence independent recognition of amphiphilic α-helices. Trends Biochem Sci 15(2):59–64 (1990)

[oGMS05] of General Medical Sciences NI. Protein Structure Initiative (2005). http://www.nigms.nih.gov/Initiatives/PSI/

[OKT+06] Okazaki K, Koga N, Takada S, Onuchic JN, and Wolynes PG. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. Proc Natl Acad Sci USA 103(32):11844–11849 (2006)

[OLSW97] Onuchic JN, Luthey-Schulten Z, and Wolynes PG. Theory of protein folding: the energy landscape perspective. Annual Review of Physical Chemistry 48:545–600 (1997)

[PB02] Price DJ and Brooks CLI. Modern protein force fields behave comparably in molecular dynamics simulations. J Comput Chem 23(11):1045–1057 (2002)

[PC03] Ponder JW and Case DA. Force fields for protein simulations. Adv Protein Chem 66:27–85 (2003)

[PFNG06] Project E, Friedman R, Nachliel E, and Gutman M. A molecular dynamics study of the effect of Ca^{2+} removal on calmodulin structure. Biophys J 90(11):3842–3850 (2006)

[PFO99] Peters GH, Frimurer TM, and Olsen OH. Molecular dynamics simulations of protein-tyrosine phosphatase 1B. I. Ligand-induced changes in the protein motions. Biophys J 77(1):505–515 (1999)

[PHE+06] Prentiss MC, Hardin C, Eastwood MP, Zong C, and Wolynes PG. Protein structure prediction: The next generation. J Chem Theory Comput 2(3):705–716 (2006)

[Pic04] Pickart CM. Back to the future with ubiquitin. Cell 116(2):181–190 (2004)

[PKL01] Palmer AGI, Kroenke CD, and Loria JP. Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. Methods Enzymol 339:204–238 (2001)

[PM95] Pedersen JT and Moult J. Ab initio structure prediction for small polypeptides and protein fragments using genetic algorithms. Proteins 23(3):454–460 (1995)

[PMDS+07] Praprotnik M, Matysiak S, Delle Site L, Kremer K, and Clementi C. Adaptive resolution simulation of liquid water. J Phys: Condens Matter 19:292201 (2007)

[Pri86] Primrose EJF. On the input-output equation of the general 7R- mechanism. Mech Mach Theory 21:509–510 (1986)

[PS91] Palmer KA and Scheraga HA. Standard-geometry chains fitted to x-ray derived structures : Validation of the rigid-geometry approximation. i. chain closure through a limited search of "loop" conformations. J Comput Chem 12(4):505–526 (1991)

[PSB98] Plaxco KW, Simmons KT, and Baker D. Contact order, transition state. placement, and the refolding rates of single domain proteins. J Mol Biol 277(4):985–994 (1998)

[PSCK07] Plaku E, Stamati H, Clementi C, and Kavraki LE. Fast and reliable analysis of molecular motions using proximity relations and dimensionality reduction. Proteins: Struct Funct Bioinf 67(4):897–907 (2007)

[PUE+04] Papoian GA, Ulander J, Eastwood MP, Luthey-Schulten Z, and Wolynes PG. Water in protein structure prediction. Proc Natl Acad Sci USA 101(10):3352–3357 (2004)

[QPKM01] Quik M, Polonskaya Y, Kulak J, and McIntosh JM. Vulnerability of ^{125}I-α-conotoxin MII sites to nigrostriatal damage in monkey. J Neurosci 21(15):5494–5500 (2001)

[RDCB97] Rufino SD, Donate LE, Canard LHJ, and Blundell TL. Predicting the conformational class of short and medium size loops connecting regular secondary structures: Application to comparative modeling. J Mol Biol 267(2):352–367 (1997)

[RF99] Rapp CS and Friesner RA. Prediction of loop geometries using a generalized born model of solvation effects. Proteins 35(2):173–183 (1999)

