CONFORMATIONAL FOLDING STATUS AND FOLDING LEVELS BASED ON GLOBAL PROTEIN PROPERTIES: A COMPUTATIONAL APPROACH

Luis Garreta

Universidad del Valle

School of Computer Science EISC

October 2015

CONFORMATIONAL FOLDING STATUS AND FOLDING LEVELS BASED ON GLOBAL PROTEIN PROPERTIES: A COMPUTATIONAL APPROACH

A Dissertation

Presented to the School of Computer Science EISC

of Universidad del Valle

in Partial Fulfillment of the Requirements for the Degree of

Doctor in Engineering

(Emphasis in Computer Science)

Advisor: Irene Tischer PhD

by

Luis Garreta

October 2015

CONFORMATIONAL FOLDING STATUS AND FOLDING LEVELS BASED ON GLOBAL PROTEIN PROPERTIES: A COMPUTATIONAL APPROACH

Luis Garreta, Doctor in Engineering
Universidad del Valle 2015

ABSTRACT

Proteins during the folding process undergo many events that continuously modify their structure adjusting their structural elements, and so altering their physical properties. These events are hard to observe by experimental methods, but computer simulations can generate valuable information about the conformational states by which proteins assume their native state. The information can be used to describe and understanding the folding process, hence, researchers in the protein folding use a single or a combination of physical properties to describe how a protein folds or what degree of folding it has. But their results differ, depending on the selected properties, the particular context, and the aimed generality.

We hypothesize that folding can be described by a few features that determine how a protein folds and that are associated with the stages of folding it passes through on the way, from unfolded to folded and intermediates. The set of folding features is what we will refer to as folding status, and we propose here a computational method to combine a wide range of different physical structural and energetic properties to obtain a concise representation of a conformations folding features. Also, as the folding status is associated with the stages of folding, we propose a mechanism for determining the most plausible conformational stages assumed by a protein conformation or what we refer to as folding levels. Therefore, the objective of this thesis is to derive a computational definition of the folding status of a protein and the associated folding levels, based on physical properties of protein structures.

We use protein folding pathways generated by the Probabilistic Roadmap Method that provides us with a set of very variable protein conformations. On these pathways, we apply a Principal Component Analysis to define Inherent

Conformational Features that allow to describe a protein's folding status in a compact way. The obtained features summarize the individual properties of a protein conformation and associate with general folding characteristics of stability, compactness, and native-likeness. The features can be used to compare the conformations of a pathway or trajectory, or to compare the folding status of the conformations of different proteins. From the features, we derived the Inherent Conformational Feature Score, which condenses the three-dimensional feature status to a one-dimensional numeric value: the higher the score, the more folded is a protein conformation.

The features allow to deduce folding levels and characterize their respective conformations in a computational way. Clustering our selected conformations, we obtain four well-defined groups that we associated with four main folding levels: unfolded, early intermediate, late intermediate and folded. These folding levels agree with experimentally observed folding states of proteins. And moreover, they reflect concisely the dynamic behavior of a pathway when we represent the folding process in terms of its dynamic transitions between folding levels.

The process of evaluating the selected properties on a large set of conformations is computationally highly expensive. With the aim of offering our methods to researchers who do not dispose of a sophisticated computing system, we developed a distributed framework that runs on personal computers using a cloud service for communication. We added a toolkit for protein analysis to the framework that allows to execute all above stated tasks.

This document is dedicated to my son, Kamilo.

ACKNOWLEDGEMENTS

I would like to thank my wonderful advisor, Irene Tischer, for her support and direction, not only as my thesis advisor, but also as a great person.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Proteins are biomolecules composed of a chain of amino acid residues formed by different atoms, mainly carbons, hydrogens, nitrogens, and oxygens. Proteins perform fundamental functions in the living organisms, and these functions are directly related to the three-dimensional (3D) structure they assume when their atoms interact with each other during their *folding process*. Different forces applied to atoms cause interactions that form and break various bonds until the protein reaches its 3D structure with a stable state and with minimal energy known as its *native state*.

In order to reach the native state, a protein follows a myriad of conformational changes or folding events starting from an unfolded state, passing through different intermediate states, and ending in the native or folded state. But the presence of these intermediates has been widely debated [Privalov, 1996, Roder and Colón, 1997, Brockwell and Radford, 2007]. Anfinsen proposed them in the middle '70s, after demonstrating that proteins do not fold randomly, but following a sequence of ordered events or *protein folding pathway* [Anfinsen and Scheraga, 1975]. In the subsequent years, intermediates were the subject of many studies, but experimental works could not always provide evidence of their existence, reducing researchers' attention to their exploration [Jackson and Fersht, 1991, Jackson, 1998].

Since the ability to predict pathways helps to elucidate the folding process itself and to improve structure prediction methods, folding is studied by several techniques as for instance classic approaches using molecular dynamics simulations ([Pande et al., 2008]); sampling of the protein's conformational space with motion planning techniques [Song and Amato, 2001],[Apaydin et al., 2003]; and also unfolded approaches, as an inverse folding by breaking up successively the supersecondary structures of the native structure [Zaki et al., 2004].

Nowadays, intermediates are studied with renewed interest as improved experimental folding techniques shows their presence in the folding of many proteins [Roder et al., 2006, Lajtha et al., 2007, Udgaonkar, 2008], suggesting that folding occurs in stages [Uversky, 2002, Chalikian and Breslauer, 1996] Also,

since the ability to predict pathways helps to elucidate the folding process itself and to improve structure prediction methods, folding is studied by several techniques as simulations using molecular dynamics [Pande et al., 2008]); conformational sampling with motion planning techniques [Song and Amato, 2001]; and inverse folding by breaking up successively the supersecondary structures of the native structure [Zaki et al., 2004].

But fundamental questions about intermediates remain unsolved. It is still unclear what role they play in the protein folding or how they are characterized. Resolving these questions can potentially help to understand the factors influencing structural properties of folding, and provide a vision of how proteins fold or inclusively how they misfold, achieving a wrong function and causing degenerative diseases as the Parkinson, Alzheimer and others.

Therefore, it is important to develop methods to characterize protein conformations structurally according to these stages. In this sense, computational work can propose new elements that can help to elucidate some of the above questions. Global properties such as native contacts, radius of gyration, accessible surface area, and other structural and energetic properties can be measured on protein conformations to characterize their folding stages. For example, conformations appearing at the beginning of the folding have different characteristics than the ones appearing at the last or intermediate states.

Here, we present an approach to the computational characterization of protein folding stages based on the idea of measuring a reduced set of 'hidden' folding characteristics that result from evaluating a high-dimensional set of 'observable' structural and energetic properties of proteins. The approach follows a dimensionality reduction process to map the data into a new low-dimensional space that explains relevant information contained in the original data, and evaluates and analyzes distributively the large amount of conformations present in a set of protein folding pathways.

1.1 Thesis Statement

The different events that modify the structure of a protein during its folding process, change its physical properties. These properties, most of them directly observable in the protein structure (measurable), are associated with features more inherent to protein folding, but not always visible or directly quantifiable by experimental methods. However, computer folding simulations can generate for a protein its plausible conformational states assumed during its folding, and a set of physical properties calculated on them could derive associated inherent-folding features. These features can be used to define theoretical and computational elements to help describe and understand the folding process and to get insight in debated folding issues as pathways and intermediates.

1.2 Research Goal

Our research goal is to develop a set of theoretical and computational elements intended to define and quantify a set of folding features derived from physical properties of proteins that help to describe how proteins fold. The set of folding features is what we will refer to as folding status, and we propose here a computational method to combine a wide range of different physical structural and energetic properties to obtain a concise representation of a conformation's folding features. Also, as the folding status is associated with the stages of folding, we propose a mechanism for determining the most plausible conformational stages assumed by a protein conformation or what we refer to as folding levels.

1.3 Hypothesis

The hypothesis of this research is that by measuring observable physical properties in the structure of a protein in its route to folding, we can identify and quantify not directly measurable inherent folding features that can be used to examine protein folding from another perspective and to formulate new theoretical and computational elements.

1.4 Research Questions

Specifically, the present project addresses the following research questions:

1. Which structural and energetic properties are important to describe a protein in terms of the progress of folding? How can they be combined to obtain a concise representation of a conformation's folding features?
2. Do these folding features allow to describe the changing folding status of protein conformations along a pathway? Are they sufficiently general to compare the folding status of conformations even from different proteins?
3. Using global folding features, is it possible to derive different folding states that share common folding features? How can we describe these folding states?
4. How can we deal with the computational highly expensive processes related to evaluation and analysis of protein conformations and pathways based only on interactions between a small number of regular personal computers?

1.5 Objectives

Our research objective is to develop theoretical and computational elements to derive a computational definition of the folding status of a protein and the associated folding levels, based on physical properties of protein structures.

- Evaluate global energetic and structural properties of protein folding and define the appropriate method to extract relevant folding features.
- Define an efficient and computational description of a protein's folding status based on these properties and construct the associated similarity measure for comparing statuses of different conformations.

- Define folding levels based on the similarity measure and characterize them with respect to their folding features and their dynamic behavior in pathways.
- Develop a distributed computing framework for evaluating global properties and analyzing folding levels of conformations of pathways.

1.6 Related Work

In this section, we give an overview of related work in the areas related to characterization of protein folding, reaction coordinates, dimensionality reduction, and some of the distributed computing strategies used in protein folding. Next chapter gives more detailed background information.

1.6.1 Characterization of protein folding states

The characterization or description of the properties of protein folding states is critical for understanding how protein folding occurs. Protein folding states can be characterized structurally, energetically, and topologically. Traditionally, the characterization has involved experimental methods ([Brockwell and Radford, 2007]), but we are interested in characterizing computationally these states according to global protein properties.

A recent study reports results of molecular dynamics simulations for 12 structurally diverse proteins [Lindorff-Larsen et al., 2011] in which a measure of folding (Q) is defined to characterize the conformations as folded, unfolded, and transition states (appearing when the system transitions entirely between the previous two states). The measure is based only on the long-range native Cα-Cα contacts. They also *individually* tried other properties as the radius of gyration, the accessible surface area, and the amount of secondary structure, but the results were not as expected, differing from the ones obtained with Q. In contrast to this approach, we consider that global properties differently describe the folding process, but for a general measure of folding, they need to be

addressed collectively and not individually, as it was done. It means that we can construct a similarity measure combining not only the information of the native contacts but also the information of various global properties that may give a better measure and so a more detailed characterization.

An experimental study on a set of single-domain globular proteins characterized four conformational states: native (N), compact intermediate (CI), partially unfolded (PU), and fully unfolded (FU) by deriving a measure from two global properties: the solvent-accessible surface area (ASA) and the volume of water-inaccessible core (VIC) [Chalikian and Breslauer, 1996]. The native state was the most compact, exhibiting the lowest ASA, the largest VIC, and the most tightly packed core. An increase in the ASA and decrease of the VIC occurred with the transition from the native to the compact intermediate state ($N \rightarrow CI$). Which also happened when the native was transitioned to the other states: $N \rightarrow PU$, and $N \rightarrow FU$, with a substantial increase in the ASA, and a low decrease in the VIC. The above study provides an experimental evidence that global properties can characterize states of proteins, and more specifically four conformational states (or folding stages) with two of them as intermediates. And also, it suggests that other global properties may behave in a similar way and, therefore, may be helpful to determine the folding stage of a protein with more detail.

1.6.2 Intermediate States

Improvements in experimental methods have indicated that the presence of intermediates becomes more evident as part of the folding process. Furthermore, they have now been detected in some proteins before considered without intermediates as the called two-state folding [Udgaonkar, 2008, Brockwell and Radford, 2007]. Mainly, two types of intermediates have been observed in distinct moments of the folding process: *early and late* folding intermediates [Baldwin and Rose, 2013, Udgaonkar, 2008, Roder et al., 2006, Uversky, 2002]. The early intermediate appears in the initial stages of folding, closer to the unfolded state. And the late intermediate appears at the end of the stages of folding, closer to the native state.

Considering the previous evidence, we conjecture that folding process can

be viewed as formed by conformations in four different stages or folding levels [Uversky, 2002, Chalikian and Breslauer, 1996]: unfolded, early-intermediate, late-intermediate, and folded. Although experimentally have been hard to characterize them, mainly the intermediates, we consider that computationally may be carried out. If we measure different folding features to a large set of protein conformations about which we know their folding stage, then we could analyze the behavior of these features on each stage and get common parameters that differentiate from each other.

1.6.3 Global Properties of Proteins

Various structural and energetic properties have been studied to characterize different elements of the protein folding process. In this work, we selected a set of 16 properties that literature associates with diverse protein folding features not directly measurable from the protein structure as stability, compactness, native-likeness, and others. It suggests that the properties may be implicitly measuring these features, so they can derived measures that evaluate on a protein structure some of its implicit folding features.

The selected properties are: the native contacts, contact order, radius of gyration, accessible surface area, hydrogen bonds, voids, root-mean-square deviation (RMSD), local RMSD, residues in correct (any) secondary structure, structural score, potential energy, and dipole moment. Also, we selected properties derived from the rigidity analysis of proteins as the degrees of freedom, rigid clusters, and stressed regions.

Specifically, native contacts have been associated with native-likeness and stability [Lindorff-Larsen et al., 2011, Song and Amato, 2001, Apaydin et al., 2003, Zaki et al., 2004]. Contact order with the topological and the stability of native states [Shi et al., 2008]. The radius of gyration (RG) and accessible surface area (ASA), with the level of compaction of a protein [Lobanov et al., 2008, Liang et al., 1998]. The ASA also has been associated with protein conformational stability, and RG have been used to trace protein conformational changes [Duan and Kollman, 1998, Liang et al., 1998]. Hydrogen bonds have been found partially responsible for the stability of folded proteins [Whitford, 2005,

Fleming et al., 2006, Gong et al., 2005]. And voids were found related to the stability [Nomura et al., 2011, Cuff and Martin, 2004], and was used to assess the packing quality of proteins [Liang and Dill, 2001].

The RMSD, a measure that uses as reference the native structure, has been widely used to compare protein structures [Moult, 2006], although it has been considered an "imperfect measure" unable to evaluate similarities in local structures. So, we define local measures to compare secondary structures elements between a pair of protein conformations. They use the RMSD, the DSSP [Kabsch and Sander, 1983], and structural information using protein blocks (reduced alphabets) [M. Tyagi A. de Brevern and Offmann, 2008] . Other three measures from the rigidity analysis: degrees of freedom, rigid clusters, and stressed regions, have been linked with the analysis of rigid and flexible regions of a protein and associated with protein stability [Jacobs et al., 2002, Rader et al., 2002].

1.6.4 Dimensionality Reduction

The main goal of dimensionality reduction is to map an original high dimensional space into a space of only a few interesting dimensions [Eitrich et al., 2007]. It has been used lately in the study of protein folding simulations to find underlying structures and information, as low-dimensional free-energy landscapes [Das et al., 2006], reduced molecular dynamics simulations [Rajan et al., 2010], relevant molecular motion information [Plaku et al., 2007], and critical features of unfolding simulations [Toofanny et al., 2010a]. Dimensionality reduction has allowed to get a better interpretation and analysis of the simulation data and use them to find new knowledge about folding, while preserving as much relevant information as possible.

Rajan et al. [Rajan et al., 2010] use two techniques: principal component analysis (PCA) and a non-metric multidimensional scaling method (nMDS) to reduce the molecular dynamics trajectories of the villin headpiece protein, and they get better results than classical clustering. Although their work deals with reducing trajectories, we consider that the techniques can be applied to select and reduce a large set of global properties of proteins, a task this present work

has to deal with. We are planning to evaluate global properties on conformations of pathways and obtain common parameters that allow to describe the folding status of a protein structure and to define a similarity measure between folding statuses of different structures.

1.6.5 Distribute Computing Frameworks

Simulations of protein folding involves a huge amount of data corresponding to the myriad of conformations calculated for every time step in the simulation trajectory. Similarly, much computer time will be needed for the task of evaluation and analysis of every conformation. It is unfeasible to achieve this task in a reasonable amount of time on a single processor. Although parallel computing is a good option for this kind of tasks, it may be expensive as it can involve specialized hardware or high speed networks to connect clusters of computers.

To compensate these disadvantages, several projects have been developed on the basis of harnessing the spare clock cycles of idle machines to simulate a dedicated computing cluster. Seti@Home [Korpela et al., 2001], Genome@Home and Folding@Home [Larson et al., 2002] are large projects that follow this approach. The latter is an example for simulating distributively the folding of a protein by using 'friend' computers around the world. A server computer, having the problem data, splits it into small work units and distributes them to the 'friend' computers. These ones in turn get the data, execute the process, and return results to the server. A particular characteristic in this approach is that 'friends' have to download previously a small piece of software that runs in the background and allows synchronization with the server.

However, many of these systems are only designed to run a small and particular piece of software with one application in mind, whereas there are several and heterogeneous software for bioinformatics. A few of them are BLAST for sequence alignment; DSSP for secondary structure assignment; Gromacs for molecular dynamics simulation; naccess for calculating the accessible surface area; the bio3D library for protein structure analysis, and many others.

Because creating a small piece of software for multiple and heterogeneous

tasks may be complicated, many of these tools are developed for particular environments and may need a previous configuration. So, it is required to encapsulate all of this environment in a special piece of software. *Virtual appliances* (VP) are a suitable option proposed by VMWare and used nowadays by various virtualization systems as VirtualBox, VMWare, and Xen [Macdonell and Lu, 2006]. A VP packs one or more software applications with a reduced or adjusted ready to use operating system, and it is run optimally by a virtualization system that executes the guest appliance on a host system [David et al., 2003].

Synchronization is a major issue on a distributed system. Server and clients must communicate among themselves data packages and control tables. Recently, various companies of the cloud computing as Google, Dropbox, Copy, Ubuntu, and others, are offering different services and applications. Cloud storage is one of these services, where files are synchronized between regular PCs and the cloud servers. We have a new idea of distributed synchronization using these services as a core layer for general data communication and add over this one a particular layer for controlling and synchronizing the data between server and client computers. We have not found in our reviewed literature a similar approach, so the design of the algorithms can be based on ideas of classic algorithms in other distributed systems.

1.7 Contributions

The main contributions of the present research project are:

1. We provide a process to transform a large set of physical (structural and energetic) protein properties into a reduced set of new variables associated with protein folding characteristics.
2. The dimensionality reduction process results in three folding features that we associated with the general characteristics of stability, compactness, and native-likeness. We provide an alternative representation of a protein conformation, its folding status, using the values of its folding features as vectors in a 3-dimensional space. Through a further reduction process, we

obtain a 1-dimensional folding score that allows to represent the folding behavior of a pathway in a resumed way.

3. From this representation we canonically deduce a similarity measure between two conformations: the 3-dimensional Euclidean distance between their folding statuses. The similarity measure allows to compare the folding statuses of two conformations and to analyze protein folding pathways and trajectories from a perspective different from the traditional atomic coordinates and dihedral angles (ϕ, ψ).
4. Through a clustering process based on the similarity measure we obtain four general folding levels of protein conformations which can be associated with the experimentally observed folding stages. Assigning the conformations to their corresponding folding level opens a new perspective for the analysis of pathways or folding trajectories.
5. We provide a computer framework for evaluation and analysis of folding trajectories that includes three major components: a software toolkit for protein analysis with functions for calculating different protein measures; a distributed platform for evaluating distributively large sets of protein conformations; and methods for analysis of trajectories based on our similarity measure.

CHAPTER 2

BACKGROUND

In this chapter, we introduce some basic concepts of protein structure and protein folding. We describe the two views of the protein folding problem, the one related to the prediction of the structure, and the other related with the explanation of the inherent protein folding process. Next, we present the concept of pathways and intermediates. And finally, we describe three computational techniques emerged to study and analyze molecule motion. The next chapter will describe the data (pathways and trajectories) generated from two of these techniques and used in our work for evaluation and analysis of folding states.

2.1 Protein Structure

Proteins are linear chains of amino acids residues (residues henceforth), that adopt unique three-dimensional (3D) structures and that are involved in nearly every function of all living organisms. They are associated with functions ranging from DNA replication to muscle motion, transport of molecules inside and outside the cell, and catalysis of important reactions, among many others.

Most of the residues have a common structure of a central backbone with a main alpha carbon atom ($C\alpha$) with four parts attached to it: an amino group; a carboxyl group; a hydrogen atom; and the R group or side chain that distinguishes one amino acid from another (see Figure 2.1). Each pair of residues is connected together by forming a peptide bond that establishes the main chain or backbone of the polypeptide. Some of these bonds, the single bonds where the $C\alpha$ is participating, allow rotations that gives the protein's flexibility and so the adoption of different protein shapes.

But this view of a protein as a simple linear sequence of residues or *primary structure* is only one of the four abstractions used to describe protein's structure (see Figure 2.2). The others three: secondary, tertiary, and quaternary, involve the regularities and underlying principles that ultimately leads to their final shape and characteristic functional properties. The *secondary structure* is

Source:The author

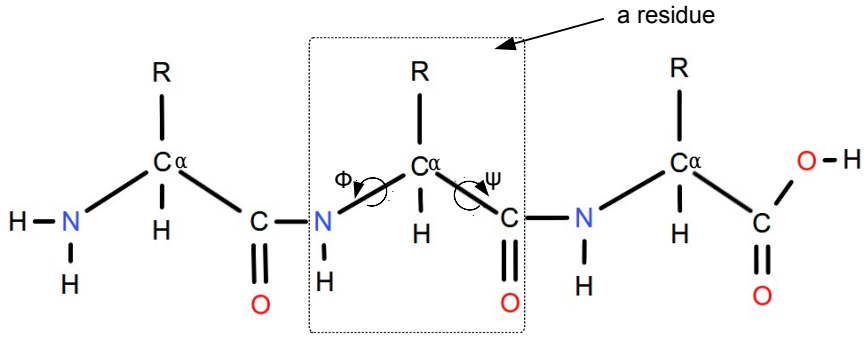


Figure 2.1: **A polypeptide as chain of amino acids residues.** Three residues linked together. The rotational freedom can be describe by torsion angles known as dihedral angles. The two major rotations associated with each residue are given by the ϕ angle (*phi*, representing the rotation along the $N - C\alpha$ bond), and the ψ angle (*psi*, representing the rotation along the $C\alpha - C$ bond).

associated with local conformations that form regular structures or secondary structure elements (SSEs) as the α helix, the β strand, and turns. The *tertiary structure* describes how these SSEs are organized spatially by folding itself the protein's final 3D shape called the *folded state* or *the native state*. And the *quaternary structure* describes the arrangement of several protein chains in a protein complex.

Source:The author

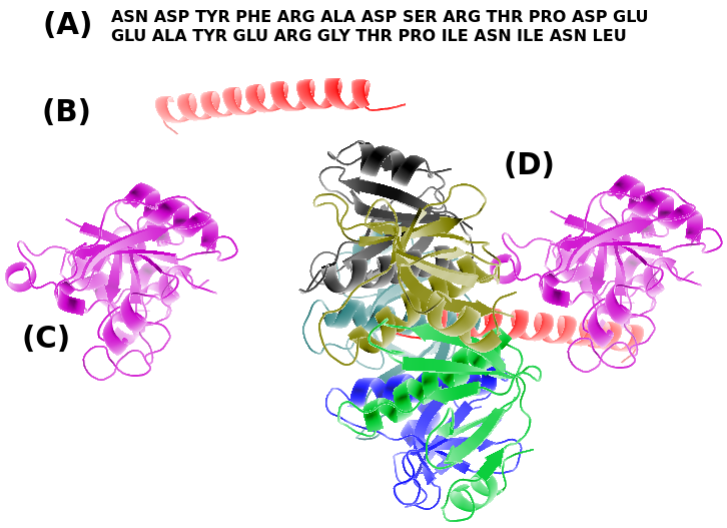


Figure 2.2: **The four abstractions of the protein structure.** (A) Primary structure (residues of the polypeptide chain). (B) Secondary structure (An alpha helix). (C) Tertiary structure (A complete protein chain) (D) Quaternary structure (Multiple chains)

2.2 The Protein Folding Process

The native state of a globular protein characterizes by a chain of amino acid residues organized into a stable and folded 3D structure exhibiting native-like secondary structures with tertiary interactions. Hydrogen bonds stabilize the structures; the internal core is mostly formed by hydrophobic residues, held together by van der Waals forces, whereas mostly polar and charged side chains form its surface. Also, it has a very low entropy because the rich residue packing, with side chains fixed that reduce the flexibility of the overall protein structure.

To reach the native state, a protein folds from a non structured chain (unfolded state), with high free energy, to a final stable state (native state), with low or minimum free energy. During this process, called *the folding process*, a protein experiments successive interactions among the atoms of its residues (and also the atoms of the surrounding solvent) until reaching a stable state that strongly determines its biological function. These interactions or *forces* (see next section) can be of different ways: short interactions between residues close in sequences, and large interactions involving parts of the protein that are very distant in sequence.

2.3 Forces Determining Protein Folding

According to the above definition, there are various molecular forces that contribute to determining the folding of a protein and so its functional 3D structure or native state. However, researchers identify a set of interactions that mainly affect the folding, these are: the hydrophobic effect, the energy of hydrogen bonds, the energy of electrostatic interactions, and the loss of conformational entropy [Lajtha et al., 2007, Dill et al., 2008].

The hydrophobic effect refers to the poor solubility of nonpolar molecules in water compared to polar molecules which results in a core due to the burial of the hydrophobic residues of the protein. It is considered to one of the major driving force for the folding of globular proteins.

The hydrogen bonds (HB) occur between atoms bound to positively polarized hydrogen atoms (donors) and negatively polarized (acceptors). They are key components of all secondary structures participating in both, stabilization of helices and interaction between strands to form sheets. The importance of HB is considered as important as the hydrophobic effect.

Electrostatic interactions, as the van der Waals, occur between adjacent atoms (uncharged and non-bonded) and can be described as the sum of the attractive or repulsive forces between molecules. They are of minor importance because most proteins have relatively few charged residues [Dill et al., 2008].

Finally, the loss of conformational entropy, as a result of the restricted motion of the main and the side chains, is the major destabilizing force to the stability of the folded state, and consequently a decreasing of the conformational flexibility of the unfolded chain lead to an increase in the stability of the overall protein structure. Although this force is mainly driven by the hydrogen bond formation and the hydrophobic interactions, its importance is still not fully understood [Lajtha et al., 2007].

2.4 The Protein Folding Problem

The Protein Folding Problem focuses on the way how a protein folds from its amino acid sequence into a stable 3D structure. It is one of the most important, complex, and fascinating puzzles in the history of science. Many areas of science have joined to search for a solution, but until today only approximations to the 3D structure of proteins are obtained.

The folding process is relevant for two different but related problems [Ozkan et al., 2007, Bystroff and Shao, 2003]: the prediction of the 3D structure from its amino acid sequence and the understanding of the underlying process of folding. In the former the main interest lies only in the prediction of the structure of an unknown target protein. In the latter, the focus is on understanding the mechanisms, forces and interactions which drive proteins to reach a final stable state.

2.4.1 Protein Structure Prediction

Predicting the protein structure has been carried out mainly by comparison and simulation techniques. The comparison is useful when knowledge of other proteins, with similar properties as the target protein, exists. If there is no similar protein, the main theoretical technique is simulating the folding process. These simulations are mainly achieved by molecular dynamics which is a computational method that allows prediction of the static and dynamics properties of the protein from the underlying interactions between its molecules. Additional to this technique, there are other methods for simulating and studying the protein folding, as the Probabilistic Roadmap Method that constructs the protein folding pathway or sequence of events that the protein probably follows to reach its native state (see a detailed description of these methods in section 2.7).

2.4.2 Protein Folding Models

For the second problem, several models have emerged in order to explain and understand the mechanism of protein folding [Ozkan et al., 2007, Daggett and Fersht, 2003a], three representatives are the framework model, in which secondary structures form first and then rapidly diffuse and collide to propagate to the tertiary structure; the nucleation model, proposing that an initial nucleus of local secondary structure are formed first and then it propagates and grows to form the tertiary structures; and, the hydrophobic collapse model, in which the hydrophobic collapse drives the folding by forming a compact nucleus or molten globule that rearranges to form the tertiary structure.