[RL68] Rhoads DG and Lowenstein JM. Initial velocity and equilibrium kinetics of myokinase. J Biol Chem 243(14):3963–3972 (1968)

[ROW+07] Roe DR, Okur A, Wickstrom L, Hornak V, and Simmerling C. Secondary structure bias in generalized born solvent models: Comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation. J Phys Chem 11(7):1846 –1857 (2007)

[RR89] Raghavan M and Roth B. Kinematic analysis of the 6R manipulator of general geometry. In Int. Symp. Robot. Res., 314–320 (1989)

[RRS63] Ramachandran GN, Ramakrishnan C, and Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol 7:95–99 (1963)

[RSA93] Ren J, Stuart DI, and Acharya KR. Alpha-lactalbumin possesses a distinct zinc binding site. J Biol Chem 268(26):19292–19298 (1993)

[RSG04] Rayan A, Senderowitz H, and Goldblum A. Exploring the conformational space of cyclic peptides by a stochastic search method. J Mol Graphics Modell 22(5):319–333 (2004)

[RT93] Rao U and Teeter MM. Improvement of turn prediction by molecular dynamics: A case study of a1-purothionin. Protein Eng 6(8):837–847 (1993)

[RVS04] Ripoll DR, Vila JA, and Scheraga HA. Folding of the villin headpiece subdomain from random structures. Analysis of the charge distribution as a function of pH. J Mol Biol 339(4):915–925 (2004)

[SBK91] Sjöbring C, Björck L, and Kaster W. Streptococcal protein G - gene structure and protein binding properties. J Biol Chem 266(1):399–405 (1991)

[SC04] Sun L and Chen ZJ. The novel functions of ubiquitination in signaling. Curr Opinion Struct Biol 16(3):119–126 (2004)

[Sch44] Schroedinger E. What is life? Cambridge University Press (1944)

[Sch58] Schellman JA. The factors affecting the stability of hydrogen-bounded polypeptide structures in solution. J Phys Chem 62(12):1485–1494 (1958)

[SCK06] Shehu A, Clementi C, and Kavraki LE. Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations. Proteins: Struct Funct Bioinf 65(1):164–179 (2006)

[SCK07] Shehu A, Clementi C, and Kavraki LE. Sampling conformation space to model equilibrium fluctuations in proteins. Algorithmica 48(4):303–327 (2007)

[SDW04] Schnell JR, Dyson HJ, and Wright PE. Structure, dynamics, and catalytic function of dihydrofolate reductase. Annu Rev Biophys and Biomolec Struct 33(1):119–140 (2004)

[SGE00] Schulze BG, Grubmueller H, and Evanseck JD. Functional significance of hierarchical tiers in carbonmonoxy myoglobin: conformational substates and transitions studied by conformational flooding simulations. J Am Chem Soc 122(36):8700–8711 (2000)

[SK90] Summers NL and Karplus M. Modeling of globular proteins. A distance-based data search procedure for the construction of insertion/deletion regions and Pro - non-Pro mutations. J Mol Biol 216(4):991–1016 (1990)

[SK97] Sawaya MR and Kraut J. Loop and domain movements in the mechanism of E. Coli Dihydrofolate Reductase: Crystallographic evidence. Biochemistry 36(3):586–603 (1997)

[SKC95] Skelton NJ, Kördel J, and Chazin WJ. Determination of the solution structure of apo calbindin D_{9k} by NMR spectroscopy. J Mol Biol 249(2):441–462 (1995)

[SKC07] Shehu A, Kavraki LE, and Clementi C. On the characterization of protein native state ensembles. Biophys J 92(5):1503–1511 (2007)

[SKC08a] Shehu A, Kavraki LE, and Clementi C. Multiscale characterization of protein conformational ensembles. Structure (2008). Submitted

[SKC08b] Shehu A, Kavraki LE, and Clementi C. Unfolding the fold of cyclic cysteine-rich peptides. Protein Sci 17(3):482–493 (2008)