In the last years, a 'new view' of protein folding known as *the folding-funnel theory* has emerged. It visualizes the folding process as a kind of free-energy funnel, in which several unfolded conformations of a protein at the rim, characterized by a high-free energy and high degree of conformational entropy, can fold following many paths, until reach its folded state (native) in a single global minimum at the bottom (see Figure 2.3). This theory gives another view of protein folding without invalidating the above models. As a protein folds, the trajectories of motion lie in narrower funnels with reduced presence of conforma-

tional states. A protein can follow many paths passing through intermediates and transition states present along the funnel.

Source: Modified from [Dobson et al., 1998]

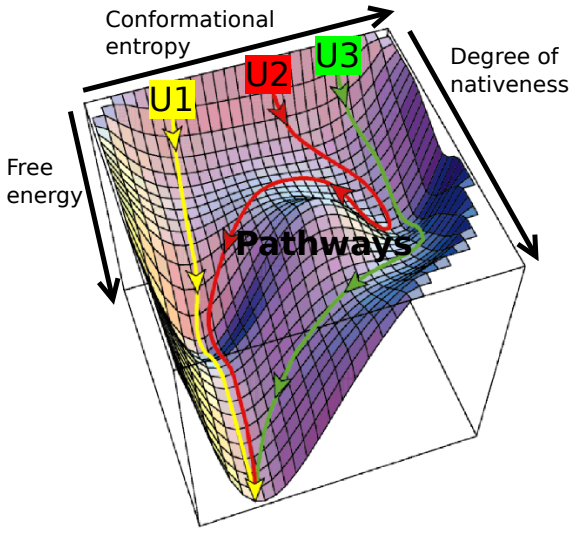


Figure 2.3: **The folding-funnel view.** As the protein folds, its free-energy decreases and its proportion of native contacts grows (Degree of nativeness). The width of the funnel (Conformational entropy) represents the conformational freedom of the chain. Several unfolded conformations U1, U2, and U3 can follow many paths passing through intermediates and transition states which can be present along the free-energy funnel. These states can slow down or accelerate the folding route or *pathway*, redirecting the protein to a steepest path (left-yellow, 'fastest' route) or to round-about paths (red and green, 'slower' routes))

2.5 Protein Folding Pathways

Many proteins fold to their native conformations in less than a few seconds because the protein folding is not a random search of the native among all possible conformations. This affirmation was proved by Levinthal in the late '60s [Levinthal, 1968] with the argumentation that proteins follow specific *protein folding pathways* that are sequences of events (folding intermediates) that fold the protein from the unfolded to the native state.

The view of a unique folding pathway formulated by Levinthal was debated when new theories, as the folding funnel, proposed the existence of multiple pathways. In this theory (described above in section 2.4.2) proteins fold, moving from the top to the bottom of the funnel, by a general bias of the energy surface that reduces the search problem to biologically relevant timescales without the

need to follow a specific path but with the possibility to reach the native by many of them [Lazaridis and Karplus, 1997].

Nevertheless, the classic concept of a *folding pathway* as a sequence of events followed by proteins, continues valid in this 'new view' of folding. The multiple and microscopic paths that a protein can follow are only covering a more general and macroscopic pathway [Pande et al., 1998, Udgaonkar, 2008]. The states a protein passes through are interpreted in terms of ensembles of conformations. And the sequence of events followed by proteins corresponds to finding one of the many members of the ensemble, resulting at the end in many possible pathways.

2.6 Folding Intermediates

In the route to reach its native state, a protein passes through intermediate states forming intermediate structures or *folding intermediates*, as Levinthal argued with the protein folding pathways [Levinthal, 1968] (see above). But their presence was subject to much controversy [Brockwell and Radford, 2007] given that they could not be detected in many proteins, as for instance the large group of proteins with simple two-state folding or direct transition from unfolded to folded state without population of intermediates. Now, that experimental folding methods have improved, their presence is more evident and the debate has changed to the questions of what role they play in the folding, and how they are characterized [Chen et al., 2008, Udgaonkar, 2008, Brockwell and Radford, 2007].

Although the importance of intermediates in the folding and functions of proteins has long been recognized, fundamental questions remain open about the role they play in folding [Brockwell and Radford, 2007, Udgaonkar, 2008]. Can they be considered as merely kinetic milestones indicating the pathway to follow by the protein? Are they responsible for misfolded states and so be the progenitors of human diseases? Or on the contrary, does their presence serve to redirect the protein to the 'correct' pathway and so that any accumulated errors can be corrected?

Due to the above reasons, currently there is no established definition of what a folding intermediate is, however, many authors coincide in common characteristics about them:

> *Intermediates are meta-stable partially folded structures, they exhibit native-like secondary structure with weak or absent tertiary interactions, and a substantial entropy because the poor residue packing as side chains are not fixed, and the hydrophobic core residues remain partly solvated [Udgaonkar, 2008, Baldwin and Rose, 1999, Brockwell and Radford, 2007, Lajtha et al., 2007, Chalikian and Breslauer, 1996].*

In the same sense, there is no unique type of intermediates as they may present different characteristics according to the stage of folding they exhibit. As native secondary and supersecondary structures can range from partial to complete, and residue packing can be weak or strong, many authors have identified two types of intermediates [Baldwin and Rose, 2013, Udgaonkar, 2008, Roder et al., 2006, Uversky, 2002, Duan et al., 1998, Roder and Colón, 1997]. First, *early intermediates*, closer to the unfolded state, with a lack of secondary structure elements and very weak or completely absent of tertiary interactions, and evidencing a very loosely packed hydrophobic core. And second, *late intermediates*, closer to the native state, and characterized as a meta-stable, semi-compact, partially folded structure with native-like secondary structure but lacking of complete tertiary interactions, and evidencing a sizable core of non-polar groups because of poor residue packing.

2.7 Computational Protein Folding

In this section, we will describe three techniques from different approaches that shows diverse views of studying protein folding: one complete atomistic technique as Molecular Dynamics (MD); and two reduced techniques: a sampling technique called Probabilistic Roadmap Method (PRM), and a fragment approach. Each one has its strengths and weaknesses and provides a different

kinetic information. The MD and PRM will be first discussed and with more detail given that we use simulated data from them for our evaluations and analyzes (see chapter 4).

Although experimental techniques for studying protein folding have improved in the last years, they still seem to be limited as far as the details of the process are concerned. In this sense, several computational strategies have emerged as complementary approaches, and in many cases as the only way to get a good approximation of the details of protein dynamics.

These computational approaches, ranging from reduced models to complex computer simulations methods of protein folding, try to explain the folding/unfolding of proteins, and become an open and flexible laboratory to create, modify and test theoretical models. Models can be tested under different conditions and simulations can be repeated easily with new conditions. At the end, their results can be used to develop more realistic models of protein folding that shed light on the intrinsic folding mechanisms.

2.7.1 Molecular Dynamics Simulations (MD)

Actually, molecular dynamics simulations (MD) is one of the principal tools for studying biological molecules as it produces highly physically-realistic motions of atoms and trajectories of protein folding [Pande et al., 2008]. One of its main results is a series of 'snapshots' of the positions and velocities of the atoms as a function of time representing a *trajectory* of the system which correspond to a single, high resolution transition pathway.

MD tracks the positions and velocities of the protein's atoms as they interact with each other and respond to external forces. MD calculates the motion for all particles in the protein structure model by numerically integrating Newton's equations of motion simultaneously and repeatedly in small time steps during a defined period. The equation in its most simplistic form states that:

$$\text{Force} = \text{Mass} \times \text{Acceleration}$$

where the force on a given atom depends on the interactions with all others.

Modeling these forces requires the definition of a detailed and accurate potential function in form of a set of equations and parameters that describe the interactions between particles of a system and returning its potential energy as function of its conformation.

Although MD produces physically realistic simulations, it is computational expensive and only small proteins can be simulated in feasible times, becoming a great computing challenge the simulation of larger proteins on biological timescales on the order of milliseconds. To overcome this challenge, other approaches for folding simulations have appeared, some of them combines ideas of MD (see a review in [Pande et al., 2008]) and others propose novel approaches (see below).

2.7.2 Probabilistic Roadmap Method (PRM)

The Probabilistic Roadmap Method (PRM) is a technique adapted for protein folding developed for robotic motion planning in which a 'robot' at a starting position finds a feasible path to reach a goal position [Song and Amato, 2001, Apaydin et al., 2003]. In the protein folding context, it can be interpreted as the problem of finding a feasible pathway that a protein can follow from an unfolded state (starting conformation) to its native state.

The PRM models the energy landscape of a protein as a *roadmap* defined as a graph of valid motions in that landscape. Nodes represent states (protein conformations), and edges represent feasible transitions (changes of state). PRM starts sampling 'all' valid conformations (the conformational space) of the protein and adding them to the roadmap. Then, it connects neighboring states having feasible transitions. And finally, it extracts folding paths from the roadmap using a graph search algorithm (e.g. Dijkstra's algorithm) which finds the sequence of possible intermediate states that connects the start and goal conformations, *the protein folding pathway*.

There are two main considerations of this method: first, sampling 'all' valid

conformations is practically impossible unless the search space is reduced. A possible method is creating conformations by altering the protein's dihedral angles ϕ and ψ only in allowed ranges, which greatly changes the shape of the entire structure. And second, connecting states having feasible transitions implies determining the energetic feasibility to change the protein from one state to another. It can be done by defining a potential energy function that evaluates each conformation in the roadmap, and selects the ones in which moving from one state to another result in a considerable energy reduction.

Although the PRM is less physically realistic than molecular dynamics, it is useful to analyze and characterize feasible folding intermediate. Also, PRM is less computational expensive than molecular dynamics and so it has been used to study high-dimensional problems in molecular domains [Thomas et al., 2013, Chiang et al., 2010, Tapia et al., 2010, Thomas et al., 2007, Chiang et al., 2006].

2.7.3 Fragment Assembly Methods

The motivation for these methods is that local sequences in the protein have selected preferences for specific local structures, and the full protein structure is the result of the interplay of these ones and the non-local interactions between them. The Rosetta Ab Initio method is one of the most successful examples of this approach [Bonneau et al., 2002]. It gets its local structures from a library of known structures observed in resolved proteins (e.g. the Protein Data Bank), and it combines them randomly to assemble individual conformations. A Monte Carlo simulated annealing search is used to combine the local structures, and a scoring function evaluates their fitness through a function based on conformational statistics of known structures [Rohl et al., 2004].

CHAPTER 3

GLOBAL PROPERTIES OF PROTEIN CONFORMATIONS

This chapter describes the full set of 16 structural and energetic properties of proteins that we selected from the literature as others developed of our own. They will be used in the next chapters to derive a representation of a protein in terms of folding features. The properties described here have been used in other studies to characterize different aspects of the protein folding process, and they have been associated with some protein folding features not directly observable from the protein structure. We start presenting a brief summary of the properties in the table 3.1 and then we describe them individually by presenting their definition, their importance associated with protein folding; and the algorithms or methods used to calculate their values.

3.1 Calculation of Metrics

We created a software toolkit for protein analysis offering a set of functions ready to use for calculating the properties described here (see chapter 8). Additionally, as these calculations will be run on complete folding pathways or trajectories with hundred or thousands of protein conformations, we created a own distributed framework to evaluate them distributively in simple desktop PCs (see chapter 8).

Now we present a summary of the different tools and parameters used to calculate each property, and in the following sections we present a more detailed description of each of them:

- The **NC** uses contact maps created with the *cmap* function from the *bio3d* package of the R system [Grant et al., 2006], it defines a contact cutoff of 7 Å and cutoff neighbor of 3 atoms.
- **CO** uses the code developed by Baker and co-workers [Plaxco et al., 1998], with a contact cutoff of 6 Å.

- **RG** uses the *g_gyrate* from the Gromacs package for molecular dynamics simulation [van der Spoel et al., 2005a].
- **HB** uses the *hbplus* program [McDonald and Thornton, 1994], identifying H-bonds within a maximum distance of 2.6 and 3.9 Å, for donor-acceptor and hydrogen-acceptor, respectively.
- **AS** uses the *naccess* program *[Hubbard et al., 1991],* with a probe of 1.4 Å to calculate the total polar and non-polar area.
- **VD** uses the *AVP* program [Cuff and Martin, 2004] for computing the volume of voids.
- **RM** uses the *rmsd* function from the *bio3d* package.
- **LR** uses the *DSSP* program [Kabsch and Sander, 1983] to get the secondary structure elements (SSEs) from the proteins.
- **PE** and **DM** uses the *analyze* program from the *tinker* package for molecular dynamics simulation [Pappu et al., 1988], with an *amber96* force field. **RC** and **RA** uses the *DSSP* to get the assigned secondary structures elements of each residue.
- The **SS** encodes the target and the native protein structure as 1D sequence of protein blocks using a modified version of the method PB-ALIGN [M. Tyagi A. de Brevern and Offmann, 2008].
- And **RG, SR,** and **DF**, uses the *FIRST* tool for rigidity analysis [Jacobs et al., 2001].

Table 3.1: **Summary of the selected properties to describe folding** The table shows the name of the property, the Id that will be used as a short name throughout this document, a short description, and its association with protein folding.

Name	Id	Description	Association
Native contacts	NC	Contacts formed by a protein conformation which also are present in the native structure.	Nativelikeness. Quantity of native structure.
Contact order	CO	Average sequence separation between contacting residues in the native structure.	Protein folding kinetics properties
Radius of Gyration	RG	How much the protein structure spreads out from its center.	Level of compaction of the structure.
Hydrogen Bonds	HB	Number of hydrogen bonds formed by the structure	Formation and stabilization.
Accessible Surface Area	AS	Accessible surface area accessible to the solvent.	Compactness, and conformational stability.
Voids	VD	Unfilled spaces inside of a protein which are not accessible from the solvent.	Packing and stability of a protein.
RMSD	RM	Ave rage distance in positions between the atoms of two superimposed proteins (here the distance to the native is used).	Proteins similarity.
Local RMSD	LR	Average of the RMSD between the secondary structure elements (developed for this project).	Protein topology and structural similarity.
Potential Energy	PE	Energy associated with forces of attraction and repulsion between protein atoms.	Thermodynamically stability.
Dipole moment	DM	The product of the charges' magnitude and their distance of separation in a protein.	Protein shape. Electrical properties.
Residues in Correct Secondary Structures	RC	Residues forming the same secondary structure elements as the native conformation (developed for this project).	Structure similarity between proteins.
Residues in Any Secondary Structures	AN	Residues in the target protein forming any secondary structure element (developed for this project).	Structure similarity between proteins.
Structural Score	SS	Residues forming 'correct' short local motifs or protein blocks (developed for this project).	Structural similarity between proteins.
Rigid Clusters	CL	Groups of atoms coupled to each other via rigid bonds.	Mechanical stability of a protein.
Degrees of Freedom	DF	Related with the number of constraints necessary to rigidify the protein structure.	Mechanical stability of a protein.
Stressed Regions	SR	Regions with over-constraints reducing protein's degrees of freedom.	Mechanical stability of a protein.

Source: The author

3.2 Native Contacts

The number of native contacts (NC) of a protein can be considered as a global measure of the 'quantity' of structure it contains concerning its native [Song, 2004]. Two distinct protein residues form a contact when the maximum inter-residue Euclidean distance of a pair of heavy atoms (C, O, S, or N) is within a threshold value (e.g. 6 or 7). Other restrictions can be a minimal Euclidean distance between residues (e.g. 3); a sequentially separation of at least 3 residues away from each other; or distance between specific heavy atoms as C-α or C-β.

The NC has been used to evaluate the "native-likeness" of the conformations–how much structure a conformation has–in several studies related to protein folding. For example, in the construction of pathways using probabilistic roadmap methods [Song and Amato, 2001, Apaydin et al., 2003]; in the study of the early stages of folding using molecular dynamics simulations [Duan et al., 1998]; and in the pathway construction using unfolding events [Zaki et al., 2004].

The NC is calculated by counting the contacts of the conformation that are also in the native structure. As defined above, two atoms of two different amino acids a_i and a_j are in contact, if their inter-residue distance $\delta(a_i, b_j) \leq \delta^{max}$, where

$$\delta(a_i, b_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

is the Euclidean distance between two atoms (e.g. C-α or C-β carbons) of the two different residues with 3D coordinates x_i, y_i, z_i and x_j, y_j, z_j; and δ^{max} is some maximum allowed distance threshold (e.g. 6 or 7) [Zaki et al., 2004]. In this work, we use a $\delta^{max} = 7$Å.

3.3 The Contact Order

The topological complexity of a protein is related to the type of interactions – local or non-local– that drives its folding [Plaxco et al., 1998]. The contact order

(CO) is defined as a measure of the relative number of local interactions compared to non-local interactions [Jackson, 1998]. Structures dominated by local interactions as highly α helical proteins correspond to low CO and fold more quickly than structures dominated by long-range interactions–needed for their stability–as α/β and β proteins with higher CO.

The CO has been used to quantify the topology and the stability of proteins in their native states [Shi et al., 2008]. It has correlated well with properties of protein folding kinetics as protein folding rates and the placement of the protein transition state [Plaxco et al., 1998]. Also, it has been used in *ab initio* protein structure programs to help to select best candidate decoys for a posterior refinement stage [Bonneau et al., 2002]. In proteins with simple kinetics–two-state folding–it has described the type of protein interactions [Plaxco et al., 1998, Jackson, 1998].

For a protein with L residues and N contacts, the CO is calculated as follows: [Plaxco et al., 1998]:

$$CO = \frac{1}{LN}\sum_{i,j}^{N} \triangle S_{i,j}$$

where $\triangle S_{i,j}$ is the sequence separation, in residues, between contacting residues i and j.

3.4 Radius of Gyration

The radius of gyration (RG) is a measure of the shape or dimensions of a protein conformation. It describes its size and compactness measuring how much the protein structure spreads out from its center. In protein analysis, the RG helps to infer how folded or unfolded a protein is, as it is indicative of the level of compaction of the structure [Lobanov et al., 2008]. Also, it helps to trace the conformational changes of a protein during its folding trajectory when it is plotted as a function of time [Duan and Kollman, 1998]. Additionally, it has been shown that the protein architectures ($\alpha, \beta, \alpha/\beta$, and $\alpha + \beta$) have a characteristic RG [Lobanov et al., 2008].

The radius of gyration of a conformation is calculated as the root-mean-square, mass-weighted distance of protein atoms from the center of mass [van der Spoel et al., 2005b]:

$$R_g = \sqrt{\left(\frac{\sum_i \|r_i\|^2 m_i}{\sum_i m_i}\right)}$$

where m_i is the mass of atom i and r_i the position of atom i with respect to the center of mass of the molecule.

3.5 Hydrogen Bonds

Hydrogen bonds (HB) play a central role in protein folding influencing the final 3D shape of a protein by contributing to the formation of its secondary and tertiary structures. A hydrogen bond is a type of attractive force that exists between a hydrogen attached to an electronegative atom of one molecule and an electronegative atom of either the same or a different molecule.

HB participate in both, stabilization of α helices and interaction between β strands to form β sheets, so they are partially responsible for the stability of the final folded protein structure [Whitford, 2005]. Several studies have used the HB to assess stability [Fleming et al., 2006], and to indicate organization in proteins [Gong et al., 2005]. Commonly, this type of bond is stronger than other intermolecular forces as the van der Waals but weaker than the ionic and the covalent bonds [Stickle et al., 1992].

To determine the number of hydrogen bonds formed in a protein structure, the pairs of atoms–donors and acceptors– that are within a threshold value are counted. The thresholds used in our work were 2.6 and 3.9 Å, for donor-acceptor and hydrogen-acceptor, respectively.

3.6 Accessible Surface Area

The accessible surface area (ASA) of a protein contributes to determining the size and shape of a protein. The ASA is a concept that corresponds to the surface area of a protein that is accessible to the solvent [Lee and Richards, 1971].

It plays an important role in various properties of proteins as compactness, protein folding, and conformational stability [Liang et al., 1998]. The ASA has been used to characterize the compactness of a protein structure by defining a ratio of the protein's ASA to the surface area of the ideal sphere of the same protein's volume [Lobanov et al., 2008].

The general idea to calculate the ASA is using a "probe", corresponding to a sphere of the solvent, which is rolled over the surface of the protein measuring the area accessible [Shrake and Rupley, 1973]. The radius of the probe corresponds to the radius of a molecule of the solvent, as 1.4 for the radius of a water molecule.

3.7 Voids

Voids (VD) are unfilled spaces or empty cavities, inside of a protein, not accessible to the solvent. They remain empty, but they may be large enough to be filled with other atoms or molecules such as water.

They have been related to the stability of proteins, as filling cavities–in protein engineering–in some tetramers can stabilize at large their structure [Nomura et al., 2011]. Also, voids are related to the assessment of the packing quality of proteins [Liang and Dill, 2001]. Few cavities of voids inside a protein imply it is more packed, whereas high packing densities in protein cores are often considered to be more like solids than liquids [Liang and Dill, 2001].

Voids are calculated using a grid-based void-finding method mapping a protein onto a 3D grid and assigning each grid point to protein, bulk solvent or cavity according to its location in the protein structure [Cuff and Martin, 2004]. Two types of probes with a small (e.g. 0.0) and a large radius (e.g. 1.4) are

used to detect voids and to delimit the solvent accessible regions, respectively. The larger probe delimits the solvent accessible regions, whereas the smaller one identifies voids without merging them with the bulk solvent.

3.8 Root-Mean Square Deviation

The root-mean-square deviation (RMSD) reflects the average distance in positions between the atoms of two superimposed proteins. Similar structures have low RMSD values, in the order of 1 to 3 Å, while larger values correspond to more dissimilar ones [Bergeron, 2002].

The RMSD has been used as a measure of the degree of structural similarity between proteins. The CASP experiment used it for evaluating the prediction protocols used by experts to predict the 3D structure of proteins [Moult, 2006]. However, new measures as the global distance test (GDT) have appeared to replace or complement the RMSD due to some imperfections in its evaluation, mainly when large deviations dominate the overall level of the RMSD. Instead of evaluating the deviations between all atoms as the RMSD does, the GDT uses only a percent of the residues to search for the largest (not necessary continuous) set of similar residues under given distance cutoffs [Zemla, 2003].

The RMSD is calculated from the coordinates of the target i and a reference j structures as follows:

$$RMSD(i, j) = \sqrt{\frac{1}{N}\sum_{k=1}^{N} \left| r_k^{(i)} - r_k^{(j)} \right|^2}$$

where i and j are two optimally superposed protein structures of equal size . N is the number of atoms in each structure, k is an index over these atoms, and $r_k^{(i)}, r_k^{(j)}$ are the Cartesian coordinates of atom k in conformations i, j.

The value of RMSD results minimal because it is determined for the superposition of the structures that minimizes their distance. In our project, the RMSD is always used to compare a given protein conformation with its native one.

3.9 Local Root-Mean Square Deviation

The RMSD mentioned above is a global measure unable to evaluate similarities in local structures, and considered as an 'imperfect' to compare intermediate conformations as their structural elements can be present but not still spatially organized [Fleming et al., 2006].

Therefore, we define for the present project the *local root-mean-square deviation* (LRMSD) that compares a reference protein i with a target one j using the RMSD on each secondary structure element (SSE) of the reference and averaging the obtained values.

We related the LRMSD to the topology of a protein and it contributes to measure the structural similarity between two proteins by comparing locally their SSEs. It is calculated by extracting the amino acid positions of the SSEs from the reference and target structures, using the DSSP program [Kabsch and Sander, 1983], fitting each pair of SSEs of both structures, and calculating their RMSD. The final value is determined as the average of all SSEs in both structures:

$$LRMSD(i, j) = \frac{1}{S}\sum_{s=1}^{S} RMSD(i_s, j_s)$$

where i, es la reference conformation and j the target conformation . S is the number of SSEs in the reference structure. And $RMSD(i_s, j_s)$ is the RMSD value between the SSE number s in reference conformation i and the corresponding atoms in the target conformation.

In our project, the reference structure is always the native one.

3.10 The Potential Energy

If a protein is taken as a static molecule with many interacting atoms, its potential energy (PE) can be defined as the energy given by its three-dimensional structure. PE is related to the thermodynamical stability, considering that a protein folds into the state of lowest energy by decreasing of its PE and reaching

the thermodynamical stability in its native state [Dokholyan, 2004].

If the interactions between protein's atoms are considered as elastic forces, then they can be written in terms of equations known as potential energy functions or force fields. In these functions the total energy is given by the sum of both the bonded and non-bonded terms describing atoms linked by both the covalent bonds and long-range interactions, respectively. A general PE function is presented below, for a more specific functions see [Ponder and Case, 2003]:

$$E_{total} = E_{bonds} + E_{angles} + E_{torsions} + E_{van-der-waals} + E_{electrostatic}$$

where the first three energies correspond to bonded terms describing bonds, angles, and bond rotations in the protein. And, the last two terms describe interactions between non-bonded atoms or atoms separated by 3 or more covalent bonds, the van der Waals interaction energy and the electrostatic energy.

3.11 The Dipole Moment

The dipole moment (DM) can be defined as the product of the magnitude of the charge and the distance of separation between the charges of a molecule. The DM is related to the geometrical and electrical structure of a molecule. Its magnitude may affect the protein in various ways: its shape, its intermolecular interaction, its biomolecular solvation, and its biochemical activities [Zhao and Ji-Jun,].

In proteins, the DM is calculated as the vector sum of the dipole moment of three components: fixed charges on the surface, polar groups in main and side chains and the mobile ion fluctuations [Takashima, 1993].

3.12 Residues in Correct and Any Secondary Structures

Helices alpha, beta strands forming beta sheets, and other *secondary structures elements (SSEs)* are essential contributors to the stabilization of the overall protein

fold. These structures are segments of the protein chain in which their backbone's torsion angles ϕ and ψ repeat in regular patterns.

We have created two measures to quantify the proportion of SSEs in a protein conformation. The first one, *residues in correct secondary structures* (RC) evaluates the proportion of residues of a target conformation that coincides with their SSEs type with the native SSEs. The second one, *residues in any secondary structure* (AN) evaluates the proportion of residues forming any SSEs in a target conformation.

Both measures are calculated by getting and comparing the SSEs assigned by the DSSP program [Kabsch and Sander, 1983] for the protein conformations: the target and native for the first measure (RC) whereas only the target for second one (AN).

3.13 Structural Score

We create the Structural Score (SS) for the present work, and it is related with another PhD thesis of our research group. It is based on structural similarity between proteins but comparisons are performed using short local motifs or protein blocks (PBs) [Kolodny and Levitt, 2003] instead of SSEs. Representing proteins with PBs gives more local structural information than SSEs because PBs highlight subtle variations and structural conservations of the structure [M. Tyagi A. de Brevern and Offmann, 2008, Kolodny and Levitt, 2003].

Protein's 3D structure was represented as a one-dimensional chain of PBs from a library of 40 that were constructed by using fragments of 5 residues consisting of the 3D coordinates of their C-α atoms. For its computation, both protein structures: target and native are encoded transforming their 3D representation to 1D sequence of PBs. Then, the SS is calculated as the proportion of PBs of the target also present in the native.

3.14 Rigid Clusters, Stressed Regions, and Degrees of Freedom

These measures are taken from the rigidity analysis that is concerned with analyzing rigid and flexible regions of a protein [Jacobs et al., 2002, Schirf,]. Rigidity can be related with the mechanical stability of a protein [Jacobs et al., 2002], which is explained as the resistance of the protein to unfold in response to an external force [Wang, 2009].

Rigid clusters (CL) are the parts of the protein that behave as rigid bodies (i.e., distances between all atoms remain fixed) [Chubynsky, 2008]. *Degrees of freedom* (DF) is related to the number of constraints necessary to rigidify the protein structure [Schirf,]. And *stressed regions* (SR) are regions that contain over-constraints [Schirf,], as distance constraints between atoms imposed by bonding forces, which reduce the total number of degrees of freedom available to the protein [Thorpe et al., 2001].

These measures are calculated using the FIRST software that implements the pebble game algorithm [Rader et al., 2002]. It generates a directed graph to model physical forces of constrains due to bonding present within the protein. Then FIRST checks every bond to determine if it is part of two types of rigidity elements: a rigid cluster or an undersconstrained region forming a flexible connection.

CHAPTER 4

DATA AND METHODS

This chapter contains information about the general approach that we followed to get our results, the protein folding data and methods on which this work is based (proteins, folding pathways and trajectories, and the different statistical analyses carried on these data). Specific details of the data handling and application of methods are described in the corresponding chapters where they are used.