[SKO97] Skolnick J, Kolinski A, and Ortiz AR. MONSSTER: a method for folding globular proteins with a small number of distance restraints. J Mol Biol 265(2):217–241 (1997)

[SLB99] Singh AP, Latombe JC, and Brutlag DL. A motion planning approach to flexible ligand binding. In Schneider R, Bork P, Brutlag DL, Glasgow JI, Mewes HW, and Zimmer R, eds., Proc Int Conf Intell Sys Mol Biol (ISMB), vol. 7, 252–261. AAAI, Heidelberg, Germany (1999)

[SP00] Shirts M and Pande VJ. COMPUTING: Screen savers of the world unite! Science 290(5498):1903–1904 (2000)

[SPS96] Schlauderer GJ, Proba K, and Schulz GE. Structure of a mutant adenylate kinase ligated with an ATP-analogue showing domain closure over ATP. J Mol Biol 256(2):223–227 (1996)

[SQH07] Snow C, Qi G, and Hayward S. Essential dynamics sampling study of adenylate kinase: comparison to citrate syynthase and implication for the hinge and shear mechanisms of domain motions. Proteins: Struct Funct Bioinf 67(2):325–337 (2007)

[SSB05] Smith GR, Sternberng MJE, and Bates PA. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. J Mol Biol 347(5):1077–1101 (2005)

[SSP04] Singhal N, Snow CD, and Pande VS. Using path sampling to build better markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. J Chem Phys 121(1):415–425 (2004)

[STF92] Svensson LA, Thulin E, and Forsén S. Proline cis-trans isomers in calbindin D_{9k} observed by X-ray crystallography. J Mol Biol 223(3):601–606 (1992)

[STHH90] Still WC, Tempczyk A, Hawley RC, and Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 112(16):6127–6129 (1990)

[SV04] Shepherd CM and Vogel HJ. A molecular dynamics study of the Ca^{2+}-Calmodulin: Evidence of interdomain coupling and structural collapse on the nanosecond timescale. Biophys J 87(2):780–791 (2004)

[SVW04] Sommese AJ, Verschelde J, and Wampler CW. Advances in polynomial continuation for solving problems in kinematics. J Mech Design 126(2):262–268 (2004)

[SW84] Stillinger FH and Weber TA. Packing structures and transitions in liquids and solids. Science 228(4666):983–989 (1984)

[SW86] Swendsen RH and Wang JS. Replica Monte Carlo simulation of spin-glasses. Phys Rev Lett 57(21):2607–2609 (1986)

[SYF+87] Shenkin PS, Yarmush DL, Fine R, Wang HJ, and Levinthal C. Predicting antibody hypervariable loop conformations. i: Ensembles of random conformations for ring-like structures. Biopolymers 26(12):2053–2085 (1987)

[SYTK07] Shin SY, Yoo B, Todaro LJ, and Kirshenbaum K. Cyclic peptoids. J Am Chem Soc 129(11):3218–3225 (2007)

[Tai04] Tai K. Conformational sampling for the impatient. Biophys Chem 107(3):213–220 (2004)

[TB97] Tjandra N and Bax A. Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. Science 278(5340):1111–1114 (1997)

[TBHM02] Tossato CE, Bindewald E, Hesser J, and Maenner R. A divide and conquer approach to fast loop modeling. Protein Eng 15(4):279–286 (2002)

[TFPB95] Tjandra N, Feller SE, Pastor RW, and Bax A. Rotational diffusion anisotropy of human ubiquitin from ^{15}N NMR relaxation. J Am Chem Soc 117(50):12562–12566 (1995)

[TL92] Tramontano A and Lesk A. Common features of the conformations of antigen-binding loops in immunoglobulins and application to modeling loop conformations. Proteins 13(3):231–245 (1992)

[TME+93] Topham CM, McLeod A, Eisenmenger F, Overington JP, Johnson M, and Blundell T. Fragment ranking in modeling of protein structure: Conformationally constrained environmental amino acid substitution tables. J Mol Biol 229(1):194–220 (1993)

[TNM92] Tanner JJ, Nell LJ, and McCammon JA. Anti-insulin antibody structure and conformation. ii. molecular dynamics with explicit solvent. Biopolymers 32(1):23–31 (1992)

[TP04] Tousignant A and Pelletier JN. Protein motions promote catalysis. Chem Biol 11(8):1037–1042 (2004)

[TSC01] Trabi M, Schirra HJ, and Craik DJ. Three-dimensional structure of RTD-1, a cyclic antimicrobial defensin from Rhesus macaque leukocytes. Biochemistry 10(40):4211–4221 (2001)

[TUG75] Taketomi H, Ueda Y, , and Go N. Studies on protein folding, unfolding and fluctuations by computer simulation: The effect of specific amino acid sequence represented by specific inter-unit interactions. Int J Peptide Prot Res 7(6):445–459 (1975)

[TV77] Torrie GM and Valleau JP. Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. J Comput Phys 23(2):187–199 (1977)

[TYG+99] Tang YQ, Yuan J, George O, Osapay K, Tran D, Miller CJ, Ouellette AJ, and Selsted ME. A cyclic antimicrobial peptide produced in primate leukocytes by the ligation of two truncated α-defensins. Science 286(5439):498–502 (1999)

[TZM+97] Thanki N, Zeelen JP, Mathieu M, Laenicke R, Abagyan RA, Wierenga RK, and Schliebs W. Protein engineering with monomeric triosephosphate isomerase (monotim): the modelling and structure verification of a seven residue loop. Protein Eng 10(2):159–167 (1997)

[UM93] Unger R and Moult J. Finding lowest free energy conformation of a protein is an NP-hard problem: Proof and implications. Bull Math Biol 55(6):1183–1198 (1993)

[URDB03] Ulmer TS, Ramirez BE, Delaglio F, and Bax A. Evaluation of backbone proton positions and dynamics in a small protein by liquid crystal NMR spectroscopy. J Am Chem Soc 125(30):9179–9191 (2003)

[VCD94] Vasmatzis G, CBrower R, and DeLisi C. Predicting immunoglobulin-like hypervariable loops. Biopolymers 34(12):1669–1680 (1994)

[vdBLLD05] van den Bedem H, Lotan I, Latombe JC, and Deacon AM. Real-space protein-model completion: an inverse-kinematics approach. Acta Crystallogr D61(1):2–13 (2005)

[vGBB+06] van Gunsteren WF, Bakowies D, Baron R, Chandrasekhar I, Christen M, Daura X, Gee P, Geerke DP, Glättli A, H HP, Kastenholz MA, Oostenbrink C, Schenk M, Trzesniak D, van der Vegt NF, and Yu HB. Biomolecular modeling: Goals, problems, perspectives. Angew Chem Int Ed Engl 45(25):4064–4092 (2006)

[VKBC87] Vijay-Kumar S, Bugg CE, and Cook WJ. Structure of ubiquitin refined at 1.8 Å resolution. J Mol Biol 194(3):531–544 (1987)

[VPDK03] Vendruscolo M, Pacci E, Dobson C, and Karplus M. Rare fluctuations of native proteins sampled by equilibrium hydrogen exchange. J Am Chem Soc 125(51):15686–15687 (2003)

[vVK97] van Vlijmen HWT and Karplus M. PDB-based protein loop prediction: Parameters for selection and methods for optimization. J Mol Biol 267(4):975–1001 (1997)

[Wal04] Walker R. Amber 8 tutorial (2004). http://amber.scripps.edu/tutorial/

[WB06] Wagoner JA and Baker NA. Assessing implicit models for nonpolar mean solvation forces: The importance of dispersion and volume terms. Proc Natl Acad Sci USA 103(22):8331–8336 (2006)