4.1 General Approach

We summarize the main steps of our approach in the flowchart of the Figure 4.1. It shows three main process: the selection of properties, the definition of the folding status, and the characterization of the folding levels. The primary inputs for all the analyzes are the set of properties and the conformations from the dataset of protein folding pathways (parallelograms at the top-left). Next, we describe each process briefly, and a detailed explanation is given in the corresponding sections throughout the document.

The selection of properties starts by performing a preliminary evaluation and analysis (section 6.2). The evaluation computes a set of 16 physical properties on the 3D structure of each of the protein conformations taken from a dataset of protein folding pathways. The analysis involves three techniques (a multidimensional scaling, properties profiles, and a principal component) performed on the previous evaluations. In the end, we retain 11 properties to be used for all the subsequent analyzes.

The definition of the folding status is a process that takes the previously selected properties and their corresponding evaluation to execute on these values a final principal component analysis (PCA) that reduces the 11 properties to three components (section 6.3). Then, we analyzed the connection among the properties grouped for each one and associated them with the folding features of stability, compactness, and native-likeness (section 6.4). Each feature was ex-

Source: The author

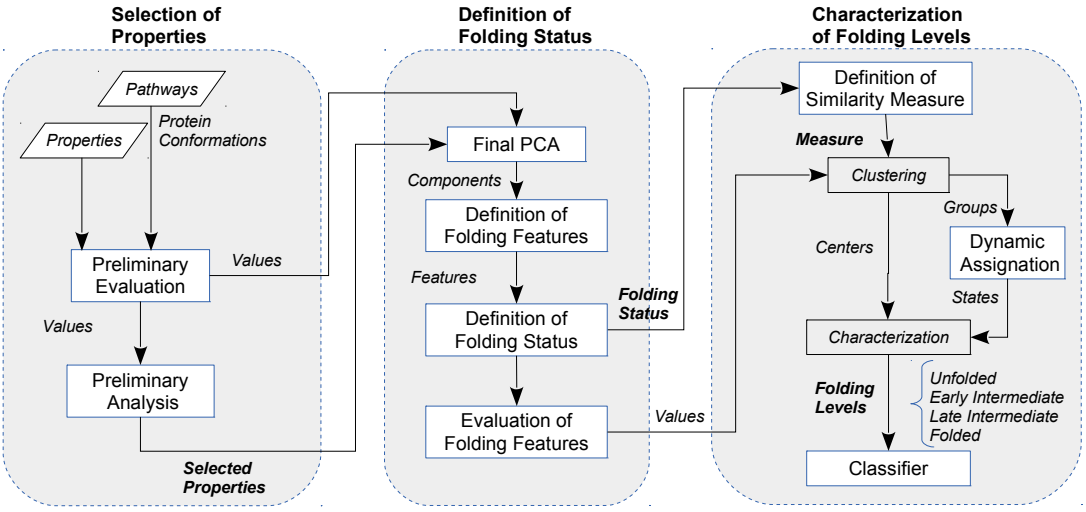


Figure 4.1: **A flowchart of the general steps of the approach.** The ellipses show the three main processes, each one with activities (rectangles) taking or producing different inputs and outputs (arrows), respectively.

pressed mathematically in a manner that they can be calculated for any protein structure (section 6.4.2), and we defined the folding status of a protein as the 3-dimensional vector composed of the three features (section 6.5). As a direct application of the folding status, we defined a folding score that measures the degree of folding of a protein conformation (section 6.5.2). In the end, we compute the values for the three features for each conformation from the protein folding pathways.

The final process is the characterization of folding levels. Having defined the features and the folding status, we used them for creating a similarity measure that compares the folding status of a pair of protein conformations (section 7.2). Using this measure, we performed a clustering analysis of all the conformations from the pathways, but now comparing their folding statuses (section 7.3). The clustering resulted in four groups which dynamic assignation in the entire set of pathways (section 7.4), modeled with a Markov chain, showed four sequential states that we interpreted as stages of folding and named as folding levels (section 7.5): unfolded, early intermediate, late intermediate, and the folded level. To complete the process, we used the clustering's parameters for each level and defined a classifier that takes as input a protein structure and assigns to it the corresponding folding level (section 7.6), according to its folding features calcu-

lated from its physical properties.

4.2 Protein Folding Data

Several computational approaches for simulating protein folding and molecular motion have been applied to protein folding [Pande et al., 2008, Javidpour, 2012], from atomistic representations as Molecular Dynamics (MD) to reduced representations based on sampling techniques as the Probabilistic Roadmap Method (PRM) (see chapter 2 for a description of both methods). In our research, we used mainly the folding data from two projects: the MD trajectories of the villin-headpiece protein from the folding@home project; and the folding pathways of several proteins generated with the PRM implemented in the Parasol Folding Server.

4.2.1 Villin-headpiece protein

For our preliminary analysis (chapter 5) we used the molecular dynamics trajectories (see below) from the fast-folding double mutant of the chicken villin-headpiece subdomain protein, known as HP-35NleNle or LYS24/LYS29. This protein folds quickly into a globular structure, in the order of 3.7 to 4.9 μs at 300 K [Wei et al., 2010], when norleucines replace residues LYS24 and LYS29 that stabilize and increase its folding rate. It is one of the fast-folding proteins most intensely studied by experiments, theory, and simulations, and it consists of 35 residues bundled in three right helices and four coils with a hydrophobic core (see Figure 4.2).

4.2.2 Folding@home Molecular Dynamics Trajectories

Folding@Home is a distributed grid computing system performing various molecular dynamics simulations of the villin-headpiece (HP-35NleNle) in the

Source: The author

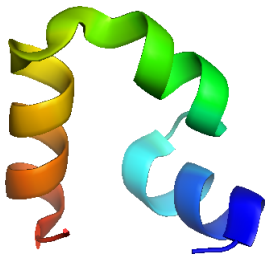


Figure 4.2: **Structure of the chicken villin-headpiece subdomain protein.** This protein folds into a three α-helices held together by a loop, a turn, and a compact hydrophobic core surrounded by the α-helices.

project called PROJ3036 [Ensign et al., 2007]. Each simulation run a full molecular dynamics simulation on a modified version of the Gromacs package with all-atom, explicit solvent, and the Amber2003 as the force field. The project PROJ3036 contains nine runs, each one starting from nine different non-folded configurations, and consisting of 100 trajectories or clones (clone0 to clone99). Each clone starts with the same configuration but with different random velocities. These trajectories were subdivided into frames corresponding to 20000 ps of simulation time with a time step of 50 ps, totaling around of 400 snapshots (PDB structures) or time-points for each trajectory.

In the preliminary analysis presented in chapter 5, we use one of these simulations, the molecular dynamics trajectories of the project PROJ3036 (run 5, clone 1) retrieved from the villin folding repository (`http://simtk.org/home/foldvillin`). Specifically in the run 5, clone 1 the simulations started from a non-folded configuration that has none of the structural characteristics of the native while in other runs the starting configuration contained some of its secondary structure.

To get the final PDB files used in our evaluations, we preprocessed the data as follows: First, frames were cleaned by removing PDB files (snapshots) with duplicated starting times. Then, the 37 frames of run5, clone 1 were concatenated into a large file, and the starting time for each frame was specified: 0 for the first frame, 20000 for the second, and so on, getting 15,200 conformations approximately.

4.2.3 Parasol-Server Protein Folding Pathways

We obtained a dataset of 37 protein folding pathways from the Parasol folding server (`http://parasol.tamu.edu/foldingserver`) that implements the Probabilistic Roadmap Method (PRM) [Song and Amato, 2001]. One of the main characteristics of these pathways is that they include the main folding states of the folding process, according to the PRM. The pathways comprise proteins with diverse topologies and sizes that we used to develop the inherent folding features and to define the folding levels in the chapters 6 and 7. See a summary of theses pathways in the tables 6.2 and 6.3 in chapter 6.

A protein folding pathway corresponds to the sequence of events or folding intermediates that a protein follows from an unfolded state to the native structure. A protein may have more than one pathway, and each pathway consists of several 3D structures or conformations (snapshots) of the protein. As an example of these PRM pathways, the Figure 4.3 shows the one constructed for the protein Cytochrome C551.

Source: The author

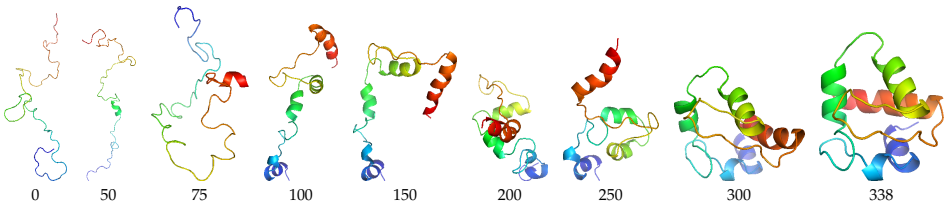


Figure 4.3: **A PRM protein folding pathway for the Cytochrome C551.** The sequence of 3D structures for the pathway produced by the PRM method for the protein Cytochrome C551. Below each structure the folding step or the PRMs sequential ordering of protein's formation is shown. The pathway contains 338 conformations (here, we show only 9 of them) starting from an unfolded state, folding step 0, to the folded or native one, folding step 338.

4.3 Properties Evaluation

Two different sets of physical properties were evaluated on conformations of both the trajectories from the molecular dynamics simulations of the villin-headpiece protein (see preliminary analysis in chapter 5), and on the conformations of the whole pathways from the Parasol server (see definition of com-

ponents in chapter 6) . The chapter 3 describes the physical properties and the methods used for their calculation

For the implementation of the methods for calculating the properties, we developed our toolkit for protein analysis. The toolkit offers Python functions for each property that can be called as standalone programs or as libraries from external programs. Many of these functions are implemented directly in a computer language and others wraps the functionality of well know programs, libraries, and frameworks that offers the functionality indirectly through more complex processes from which our function extracts the necessary information to return the appropriate result. The chapter 8 presents a more detailed description of the toolkit and its functions.

Since the evaluation of the properties is performed for a lot of protein conformations, this is a time-consuming task taking long times in uniprocessor systems as the ones available in most of our computing laboratories. Then we developed a distributed framework that executes the evaluations distributively reducing considerably the total execution times. The framework is based on the publisher/subscriber approach [Eugster et al., 2003] of distributed computing and uses as a middleware the Dropbox cloud storage services [Drago et al.,] to communicate any desktop PC or server inside or outside of our institution. The chapter 8 present the details of this framework.

4.4 Data Preparation

All protein conformations were energetically minimized to remove steric restraints by a conjugated gradient minimization algorithm from the Gromacs package [van der Spoel et al., 2005a]). The evaluations performed on both preliminary analysis (chapter 5) and definition of components (chapter 6) were scaled to use the same range of values for all the properties by a min-max normalization [Jain et al., 2005], unitizing from 0 to 1 all the properties. Next, they were averaged by a modified average moving filter to smooth out noisy values that may result from the evaluation of conformations with sharply folding events producing extreme-valued protein properties.

In the definition of components (chapter 6), the full set of PRM pathways was split into two datasets: one for training and the other for testing. The training set was conformed by 5399 conformations from 29 pathways (80%); it was used to filter and define final properties, components, groups, distributions, and the parameters for the definition of a similarity measure (see chapter 6). The testing set was conformed by 2610 conformations from 7 pathways (20%) and it was used for assigning the folding level to their conformations and analyzing the results (see chapter 7).

4.5 Statistical Analysis

4.5.1 Multidimensional Scaling

Multidimensional Scaling (MDS) is a technique for exploratory data analysis which represents relations among objects (similarity or dissimilarity data) as distances between pairs of points in a lower dimensional space, making data accessible to visual inspection and exploration [Borg and Groenen, 2005]. Two MDS techniques are differentiated according to the way that proximity data is generated (distances between points): classical metric and non-metric. If it displays metric properties, like distances (e.g. Euclidean distance) it belongs to the former; but if it only assumes that order of the proximities is meaningful (e.g. rank order), then it belongs to the latter [Wickelmaier, 2003].

We performed two classical (metric) Multidimensional Scaling (MDS) analyzes (described below) with data from the evaluated PRM pathways (chapter 6). The first MDS selected more relevant properties, and the second found a preliminary structure of interrelated properties. Both MDSs used as input a matrix of data of evaluated PRM pathways, the former with the 16 properties and the latter with a subset of 11 of them. Also, they were using the inter-property dissimilarities calculated with a Euclidean distance to approximate a set of points representing the protein properties inter-point distances. The resulting points were plotted in a 2D graphic (XY) from which we interpret the solution. We used for both MDSs the *cmdscale* function from the R stats package.

4.5.2 Principal Component Analysis and Factor Analysis

Principal component Analysis (PCA) and factor analysis (FA) are two methods from the multivarate analysis that are particularly used to understand what data mean when the problem involves many variables. Although both methods try to reduce the number of manifest variables to a small set of latent variables, they differ in their primary objectives and approaches.

While PCA identifies groups of manifest variables interrelated via an underlying structure which explains or implies new variables, called components; FA tries to discover a hidden structure given by the factors that explain the manifest variables. So, PCA defines the components as linear combination of the original variables and analyzes all of the observed variance, and FA defines the original variables as linear combinations of the factors but analyzes only the shared variances.

Most of the physical properties of proteins used in our study have heterogeneous measurenments, meaning that the units matter (e.g. Å and number of hydrogen bonds) and so there is no ”natural scaling” to adopt. Then, unlike factor analysis, PCA is not scale invariant [Harner, 2012]. Also, in order to enhnace interpretability and to seek a simple structure, principal components are often rotated according to different methods (see [Dean, 2009a]). A Varimax method obtains orthogonal components that preserve their independence. Other methods as the Oblimin or Promax, break this independence and allow components to correlate.

4.5.3 Conducting PCA

General steps in conducting PCA are an initial extraction of components; determination of the number of components; rotation to a final solution; interpretation of the rotated solution, and calculation of the component scores.

We performed three Principal component analysis (PCA). The first one is the preliminary analysis presented in the chapter 5 with nine properties evaluated on the data from the villin-headpiece protein's trajectory. The second and third

ones are used in the chapter 6 for developing folding features from 11 properties evaluated on the data from the PRM pathways. We use a general method to perform the three PCAs called *spectral decomposition or eigendecomposition* that examines the covariances and correlations among variables and identifies independent components explaining the maximum amount of mutual correlation (variation) measured by its *eigenvalues* [Burstyn, 2004]. The first component has the largest value; the second has a smaller values, and so on.

We use three different functions from the R system for the three PCA. In the first one, we used the function *princomp*, as it is simply and doesn't rotate components (see below). For the second, we used the function *principal* of the *psych* package to get a rotated solution and select the number of components; and for the final PCA we implemented our function, based on eigenvalues, that gives us more information than the others and makes additional calculations.

In all of the PCA analyses we followed a combination of three methods for deciding how many components will be retained (selected) to explain as much the information as possible in the original data: eigenvalue-one criteria, percent of cumulative variance, and non-trivial factors rule. The first one proposes to select a component when its eigenvalue is greater than 1.0; in the second one, the first N components are retained if they can explain a specified percentage of the data variance (usually 70% or 80%); and the third one, defines a grouping criterion in which components containing only one variable are removed given that they do not contribute to the goal of finding grouping patterns [Dean, 2009a, Hanebutte et al., 2003].

After selecting the number of components, we used an orthogonal Varimax rotation to make the solution more interpretable [Dean, 2009b, Costello and Osborne, 2005]. A rotation refers to perform arithmetic to obtain a new set of factor loadings from a given set, corresponding the factor loadings to the weights (correlations) between original variables and the new components.

4.5.4 Hierarchical Clustering

Hierarchical clustering works in a bottom-up approach starting with each observation as a distinct singleton cluster and merging the two closest one successively at each level until one cluster remains [Bien and Tibshirani, 2011]. A binary tree, called dendrogram, is formed at each step as two clusters merge into a new one. Leafs of the dendrogram (individual observations) have heights or level 0, and inner nodes resulting from merging have the height equal to their intercluster distance. There are four common clustering strategies (or linkages) to measure this distance: single, takes the minimum distance; complete, the largest distance; average, the average distance; and centroid, takes the centroids of the clusters.

We performed a hierarchical clustering on the conformations of the protein folding pathways from the Parasol server with the objective to discover groups that share similar folding characteristics (see chapter 7). A complete linkage method to merge two clusters based on a squared Euclidean distance [Bien and Tibshirani, 2011] was used as the clustering strategy. It selects the longest possible intercluster distance (most dissimilar members) to ensure that more conformations in a cluster are within some maximum distance to one another, producing well-distributed clusters.

To determine the number of clusters, we applied the criterion of inconsistency coefficient [Pardalos et al., 2012]. It is calculated by comparing the height of each link in a cluster with the average heights of neighboring links below it. The highest value of the inconsistency coefficient's value determines the height where to cut the dendrogram and the resulting clusters.

4.5.5 Clustering Validation

Although there is no a general process or method to validate clustering solutions, three main approaches are commonly used: external, internal, and relative validations (see below). In our work, the clustering solution obtained in chapter 7 was validated by both an internal and relative approaches.

The internal validation uses only inherent features of the dataset for measuring the solution's quality, as cohesion and separation of obtained groups. And the relative validation compares the cluster structure with other cluster structures resulting with the same algorithm but with different parameters [Sivogolovko and Novikov, 2012]. For the former, we used the Silhouette index, and for the latter we used a resampling method (both explained below).

External validation was not used as it needs some predefined knowledge as for instance the correct class labels, and our clustering applied an unsupervised approach with no prior biological knowledge. Specifically, our process intends to *define* the folding level of protein conformations, not to assign known levels.

Silhouette index

For each group, we calculate the Silhouette index Si (chapter 7) by averaging the Si of each of its protein conformations. At the end, an average Si was calculated for the complete clustering by averaging the groups' Si. The index Si for the ith protein conformation was calculated as $Si = (bi - ai)/max(ai, bi)$, where ai is the average distance from the ith conformation to the other conformations in the same cluster as i, and bi is the minimum average distance from the ith conformation to the ones in a different cluster [Maechler et al., 2013].

The Silhouette index combines ideas of both cohesion (how closely related are objects in a cluster) and separation (how distinct a cluster is from the others). The value of this index is obtained by comparing how close the object is to objects in its cluster compared to objects in other clusters [Rousseeuw, 1987]. The index ranges from 1 to -1, with values close to 1 for a correct classification, close to -1 bad classification, and near to zero indicating that the objects can be equally classified into any other group.

Validation by Resampling

This cluster validation was adapted from the resampling approach proposed by [Levine and Domany, 2000] in which the basic criterion is founded on the

fact that if a clustering is valid then each of its subsets should be valid as well [Abul et al., 2003]. In general, for the unsupervised estimation of the cluster validity, the available data is resampled, and a figure of merit is used to measure the stability of the solution against the one given by the resampling solutions.

In our cluster validation (chapter 7), we used the complete set of pathway conformations as a full dataset split in various sub-datasets. The clustering procedure described above (section 4.5.4) was applied to both the full and each sub-datasets. And to compare the solutions, we created their respective connectivity matrix, a symmetric matrix formed by the conformations of each dataset. The matrices were marked with 1's and 0's; 1 if the conformation of the column i was in the same cluster as the one in the row j; 0 otherwise. Then, the matrix of the full dataset was compared with each matrix of the sub-datasets obtaining a measure of agreement between the cluster solutions, ranging from 0 to 1 (0% to 100% agreement). Finally, the individual results were averaged to obtain an overall average of agreement.

CHAPTER 5

PRELIMINARY ANALYSIS OF A PROTEIN FOLDING TRAJECTORY

In this chapter, we perform an evaluation and analysis of a set of nine physical properties on the conformations of the folding trajectory of the villin-headpiece protein. Our goal is to observe if the properties measure hidden folding features and if their values are different for each stage of folding (unfolded, intermediates, and folded).

We took the properties from works related to the influence of structural and energetic protein properties in different issues of protein folding as topological determinants of folding [Dokholyan et al., 2002], structural determinants of folding type [Ma et al., 2007], and determinants of protein globularity [Costantini et al., 2007, Mucherino et al., 2008].

The analysis methods used in this preliminary analysis were: property profiles, correlations, factor analysis, and the principal component analysis. While the former two methods determine the dynamic behavior of properties and evaluate their relations, the latter ones allow to reduce the original set of variables to a smaller number that are called factors, and components respectively. Factors and components represent hidden (unmeasurable) variables that explain variability and dependence of the original visible (measurable) variables.

5.1 Data and Methods

5.1.1 Villin-headpiece Protein and Folding Simulations

The villin-headpiece subdomain is one of the most studied proteins as it has with only 35 residues and is fast folding. It folds into a three α-helices held together by a loop, a turn, and a compact hydrophobic core surrounded by the α-helices (see details in chapter 4).

We are using the molecular dynamics trajectories of the villin-headpiece protein (HP-35NleNle) collected from the Folding@home project (`http://www.`

simtk.org). Specifically we are using the trajectories from the project PRJ3036, run5, clone 1, with 16000 protein conformations (PDB structures), approximately (see chapter 4 for more details).

5.1.2 Properties Selection and Evaluation

We selected nine different structural and energetic properties used in evaluations of folding trajectories mainly for measuring the folding status or folding stage of a protein conformation [Das et al., 2006]. But there is not an individual property that can describe all the protein folding stages by itself [Pande et al., 1998], so the study of several of them can help to improve the understanding of the folding process.

The evaluated proteins were the native contacts, contact order, radius of gyration, hydrogen bonds, accessible surface area, root-mean-square deviation, potential energy, voids, and dipole moment (see Figure 6.1). For a full description of the properties see chapter 3.

Table 5.1: **Set of nine properties evaluated on the villin protein folding trajectory.** First column for the short name (Id) that will be used through the rest of this chapter.

Id	Properties
NC	Number of Native contacts
CO	Contact Order
RG	Radius of Gyration
HB	Number of Hydrogen Bonds
AS	Accessible Surface Area
RM	Root-mean square deviation (RMSD)
PE	Potential Energy
VD	Voids
DM	Dipole moment

We evaluated the nine properties on each conformation of the villin trajectory by our distributed framework for protein analysis (see chapter 8) that includes a toolkit implementing and wrapping different algorithms and functions to measure different protein properties given the 3D structure of the conformation. The evaluations were normalized, and a moving average is applied to smooth out short-term fluctuations, as described in chapter 4.

5.1.3 Analysis Methods

We applied four types of analysis methods on the evaluated conformations of the villin folding trajectory. First, we evaluated the properties on the conformations and created a data matrix with the properties as columns and all the conformations as rows. Then we created pathway profiles for each of the properties, which showed the evolution of an individual property on the protein conformations during the simulation time.

Correlations were calculated using the Pearson's correlation coefficient on the data; factor analysis performed maximum-likelihood estimation on the data [Costello and Osborne, 2005]; and principal components were calculated computing the eigenvalues on the correlation matrix, and without rotation of the components [Burstyn, 2004]. For both, factors analysis and principal components, the selection of the number of factors and components respectively, was determined by the criterion of the cumulative proportion of variance [Dean, 2009a]. This criterion observes the value of the cumulative variance and selects the number of components that explains at least 70% or 80% of the information on the data. See chapter 4 for a more detailed description of these methods.

5.2 Results

5.2.1 Properties profiles

Figure 5.1 shows the profiles for the nine properties. The profiles show the values of each property on each time step of the protein's trajectory. The simulation time was 760 ns (approx.) and each conformation corresponded to the protein's structure at every 50 ps, resulting in 15201 conformations. We can see that in most profiles, values of each property fluctuate in different ranges through all the simulation time. Some profiles share a visually distinguishable common behavior but for others this is not obvious. For example, the profiles for native contacts and hydrogen bonds have common behavior but others are completely

different as for instance the radius of gyration and RMSD.

It is important to note that although some properties show a trend during the simulation (e.g. RMSD, Hydrogen Bonds), after a certain time step (about 400 ns) this trend is lost (RMSD grows, Hydrogen Bonds decreases). This behaviour is given by the particular characteristics of the simulation in which the entire protein never gets close to the native state, but only some elements of it at specific time steps [Ensign et al., 2007]. Contrary, other properties do not suggest any trend (e.g. voids, potential energy). In the case of voids, although this property has been used to evaluate packing quality in proteins which native structure is usually well-packed, studies show that packing is not a major indicator of stability [Cuff and Martin, 2004]. And due to the dynamics of the folding process, voids can appear and disappear in the different conformational states of the protein and so their valuation may not follow a particular trend.

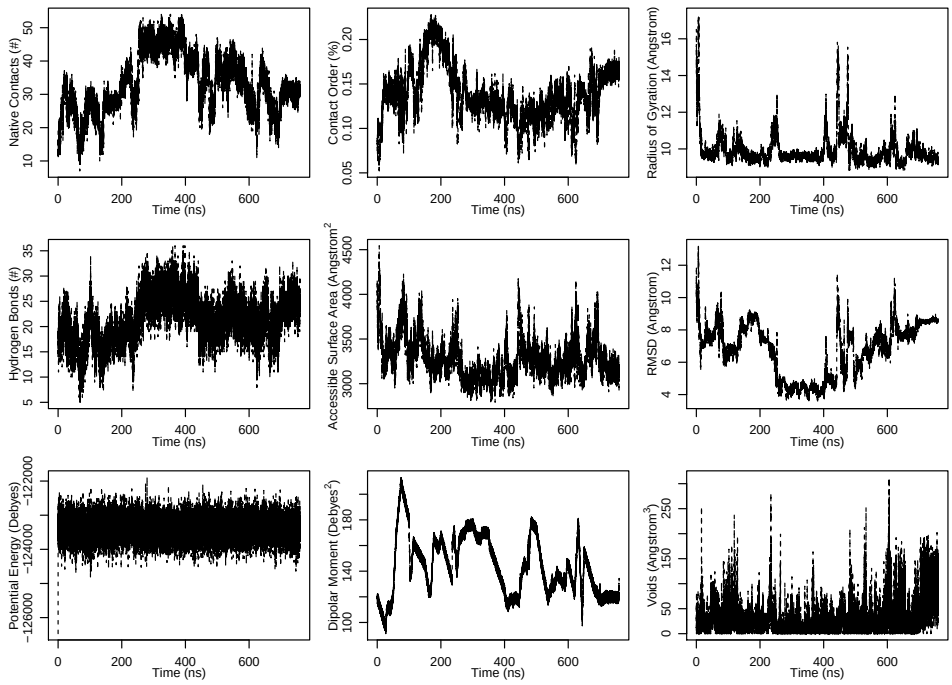


Figure 5.1: **Profiles for nine properties evaluated on the villin trajectory.** Horizontal axis for trajectory conformations appearing every 50 ps. And vertical axis for property values for each conformation. We can see that some profiles follow a common behavior, while others are not easy to compare or present specific patterns.

5.2.2 Correlations

We used the Pearson coefficient to calculate the correlation between properties. In general, correlations were not strong (< 0.7 in absolute value, save native contacts and RMSD with a correlation of -0.78), and some were weak (< 0.4 in absolute value, as voids and dipole moment). We took a threshold value > 0.5 to decide whether two properties were correlated. Table 5.2 shows the results of correlations and the Figure 5.2 presents a graphical view of the relations between pairs of properties.

As it can be seen from the figure and checked in the table, there were two groups of properties: a correlated set, and another of uncorrelated ones. In the former group, native contacts correlated well with hydrogen bonds, accessible surface area, and RMSD; while contact order correlated weakly with radius of gyration, which also correlates with accessible surface area. In the latter one, potential energy, voids, and dipole moment don't correlate with any other property.

Table 5.2: **Correlation matrix for the properties evaluated on the villin trajectory.** Properties IDs on top. Correlations with absolute values > 0.5 are marked in bold. The only two properties that consistently present very weak correlations are voids (VD) and dipolar moment (DP).