[WC91] Wang LT and Chen CC. A combined optimization method for solving the inverse kinematics problem of mechanical manipulators. IEEE T Robotic Autom 7(4):489–499 (1991)

[WCB+94] Wendy DC, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, and Kollman PA. A second gener-

ation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117(19):5179–5197 (1994)

[WD03] Wang G and Dunbrack RL. Pisces: a protein sequence culling server. Bioinformatics 19(12):1589–1591 (2003)

[Wet73] Wetlaufer DB. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc Natl Acad Sci USA 70(3):697–701 (1973)

[WLH01] Wrabl JO, Larson SA, and Hilser VJ. Thermodynamic propensities of amino acids in the native state ensemble: Implications for fold recognition. Protein Sci 10(5):1032–1045 (2001)

[WM89] Wampler C and Morgan A. Solving the 6R inverse position problem using a generic-case solution methodology. Mech Mach Theory 26(1):91–106 (1989)

[WMH04] Weaver T, Maurer J, and Hayashizaki Y. Sharing genomes: an integrated approach to funding, managing and distributing genomic clone resources. Nat Rev Genet 5(11):861–866 (2004)

[WS99] Wedemeyer WJ and Scheraga HJ. Exact analytical loop closure in proteins using polynomial equations. J Comput Chem 20(8):819–844 (1999)

[XA03] Xie D and Amato NM. A kinematics-based probabilistic roadmap method for high DOF closed chain systems. In Proc. IEEE Int. Conf. Robot. Autom., vol. 1, 473–478 (2003)

[XH01] Xiang Z and Honig B. Extending the accuracy limits of prediction for side-chain conformations. J Mol Biol 311(2):421–430 (2001)

[XJ99] Xu H and J BB. Multicanonical jump walking: a method for efficiently sampling rough energy landscapes. J Chem Phys 110(21):10299–10306 (1999)

[YLK01] Yakey J, LaValle SM, and Kavraki LE. Randomized path planning for linkages with closed kinematic chains. IEEE T Robotic Autom 17(6):951–959 (2001)

[ZB97] Zhou R and Berne BJ. Smart walking: a new method for Boltzmann sampling of protein conformations. J Chem Phys 107(21):9185–9196 (1997)

[Zem03] Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucl Acids Res 31(13):3370–3374 (2003). http://as2ts.llnl.gov/AS2TS/LGA/lga.html

[ZJZ07] Zhang BW, Jasnow D, and Zuckermann DM. Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin. Proc Natl Acad Sci USA 104(46):18043–18048 (2007)

[ZK02] Zhang M and Kavraki LE. Finding solutions of the inverse kinematics problem in computer-aided drug design. In Florea L, Walenz B, and Hannenhalli S, eds., Currents in Computational Molecular Biology, TR02-385, 214–215. ACM Press, Washington, DC (2002)

[ZKS02] Zhang Y, Kihara D, and Skolnick J. Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins: Struct Funct Bioinf 48(2):192–201 (2002)

[ZLT+05] Zhao E, Liu HL, Tsai C, Tsai H, Chan CH, and Kao CY. Cysteine separations profiles on protein sequences infer disulfide connectivity. Bioinformatics 21(8):1415–1420 (2005)

[ZRDK94] Zheng Q, Rosenfeld R, DeLisi C, and Kyle DJ. Multiple copy sampling in protein loop modelling: computational efficiency and sensitivity to dihedral perturbations. Protein Sci 3(3):493–506 (1994)

[ZRVD93] Zheng Q, Rosenfeld R, Vajda S, and DeLisi C. Determining protein loop conformation using scaling-relaxation techniques. Protein Sci 2(8):1242–1248 (1993)

[ZWW+04] Zhang M, White RA, Wang L, Goldman R, Kavraki LE, and Hasset B. Improving conformational searches by geometric screening. Bioinformatics 21(5):624–630 (2004)