ID	Property	NC	CO	RG	HB	AS	RM	VD	PE	DM
NC	Native Contacts	-								
CO	Contact Order	-.11	-							
RG	Radius of Gyration	-.36	-.46	-						
HB	Hydrogen Bonds	**.67**	-.08	-.27	-					
AS	Accessible Surface Area	**-.70**	-.31	**.68**	**-.62**	-				
RM	RMSD	**-.78**	.28	.24	**-.51**	.45	-			
VD	Voids	-.15	.13	-.14	-.02	-.08	.22	-		
PE	Potential Energy	-.50	**.57**	-.05	-.31	.23	**.56**	.23	-	
DM	Dipolar Moment	.04	-.02	.05	-.03	.10	-.21	-.13	-.13	-

Source: The Author

5.2.3 Factor Analysis

A factor analysis was performed selecting five factors according to a percent of cumulative variance criterion (see Table 5.3). Although the analysis indicated a solution with five factors (88.2% of total variance), only the first two factors were relevant as they loaded with more than one property. In contrast, the others

Source:The Author

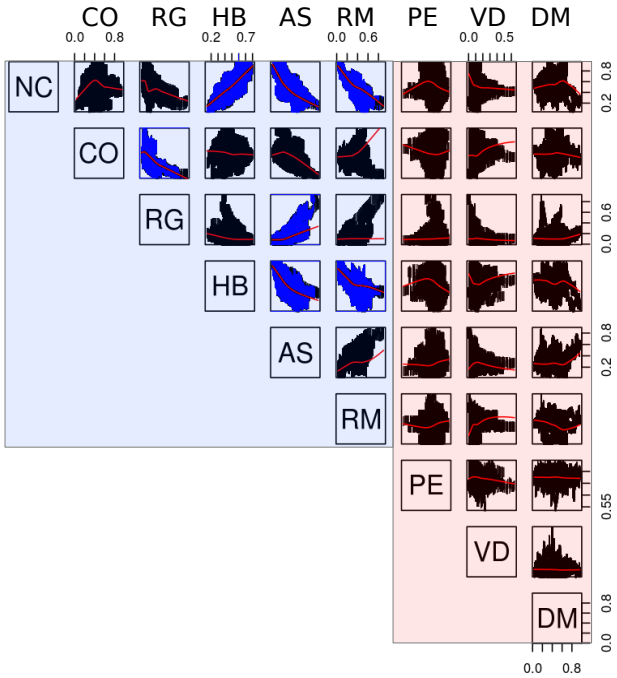


Figure 5.2: **Graphical view of the relations between properties evaluated on the villin trajectory.** Colored in blue are the pairs of properties that correlate with absolute values > 0.5. Two groups of properties: correlated (left blue rectangle) and uncorrelated (right red rectangle). The cloud of points are distributed near to the diagonals in correlated properties and dispersed in the others.

were trivial factors, that loaded dominantly with only one property, and are considered non-meaningful for the analysis [Dean, 2009a, Hanebutte et al., 2003] (see section 4.5.3).

Most properties have a low uniqueness; exceptions are contact order (CO) and voids (VD) with values over to 0.5. Properties with small uniqueness are relevant for the factor model given that they share variance with the others, as reflected in the factors. But factors with larger uniqueness do not share variance, and so they constitute own factors, with low relevance in the factor model.

We performed a *Null hypothesis Test* affirming that the retained five factors predict the model adequately versus the alternative that this is not true. A Chi-square goodness-of-fit resulted in a value of 25359.09 and a p-value of 0, meaning that we must reject the null hypothesis because the factors poorly predict the data (p-value << 0.1), from the statistics perspective.

Table 5.3: **Factor analysis for the data evaluated on the villin trajectory** The solution suggests five factors, but only the first two ones loaded with more than one property. Uniqueness represents the variance that is "unique" to the property. The factors matrix (loadings) shows the weights and correlations between factors and properties.

			Factors matrix (Loadings)				
ID	Property	Uniqueness	**Factor1**	**Factor2**	Factor3	Factor4	Factor5
NC	Native Contacts	.07			.96		
CO	Contact Order	.52	**.46**				.46
RG	Radius of Gyration	.05				.96	
HB	Hydrogen Bonds	.03		**.97**			
AS	Accessible Surface Area	.005	**.99**				
RM	Root Mean Square Deviation	.11		**.93**			
VD	Voids	.55		**.66**			
PE	Potential Energy	.11		**.93**			
DM	Dipolar Moment	.09	**.97**				
	Proportion of Variance		**.21**	**.19**	.19	.15	.13
	Cumulative Variance		**.21**	**.40**	.59	.74	.88

Then, what is the adequate number of factors? We tried other analyzes with 4, 6, and 7 factors but the *Null hypothesis test* continued with negative results. With 8 factors, we got an exact fit, what is a consequence of the fact that the models improve if more variables are considered. But this is contrary to the goal of FA to reduce variables and find an underlying structure in data.

5.2.4 Principal Component Analysis

Principal components analysis (PCA) performed on the evaluated conformations of the villin-headpiece showed that the set of nine properties can be reduced to three components (74% of the total variance) according to the cumulative proportion of variance (see Table 5.4).

Components*.	Cumulative Proportion
C1	.38
C2	.60
C3	**.74**
C4	.87
C5	1.0
...	...

*Only 5 of 9 components are showed

Table 5.4: **Selection of components for the evaluation on the villin trajectory.** Cumulative proportion of variance suggests three components are enough, explaining 74% of the total variance.

However, only two loaded with more than one property, the others were *trivial components*, in which only one property loads highly, generally meaningless for the analysis [Dean, 2009a, Hanebutte et al., 2003] (see section 4.5.3). Thus, the properties potential energy, voids and dipole moment, that loaded on the trivial components C3, C4, and C5, presented an unclear tendency, as we can observe from the profiles of the protein properties in the Figure 5.1. Therefore, we center our attention only to the first two components which explain 60% of the total variance.

Table 5.5: **Summary of principal component analysis on the villin trajectory.** The table shows only coefficients with absolute values greater than 0.5, and only the first five components from nine as the others were non-meaningful for the analysis. We selected only the first two components that loaded with more than one property.

		Components				
ID	Property	**C1**	**C2**	C3	C4	C5
NC	Native Contacts	**.51**				
CO	Contact Order		**.62**			
RG	Radius of Gyration		**-.45**			
HB	Hydrogen Bonds	**.44**				
AS	Accessible Surface Area	**-.49**				
RM	Root Mean Square Deviation	**-.43**				
PE	Potential Energy				-.98	
VD	Voids					.8
DM	Dipolar Moment			.78		

5.3 Discussion

5.3.1 Property Profiles and Factor Analysis

The visual analysis of the property profiles (see Figure 5.1), showed that some properties have relatively similar profiles what suggests similar behavior. But there were other properties in which similarities were not obvious in a visual inspection. Their profiles do not permit to establish concrete relations between properties, and, therefore, we use more quantitative analyses.

The results of the correlation analysis showed two groups: one of correlated and another of uncorrelated properties (see Figure 5.2 and Table 5.2). A first con-

jecture is that correlated properties may be associated with some related folding characteristics; and the uncorrelated may be either associated with 'special' folding characteristics with complex behavior or we need specialized analysis methods to uncover a hidden structure.

We tried to discover this hidden behavior by factor analysis but it proved to be unsuccessful. Only Factor1 and Factor2 loaded with more than one property (see Table 5.3), meaning that they can represent an unobservable characteristic that is not directly measured with the current properties. But the individual loadings on the other 3 factors make them uninteresting for the analysis of hidden structures as they are trivial factors, non-meaningful for the analysis [Dean, 2009a, Hanebutte et al., 2003], evaluating the properties themselves (see section 4.5.3). Resulting factors did not reproduce the relationships of the original variables as best as possible, which we explained with the fact that this type of analysis only evaluates the shared variance. The principal components analysis, in contrast, evaluates the total observed variance and combines properties in a way that allows to associate components with folding features and folding stages, as shown in the following sections.

5.3.2 Components associate with folding features

The principal component analysis (see Table 5.5) allows to state that the observable properties used in the evaluation are grouped into two components C1 and C2, suggesting that they are measuring or representing inherent folding features that are not directly measurable, but can be calculated from the physical properties.

According to the literature, the properties of the component C1 have been associated with the folding features of stability and native-likeness: hydrogen bonds and accessible surface area have been related to the stability in proteins [Hespenheide et al., 2002]; and native contacts and the RMSD have been related to the native-likeness. The properties of the component C2: radius of gyration [Lobanov et al., 2008] and contact order [Ivankov et al., 2009] have been related to the compactness.

5.3.3 Components differentiate types of conformations

Are components capable to distinguish folding stages? To answer this question in the case of our villin-headpiece trajectory, we split the conformations in three sequential groups according to the time interval in which they appear: the initials in the first period; the intermediates after the initials; and the final at the last period. Then we constructed a biplot of component C1 vs. C2, coloring the conformations according to the time-related groups with red, blue and green colors respectively (see Figure 5.3).

As we observe from the figure, the two components were capable of differentiating the three groups, suggesting the existence of folding stages: if we interpret the groups of conformations according to the three periods of time as folding stages (unfolded, intermediates, and folded) then folding features of conformations in the same stage have similar values of, and their values differ from one folding stage to another.

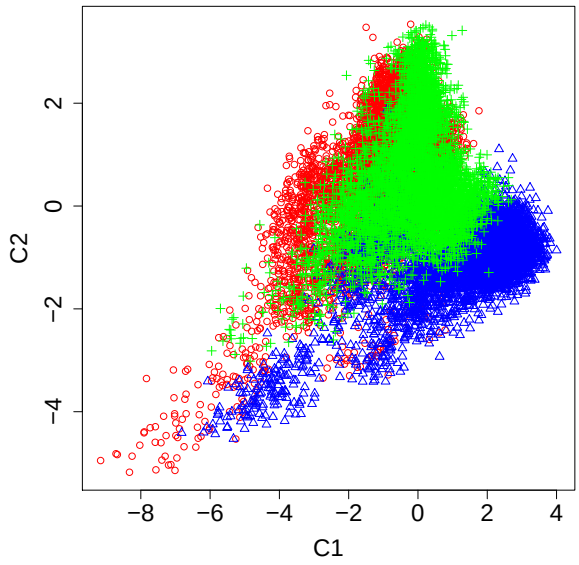


Figure 5.3: **Biplot of C1 vs. C2 differentiating groups of conformations.** The villin trajectory was split into three groups according to the simulation time interval: red circles for conformations at starting times (0 to 2533 ns); blue triangles for the ones at middle times (2534 to 5067 ns); and green crosses for conformations at last times (5068 to 7600 ns).

5.3.4 Concluding Remarks

The preliminary analysis presented in this chapter allowed us to observe the performance of various analysis methods in the case of the villin-headpiece simulation and helped to decide our further proceedings. Our main findings are the following:

1. The principal component analysis is promising in two aspects: first, it allows to reduce the set of folding related properties to few components and permits to interpret them as folding features. Second, the time-related grouping of components is compatible with the feature values that are similar within groups and different between groups, motivating us to associate these groups with folding stages,
2. To obtain a more detailed description of folding features, it is desirable to increase the number of folding properties that enter to the PCA. For this reason, we expand the number of properties to the set of 16, described in chapter 6. We expect that a more refined set of folding features contributes to a more concise description of folding states.
3. The villin-headpiece simulation data are not enough to obtain the generality of results we seek: the villin-headpiece protein is a short sequence of residues which folds basically in helical structures. That's why in the next chapter, we select for our global approach a dataset of 37 protein folding pathways obtained from the Parasol folding that comprises proteins with diverse topologies and of different sizes.

CHAPTER 6

REPRESENTATION BY INHERENT CONFORMATIONAL FEATURES

This chapter describes the process we defined to characterize a protein conformation in terms of a hidden set of folding features that are extracted from a large set of observable physical properties evaluated on the protein structure. We define a score that combines these features and use it to evaluated the progress of folding for the conformations of various protein folding pathways and so to measure the quantity of folding or folding degree of a protein conformation.

6.1 Approach

We performed a number of statistical analyzes to transform a set of 16 properties into a new reduced set of folding features applying the following process: first, we evaluated a set of 16 physical properties on every conformation of a dataset of protein folding pathways. Second, we carried out an exploratory analysis in which we performed a preliminary principal components analysis (PCA) and a non-metric multidimensional analysis (MDS analysis) to evaluate two important criteria: the most relevant properties, and the adequate number of components that reduces the properties set. Third, we used the deduced values to perform a complete PCA, and associate the resulting important components with folding features. And, finally, we created a score that integrates the found features to produce a single value that give us an approximation of the folding degree of a protein conformation.

6.2 Data Preparation

6.2.1 Physical Properties

From the literature, we compiled a set of 16 protein properties and evaluated them on an input dataset of protein folding pathways. In the next sections, we

show how our different analyzes use these evaluations for deriving the folding features. The properties correspond to structural and energetic physical properties used to study and characterize protein folding [Mucherino et al., 2008, Plaxco et al., 1998, Lindorff-Larsen et al., 2011, Lobanov et al., 2008]. Table 6.1 presents a summary of these properties which were detailed in chapter 3.

Table 6.1: **Set of physical protein properties used in this work.** The table shows the IDs of the properties and their names. These IDs will be used throughout the rest of the document.

Id	Properties
NC	Number of Native contacts
CO	Contact Order
RG	Radius of Gyration
HB	Number of Hydrogen Bonds
AS	Accessible Surface Area
VD	Voids
RM	Root-mean-square distance (RMSD)
PE	Potential Energy
DM	Dipole moment
LR	Local RMSD
RC	Residues in correct secondary structures
RA	Residues in any secondary structures
SS	Structural Score
CL	Rigid Clusters
DF	Degrees of Freedom
SR	Stressed Regions

Source: The author

6.2.2 Protein Folding Pathways

As input, we used 37 protein folding pathways simulated with the Probabilistic Roadmap Method (PRM) (see chapter 4). We split the set of pathways into two datasets: one for training and the other for testing. 5506 conformations formed the training dataset, from 28 pathways (80%), and they were used to define the characterization presented in this chapter. And 1446 conformations formed the testing dataset, from 9 pathways (20%) and they were used as test cases for the general validation of our approach. Tables 6.2 and 6.3 show a summary of theses two datasets.

Table 6.2: **PRM Protein Folding Pathways for Training.**

#	Protein	PDB	Topology	Conformations
1	Diadenosine Tetraphosphate Hydrolase	1KTG	4a+7b	305
2	Proteinase inhibitor (trypsin)	4PTI	1a+2b	152
3	Albumin-binding protein	1PRB	4a	248
4	Hydrolase	1O6X	2a+2b	238
5	Chromosomal protein	1UBQ	3a+5b	80
6	Glycosidase inhibitor	1HOE	6b	103
7	Proteinase inhibitor (chymotrypsin)	2CI2	1a+4b	76
8	Blood coagulation inhibitor	1TCP	2a+2b	112
9	Protein g variant nug2	1MI0	1a+4b	185
10	Protein g variant nug1	1MHX	1a+4b	231
11	Matrix metalloproteinase-19 (MMP-19)		3a+4b	259
12	DNA-binding protein SATB1	2O49	6a	415
13	Protein L - Variant L4	2PTL	1a+4b	147
14	Protein L - Variant L3	2PTL	1a+4b	278
15	Protein L - Variant L2	2PTL	1a+4b	105
16	Alpha-amylase inhibitor	2AIT	6b	90
17	Hydrolase (nucleic acid, RNA)	1RBX	3a+7b	140
18	Chromosomal protein (synthetic)	1UBI	3a+5b	74
19	Growth factor	2AFG	10a+4b	71
20	Cardiotoxin	2CRS	5b	93
21	Hydrolase(o-glycosyl)	2EQL	8a+5b	183
22	Drosophila argonaute2 paz domain	1VYN	5a+8b	208
23	Hydrolase(endoribonuclease)	2RN2	4a+5b	429
24	Cytochrome c	2YCC	3a	268
25	Electron transport	351C	4a	338
26	Signal transduction protein	3CHY	5a+5b	283
27	ARKADIA_120 (Designed)		5a+2b	198
28	Fis mutant k36e	1F36	4a+2b	197
TOTAL				5506

Table 6.3: **PRM Protein Folding Pathways for Testing.**

#	Protein	PDB	Topology	Conformations
1	B domain of staphylococcal protein A	1BDD	3a	254
2	Bovine pancreatic trypsin inhibitor	1BPI	2a+2b	108
3	Serine protease inhibitor	1COA	1a+4b	107
4	Growth factor	1EGF	3\betha	81
5	Protein G	1GB1	1a+4b	153
6	Protein L - Variant S2	2PTL	1a+4b	179
7	Protein L - Variant S1	2PTL	1a+4b	136
8	Glycoprotein	1HCC	7a	135
9	Calcium-binding protein	1HML	2a+3b	293
TOTAL				1446

Source: The author. Protein's information was obtained from the EMBL-EBI website `http://www.ebi.ac.uk/pdbsum/`

6.2.3 Evaluation of Properties on Pathway Conformations

All conformations were minimized with a conjugated gradient algorithm from the Gromacs package [van der Spoel et al., 2005a] to remove steric clashes. Then, all 16 properties were evaluated in all conformations of the PRM pathways, resulting in a matrix of data of approximately 7000 rows and 16 columns. Each property was then normalized by scaling its values from 0 to 1 with a min-max unitization algorithm [Jain et al., 2005]. Noisy values were removed

by averaging each of the properties using a moving average filter of length 5. Values were considered noisy when folding produced sharply events that evaluated extreme values on the properties.

Evaluations were performed using our framework for evaluation and analysis of protein folding pathways (described in the chapter 8) that contains, among other features, a toolkit for protein analysis implementing the above-named functions that calculate the properties' values for an input protein conformation.

6.3 From Properties to Components

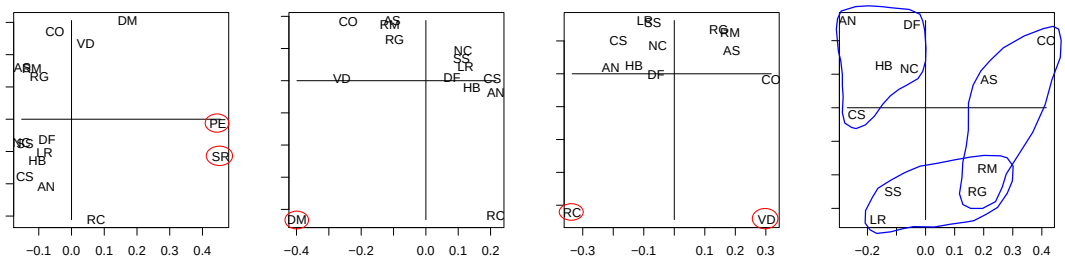
Here, we carried out two analyzes, the first to get an idea of how the properties are interrelated, and the other to obtain the appropriate number of components to use in subsequent analyzes (next section) that define our proposed folding features from the properties.

6.3.1 Selection of Properties

We conducted a series of multidimensional scaling analyzes (MDS) to transform the set of 16 properties into a 2-Dimensional space (2D) or MDS plot and to observe how they were interrelated. The Figure 6.1 shows the results of four MDSs with some properties distant from the others, as in the case of the first plot with the potential energy (PE) and stressed regions (SR) (Figure 6.1.a). These distant properties presented an unclear behavior during folding as we see after evaluating their folding profile during the simulation (see the discussion in this chapter section 6.6.1). So, we took them as 'noisy' properties and remove them one-by-one (Figures 6.8.b,c) until arriving to a structure of three groups (Figure 6.1.d). And, as we will see in the next section, these properties resulted associated with five components non-meaningful for the analysis.

The 11 retained properties were the native contacts, contact order, radius of gyration, hydrogen bonds, accessible surface area, root-mean-square deviation,

Source:The author



(a) 16 properties (b) 14 properties (c) 13 properties (d) 11 properties

Figure 6.1: **Preliminary structure suggested by a series of MDS analyzes.** Properties rounded by red ovals are visually distant from the others, and we removed them for the subsequent analyzes. (a) MDS with all the 16 properties. (b) MDS without the PE and SR. (c) MDS without DM. (d) Final MDS without RC and VD. In the end, the MDS analysis shows 11 properties with a possible structure of three groups (blue ovals).

local root-mean-square deviation, residues in correct secondary structure elements (SSEs), residues in any SSEs, structural score, and degrees of freedom (NC, CO, RG, HB, AS, RM, LR, RC, RA, SS, DF, respectively). And the 5 removed properties were the potential energy, dipole moment, voids, rigid clusters and, the stressed regions (PE, DM, VD, CL, and SR, respectively).

All MDSs used dissimilarities to approximate the distance between properties. The dissimilarities were calculated from the correlations of the different datasets (16, 14, 13, and 11 properties), in the following form:

$$dissimilarities = 1 - |correlations(M)|$$

where M is the matrix of of data of the different datasets.

In the end, properties were represented as points in a 2D plot with their inter-point distance representing the approximated dissimilarities between them (see details in the chapter 4).

6.3.2 Selection of Components

We selected the adequate number of components by performing a preliminary PCA without rotations over the matrix of evaluated properties. Initially, we

retained the first four components as they have eigenvalues greater than 1.0 and explain 84% of the information the original data had (total variance) (Table 6.4).

Table 6.4: **Statistics for selecting the number of components.** Components PC1, PC2, PC3, and PC4 accounted for 84% of the total variance. Note that a solution with three components is also acceptable (78% of the variance).

Source:The author

Components*	Eigen-values	Proportional Variance	Cumulative Variance
PC1	**10.00**	.63	.63
PC2	**1.30**	.09	.72
PC3	**1.10**	.06	.78
PC4	**1.02**	.06	**.84**
PC5	.69	.05	.89
PC6	.52	.03	92
PC7	.38	.02	.94
PC8	.30	.03	.97
...	...	...	...

*Only shown 8 (from 16) components

To get a better interpretation of the preliminary PCA, we rotated the components using a *varimax* method and obtained an improved solution whose matrix of loadings is shown in the Table 6.5. In this solution, we found that only the first three components PC1, PC2, and PC3 were loaded highly with multiple properties while component PC4 was loaded highly with *only one* property.

We removed the last component, PC4, from the analysis as it is considered a trivial component that represents only the property itself and generally, this kind of components are non-meaningful for the analysis [Dean, 2009a, Hanebutte et al., 2003] (see section 4.5.3). In addition, the potential energy (PE) had an unclear trend when was evaluated in the folding pathway of most proteins (see discussion in this chapter, section 6.6.1). Hence, we considered PC4 as a noisy property retained only the first three components PC1, PC2, and PC3, which explain 78% of the total variance (see Table 6.4).

6.4 From Components to Features

Using the above results, now we performed a final PCA to define the final distribution of the 11 properties on the three components (Figure 6.2). This new PCA

Table 6.5: **Loadings of a preliminary PCA with 16 properties.** We selected as main components to PC1, PC2, and PC3 (blue values), which contains more that one property, whilst the remainder (PC4 to PC8) contains only a single one (red values). Combined, PC1, PC2, and PC3, accounted for 78% of the total variance.

Only shown the first 8 components and values greater than 0.5.

Properties	Components							
ID	**PC1**	**PC2**	**PC3**	PC4*	PC5*	PC6*	PC7*	PC8*
NC	**.60**							
CO		**.92**						
RG		**-.51**	**.59**					
HB	**.58**							
AS		**-.76**						
RM		**-.60**	**-.57**					
LR			**-.87**					
RC	**.64**		**-.62**					
RA	**.90**							
SS			**.83**					
DF	**-.58**							
***PE**				1.0				
***DM**						.94		
***VD**							.78	
***CL**					.83			
***SR**								.99

Source: The author

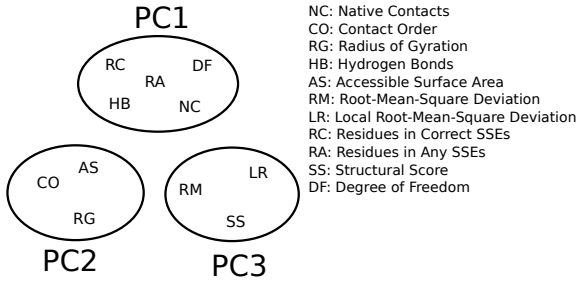
showed that component PC1 mainly was loaded with five properties; while PC2 and PC3, were mainly loaded each with three properties, as shown by the matrix of loadings in the Figure 6.2.a. It is worth noting that the structure of the three components with their included properties agrees with the three groups given by the MDS analysis (see section 6.3.1).

Component PC1 includes native contacts, hydrogen bonds, residues in correct secondary structures, residues in any secondary structure, and degrees of freedom (NC, HB, RC, RA, and DF, respectively). Component PC2 includes contact order, radius of gyration, and accessible surface area (CO, RG, and AS, respectively). And component PC3 includes root-mean-square deviation (RMSD), local RMSD, and structural score (RM, LR, SS, respectively).

Source:The author

Properties	Components		
ID	PC1	PC2	PC3
NC	.72		
CO		-.93	
RG		.64	
HB	.72		
AS		.82	
RM			-.61
LR			-.84
RC	.73		
RA	.89		
SS			.81
DF	-.73		

(a) Matrix of Loadings Final PCA.



(b) Component as Groups of Properties.

Figure 6.2: **Final distribution of properties on components.** (a) The matrix of loadings of the final PCA with the main properties (high weight) for each component. (b) A view of the components as groups of interrelated properties showing some hidden structure. Note that the three components and their main properties agree with the structure of three groups resulted from the MDS analysis (see Figure 6.1).

6.4.1 Interpretation of Components

Inspecting the principal properties that loaded on each of the three retained components (Figure 6.2.b), we associated them with three hidden or not directly measurable folding features: *stability, compactness, and native-likeness* (see Figure 6.3). The ***stability component*** was mainly influenced with by the property related to hydrogen bonds (HB) whose interactions stabilize the protein structures. Its other properties show consequences of these interactions, as the number of native contacts, the residues in correct or any secondary structure, and the degrees of freedom. These properties are reduced as hydrogen bonds impose constraints on bond rotations [Hespenheide et al., 2002]. The ***compactness component*** was influenced by the radius of gyration and the accessible surface area, two properties related with the compactness of proteins [Lobanov et al., 2008]. And the ***native-likeness component*** was influenced by properties related with the structural similarity between a target structure and the native one, described by RMSD, local RMSD and the structural score.

Source:The author

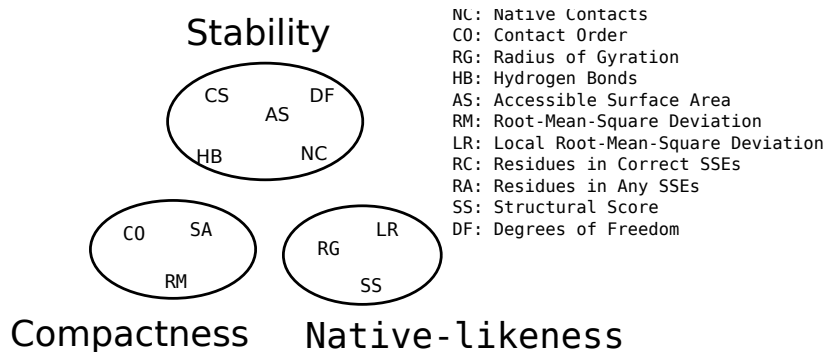


Figure 6.3: **Interpretation of Components.** According to the properties that loaded on the three components, they were associated with three intrinsic features of protein folding: stability, compactness, and native-likeness.

6.4.2 Definition of IC Features

The above interpretation leaded us to propose the components as inherent conformational features of protein conformations, which we called ***IC features***. The values of each feature in terms of its physical properties are calculated following the general form used in PCA to calculate the component scores [Hatcher, 1994]: Each feature is taken as a weighted sum of the protein properties, with loadings (weights) proportional to the strength of the relationship between the feature and the protein properties (see Figure 6.4).

Source:The author

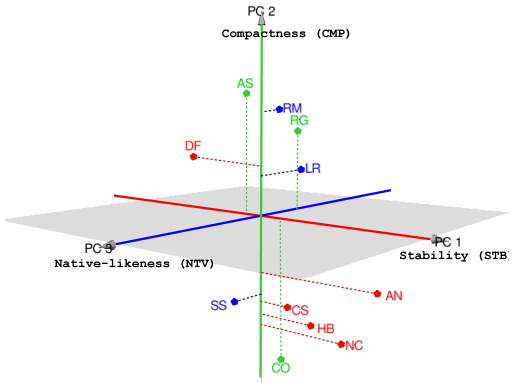


Figure 6.4: **Representation of the IC features in terms of their properties.** The three components define a 3D space that we interpreted as our three IC features: stability (PC1, red axis), compactness (PC2, green axis), and native-likeness (PC3, blue axis). A numerical value can represent them in terms of the distance of their physical properties from the origin (dotted lines) along the appropriate IC feature axis.

The vector of three features $[F_1, F_2, F_3]$ for a protein conformation corre-

sponding to the main components PC1, PC2, and PC3, is calculated by multiplying the vector of its 11 properties $[p_1, ..., p_{11}]$ by the matrix of loadings calculated for the 3 components $[[L_{1,1}, ...L_{1,11}], [L_{2,1}, ...L_{2,11}], [L_{3,1}, ...L_{3,11}]]$ as:

$$\begin{aligned} F_1 &= p_1 * L_{1,1} + ... + p_k * L_{1,k} + ... + p_{11} * L_{1,11} \\ F_2 &= -(p_1 * L_{2,1} + ... + p_k * L_{2,k} + ... + p_{11} * L_{2,11}) \\ F_3 &= p_1 * L_{3,1} + ... + p_k * L_{3,k} + ... + p_{11} * L_{3,11} \end{aligned} \tag{6.1}$$

where F_1, F_2, and $F3$ correspond to our IC Features of stability, compactness, and native-likeness, respectively. The negative sign in the definition of the second feature (F_2), corresponding to the compactness, was introduced to unify the tendency of this feature to increase on a pathway that approaches the native state. Its principal component PC2 decreases as the folding progress, contrary to the behaviour of other two components.

6.5 From Features to Folding Status

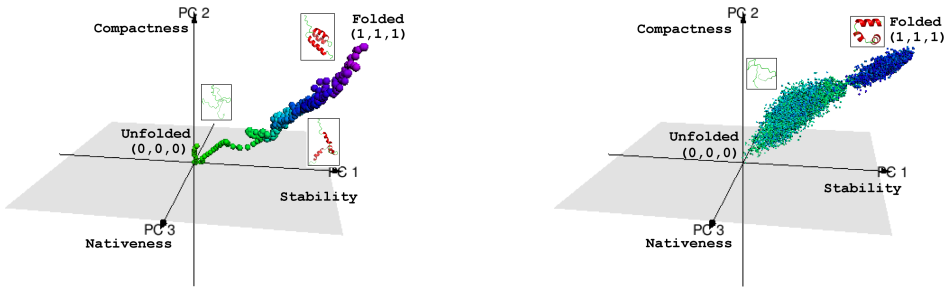
6.5.1 The Folding Status

Now we can describe any protein conformation in terms of the three IC features stability, compactness and native-likeness, rather than using the large set of physical properties; and these IC features became observable properties that can be measured. The three-dimensional feature vector composed of stability, compactness, and native-likeness we refer to as *folding status* of a conformation. The folding status offers a concise, compact representation of the degree a given conformation is folded, summarizing its individual properties and does not depend on its size or topology of the conformation.

As an example of this new description, in the Figure 6.5 we plotted, in terms of the three features, all the conformations of two protein folding pathways: one for a protein with visible intermediate states; and the other for a protein presenting a two-state folding.

In the case of a protein with the presence of intermediates (Figure 6.5.a), our

Source:The author



(a) Pathway three-state folding (b) Pathway two-state folding protein

Figure 6.5: **3D space of features for two pathways.** (a) For the protein with PDB 1BDD proceeding through intermediates (three-state folding). (b) For the villin protein (PDB 2FK4, Anton simulation) proceeding without intermediates (two-state folding). For the two cases, small feature values for conformations near to unfolded state, whilst high feature values for conformations near the folded state.

IC Features show a continuous route from unfolded to folded. But in the event of protein folding without intermediates (Figure 6.5.b), the IC Features present a path with a leap between two groups of accumulated conformations (unfolded and folded). In both cases feature values were small for conformations near to the unfolded states (center of the box, coordinate 0,0,0); while they were high for the others close to the folded state or native (top right of the box, coordinate 1,1,1).

6.5.2 The Score of Folding Degree

The above findings suggested that the features are measuring in some way the quantity or degree of folding for the conformations of the two pathways. To verify this result, we plotted the pathways for multiple proteins (Figure 6.6), and we observed that the features behave in a comparable manner: small values for conformations near to the unfolded state, medium values for the ones at intermediate folding steps, and high values for the ones near to the folded (native) state. Then, we create a score using the three features to get an approximation of the degree of folding (see below). The score defined was named *inherent conformational feature score*, or shortly as ***the ICF score*** and it gives an approximation of the folding degree of a protein conformation by calculating the numerical information of the three features of stability, compactness, and native-likeness. The

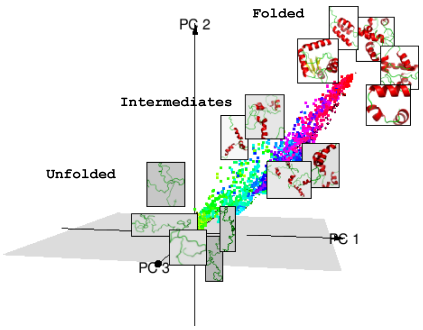


Figure 6.6: **Multiple pathways from diverse proteins.** A sample of five pathways plotted in the 3D space of the IC features. The cloud of points stand for the conformations of the pathways. Color graduation shows folding progress from initial steps (unfolded) to final ones (folded). IC Features describe how the folding proceeds from the unfolded to the folded state.

score takes into account the importance of the features according to the variance of each component, and it is calculated as the weighted sum of the individual IC feature values, as shown in the formula below:

$$ICF(c) = \frac{v_1 * F_1(c) + v_2 * F_2(c) + v_3 * F_3(c)}{V} \tag{6.2}$$

where v_1, v_2, and v_3 are the individual variances for the components PC1, PC2, and PC3, respectively (table 6.4). F_1, F_2, and F_3 are the calculated values for the three features of stability, compactness, and native-likeness, respectively. And V is the total variance explained by the tree components (table 6.4).

The ICF score can be used as a measure of folding progress by verifying how the conformations of a protein folding pathway are accumulating conformational changes to reach the native state. When we compared the ICF score with other three properties used for this purpose (see Figure 6.7), we found that there are differences between properties at various folding steps, but our ICF score was acting as a filter of these differences. It 'averages' part of these sudden changes, suggesting that our score is more informative and robust as it involves multiple properties (see discussion section).

6.6 Discussion

As we derived in this section, a large set of protein properties can be represented by the reduced set of three protein features, the *inherent conformational*

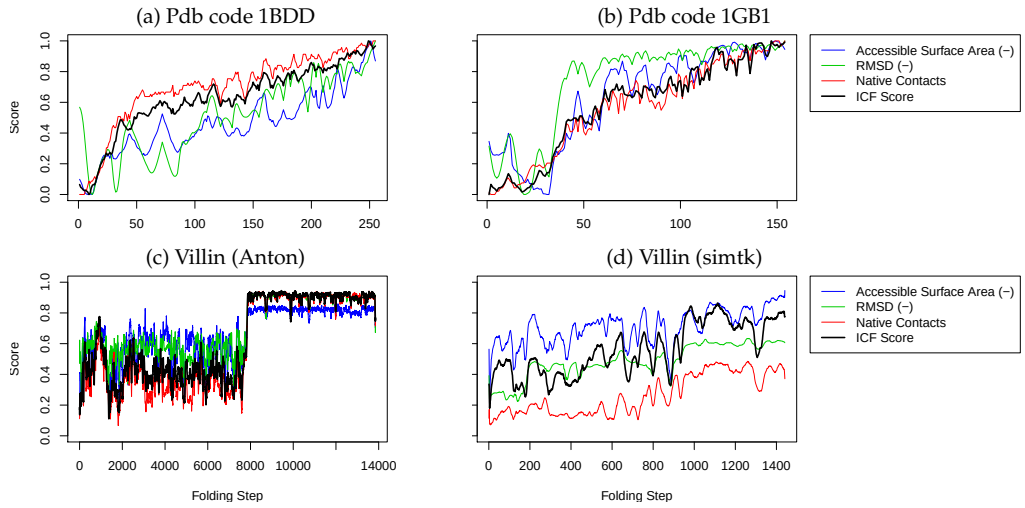


Figure 6.7: **The ICF Score as a measure of Folding Progress.** The ICF score and three measures evaluated on four pathways. (a, b) The score describes a steadily increasing route. (b) The score describes a two-folding state of the villin protein (Anton simulations [Lindorff-Larsen et al., 2011]). (d) All the measures show rugged routes in the villin of the folding@home project, but the ICF score indicates an average of them.

features (or IC features) that are related to the stability, the compactness, and the native-likeness of a protein conformation. The major significance of this finding is that we can 'view' a conformation from another perspective that is more related with underlying characteristics of protein folding than simple physical properties. Consequently, this new representation derived a macroscopic measure of the folding status of conformations, which can be used to characterize protein folding pathways (see next chapter), a fact we consider one of the main contributions of this chapter.

In the following, we discuss other aspects related to the IC features representation of a conformation and its corresponding score.

6.6.1 Noisy Properties

The selection of properties showed that five of them were 'special' (section 6.3.1): the potential energy, dipole moment, voids, rigid clusters and, the stressed regions (PE, DM, VD, CL, and SR, respectively). Whereas the MDSs showed them distant from the others, the PCA associated each one with a trivial component, non-meaningful for the analysis. In addition to this, their folding

profile–evaluating the property value at each folding step–did not indicate any trend for most of the proteins. Such is the case of the potential energy shown in the Figure 6.8, where we expected that its values decrease as a protein is close to its native state at the final path steps of the pathway. Therefore, we determined to remove them from the rest of our analyzes and work with the others eleven properties.

We explain the behavior of these properties as caused by different events. For the potential energy and dipole moment, the functions for calculating their values depend on a potential function or force field, which, in our case, is the Amber96, used in all our calculations. However, force fields are designed or tuned to favour some protein configurations as the ones with significant helical or beta content. But we are using proteins with different topologies for which this force field could be inadequate and, therefore, resulting the random like values.

In the case of voids, although they are related to packing in protein structures, packing by itself is not a major indicator of stability and so of folding [Cuff and Martin, 2004]. Then, voids can be appearing and disappearing dynamically during the folding of a protein and so their evaluations can present high and low values without a clear trend during its folding. The last two properties: rigid clusters and stressed regions, have also been associated with packing [Jacobs et al., 2002], and so their resulting values not necessarily reflect the folding of the proteins.

6.6.2 IC Features

The final PCA showed that the arrangement of properties for components outline global characteristics of protein folding that we interpreted as three folding features: stability, compactness, and native-likeness. Next, we discuss each of these features.

Source:The author

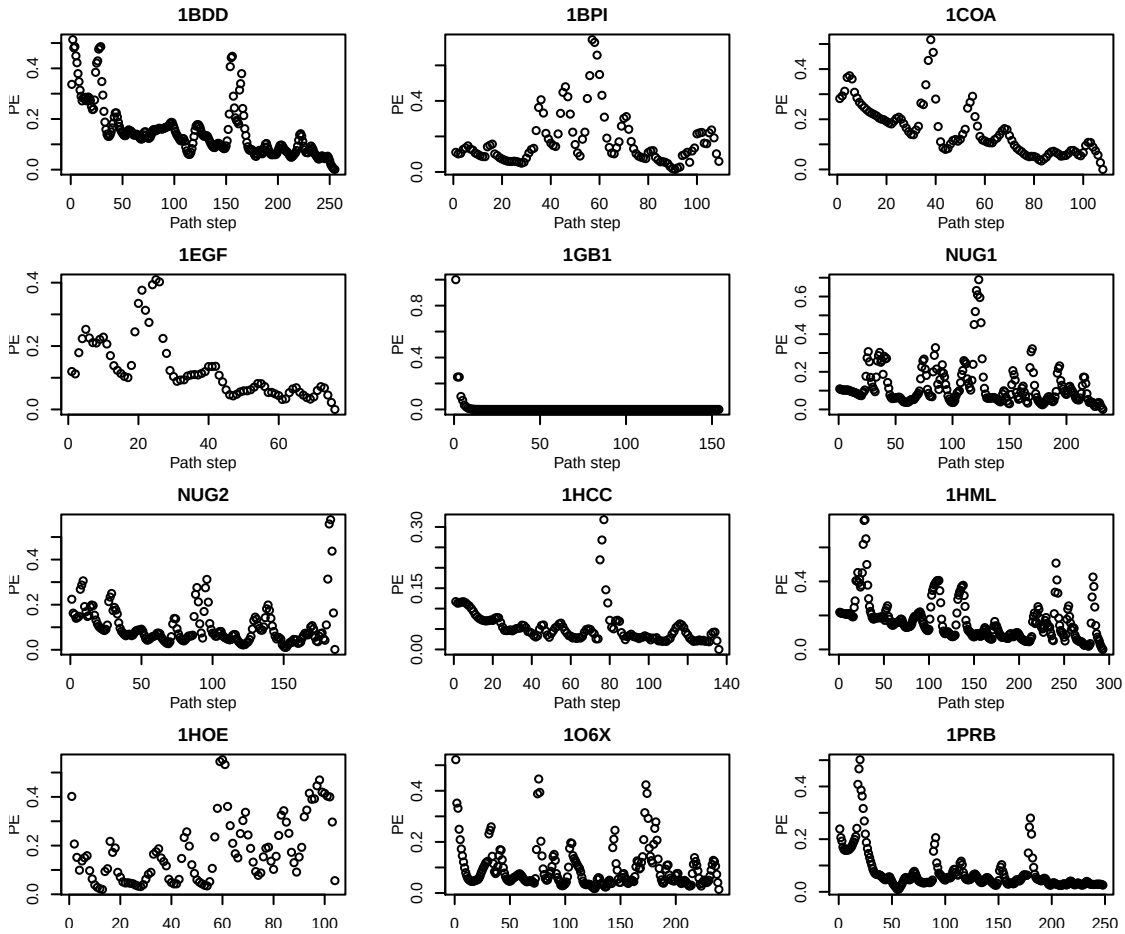


Figure 6.8: **Potential energy without a clear tendency in protein folding pathways.** Profiles of the potential energy (PE) for a sample of 12 protein folding pathways in which the values of PE for the conformations for each pathway are very noisy. We also observed a similar behaviour for the other four 'special' properties removed from the analysis (DM, VD, CL, and SR).

The Stability Feature

We named this component ***stability feature*** (component PC1) as it was represented by five properties associated with the stability, understood this characteristic as the maintenance of a defined structure under folding conditions. According to our analysis, stability is the most important folding feature for determining the folding of proteins. The explanation of this result is related to the main of the five properties integrated in the stability feature: the formation of hydrogen bonds (HB), which have played a critical role for the stability of the final folded structure [Murphy, 2001, Onuchic and Wolynes, 2004].

Hydrogen bonds promote the organization in proteins in different ways [Gong et al., 2005], for example contributing to the stabilization of α helices and β sheets [Whitford, 2005]. Also, two properties, integrated in the stability feature, can be interpreted as a consequence of the formation of hydrogen bonds: first, the quantification of secondary structures elements (SSEs), which refers to forming complex structures constraining the protein and incrementing its stability [Fleming et al., 2006]. Second, the formation of small networks of native contacts (NC), is also related with the stabilization of proteins [Faísca et al., 2010], and its quantification has been a common measure of the progress a protein acquires during its folding [Beck and Daggett, 2007]. Finally, the degrees of freedom (DF), concerned with the rigidity of a protein structure [Schirf,], correspond to a decrease in conformational entropy and in this way to the stabilization of proteins [Jaenicke and Böhm, 1998].

The Compactness Feature

We understand compactness as the ratio of accessible surface area of a protein to the surface area of the ideal sphere of the same volume [Lobanov et al., 2008]. We found the second feature (component PC2) was loaded by three properties related with the characteristic of compactness in proteins: radius of gyration (RG), accessible surface are (AS), and contact order (CO). So we interpreted this feature as a measure of compactness calling it ***compactness feature***. This feature is less weighted than the previously discussed stability feature, contributing with only 9% of the variance (Table 6.4). According to our data, compactness is an important characteristic to determine the folding of a protein, but it is not as crucial as the stability. We support it by experimental works about intermediates in folding that have observed the presence of compact intermediates in the early and late stages of the folding, some of them acting as milestones helping to follow the correct or quickest pathway [Roder et al., 2006, Chen et al., 2008, Udgaonkar, 2008, Chalikian and Breslauer, 1996]. Thus, although the presence of these compact intermediates may be relevant to search and converge quickly to native states, proteins may fold without their explicit presence.

The Native-likeness Feature

The third feature (component PC3) contains properties associated with the characteristic of native-likeness, understood as the degree of likeness of a protein conformation with its native structure. Therefore, we called it the ***native-likeness feature***. It is the least important of the inherent features as it only explained 6% of the variance (Table 6.4). This result can be explained by one of its integrated extrinsic measure, the RMSD that depends on the native structure. But, the it does not take into consideration structural elements that may not still spatially organized, although they can be present in the conformation as well as in its native [Fleming et al., 2006, Mucherino et al., 2008]. We balance this limitation of the RMSD by our definition of two new properties: the local RMSD and the structural score (see chapter 3). The former measures the structural similarity between two protein structures by comparing locally their secondary structure elements (SSEs). The later compares local structures using short local motifs or protein blocks (PBs) [Kolodny and Levitt, 2003] instead of SSEs. Consequently, these two properties that complement the RMSD were integrated into the native-likeness feature by the PCA.

6.6.3 IC Feature Score

Our ICF score combines the three IC features to measure how folded a conformation is, or to what extent the conformation is folded, its folding degree. The information given by the score will allow us (as we will see in the next chapter) to compare pairs of conformations of a same protein and so relate them with possible conformational states of folding. Observing the ICF score on a pathway (Figure 6.4), it allows to measure how a protein evolves in its route to the native. Therefore, the score can be used as global measure of the folding progress similar to the way that other individual measures are used in simulations, as for instance RMSD, radius of gyration (RG), or other native-contact derived measures [Lindorff-Larsen et al., 2011, Duan et al., 1998]. Nevertheless, the combination of more informative IC features turns the ICF score in a more complete, robust, and multi-dimensional measure of folding progress, conversely to using an isolated property, which may work well when the property is related with

the conformational characteristics of a conformation, but may give unexpected results in other cases.

As an example, we evaluate the ICF score along with some separated properties on four protein folding pathways from our dataset and from the molecular simulation of the villin headpiece (Figures 6.5 and 6.6). Our ICF score describes the folding progress as the other three measures, but in a more subtle and sophisticated way. Its values present in some folding steps lesser sudden changes than the other measures , as can be observed in the evaluation of the villin from the folding@home project (Figure 6.7.d), in which all the measures but our score describe extreme situations (accessible surface area arrives near to the native, but native contacts suggests the opposite). On the contrary, our ICF score shows a rugged route that approximated to the native, but it is never reached, in agreement with the strategy the authors of the simulations applied [Ensign et al., 2007] determining that the protein is close to the native by comparing parts instead of the entire protein structure.

The above means that the score is more robust to these type of changes where the other measures are more sensitive. The ICF score has the advantage of being a more informative measure because it involves multiple properties giving it the capacity to compensate deficiencies of some measures with the strengths from the others. Similar problems have been highlighted in other works [Dokholyan et al., 2002, Beck and Daggett, 2007, Das et al., 2006, Toofanny et al., 2010b], where the use of various properties is proposed in order to obtain a more informative measures of folding progress, considerations that motivated the present work.

6.6.4 Limitations

Concerning the Input Dataset

In the calculation of the ICF features we used a dataset of 37 folding pathways for different proteins constructed by the Probabilistic Roadmap Method (PRM). The PRM samples a large conformational space and constructs the pathways by

selecting the more plausible conformations according to a defined energy function. But, it is well known that the complete conformational space of a simple molecule is impossible to construct, and therefore the PRM is only sampling a limited conformational space. But instead of taking this fact as a limitation, we see as an advantage in the sense that the PRM is intended to encode the main folding events (on-pathway intermediates), which allows us to focus on their conformational information and get the general parameters to construct our model of IC Features. So, the main product that we take from the PRM pathways is the knowledge that we can extract from the large and diverse set of protein conformations that the PRM is able to sample.

Another limitation about the PRM pathways is that could be seen in the fact that they do not have specific information about time, in contrast to the trajectories generated by Molecular Dynamics (MD) simulations. But this is not relevant for our approach as we are not using time related information; our focus is on the conformational properties of the proteins. Nevertheless, as shown when we used the IC Features to describe folding (Figure 6.5), the IC Features revealed an implicit temporal ordering (from unfolded to folded states) according to the folding steps of the pathways, which in turn establishes a certain 'folding time' in terms or progress of folding instead of time of simulation.

Concerning the Selected Properties

We are evaluating a set of 11 very general physical protein properties. We carefully selected the included properties with the primary goal of including general characteristics that can be evaluated in all the proteins. So, properties as number of hydrogen bonds, radius of gyration, accessible surface, degrees of freedom are general characteristics of all proteins. In the same way the content of secondary structure elements (SSEs) as the number of residues in any or in correct SSEs are properties that are independent of the kind of SSEs (e.g. α or β contents). We are omitting more specific properties such as the content of α-helices or β-sheets, that in some cases may be adequate for specific topologies of proteins, but could put at risk the generality of our approach.

CHAPTER 7

CHARACTERIZATION OF FOLDING LEVELS

Protein conformations can assume different conformational states, from unfolded, passing through intermediate states, until reaching the folded state or native. From these states, the intermediates are of great importance as they still present many questions to solve. For example, it is unclear which role they play, or inclusive if they truly exist during the folding process. But although improved experimental techniques have proved their presence in many proteins, their characterization is still a difficult task as experimental information is at present somewhat limited.

However, protein folding simulations gives much detailed information about each folding step of the process that we can use it to construct a computational characterization of the intermediates. So, in this chapter we analyze the conformations of many simulated protein folding pathways with the aim to discover a hidden folding structure that allows us to define a mechanism for determining the most plausible conformational stage assumed by a protein conformation or what we will call as ***folding levels***. All of this was done according to the protein's folding status given by the inherent conformational features, defined in the previous chapter. The conformational stages that we are mainly interested in are the ones related to intermediates events.

7.1 Approach

The folding status defined in the previous chapter (section 6.5), indirectly measures the folding features (stability, compactness, and native-likenness) of a protein conformation through the knowledge of a set of physical properties. Now, we use this definition to search for an implicit structure given by protein conformations having similar conformational features.

We performed a hierarchical clustering on the protein conformations using their folding features calculated from their physical properties. But first, we defined a similarity measure of folding statuses between protein structures and

used it for the clustering algorithm. The resulting groups were analyzed according to the folding stages of their conformations what we call as ***folding levels***. We obtained a representative description of every level. And, finally, we determined the implicit folding dynamics presented between the levels with the goal to analyze how conformations are changing from one level to another.

7.2 A Similarity Measure for Folding Statuses

The previous transformation showed a protein conformation in terms of a reduced set of features embedding a larger set of different structural and energetic physical properties. Now, we use this transformation to define a similarity measure for comparing two protein conformations with respect to their IC features or folding statuses.

We define the inherent-features similarity measure (IF measure) for two protein structures as the Euclidean distance between the vectors of their IC features: stability (ST), compactness (CP), and native-likeness (NT). Formally, given two protein conformation a and b along with their three features scores (ST_a, CP_a, NT_a), and (ST_b, CP_b, NT_b), respectively, define the Euclidean distance between them (δ) as follows:

$$\delta(a,b) = \sqrt{(ST_a - ST_b)^2 + (CP_a - CP_b)^2 + (NT_a - NT_b)^2} \tag{7.1}$$

The measure allows us to know how far or close a pair of conformations are with regards to their folding process. Also, it permits to characterize the conformations of a protein folding pathway in terms of groups with common conformational characteristics according to their folding degree, as we will see in the next sections.

7.3 Discovering Groups

We use the structural information encoded in the dataset of pathways taken from the Parasol server (section 4.4 in chapter Methods) to perform a hierarchical clustering searching for both a hidden structure in their conformations and a mechanism to assign the folding level to any conformation.

The dataset contains 37 pathways with approximately 7000 conformations that we initially split it into two sets: one for training a classifier (80% of the pathways) and the other for testing it (20%). We did that because our aim was to use a supervised classifier from machine learning to train and use it for predicting the folding stage of protein conformations. But it was non-viable as the training needed the information of the classes, and we did not have it, that is, the specific folding stage for each protein does not exist or is not easily available. So, we used an unsupervised classifier (the hierarchical clustering) that doesn't need the classes for the conformations but it divides them into groups capturing the natural structure of the data.

To perform the clustering, we took the similarity measure for folding statuses previously defined in section 7.2. And for the input data, we calculated the folding statuses of all the protein conformations from the training set, leaving the pathways of the testing set to use as test cases. We evaluated the physical properties to the conformations and computed their folding features using the equations defined in 6.1. The Figure 7.1 shows a general description of the data transformation, from the 11 properties to the three features of stability (ST), compactness (CM), and native-likeness (NT).

$$\begin{array}{c} \text{Physical Properties} \\ \begin{bmatrix} & P_1 & P_2 & \cdots & P_{11} \\ c_1 & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_n & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \end{array} \longrightarrow \begin{array}{c} \text{Conformational Features} \\ \begin{bmatrix} & ST & CM & NT \\ c_1 & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ c_n & \vdots & \vdots & \vdots \end{bmatrix} \end{array}$$

Source:The author

Figure 7.1: **Structure of the Matrix of Conformational Features** The matrix of 11 physical properties was transformed into a matrix of conformational features by calculating for each conformation $c_1, ..., c_n$ of the training pathways the associated features ST, CP, NT corresponding to the stability, compactness, and native-likeness, respectively.

After run the clustering, we obtained a structure of four groups: g1, g2, g3, and g4 (Figure 7.2). The enumeration order agreed to the temporal appearance of the groups in the pathways (explained in section 7.4), after cutting the dendrogram at the height of 1.0 subject to the criterion used in [Pardalos et al., 2012]. Below this point the distance between the objects being joined is approximately the same as the distance they contain.

Source:The author

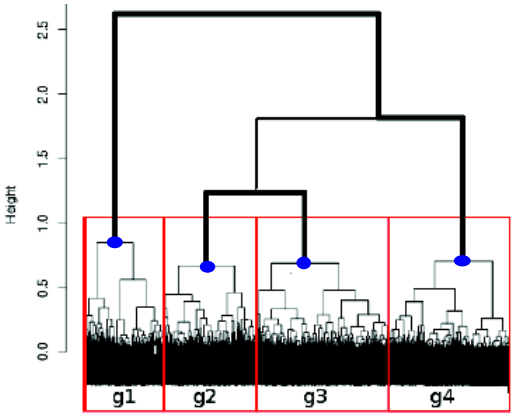


Figure 7.2: **Clustering given by the three features.** Red boxes for the four groups. Blue points show that below this point the groups are more consistent than above them. Brackets give an approximation of how groups are separated from each other: g1 and g4, most distant groups; while g2 and g3 the closest ones.

The organization of conformations in four groups led us to search for their characterization in terms of how the three conformational features: stability, compactness, and native-likeness, behave in each one. So, we plotted the distribution of the three features on the four group (boxplots Figure 7.3), and found that each group has a unique combination of feature values that makes it distinguishable from the others. In our case, the values used to represent the features were their medians, which corresponded to the coordinates for the central point for each group within the clustering.

It is worth noting the generally increasing trend of the feature values across the groups (red lines, Figure 7.3), where the values of features increase the higher the group. So the group 1 has the lowest values of stability, compactness, and native-likeness, whereas the group 4 has the highest ones. It indicates us that the groups could be associated with stages of folding, being in the group 1 the conformations whose structures exhibit a low stability, a loose core, and

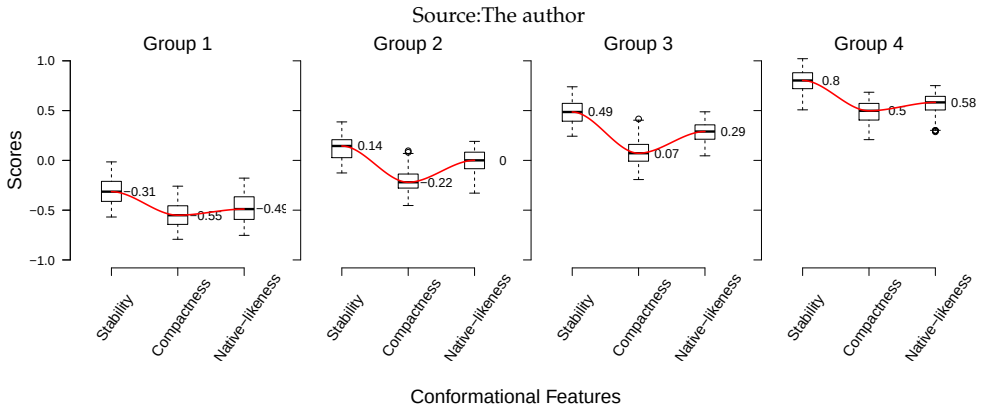


Figure 7.3: **Distribution of the features on the four groups.** Parallel boxplots showing how the three features (at the bottom) distribute to each group. The red line passing through the median value of each feature shows how each group has a particular response to its three features.

without native elements. Whereas the ones in the group 4 exhibit high stability, a very compact core, and present native elements. All this led us to consider that each group has a significant role in the process encoded in the folding of a protein, what we are going to describe in the next section.

7.4 Groups and Folding Dynamics

To observe how these groups appear dynamically in the protein folding pathways, we inspected visually the way how they are distributed along the pathways. Figure 7.4 shows a sample of three pathways with groups of conformations colored as black, red, green, and blue colors, for groups 1, 2, 3, and 4, respectively. We can observe that conformations assigned to specific groups appear at specific steps of the folding, group 1 at initial steps, group 2 and 3 at intermediate steps, and group 4 at final steps. This finding suggested us that the groups are encoding a temporal ordering that we could associate with stages of folding, when a protein folds from unfolded to folded state.

To confirm and generalize the previous visual result, we modeled the dynamic assignation in the entire set of pathways as a Markov chain, and constructed the transition diagram from the assignations of groups to pathway conformations (Figure 7.5). As shown in the diagram, there is a strong tem-

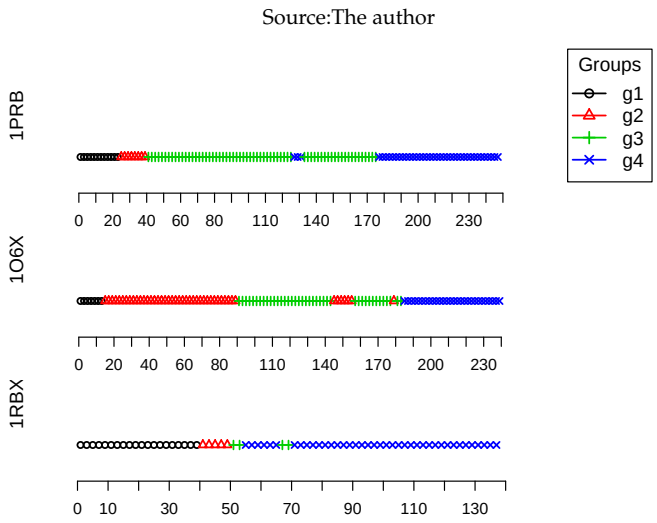


Figure 7.4: **Assignation of groups for conformations of three pathways.** Pathway profile for the proteins with PDB: 1PRB, 1O6X, and 1RBX. At the bottom, the folding steps (0 to n) from unfolded to the native state. In general, group g1 was at initial, g2 and g3 at intermediate, and g4 at final steps.

poral relation between the four groups, mostly sequential, with initial state g1, two intermediates g2 and g3, and final state g4.

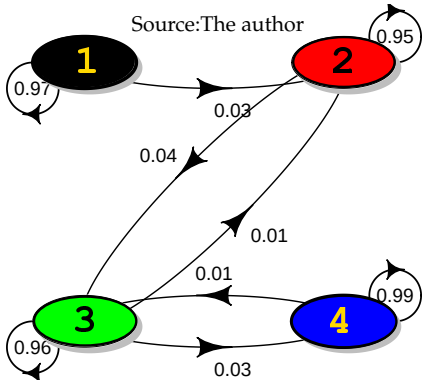


Figure 7.5: **Transition diagram for groups of conformations.** The diagram shows the probability of transitioning between states, with state 1 as the initial, 2 and 3 as the intermediates, and 4 as the final state. States are marked with colors as the ones used to identify ordering of groups in pathways (Figure 7.4).

In all considered pathways, all conformations were assumed according to the temporal ordering; the process always starts in state 1 (no back transition) and ends in the state 4 (highest probability to remain within the same state). Never, neither in the forward direction nor backward, neighboring states are skipped. The probabilities to stay in a given state always are superior to the leaving ones. There are no jumps omitting consecutive states, and the forward

probabilities are always greater than the backward ones.

7.5 Introducing Folding Levels

Justified by the characteristics of each group and the temporal arrangement of groups in pathways we interpret the groups or states in the transition diagram as ***folding levels***: Accumulating changes from the unfolded to the native conformation, a protein assumes dynamically different conformational states that share similar conformational features. Our previous analysis allows to distinguish the following four main levels: ***unfolded, early intermediate, late intermediate, and folded***, as we can see in the Figure 7.6.

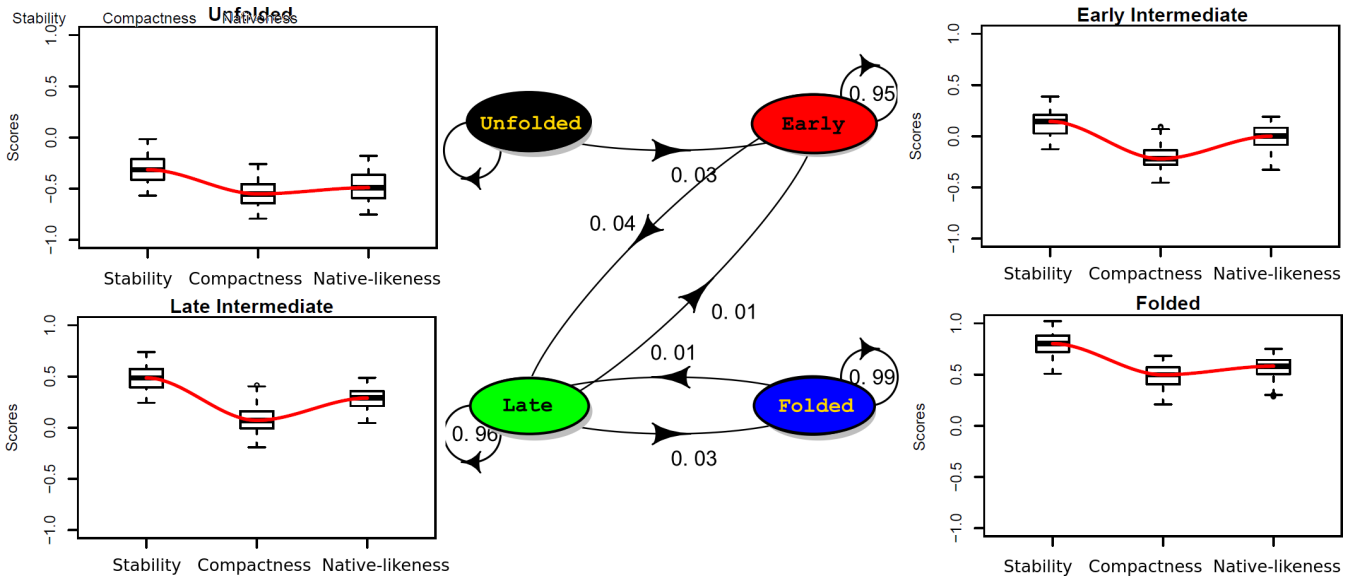


Figure 7.6: **Transition diagram for protein folding levels.** Groups of conformations can be seen as metastable states of folding or *folding levels*. Proteins start in an unfolded level, pass through early and late intermediates, and end in a folded level. Distributions for the three features for each level on the side of each state.

According to the results of the section 7.3, each folding level can be characterized by a unique response of the three conformational features of stability, compactness, and native-likeness, which improve along the folding levels (see Figure 7.3).

Using the ICF score introduced in the previous chapter (Equation 6.2), we

can summarize the folding levels by their characteristic ICF score ranges, as shown in the Figure 7.7. The unfolded level assumes a range from -0.42 to 0.02, the early intermediate from -0.09 to 0.34, the late intermediate from 0.19 to 0.61, and the folded from 0.4 to 0.8.

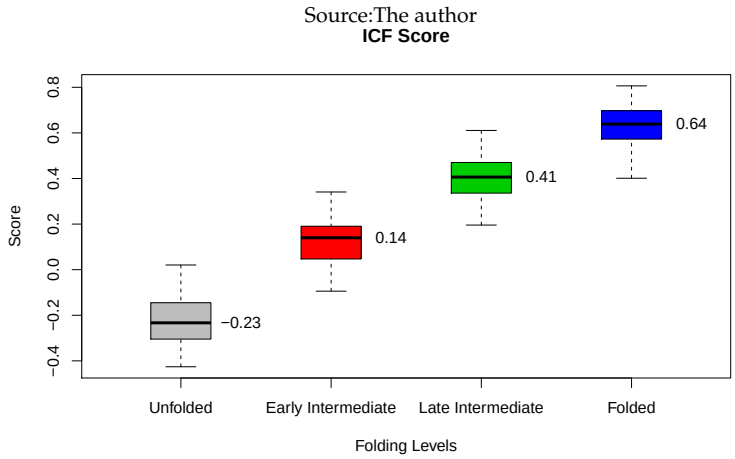


Figure 7.7: **Distribution of the ICF score for the four folding levels.** The score values grow as levels of folding pass from unfolded to folded. Each folding level is well differentiated from the others as the median value for each level shows.

7.6 Assigning Folding Levels to Conformations and Pathways

Using the similarity measure for folding statuses previously introduced (Equation 7.1), we construct a classifier that assigns the folding level to a given protein conformation: for each protein conformation c we can evaluate its distance to the central conformation g_k as $\delta(c, g_k)$. The folding level $lev(c)$ is assigned as the value of k which minimizes the distance between c and g_k.

$$lev(c) = \underset{g_k}{\operatorname{argmin}}\, \delta(c, g_k) | k = 1, ..., 4 \tag{7.2}$$

Figure 7.8 shows the assigned folding levels for conformations of four folding pathways. A pathway starts in the unfolded level and proceeds through early and late intermediates to reach the folded state, describing the folding process of a protein according to the dynamics of its IC features. Conformations with a low degree of folding are at initial folding steps, and the ones with a high degree are at the final folding steps. Our assignations follow this order,

conformations predicted as unfolded, and early intermediates are located at the beginning of the folding steps, and the ones predicted as late intermediates and folded are located at the end of the folding steps.

Also, note that in the Figure 7.8 some pathways present no-persistent leaps to neighbouring levels. We explain this behaviour due to the dynamics of the folding process itself. Proteins during folding constantly rearrange in a way that some of their structural elements can appear and disappear in trying to reach more stable states, what may generate the leaps in the pathway. But, what makes this behaviour more interesting is that our model of the dynamics of the folding levels captures it (see section 7.4). As we observed in the transition diagram for protein folding levels (see Figure 7.6), there is a high probability of the conformations in one level for transitioning forwards (e.g. early to late intermediate), but also there is much smaller probability of transitioning backwards (e.g. late to early intermediate). And when this last event happens, it is observed in the diagram of folding levels the short jumps in the assignations (Figure 7.8).

7.7 Validation

We validated both the quality of the obtained cluster and the data independence of the clustering process, using an internal validation for the former and relative validations for the latter (see chapter 4). The internal validation used the Silhouette index giving a measure of clustering quality based on inherent properties of the clustering. The relative validation used a resampling approach, which validates the global solution by comparing it with particular solutions using the same clustering process but with different parameters.

7.7.1 Internal Validation

To calculate the Silhouette index, we used all the values resulting from the clustering. The index is comparing how close a protein conformation is to others within its cluster and how close it is to conformations in other clusters. A sum-

Source:The author

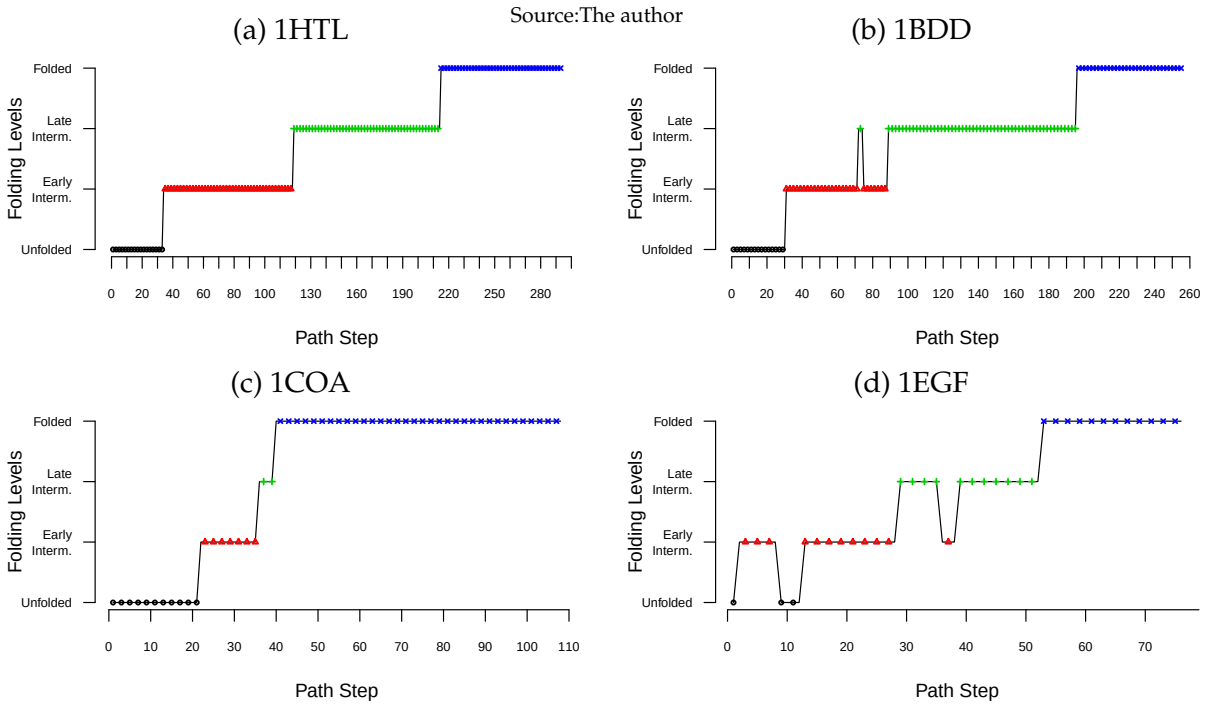


Figure 7.8: **Folding levels for four pathways.** Pathways for the proteins with PDB: 1HTL, 1BDD, 1COA, and 1EGF. Conformations arranged along the x-axis as they appear in their pathway: unfolded to folded. Assignations follows this sequential ordering: unfolded levels at initial path steps (black circles); intermediate levels (red triangles for early ones, and green line for late ones); and folded levels at final path steps (blue Xs). Some pathways present no-persisting leaps to neighbouring levels.

mary of the Silhouette index is shown in the Figure 7.9. The overall average Silhouette width was 0.53, a good value of quality as this index ranges from -1 (misclassified) to 1 (correct classified). The groups were well distributed, and few conformations with negative indexes (horizontal lines at the bottom of each group) were obtained.

7.7.2 Relative Validation

We applied the resampling method for unsupervised estimation of clustering validity proposed in [Levine and Domany, 2000]. It is based on the fact that if a global clustering is valid then each of its subsets should be valid as well (see chapter 4). For the validation, we used as input all conformations from our input dataset of pathways. Then, we run on it a global clustering using the same hierarchical clustering described in this work (see section 7.3). After that,

Source:The author

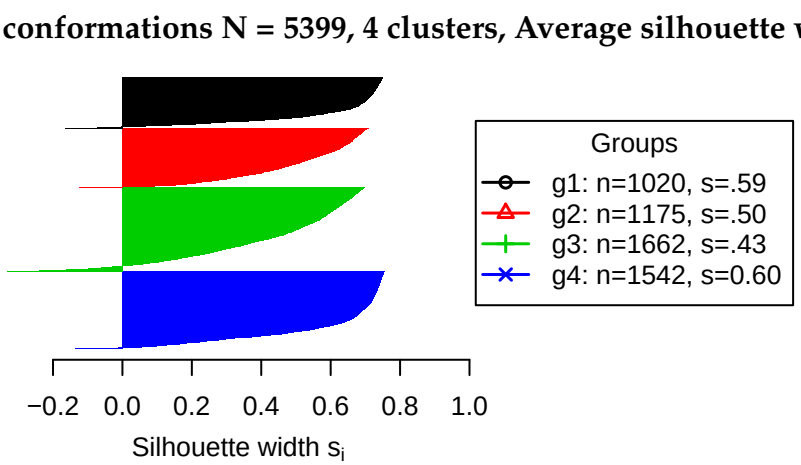


Figure 7.9: **Silhouette plot describing the quality of the clustering.** The plot shows the four groups of the clustering as colored blocks. The overall average Silhouette index was 0.53. The legend on the right side shows for each group the number of conformations (n) and the average silhouette (s). The individual groups also present good Silhouette widths.

the dataset was split into ten subsamples, and we run the same clustering for each of the subsamples. At the end, we compare the results of each subsample clustering in accordance with the results of the global clustering.

Table 7.1 shows the results that we obtained after running the clustering for the full dataset of 6000 conformations approximately, taking from it ten subsamples, each with 90% of random conformations. The method defines a *figure of merit*[1] for comparing the agreement of clustering solutions of the subsamples with the ones of the global clustering and take their average as a measure of stability. The overall average of agreement was 0.84, with subsample agreements in the range from 0.77 to 0.93.

The result the resampling method, an overall agreement of 0.84, indicates that our clustering process has a good stability and validity, given that the agreement in this method ranges from 0 (bad agreement) to 1 (full agreement).

[1] A figure of merit is a numerical quantity that represents a measure of efficiency

Table 7.1: **Results of relative validation by subsampling.** The table compares the global clustering with the solution of each subsample. The first column denoted the sample, and the second column contains the percentage of the agreement (figure of merit) of each sample. The average agreement was 0.84 that is a good validity value for our overall clustering procedure.

Sample	% of Agreement
1	0.87
2	0.93
3	0.94
4	0.82
5	0.80
6	0.79
7	0.80
8	0.93
9	0.77
10	0.77
Average agreement	**0.84**

Source: The author

7.7.3 I-sites and Folding Levels on PRM protein folding pathways

I-sites are short sequence patterns or structural motifs that have been observed repeatedly in the structure of different proteins of the Protein Data Bank (PDB) [Bystroff and Baker, 1998]. Studies show that they strongly correlate with local protein structure features, so they have been proposed as initiation sites of folding [Bystroff and Shao, 2003].

Here, we analyze the relation between the I-sites and our defined folding levels by tracing the I-sites present in the native structure of a target protein in the conformations from its folding pathway. Our hypothesis is that if the I-sites are acting as potential initiation sites, they should appear in our defined early intermediate folding level, where the protein start to exhibit some structural elements and characteristics that could guide the process of folding of the protein.

To determine which I-sites to trace in the folding pathway of a protein, we defined a general process and created an algorithm to search for the best superimposable I-sites in locations of a protein structure. Our results show that for the torsion angles of the I-sites found in the native structure of a protein, there is a pronounced decrease in the average RMS with the torsion angles of conformations in the unfolded level until reach the early intermediate level, from which

the RMS values vary slightly.

General process

To trace the I-sites in the folding pathway of a target protein, we determine first the folding level for its conformations using our approach describe in the section 7.6, and then we search for locations in the native structure of the protein that are the close to I-sites. The search is carried out using an algorithm that recursively compares I-sites with locations in the structure using the Euclidean distance of dihedral angles, both described below.

Once the protein's I-sites are determined, we trace them backwards on the folding pathway by measuring their Euclidean distance with respect to the determined locations in the conformations. It produces a vector of distances, one for each I-site, that we average to get a single value. Finally, we calculate the global average per level and plot these values in an XY plot describing the average inter distance of torsion angles between the determined I-sites and their locations in conformations of a specific folding level.

Methods

I-sites library and pathways We are using the I-sites library version 5.3 that consists of about 250 motifs (`http://www.bioinfo.rpi.edu/bystrc/Isites2/isl5.3`). The pathways for tracing the I-sites correspond to the ones generated by the Probabilistic Roadmap Method [Song and Amato, 2001] (`https://parasol.tamu.edu/foldingserver`).

Structure distance based on torsion angles To assign the most appropriated I-sites to a protein conformation, we perform comparisons between each i-isite and locations (subsequences) of a target protein structure. From the I-sites library, each I-site is represented by its dihedral angles ϕ, ψ, and ω, so we compute the root-mean square distance (RMS) of the dihedral angles between the target structure and the I-sites using the Euclidean distance between two pairs

of torsion angles (ϕ, ψ).

Particularly, the distance, D, is calculated from the dihedral angles of the I-site I and the ones of a location in the protein structure P as follows:

$$D(I,P) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\phi_{I,i} - \phi_{P,k+i})^2 + (\psi_{I,i} - \psi_{P,k+i})^2}$$

where N is the size of the I-site I and k is the position of the starting residue in the protein P from which the evaluation is done. The quantities $|\phi_{I,i} - \phi_{P,k+i}|$ and $|\psi_{I,i} - \psi_{P,k+i}|$ should be less than 180°, otherwise the 360° complementary angle is used.

Algorithm for assignation of I-sites For assigning I-sites to locations of protein conformations, we developed a recursive algorithm that starts by searching for the I-site with the best fit along the complete sequence. According to our distance measure, we decide that an I-site fits in a protein location if the RMS for dihedral angles is lower than a defined threshold (8 degrees, in our case). Each fitted I-site is assigned to the corresponding location in the sequence and causes the sequence to be split into the two subsequences to the left and to the right of the I-site. Then, the assignment algorithm is called recursively, but now for the new two subsequences. The process stops when the subsequences are empty. Then the resulting left and right solutions are joined in one sequence, and by the recursion, the global solution is returned, which corresponds to a list of fitted I-site locations.

Algorithm 1 Recursive algorithm for I-sites assignment.

```
recursive_assignation ALGORITHM (protein_sequence, isites_library):
  if EMPTY (protein_sequence):
    RETURN []

  n = LENGTH (protein_sequence)

  assignation_list = [0..n]
  best_isite = SEARCH_BEST (protein_sequence, isites_library, 0, n)
  ini, end = GET_LOCATION_ISITE (best_isite)
  INSERT (assignation_list, best_isite, ini, end)
  left_seq, right_seq = SPLIT (protein_sequence, ini, end)
  left_list  = recursive_assignation (left_seq)
  right_list = recursive_assignation (right_seq)

  assignation_list = JOIN (left_list, right_list)
  RETURN assignation_list
```

Results and Discussion I-sites and Folding Levels

We traced the I-sites on a set of different proteins with known folding pathway. For each protein, we determined the folding level for each conformation on the folding path. Then we calculated the average RMS of dihedral angles between fitted I-sites and corresponding protein conformations for each folding level. They all follow a similar pattern value which is illustrated in Figure 7.10. In the plot of six different proteins there is a pronounced decrease in the average RMS in the unfolded level until the early intermediate level is reached; from this point on the RMS values vary only slightly.

This behavior suggests that if the I-sites act as initiation sites for protein folding, then they start forming when the protein reaches our defined early intermediate level. Generalizing the results obtained by PRM simulations we hypothesize that a protein folds into I-sites early when -corresponding to the early intermediate folding level described in our work (see section 2.6), meaning that it starts forming initial folding characteristics as very loosely packed hydrophobic core, with a lack of secondary structure elements and very weak tertiary interactions.

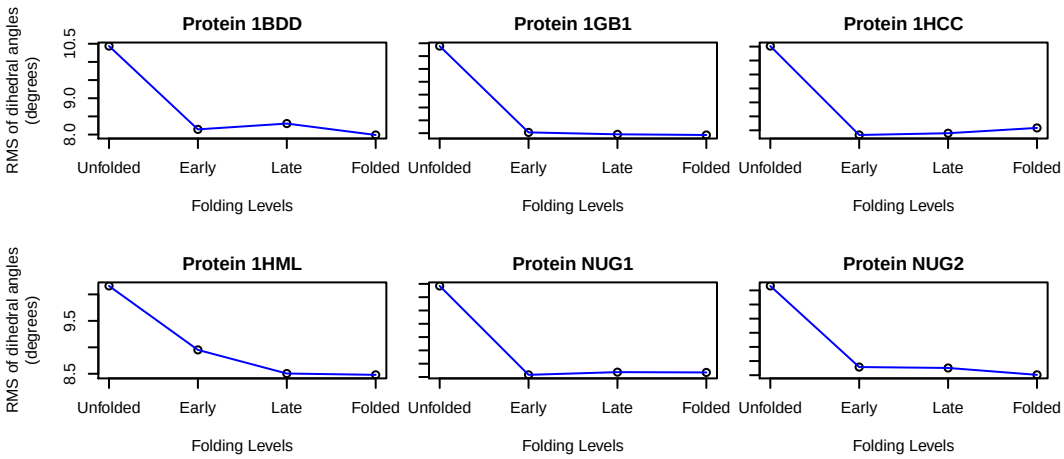


Figure 7.10: I-sites per folding level. The horizontal axis represent our four levels occurring during protein folding, and the vertical axis are the average RMS values in degrees to the locations in the protein structures of the best superimposed I-sites. It is observed that the pronounced decrease of RMS values stop in the early intermediate level from which the RMS changes changes are less pronounced .

7.8 Discussion

The results of this section indicate that proteins fold following a sequential order of events organized in a series of metastable states, well characterized, which we called *folding levels*. Specifically, we characterized four folding levels that proteins pass through on their way from the unfolded to the native state, and that we named as: *the unfolded level, the early intermediate level, the late intermediate level, and the folded level*.

7.8.1 Folding follows a sequential order through metastable levels

One of the most interesting findings derived from these results was that the classic definition of protein folding pathways [Levinthal, 1968], a sequence of specific structural events leading to fully folded protein, is partly supported by

our results but with some slight variations. As we observed from the transition diagram of folding shown in the figure 7.6, folding proceeds sequentially by well-characterized levels. When we plotted the pathway for a set of specific proteins (see figure 7.8), we found that in average, for each level there is no a single conformation but an ensemble of them each with the possibility of transitioning from one level to another. We consider these findings relevant for the study of protein folding given that it becomes a theoretic evidence of the hypothesis that folding is carried out via multiples and unrelated pathways that makes the folding robust [Udgaonkar, 2008, Krishna and Englander, 2007].

7.8.2 Folding Proceeds through Intermediates

The folding dynamics that we found from the conformational analysis in terms of the features (Figure 7.6) suggested that folding is proceeding by the accumulation of folding intermediates. We consider that this finding has important implications in the theoretical and experimental studies of folding as intermediates has been a subject widely discusses because they have not been experimentally detected in all of the proteins [Brockwell and Radford, 2007, Chen et al., 2008]. Our work supports the existence of these intermediates theoretically, as other computational and experimental studies have done [Udgaonkar, 2008, Brockwell and Radford, 2007, Roder et al., 2006, Daggett and Fersht, 2003b, Pande and Rokhsar, 1999], but most important, this work gives computational tools as the IC Features and the ICF score (see Figure 7.6, and 7.7), to help in establishing when a protein conformation from a pathway could be a possible intermediate. This result can be used in other computational studies as protein folding simulations, to classify and verify the type of conformations the trajectories are identifying and so to get and insight about how the folding is progressing.

7.8.3 There are two types of folding intermediates

Our folding dynamics showed that folding proceeds through two levels of folding intermediates: early and late (Figure 7.6), depending on whether confor-

mations accumulate near the initial or near the ending folding steps, respectively. It suggests that during the protein folding, intermediates may play important roles either at the beginning or end of folding, which agrees with some experimental works that have also identified early intermediates associated as the responsible for defining the pathway through which the folding proceeds [Roder and Colón, 1997, Udgaonkar, 2008]; while intermediates occurring during final stages of folding have been associated with some misfolded states, causing of degenerative diseases [Roder and Colón, 1997].

Therefore, the importance of this finding are twofold: it helps to support experimental studies on this subject; and it can be used to analyze the type of intermediates a protein is forming and so determine their structural and folding characteristics with the aim of analyzing how they influence the subsequent stages of folding. For example, the characteristic misfolded or abnormal transition from α to β-sheet in amyloid deposits, is related with neurodegenerative disorders as the Alzheimer's and Parkinson's diseases [Reynaud et al., 2010]; so, if the folding of a protein shows that its early intermediates are characterized by secondary structures dominated by α-helices, and the late intermediates are dominated by β-sheets, it could means that there is a possibility of a misfolded state.

7.8.4 A score of folding level for protein conformations

Our score for folding status or *ICF score* was designed to have several useful characteristics. First, the score depends on features derived from physical properties of proteins, and in this sense, it could be applied to any class of protein, independent of its topology. The values of the score are maximum when the protein is folded and minimum when it is unfolded (see Figure 7.7). It is a consequence of the features the score is representing. It is expected that a protein in a folded state has high stability, is very compact, and is native-like. In contrast, an unfolded protein has low stability, is very loose, and very dissimilar from the native.

Derived from the above, the ICF score is expected to have intermediate values between the folded and unfolded states, indicating semistable

structures partially folded, with a loose core, exhibiting native-like secondary structure with weak or absent tertiary interactions, and a poor residue packing as side chains are not fixed. Characteristics that are the ones that some authors have identified for experimental folding intermediates [Udgaonkar, 2008, Baldwin and Rose, 1999, Brockwell and Radford, 2007, Lajtha et al., 2007, Chalikian and Breslauer, 1996].

In summary, we have developed a theoretical approximation of the folding state of a protein. The definition of the IC features, the ICF score, and the classifier that assigns a folding level to a protein conformation, can be used as a tool for studying how proteins could behave by observing what folding levels are most or sparsely populated; showing how proteins are using folding intermediates in their folding process and what type they are: closer to unfolded, closer to the folded, or some intermediate; and showing how proteins fluctuate during folding when some intermediates persist or rapidly disappear.

CHAPTER 8

DISTRIBUTED FRAMEWORK FOR EVALUATION AND ANALYSIS OF PROTEIN FOLDING PATHWAYS

8.1 Introduction

Simulations of protein folding generate trajectories with thousands of conformations corresponding to 'snapshots' of the conformational changes of a protein during the folding process. Evaluating and analyzing a trajectory usually requires processing all conformations, possibly taking a long time on uniprocessor systems. An alternative is to use multiprocessor machines or dedicated clusters, but they are expensive and only available in few institutions, and they may implicate special requirements in software, hardware, and network configurations.

Here we propose a distributed computing platform using a cloud storage service like a main layer of communication and synchronization, plus our own control layer to distribute processing, check redundancy and crashing, collect results and assemble the whole solution. The platform uses local and external resources, many of them available in our institutions, such as computers and laboratory servers. To allow the easy deployment and usage for collaborators, we packaged all elements needed for the distributed computing in a virtual appliance that can be spread over local and external collaborators' devices and so simulate a dedicated computing cluster.

The appliance includes the applications and the needed environment to be an independent piece of software ready to be run on diverse host computer devices acting as processing units (e.g. desktops, laptops, servers, and others). Before it starts its normal operation, it must register itself with the cloud storage service employed to receive and send data packages. Then, it starts the programs to run and control the distributed processing. In its basic operation, each file produced by the appliance is reflected in the other devices that consume and process it and, and reflecting the results again to all the appliances.

However this basic operation presents some disadvantages for an adequate

distributed processing. A package may be processed by one or more computers at the same time which in turn may fail and never return a result. The cloud storage service may take long times sending packages to registered appliances. Therefore our control layer adds mechanisms to avoid redundancy in the processing and communication of data packages by checking the current work of computer devices, performing actions to recover from crashing and redundancy, and allowing direct communication between appliances.

8.2 Background

Data from large biological databases have been used commonly in bioinformatics processes and applications, but in the last years, biological data are also generated from biological simulations. Usually, evaluation and analysis take long computer times when data are processed on regular desktop computers as the ones mostly used in university laboratories. It could be done using supercomputers or dedicated clusters, but this alternative maybe is not feasible for some institutions due to the implied high costs and special requirements.

To overcome these disadvantages, spare clock cycles of idle machines have been used to simulate a dedicated computing cluster: Seti@Home [Korpela et al., 2001], Genome@Home and Folding@Home [Larson et al., 2002] are large projects that follow this approach. For example, Folding@Home simulates the folding of a protein distributively by using 'friend' computers around the world, instead of simulation in specialized supercomputers [Shaw et al., 2008]. This approach of distributed computing, known as Public Resource Computing, works because individuals with an Internet connection donates CPU cycles on their computer to a distribute computing project. Its main requirements are: a special software as middleware has to be provided to allow synchronization between 'friends' and server machines, and 'friend' computers have to download a small piece of software to run in the idle times.

However, this piece of software is designed with one application in mind, contrary to what happens in some bioinformatics applications where diverse and heterogeneous tools are required for a specific problem (e.g. BLAST, DSSP,

Gromacs, and others). Additionally, most of this kind of software is developed for specific environments with particular operational systems, programs, libraries, and packages. A common middleware used in this type of projects is BOINC [Anderson, 2004], but porting applications to it requires considerable effort because it requires specially designed server side functions tailored to the specific application.

In order to deal with such difficulties, we had the idea to communicate and synchronize computers using as basic middleware the straightforward and practical public *cloud storage* services; add over it a specific layer for controlling the distributed processing; and pack the middleware, additional layers, applications and working environment in a special piece of software known as *virtual appliance*.

At the present, various companies of the cloud computing (e.g. Dropbox, Copy, UbuntuOne), are offering storage services where files are synchronized between regular PCs and cloud servers by a simple and direct way of copying files to a special folder. In the same sense, virtual appliances are being used extensively as a suitable alternative for encapsulating the entire environment plus specific applications in a special piece of software that is run optimally as a guest appliance on a host system by a virtualizer software as VirtualBox, VMWare, and Xen [Macdonell and Lu, 2006].

8.3 Distributed Model

We based our distributed model in the publishers/subscriber model [Eugster et al., 2003], which the basic interaction is executed by three entities: subscribers, publishers, and a message system. *Subscribers* registering with the message system to consume specific messages. *Publishers* producing messages. And the *message system* distributing messages efficiently from publishers to subscribers, and managing the registrations.

We named our distributed model ***cloud worker/commander model***, according to its three basic entities: worker, commander, and a cloud message system (described below). Our model slightly changes with respect to the publishers/-

subscriber model (see Figure 8.1). We do not use publishers, but we divide the subscribers into several workers and one commander, both interchangeably producing and consuming any message. Furthermore, the message system evolves to a cloud message system (CMS) that besides managing registrations, acts a switch for messages by multicasting them to all registered subscribers.

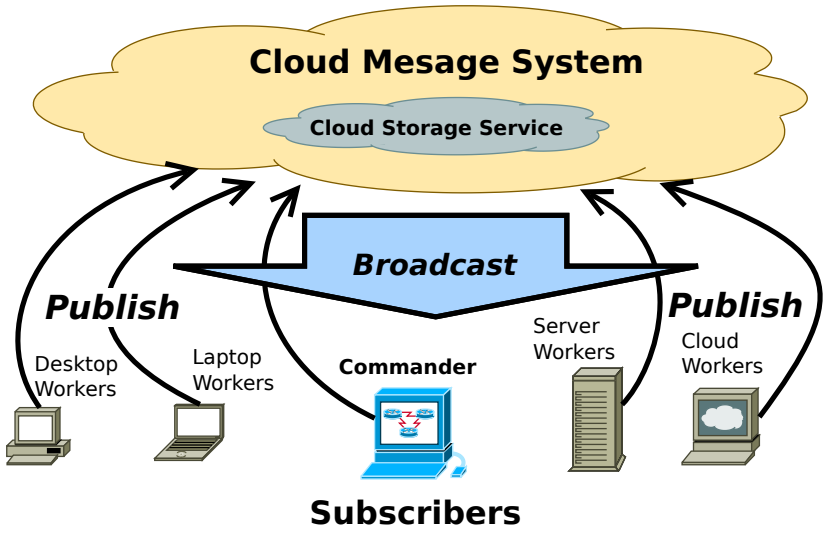


Figure 8.1: **Cloud Workers/Commander Model.** Several workers and one commander *register* with the Cloud Message System (CMS) to receive and send any messages. The CMS corresponds to a cloud storage service (e.g. Dropbox, Copy, UbuntuOne) plus a control and synchronization layer. The commander produces data to be processed distributively by the workers, which also produces results to be collected by the former. The communication, interaction, and control among the entities is managed by the CMS. In practice, subscribers correspond to real computer devices.

8.3.1 The Subscribers

We define two type of subscribers: commanders and workers (see Figure 8.2). Both can produce and consume messages interchangeably by sending messages to the CMS. The commander contains the whole data that it split into small packages, named work units. The workers process the work units and produce results, small-size data files, to be consumed by the commander.

Additionally, all subscribers have to update a *cloud schedule* structure to allow for the control and synchronization of the distributed processing. They have to register and update all events happening on the data by updating this

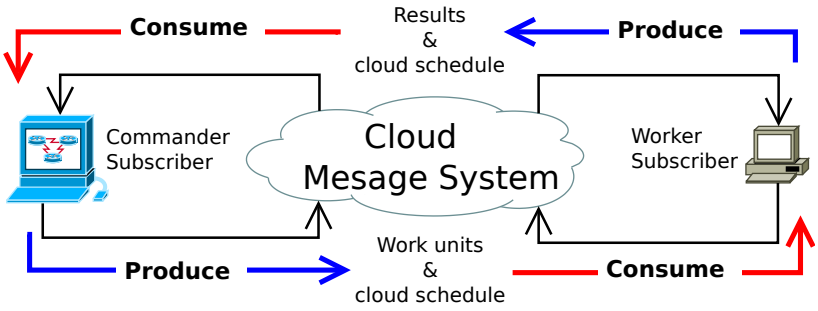


Figure 8.2: **Messaging between the Commander and workers subscribers.** The commander produces work units to be consumed and processed by workers; who produce results to be consumed by the commander. Synchronization is achieved by both subscribers by registering and updating processing information in a sync structure named *cloud schedule*.

structure with the work unit's name, its processing start time and estimated time, and its completion time (if successful).

8.3.2 The Cloud Message System

This entity is the responsible for messaging and controlling distribution and synchronization between registered subscribers. It has two layers: The first one is a basic layer of communication and synchronization of data packages given by the cloud storage service (described below). The second one is our layer implementing algorithms to control and synchronize the distribution and processing of work units (described in the section 8.4).

The cloud storage service provides the basic middleware for interchanging, controlling and synchronizing messages (data packages) among independent computer devices (subscribers). It allows storing files on a *cloud server* and updating asynchronously with the computer devices previously registered. In its basic functioning, every package produced by a device is copied to the cloud server and reflected in the others by copying the packages into a special device's local repository.

In our platform, we are using the Dropbox storage service [Drago et al.,] (`http://www.dropbox.com`) but others provide similar functionalities (e.g. Copy, UbuntuOne). In the basic usage of this service, data are interchanged between a computer device and the cloud server by copying data to a local

synchronization repository (named 'Dropbox' if using Dropbox) created during the installation of a Dropbox client software (see details in section 8.6).

8.4 Synchronization and Control Layer

We added a synchronization and control layer over the one provided by the cloud storage service (described in section 8.3.2). Its operations allow an adequate distribution and data processing by splitting data into work units (WUs); publishing the WUs into the cloud message system; processing them by subscribers, and collecting and assembling results. Finally, it also controls the optimal ending of the processing by avoiding redundant work, workers crashing, and stopping and reassigning works. In the following sections, we will describe the main elements, operations, and message types in the layer.

8.4.1 Cloud scheduler structure

This is a structure where subscribers register information of the main events happening during the distributed processing: the name of every work unit to be processed, as long as three time values related to the work unit's processing: start, expected, and completion time (if it succeeds). Name and times, safe completion time, are written when the work unit arrives, and the worker starts its processing.

Given our distributed computing model, the management of this scheduler is decentralized, meaning that each worker has its copy of the structure where to register its events. The updated structure is sent to the Cloud Message System, which replicates it to the other registered subscribers, to keep them informed of new processing events.

8.4.2 Commander operations

The *commander* has two main operations: handling and distribution of work units (WUs), as well as collecting and assembling the results. In the former, the commander creates the cloud scheduler structure and splits the whole data into WUs. Next, it copies them–one at time–into its local synchronization repository, waiting that a worker consumes the WU before copying to the next one. In the latter, the commander is checking its cloud synchronization repository continuously for results produced by workers, and when they arrive, it collects them to assemble the whole solution.

In the above operations, the commander always updates the cloud scheduler structure and removes consumed results, both actions are reflected in the CMS which also reflect them to all other worker devices.

8.4.3 Worker operations

Workers have two operations: consuming and processing WUs. In the former, a worker checks its cloud synchronization repository continuously until a new WU arrives. Then, it loads the WU into memory and removes that from the repository, and registers the processing information (name and times). In the latter, it calls the procedure that evaluates the WU and waits until it finishes, then it copies the result to the cloud synchronization repository and updates the cloud schedule structure with the completion time. All file operations are being reflected in the CMS, meaning that changes in the schedule and new results are updated and copied in the other devices.

8.4.4 Controlling operations

There are two control operations: checking redundant work and checking workers' crashing. Checking redundant work is necessary because two workers may process the same work unit at the same time (by the multicast of the cloud message system). To avoid this situation, one of the two workers will stop its pro-

cessing if it finds that another worker has also registered the work unit in the cloud schedule structure.

With regards to crashing control, it may occur that some worker never finishes processing a work unit (e.g. the worker crashes or processing takes too much time). In this case, the commander removes the work unit from the scheduler structure and produces it again to be processed by another worker. The commander detects crashing by continuously checking that the WUś expected time is not exceeded.

8.4.5 Messaging

Five types of messages are handled: *work unit copied*, *result copied*, *scheduler updated*, *work unit duplicated*, and *worker crashing*. We use an indirect mechanism for interchanging messages among subscribers based on the events happening on their cloud synchronization repositories. Given that we base our model in a cloud storage service, the main events for subscribers are associated with file management (copy, remove, or update). When a file event occurs, it is reflected in all subscribers' cloud repositories (see Figure 8.3).

8.5 Toolkit for Protein Analysis

To allow messaging, each subscriber has to check its cloud synchronization repository continuously to know what is happening and so associate the events with one of the message types. To organize these tasks, we divided a subscriber's cloud repository into three specific directories, one for each type of produced data. An *input* subfolder in which the commander puts WUs and the workers get and remove them. A *results* subfolder, in which workers put results and the commander gets and removes them. And a *control* subfolder, in which both types of subscriber read and update the cloud scheduler to synchronize their WU events.

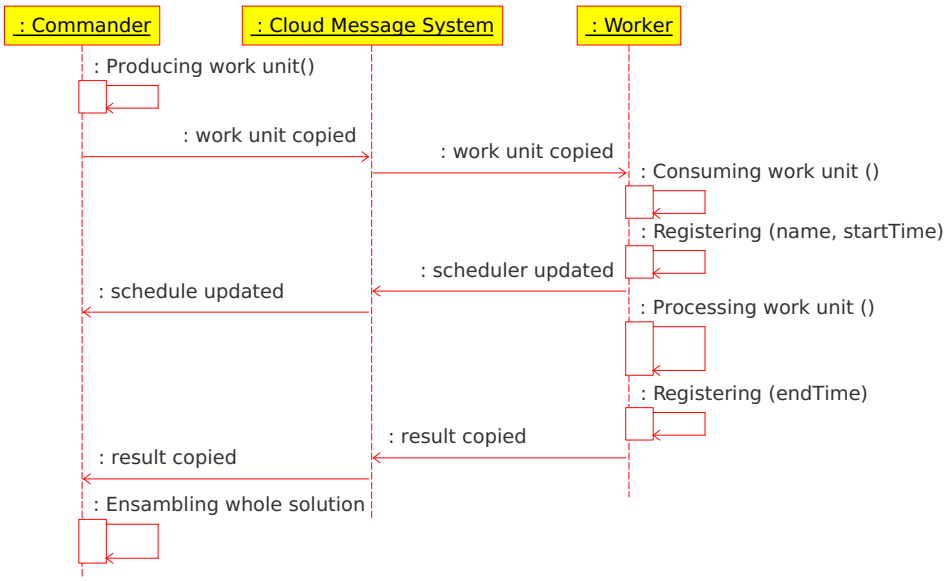


Figure 8.3: **Messaging among subscribers.** Messages are associated with file operations. The commander starts the interaction by producing a work unit and copying it into its cloud synchronization repository. This action is reflected in the Cloud Message System (CMS), as well as into the workers *(work unit copied)*. A worker consumes the work unit, registers initial information into the scheduler, processes the work unit, and, again register ending information into the scheduler. Then, the worker copies the result on its cloud synchronization repository, reflecting it to the CMS, and to the commander, which finally assembles the whole solution including the new result.

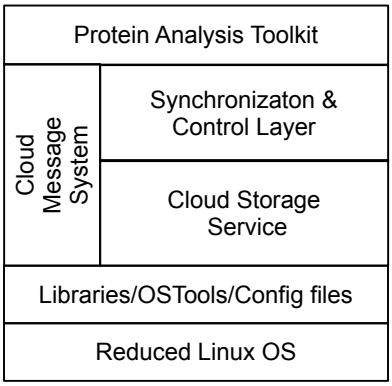


Figure 8.4: **Virtual appliance for distributed protein evaluation.** The basis of the appliance is a reduced and optimally preconfigure Linux operational system (OS) (bottom). All required libraries, tools and configuration files are installed on this reduced OS. Two elements to allow the distributing processing are added through the *Cloud Message System* (middle): the *cloud storage service* for communicating and synchronizing with the cloud server. Second, *the synchronization and control layer* with the algorithms for controlling assignation and processing of work units. Finally, the *protein analysis toolkit* (top), which contains the functions for evaluating properties of protein structures.

8.5.1 Functions

The main functions of the toolkit are presented in Table 8.1. The toolkit offers a suite of functions through python standalone programs and libraries that can be used directly from a command line interface or called from external programs, respectively.

Table 8.1: **Main functions of the protein analysis toolkit.**

Function	ID	Description
native_contacts	NC	Counts the number of contacts formed by a protein conformation that also are present in the native.
contact_order	OC	Compute the average sequence separation between residues that make contacts in the native structure
radius_gyration	RG	Compute how much the protein structure spreads out from its center.
hydrogen_bonds	HB	Counts the number of hydrogen bonds (attractive force between electronegative atoms of different–or same–molecules).
asa	AS	Computes the surface area of a protein that is accessible to the solvent.
voids	VD	Compute the unfilled spaces or empty cavities, inside of a protein, which are not accessible from the solvent.
rmsd	RM	Calculates the root-mean-square deviation (RMSD) between the atoms of two superimposed proteins
local_rmsd	LR	Calculates the average of the RMSD between the secondary structure elements.
potential_energy	PE	Calculate the energy associated with forces of attraction and repulsion between protein atoms.
dipole_moment	DM	Computes the vector sum of the dipole moment of fixed charges on the surface, polar groups in the chain, and mobile ion fluctuations.
correct_sse	RC	Counts target protein's residues forming secondary structure elements that are also present in the native structure.
any_sse	RA	As above, but it counts residues in the target protein forming any secondary structure element.
structural_score	SS	Compares secondary structure elements between two proteins by short local motifs or protein blocks (PBs).
rigid_clusters	CL	Counts the parts of the protein that behave as rigid bodies (i.e., distances between all atoms remain fixed).
degrees_freedom	DF	Related with the number of constraints necessary to rigidify the protein structure (degrees of freedom).
stressed_regions	SR	Counts the regions that contribute to reducing the total number of degrees of freedom available to the protein.

8.5.2 Implementation

The toolkit implements different algorithms and wraps functionalities included in other well-proved software tools for protein manipulation developed in the last years as programs, libraries, packages, and frameworks. The Figure 8.5 presents a view of the functions implemented in the toolkit. In the external layer, there are our functions used directly to evaluate the properties on a protein structure. In the middle, there is a layer of external programs as libraries and packages that are 'wrapped' by our functions for implementing their func-

tionality. Also, in the inner layer there are the different languages in which some of our functions or external programs are implemented.

Source:The author

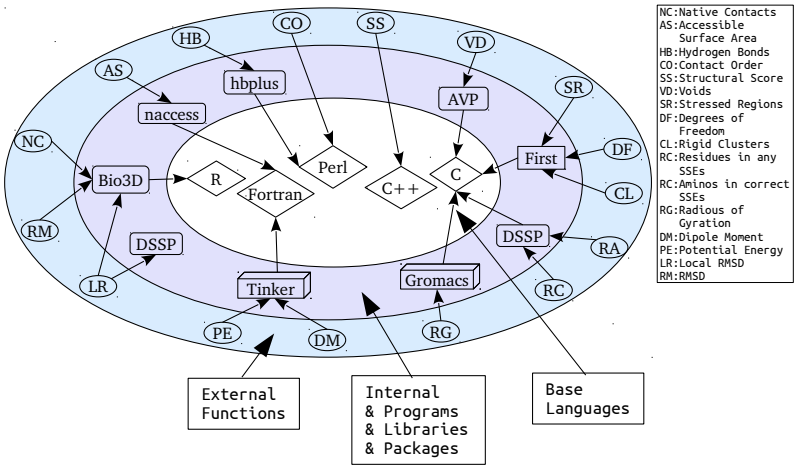


Figure 8.5: **Structure of the Toolkit for Protein Analysis.** Three layers. In the first layer, our functions to evaluate the protein properties. In the middle the external programs 'wrapped' by our functions. Also, in the inner layer, the computer languages used by the functions and external programs.

8.6 Installation and Deployment

The entire system consists of a single archive file ('vm-eval.ova') that corresponds to a virtual appliance (also known as a virtual machine). It is a complete ready-to-use software package (operating system with applications) that does not need special configuration or installation. The user only requires to download the appliance (`http://bioinformatica.univalle.edu.co/software/vm-eval`) and import it to a pre-installed virtualizer (a software as VirtualBox, VMWare, Xen).

In our project, we use the virtualizer VirtualBox (`http://www.virtualbox.com`) to construct, import and execute the appliance. It was preconfigured with basic hardware requirements that users can change according to their virtualizer software and machine capacity, which in VirtualBox are the amount of RAM allocated for the appliance; and the number of virtual CPU cores the appliance should see, and the percent of CPU usage (default is 100%).

Currently, the appliance is running on a reduced Linux OS (Ubuntu server 12.04) with a command line interface. The predefined user that executes the application is the *fm* user (*/home/fm*) that contains two subdirectories, one for running the distributed processing *(/home/fm/framework)* and the other for the user's application *(/home/fm/application)*. The cloud repository created for the cloud storage service is the */home/fm/Dropbox*. Other directories are the repository for the results */home/fm/results*, and a RAM filesystem to fast reading and writing of temporal files */home/fm/ramdisk*.

The distributed platform embedded in the appliance can be used for other applications different to the toolkit we included (see section 8.5). Change the applications is simple, only two actions are needed: first, the application and related files have to be copied to the application directory of the user's home directory */home/fm/application*, and the main application has to be named as or linked to *eval*. And second, new account in the Dropbox storage service (`http://www.dropbox.com`) has to be created and the appliance has to be relinked to this new account (see Dropbox manual).

8.6.1 Results

We used the framework to evaluated all pathways and trajectories distributively, and to perform the analyzes required to culminate the present work. It was operated on the desktop PCs of the Bioinformatics Lab in Universidad del Valle (`http://bioinformatica.univalle.edu.co/thesis/framework`). Specifically, we conducted the evaluation of 16 protein properties on three different datasets: (1) on the 16000 conformations of a protein folding trajectory of the villin headpiece subdomain from the folding@home project using by molecular dynamics (see chapter 5); (2) the 37 pathways generated with the PRM-based approach and taken from the Parasol Server with about 6000 conformations (see chapter 6); and (3) on the 630000 conformations of a protein folding trajectory of the villin headpiece subdomain achieved by the David E. Shaw's research group [Lindorff-Larsen et al., 2011] using molecular dynamics (see chapter 7). The framework operated without problems, and the execution times decreased considerably in comparison with our intents before developing

the framework.

CHAPTER 9

CONCLUSIONS AND FUTURE WORK

In this chapter, we present the main conclusions and possible future work of our present project. We begin discussing the generality and validity of the used data. We continue with the importance of the inherent conformational features and the resulting feature score. Then, we conclude about the derived folding levels, and we discuss our distributed framework. At the end, we outline what we consider possible continuations of this work.

9.1 Data

The data used in the present research, result from two different simulation approaches: molecular dynamics trajectories and PRM-based pathways.

In our preliminary analysis, we use folding trajectories from molecular dynamics simulations of the villin-headpiece subdomain from the folding@home project. Although the results show that physical properties are implicitly measuring folding features, the results can not be generalized due to missing generality of data given that only one protein is considered and many of the folding events keep repeating themselves long periods of time in the trajectory. Additionally, the trajectory starts in a semi-folded state and finishes before reaching the native state; being some conformational states not clearly present. Nevertheless, the preliminary analysis allows to confirm the validity of our proposed approach.

For the definitive analysis, we use data from the Parasol Server, which implements the PRM method obtaining pathways for a variety of proteins with different topologies and sizes. PRM is able to produce large sets of unrelated folding pathways, with conformations in different states, from unfolded to folded, providing a valuable source of folding data which others approaches, as molecular dynamics and experimental methods are still not capable to produce. There are various studies using the PRM data that show that they can generate valid results in a very general context. It means that we can generalize our results to

protein conformation, pathways or trajectories, obtained from different types of folding simulations or even experimental observed data. Although the PRM data are generated as pathway (a sequence of folding events), this specific organization is not used in our project. For our purposes, it is sufficient to use the simulation data as a set of conformations of proteins in very different states of folding.

For additional experimentation and validation of our results, we used an additional dataset of trajectories generated by the Anton supercomputer for molecular dynamics simulation. Specifically we use a trajectory of the same villin-headpiece subdomain we took from the folding@home project for our preliminary analysis.

9.2 Inherent Conformational Features

9.2.1 IC Features

A main result of our work is the definition, analysis and evaluation of inherent conformational features and the corresponding feature score in a computational way. The applied method, a Principal Component Analysis on a selected number of folding properties, is appropriate given that the explained variance is high, and the way properties group into components is consistent.

The PCA approach conduces to three relevant components that we interpreted as inherent conformational features. These features are based on global structural and energetic properties. They express the folding status of a protein conformation in a compact way. Analyzing the original properties that contributed strongly to each of the features we find that the features describe global folding properties as Stability, Compactness, and Native-likeness.

The stability feature combines mainly properties as native contacts, hydrogen bonds, residues in correct secondary structures, residues in any secondary structure, and degrees of freedom; all of these are generally contributed to the structural stability of proteins. The compactness feature is principally com-

posed of contact order, radius of gyration, and accessible surface area, all of them related to a conformation's spatial extension. And the native-likeness is given by properties as RMSD, local RMSD, and structural score, naturally associated with the similarity to the native structure of a protein. It is important to emphasize that these computationally obtained features correspond to characteristics that are measured experimentally, a fact that can be considered an indicator of their practical validity and applicability.

The IC features describe in a general and compact way the folding status of a protein. Given that the folding status of a conformation does not depend on its size or topology, we can employ it in the comparison of conformations from different proteins.

Observing the changing folding status of a protein along a pathway or trajectory, we can analyze or characterize the dynamic folding process. The IC features adopt a natural organization and dynamics beginning with low feature values in unfolded states and incrementing their values until reaching the native state.

Using the concept of folding status, we derived the IC feature score that we defined as the weighted average of its IC features, where the weights correspond to the variance each feature explains. Giving the more important features more relevance in the score. The score, condensed from the IC features which in turn are extracted from the general folding properties, captures the strengths of each feature (and indirectly, each of the original properties) in a given conformation and compensates its weaknesses. Therefore, it constitutes a robust and compact measure of the folding stages of a protein with a high information content and is in this sense, comparable to other folding measures or even with reaction coordinates.

9.3 Computational Folding Levels

To obtain a computational description of folding levels, we use hierarchical clustering, based on a similarity measure we constructed from the IC features, that allows to compare the folding status of two conformations. By means of inter-

nal and relative analyses we show, that the obtained clusters are independent of the specific dataset.

The clustering approach exhibits a natural organization in four balanced, well-differentiated groups of conformations, which we associate with four folding levels. The conformations of a specific group share common IC features, i.e. they show similar levels with respect to their stability, compactness, and native-likeness. And what is more: Analyzing the transitions between groups it becomes evident that they are organized in a time-dependent manner along the pathways even though their dynamic organization of pathways is never used in computational approach.

Therefore we are able to associate the four folding groups with time and features dependent folding levels or folding states: We introduce the unfolded, the early intermediate, the late intermediate and the folded states. The computationally derived folding states coincide with the experimentally observed states, widely described in the literature. It validates in retrospect our approach in a practical manner and increases the confidence in applications and conclusions of our results. It is interesting that our approach postulates four folding states, two of them intermediates, similar to the ones that some experimental studies have detected. At an experimental level, it is very difficult to characterize folding intermediates; here we offer a computational approach that allows to identify or classify intermediates and may complement the experimental results.

Due to the capacity of the PRM data to produce results of a very high generality, we conclude that the folding levels derived by our approach define in a general form the folding states, a protein assumes in its folding process. Therefore, we can use the obtained computational folding states to analyze folding pathways, to compare the folding process of different pathways and even different proteins. We can apply it to analyze the behavior of computer simulations and eventually to experimentally observed trajectories when they become more common. The assignment of folding states in a simulated villin-headpiece subdomain from the Anton supercomputer confirms our conclusion: The villin-headpiece subdomain is experimentally and computationally characterized as two-state folding; and accordingly, the dynamic behavior of the folding states

shows that principally two states are assumed, with very short transitions to the other ones.

9.4 Framework for Protein Analysis

Our distributed framework uses simple resources as desktop PCs and basic middleware through typical cloud storage services (e.g. Dropbox, copy, and others) to run a series of tasks distributively. In our case, the framework facilitates the evaluation of hundred of protein conformations from the different pathways and trajectories in current desktop PCs. It makes this distributed framework a valuable resource for the researchers when specialized machines are not easily available, or complex configurations are needed.

Our framework merges ideas of three current technologies: cloud computing, virtualization, and public computing. With respect to cloud computing, a cloud service storage (e.g. Dropbox) was used as a basic middleware; an additional control layer allows an application to execute a work distributively. Concerning virtualization and public computing, we packed in a virtual appliance all the needed components of our platform in a completely functional way which is *ready to download and execute* in a donor computer that is going to be part of a distributed project.

Different from other projects of public computation, we believe that our idea of packing the whole software for the distributed processing in a virtual appliance enhances the trust of the people donating cycles. The appliance acts as a sandbox, being isolated from the rest of the computer; therefore anything outside from its virtual space is protected against damaging. Additionally, the virtual appliance runs on most of the current operational systems, after executing a pre-installed virtualization software (e.g. VirtualBox, VMWare, Xen), and does not require complicated installations or configurations.

To validate our framework, we include a toolkit for protein analysis; it offers all functions to evaluate in a protein structure the physical properties we selected and used in the distinct analyzes of the present work. The toolkit either implements directly the methods for calculating the values for each property

or wraps the functionality of well-proved programs, libraries or packages that offer related resources to calculate the property.

9.5 Future Work

As an immediate continuation of the here presented work is the extension of our methods and analyzes to a greater variety of data. Using more simulated pathways and trajectories, the general validity of our approach will become even more evident. We could work with different pathways from the PRM approach, use more simulated trajectories of the villin-headpiece subdomain from the folding@home project and include other trajectories of different proteins simulated by the Anton supercomputer.

Our methodological and computational framework is easily extended to contribute to specific folding problems that arise in the protein design and to understand better the dynamics related with misfolding.

For these and other possible application it could be desirable to extend the number of folding properties which are the basic input for the construction of folding features and levels, expecting that this type of extension could lead to the detection of more specific intermediate states. Increasing the number of properties in the framework only requires adding their specific evaluations to the toolkit.

Taking advantage of the ease with which new folding properties can be integrated with our framework, we would like to apply our methodology to folding state prediction of conformations where the native structure is unknown. This implies the replacement of all properties that require the knowledge of the native structure by similar ones that can be defined without this knowledge (for instances the number of native contacts could be replaced by degrees of freedom and other properties from rigidity analysis).

We intend to evolve our distributed framework to a general framework for distributed computing in bioinformatics. Currently, it offers all necessary elements to distribute tasks independent of the kind of application. In order to

adapt it to this general approach we have to standardize some processes for the insertion of the new toolkit and to define a standard protocol for calling its functions and localizing their input parameters and writing their output values.

BIBLIOGRAPHY

[Abul et al., 2003] Abul, O., Lo, A., Alhajj, R., Polat, F., and Barker, K. (2003). Cluster validity analysis using subsampling. In *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, volume 2, pages 1435–1440. IEEE.

[Anderson, 2004] Anderson, D. (2004). BOINC: A System for Public-Resource Computing and Storage. *grid*, 00:4–10.

[Anfinsen and Scheraga, 1975] Anfinsen, C. and Scheraga, H. (1975). Experimental and theoretical aspects of protein folding. *Adv Protein Chem*, 29:205–300.

[Apaydin et al., 2003] Apaydin, M. S., Brutlag, D. L., Guestrin, C., Hsu, D., and Latombe, J. (2003). Stochastic roadmap simulation: an efficient representation and algorithm for analyzing molecular motion. *J. Comput. Biol.*, 10:257–281.

[Baldwin and Rose, 1999] Baldwin, R. L. and Rose, G. (1999). Is protein folding hierarchic ? II . Folding intermediates and transition states. volume 98, pages 77–83.

[Baldwin and Rose, 2013] Baldwin, R. L. and Rose, G. (2013). Molten globules, entropy-driven conformational change and protein folding. *Current opinion in structural biology*, 23(1):4–10.

[Beck and Daggett, 2007] Beck, D. A. C. and Daggett, V. (2007). A one-dimensional reaction coordinate for identification of transition states from explicit solvent P(fold)-like calculations. *Biophysical journal*, 93(10):3382–91.

[Bergeron, 2002] Bergeron, B. (2002). *Bioinformatics Computing*. Prentice Hall PTR.

[Bien and Tibshirani, 2011] Bien, J. and Tibshirani, R. (2011). Hierarchical Clustering With Prototypes via Minimax Linkage. *Journal of the American Statistical Association*, 106(495):1075–1084.

[Bonneau et al., 2002] Bonneau, R., Ruczinski, I., Tsai, J., and Baker, D. (2002). Contact order and ab initio protein structure prediction. *Protein Science*, pages 1937–1944.

[Borg and Groenen, 2005] Borg, I. and Groenen, P. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics, Berlin, 2 edition.

[Brockwell and Radford, 2007] Brockwell, D. J. and Radford, S. E. (2007). Intermediates : ubiquitous species on folding energy landscapes ? *Current opinion in structural biology*, 17:30–37.

[Burstyn, 2004] Burstyn, I. (2004). Principal component analysis is a powerful instrument in occupational hygiene inquiries. *The Annals of occupational hygiene*, 48(8):655–61.

[Bystroff and Baker, 1998] Bystroff, C. and Baker, D. (1998). Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol*, 281:565–577.

[Bystroff and Shao, 2003] Bystroff, C. and Shao, Y. (2003). Modeling Protein Folding Pathways.

[Chalikian and Breslauer, 1996] Chalikian, T. and Breslauer, K. (1996). Compressibility as a means to detect and characterize globular protein states. *Proc. Natl. Acad. Sci. USA*, 93(3):1012–1014.

[Chen et al., 2008] Chen, Y., Ding, F., Nie, H., Serohijos, A. W., Sharma, S., Wilcox, K. C., Yin, S., and Dokholyan, N. V. (2008). Protein folding: then and now. *Archives of biochemistry and biophysics*, 469(1):4–19.

[Chiang et al., 2006] Chiang, T.-H., Apaydin, M. S., Brutlag, D., Hsu, D., and Latombe, J. (2006). Predicting Experimental Quantities in Protein Folding Kinetics Using Stochastic Roadmap Simulation. In *LECTURE NOTES IN COMPUTER SCIENCE*, number 3909, pages 410–424.

[Chiang et al., 2010] Chiang, T.-H., Hsu, D., and Latombe, J. (2010). Markov Dynamic Models for Long-Timescale Protein Motion. *Bioinformatics [ISMB]*, 26:269–277.

[Chubynsky, 2008] Chubynsky, M. V. (2008). Characterizing the intermediate phases through topological analysis. pages 1–37.

[Costantini et al., 2007] Costantini, S., Facchiano, A., and Colonna, G. (2007). Evaluation of the structural quality of modeled proteins by using globularity criteria. *BMC Structural Biology*, 7(9).

[Costello and Osborne, 2005] Costello, A. B. and Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, 10:173–178.

[Cuff and Martin, 2004] Cuff, A. L. and Martin, A. C. R. (2004). Analysis of void volumes in proteins and application to stability of the p53 tumour suppressor protein. *Journal of Molecular Biology*, 344(5):1199–1209.

[Daggett and Fersht, 2003a] Daggett, V. and Fersht, A. (2003a). Is there a unifying mechanism for protein folding? *Trends Biochem Sci.*, 28(1):18–25.

[Daggett and Fersht, 2003b] Daggett, V. and Fersht, A. (2003b). The present view of the mechanism of protein folding. *Nature Reviews Molecular Cell Biology*, 4:497–502.

[Das et al., 2006] Das, P., Moll, M., and Stamati, H. (2006). Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction. *Proceedings of the . . .*, 103(26).

[David et al., 2003] David, C. S., Brumley, D., Chandra, R., Zeldovich, N., Chow, J., Lam, M. S., and Rosenblum, M. (2003). Virtual Appliances for Deploying and Maintaining Software. In *Proceedings of the Seventeenth Large Installation Systems Administration Conference (LISA 2003)*, pages 181–194.

[Dean, 2009a] Dean, J. (2009a). Choosing the Right Number of Components or Factors in PCA and EFA. *Shiken: JALT Testing & Evaluation SIG Newsletter*, 13(May):19–23.

[Dean, 2009b] Dean, J. (2009b). Choosing the Right Type of Rotation in PCA and EFA. *Shiken: JALT Testing & Evaluation SIG Newsletter*, 13(November):20–25.

[Dill et al., 2008] Dill, K. A., Ozkan, B., Shell, M., and Weikl, T. (2008). The Protein Folding Problem. *Annual Review of Biophysics*, 37(1):289–316.

[Dobson et al., 1998] Dobson, C. M., Šali, A., and Karplus, M. (1998). Protein Folding: A Perspective from Theory and Experiment. *Angewandte Chemie International Edition*, 37(7):868–893.

[Dokholyan, 2004] Dokholyan, N. V. (2004). What is the protein design alphabet? *Proteins*, 54(4):622–8.

[Dokholyan et al., 2002] Dokholyan, N. V., Li, L., Ding, F., and Shakhnovich, E. I. (2002). Topological determinants of protein folding. *Proc. Natl Acad. Sci. USA*, 99:8637–8641.

[Drago et al.,] Drago, I., Mellia, M., Torino, P., Munafò, M. M., and Sperotto, A. Inside Dropbox : Understanding Personal Cloud Storage Services.

[Duan and Kollman, 1998] Duan, Y. and Kollman, P. A. (1998). Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution. *Science*, 282(October):740–744.

[Duan et al., 1998] Duan, Y., Wang, L., and Kollman, P. A. (1998). The early stage of folding of villin headpiece subdomain observed in a 200-nanosecond fully solvated molecular dynamics simulation. *Proceedings of the National Academy of Sciences of the United States of America*, 95(17):9897–9902.

[Eitrich et al., 2007] Eitrich, T., Mohanty, S., Xiao, X., and Hansmann, U. H. E. (2007). Dimensionality Reduction Techniques for Protein Folding Trajectories. *From Computational Biophysics to Systems Biology (CBSB07)*, 36:99–102.

[Ensign et al., 2007] Ensign, D. L., Kasson, P. M., and Pande, V. S. (2007). Heterogeneity even at the speed limit of folding : Large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. *Journal of molecular biology*, 374(3):806–816.

[Eugster et al., 2003] Eugster, P. T. H., Felber, P. A., and Kermarrec, A.-m. (2003). The Many Faces of Publish / Subscribe. *Computing*, 35(2):114–131.

[Faísca et al., 2010] Faísca, P. F. N., Nunes, A., Travasso, R. D. M., and Shakhnovich, E. I. (2010). Non-native interactions play an effective role in protein folding dynamics. *Protein Science*, 19(11):2196–2209.

[Fleming et al., 2006] Fleming, P., H. Gong, and Rose, G. (2006). Secondary structure determines protein topology. *Protein Sci*, 15(8):1829–1834.

[Gong et al., 2005] Gong, H., Fleming, P. J., and Rose, G. (2005). Building native protein conformation from highly approximate backbone torsion angles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(45):16227–16232.

[Grant et al., 2006] Grant, B., Rodrigues, A., ElSawy, K., McCammon, J., and

Caves, L. (2006). Bio3D: An R package for the comparative analysis of protein structures. *Biinformatics*, 22:2695–2696.

[Hanebutte et al., 2003] Hanebutte, N., Taylor, C. S., and Dumke, R. R. (2003). Techniques of successful application of factor analysis in software measurement. *Empirical Software Engineering*, 8(1):43–57.

[Harner, 2012] Harner, E. J. (2012). *Modeling Multivariate Data*.

[Hatcher, 1994] Hatcher, L. (1994). *A step-by-step approach to using the SAS system for factor analysis and structural equation modeling*. SAS Publishing, 1st edition.

[Hespenheide et al., 2002] Hespenheide, B. M., Rader, A. J., Thorpe, M. F., and Kuhn, L. A. (2002). Identifying protein folding cores from the evolution of flexible regions during unfolding. *J Mol Graph Models*, 21:195–207.

[Hubbard et al., 1991] Hubbard, S., Campbell, S., and Thornton, J. M. (1991). Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. *J Mol Biol.*, 220:507–530.

[Ivankov et al., 2009] Ivankov, D. N., Bogatyreva, N. S., Lobanov, M. Y., and Galzitskaya, O. V. (2009). Coupling between properties of the protein shape and the rate of protein folding. *PloS one*, 4(8):e6476.

[Jackson, 1998] Jackson, S. (1998). How do small single-domain proteins fold? *Folding and Design*, 3(4):R81–R91.

[Jackson and Fersht, 1991] Jackson, S. and Fersht, A. R. (1991). Folding of chymotrypsin inhibitor-2.1. Evidence for a two-state transition. *Biochemistry*, 30:10428–10435.

[Jacobs et al., 2002] Jacobs, D. J., Kuhn, L. A., and Thorpe, M. F. (2002). *Flexible and Rigid Regions in Proteins*. Fundamental Materials Research. Springer US.

[Jacobs et al., 2001] Jacobs, D. J., Rader, a. J., Kuhn, L. a., and Thorpe, M. F. (2001). Protein flexibility predictions using graph theory. *Proteins*, 44(2):150–65.

[Jaenicke and Böhm, 1998] Jaenicke, R. and Böhm, G. (1998). The stability of proteins in extreme environments. *Current opinion in structural biology*, 8(6):738–48.

[Jain et al., 2005] Jain, A., Nandakumar, K., and Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern recognition*, 38(12):2270–2285.

[Javidpour, 2012] Javidpour, L. (2012). Computer Simulations of Protein Folding. *Computing in Science & Engineering*, 14(2):97–103.

[Kabsch and Sander, 1983] Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637.

[Kolodny and Levitt, 2003] Kolodny, R. and Levitt, M. (2003). Protein decoy assembly using short fragments under geometric constraints. *Biopolymers*, 68(3):278–285.

[Korpela et al., 2001] Korpela, E., Werthimer, D., Anderson, D., Cobb, J., and Lebofsky, M. (2001). SETI@home-Massively Distributed Computing for SETI. *IEEE: Computer Science and Engineering*, 3(1):77–83.

[Krishna and Englander, 2007] Krishna, M. M. G. and Englander, S. W. (2007). A unified mechanism for protein folding: Predetermined pathways with optional errors. *Protein science*, 16(3):449–464.

[Lajtha et al., 2007] Lajtha, A., Banik, N., Szilágyi, A., Kardos, J., Osváth, S., Barna, L., and Závodszky, P. (2007). *Chapter 10 - Protein Folding. Handbook of Neurochemistry and Molecular Neurobiology*. Springer US, Boston, MA.

[Larson et al., 2002] Larson, S., Snow, C., Shirts, M., and Pande, V. S. (2002). Folding@Home and Genome@Home: Using distributed computing to tackle previously intractable problems in computational biology.

[Lazaridis and Karplus, 1997] Lazaridis, T. and Karplus, M. (1997). New View of Protein Folding Reconciled with the Old Through Multiple Unfolding Simulations. *Science*, 278:1928–1931.

[Lee and Richards, 1971] Lee, B. and Richards, F. M. (1971). The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.*, 55:379–400.

[Levine and Domany, 2000] Levine, E. and Domany, E. (2000). Resampling Method For Unsupervised Estimation Of Cluster Validity.

[Levinthal, 1968] Levinthal, C. (1968). Are there pathways for protein folding? *J Chem Phys*, 65:44–45.

[Liang and Dill, 2001] Liang, J. and Dill, K. A. (2001). Are proteins well-packed? *Biophysical journal*, 81(2):751–66.

[Liang et al., 1998] Liang, J., Edelsbrunner, H., Fu, P., Sudhakar, P. V., and Subramaniam, S. (1998). Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins. *Proteins: Struct., Funct., Genet.*, 33(1):18–29.

[Lindorff-Larsen et al., 2011] Lindorff-Larsen, K., Piana, S., Dror, R. O., and Shaw, D. E. (2011). How fast-folding proteins fold. *Science (New York, N.Y.)*, 334(6055):517–20.

[Lobanov et al., 2008] Lobanov, M., Bogatyreva, N. S., and Galzitskaya, O. V. (2008). Radius of gyration as an indicator of protein structure compactness. *Molecular Biology*, 42(4):623–628.

[M. Tyagi A. de Brevern and Offmann, 2008] M. Tyagi A. de Brevern, N. S. and Offmann, B. (2008). Protein structure mining using a structural alphabet. *Proteins*, 71(2):920–937.

[Ma et al., 2007] Ma, B., Chen, L., and Zhang, H. (2007). What Determines Protein Folding Type? An Investigation of Intrinsic Structural Properties and its Implications for Understanding Folding Mechanisms. *J. Mol. Biol*, 370:439–448.

[Macdonell and Lu, 2006] Macdonell, C. and Lu, P. (2006). Pragmatics of Virtual Machines for High-Performance Computing : A Quantitative Study of Basic Overheads. *Computing*.

[Maechler et al., 2013] Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2013). cluster: Cluster Analysis Basics and Extensions.

[McDonald and Thornton, 1994] McDonald, I. K. and Thornton, J. M. (1994). Satisfying Hydrogen Bonding Potential in Proteins. *Journal of Molecular Biology*, 238:777–793.

[Moult, 2006] Moult, J. (2006). Rigorous performance evaluation in protein structure modelling and implications for computational biology. *Philos Trans R Soc Lond B Biol Sci*, 361:453–458.

[Mucherino et al., 2008] Mucherino, A., Costantini, S., di Serafino, D., D'Apuzzo, M., Facchiano, A., and Colonna, G. (2008). Understanding the role of the topology in protein folding by computational inverse folding experiments. *Comput. Biol. Chem.*, 32(4):233–239.

[Murphy, 2001] Murphy, K. (2001). *Protein structure, stability, and folding*. The Humana Press, Totowa, New Jersey, 1 edition.

[Nomura et al., 2011] Nomura, T., Kamada, R., Ito, I., Sakamoto, K., Chuman, Y., Ishimori, K., Shimohigashi, Y., and Sakaguchi, K. (2011). Probing phenylalanine environments in oligomeric structures with pentafluorophenylalanine and cyclohexylalanine. *Biopolymers*, 95(6):410–9.

[Onuchic and Wolynes, 2004] Onuchic, J. N. and Wolynes, P. (2004). Theory of protein folding. *Current Opinion in Structural Biology*, 14(1):70–75.

[Ozkan et al., 2007] Ozkan, S., Wu, G., Chodera, J., and Dill, K. A. (2007). Protein folding by zipping and assembly. *Proc Natl Acad Sci USA.*, 104:119:87–92.

[Pande et al., 1998] Pande, V. S., Grosberg, A., Rokshar, D., and Tanaka, T. (1998). Pathways for protein folding: is a new view needed? *Curr. Opin. Struct. Biol.*, 8:68–79.

[Pande and Rokhsar, 1999] Pande, V. S. and Rokhsar, D. (1999). Folding Pathway of a Lattice Model for Protein Folding. In *Natl. Acad. Sci. USA*, pages 1273–1278, Natl. Acad. Sci. USA.

[Pande et al., 2008] Pande, V. S., Sorin, E. J., Snow, C. D., and Rhee, Y. M. (2008). Chapter 8 Computer Simulations of Protein Folding. In *Protein Folding, Misfolding and Aggregation: Classical Themes and Novel Approaches*, pages 161–187. The Royal Society of Chemistry.

[Pappu et al., 1988] Pappu, R. V., Hart, R. K., and Ponder, J. (1988). Tinker: a package for molecular dynamics simulation. *J. Phys. Chem. B*, 102:9725–42.

[Pardalos et al., 2012] Pardalos, P. M., Coleman, T. F., and Xanthopoulos, P. (2012). *Optimization and Data Analysis in Biomedical Informatics*, volume 63. Springer.

[Plaku et al., 2007] Plaku, E., Stamati, H., Clementi, C., and Kavraki, L. E. (2007). Fast and reliable analysis of molecular motion using proximity re-

lations and dimensionality reduction. *Proteins: Structure, Function, and Bioinformatics*, 67(4):897–907.

[Plaxco et al., 1998] Plaxco, K. W., Simons, K. T., and Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol*, 277(4):985–994.

[Ponder and Case, 2003] Ponder, J. W. and Case, D. A. (2003). Force fields for protein simulations. *Adv Protein Chem*, 66:27–85.

[Privalov, 1996] Privalov, P. (1996). Intermediate States in Protein Folding. *J. Mol. Biol.*, 258:707–725.

[Rader et al., 2002] Rader, A. J., Hespenheide, B. M., Kuhn, L. A., and Thorpe, M. F. (2002). Protein unfolding: rigidity lost. *Proceedings of the National Academy of Sciences of the United States of America*, 99(6):3540–5.

[Rajan et al., 2010] Rajan, A., Freddolino, P. L., and Schulten, K. (2010). Going beyond clustering in MD trajectory analysis: an application to villin headpiece folding. *PloS one*, 5(4):e9890.

[Reynaud et al., 2010] Reynaud, B. E., Biotecnologia, P. D. I. D., and De, U. N. A. (2010). Protein Misfolding and Degenerative Diseases. *Nature Education*, 3(9):7–13.

[Roder and Colón, 1997] Roder, H. and Colón, W. (1997). Kinetic role of early intermediates in protein folding. *Current opinion in structural biology*, 7(1):15–28.

[Roder et al., 2006] Roder, H., Maki, K., and Cheng, H. (2006). Early Events in Protein Folding Explored by Rapid Mixing Methods. *Chemical Reviews*, 106:1836–1861.

[Rohl et al., 2004] Rohl, C., Strauss, C. E. M., Misura, K. M. S., and Baker, D. (2004). Protein structure prediction using Rosetta. *Methods in enzymology*, 383:66–93.

[Rousseeuw, 1987] Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65.

[Schirf,] Schirf, C. *Automated Protein Classification Using Rigidity Analysis*. PhD thesis.

[Shaw et al., 2008] Shaw, D. E., Deneroff, M. M., Dror, R. O., Kuskin, J. S., Larson, R. H., Salmon, J. K., Young, C., Batson, B., Bowers, K. J., Chao, J. C., and Others (2008). Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *COMMUNICATIONS OF THE ACM*, 51(7).

[Shi et al., 2008] Shi, Y., Zhou, J., Arndt, D., Wishart, D., and Lin, G. (2008). Protein contact order prediction from primary sequences. *BMC Bioinformatics*, 9(1):255.

[Shrake and Rupley, 1973] Shrake, A. and Rupley, J. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.*, 79:351–371.

[Sivogolovko and Novikov, 2012] Sivogolovko, E. and Novikov, B. (2012). Validating cluster structures in data mining tasks. *Proceedings of the 2012 Joint EDBT/ICDT Workshops on - EDBT-ICDT '12*, page 245.

[Song, 2004] Song, G. (2004). *A Motion Planning Approach to Protein Folding*. PhD thesis, Dept. of Computer Science, Texas A&M University.

[Song and Amato, 2001] Song, G. and Amato, N. M. (2001). Using Motion Planning to Study Protein Folding Pathways. *Journal of Computational Biology*, pages 287–296.

[Stickle et al., 1992] Stickle, D. F., Presta, L. G., Dill, K. A., and Rose, G. (1992). Hydrogen bonding in globular proteins. *J Mol Biol*, 226:1143–1159.

[Takashima, 1993] Takashima, S. (1993). Use of protein database for the computation of the dipole moments of normal and abnormal hemoglobins. *Biophysical journal*, 64(5):1550–8.

[Tapia et al., 2010] Tapia, L., Thomas, S., and Amato, N. M. (2010). A motion planning approach to studying molecular motions. *Communications in Information & Systems*, 10(1):53–68.

[Thomas et al., 2007] Thomas, S., Tang, X., Tapia, L., and Amato, N. M. (2007). Simulating Protein Motions with Rigidity Analysis. *Journal of Computational Biology*, 14(6):839–855.

[Thomas et al., 2013] Thomas, S., Tapia, L., Ekenna, C., Yeh, H.-Y. C., and Amato, N. M. (2013). Rigidity Analysis for Protein Motion and Folding Core Identification. In *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

[Thorpe et al., 2001] Thorpe, M. F., Lei, M., Rader, a. J., Jacobs, D. J., and Kuhn, L. a. (2001). Protein flexibility and dynamics using constraint theory. *Journal of molecular graphics & modelling*, 19(1):60–9.

[Toofanny et al., 2010a] Toofanny, R. D., Jonsson, A. L., and Daggett, V. (2010a). A comprehensive multidimensional-embedded, one-dimensional reaction coordinate for protein unfolding/folding. *Biophysical journal*, 98(11):2671–81.

[Toofanny et al., 2010b] Toofanny, R. D., Jonsson, A. L., and Daggett, V. (2010b). A comprehensive multidimensional-embedded, one-dimensional reaction coordinate for protein unfolding/folding. *Biophysical journal*, 98(11):2671–81.

[Udgaonkar, 2008] Udgaonkar, J. B. (2008). Multiple Routes and Structural Heterogeneity in Protein Folding. *Annual Review of Biophysics*, 37:489–510.

[Uversky, 2002] Uversky, V. N. (2002). What does it mean to be natively unfolded? *The Federation of European Biochemical Societies Journal*, 269(1):2–12.

[van der Spoel et al., 2005a] van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., and Berendsen, H. J. C. (2005a). GROMACS: Fast, Flexible and Free. *J. Comp. Chem.*, 26:1701–1719.

[van der Spoel et al., 2005b] van der Spoel, D., Lindahl, E., Hess, B., van Buuren, A. R., Apol, E., Meulenhoff, P. J., Tieleman, D. P., Sijbers, A. L. T. M., Feenstra, K. A., van Drunen, R., and Berendsen, H. J. C. (2005b). Gromacs User Manual version 3.3.

[Wang, 2009] Wang, H.-c. (2009). *STUDIES ON THE MECHANICAL STABILITY OF A PROTEIN BY SINGLE-MOLECULE ATOMIC FORCE MICROSCOPY*. PhD thesis.

[Wei et al., 2010] Wei, H., Yang, L., and Gao, Y. (2010). Mutation of Charged Residues to Neutral Ones Accelerates Urea Denaturation of HP-35. *The Journal of Physical Chemistry B*, 3:11820–11826.

[Whitford, 2005] Whitford, D. (2005). *Proteins: Structure and Function*. John Wiley and Sons, Ltd.

[Wickelmaier, 2003] Wickelmaier, F. (2003). *An introduction to MDS*. Aalborg Universitetsforlag.

[Zaki et al., 2004] Zaki, M., Nadimpally, V., Bardhan, D., and Bystroff, C. (2004). Predicting protein folding pathways. *Bioinformatics*, 20:386–393.

[Zemla, 2003] Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Research*, 31(13):3370–3374.

[Zhao and Ji-Jun,] Zhao, F.-Y. L. and Ji-Jun. Quantum chemistry PM3 calculations of sixteen mEGF molecules.