

## University of Southampton Research Repository ePrints Soton

Copyright © and Moral Rights for this thesis are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given e.g.

AUTHOR (year of submission) "Full thesis title", University of Southampton, name of the University School or Department, PhD Thesis, pagination

UNIVERSITY OF SOUTHAMPTON

# **The Computational Investigation of Protein/Ligand Complexes: Implications for Rational Drug Design**

Francesca Toschi

DECLARATION OF ORIGINALITY  
I hereby declare that the work presented in this thesis is my own work and that it has not been submitted for publication elsewhere in any form. I have not plagiarised any other work and I have acknowledged all sources of information used in the preparation of this thesis.

A thesis submitted for the qualification of  
Doctor of Philosophy at the University of Southampton

**Department of Chemistry**

**December 2004**



University of Southampton

ABSTRACT

FACULTY OF SCIENCE

CHEMISTRY

Doctor of Philosophy

THE STUDY OF PROTEIN/LIGAND COMPLEXES:  
IMPLICATIONS FOR RATIONAL DRUG DESIGN  
by Francesca Toschi

Ligand binding may involve a wide range of structural changes in the receptor protein, from *hinge* movements of entire domains to small side-chain rearrangements in the binding pocket residues.

While changes in the backbone of proteins are sometimes negligible, side-chain rearrangements are generally observed. Knowledge of the extent to which side-chain conformational changes occur is very important to improve drug design and docking prediction algorithms, particularly since most of these algorithms adopt a rigid receptor hypothesis.

In this thesis, the extent of side-chain motion in known apo- and holo- X-ray crystal structures is examined. Different methods to characterise side-chain conformational changes occurring in protein-ligand complexes were employed. A dataset of PDB structures was chosen and apo-/apo-, apo-/holo- and holo-/holo- protein pairs' torsion angles compared.

Side-chain conformational changes were defined on the basis of both constant and environment- and residue- dependent thresholds. Also, recently published rotamer libraries were employed and the probabilities of the different rotamers found compared.

The results of the analysis provide interesting insights into the intrinsic flexibilities of protein active sites, and the extent of side-chain conformational change on ligand binding. The general patterns and features are potentially useful for the prediction of conformational changes occurring in proteins upon ligand binding.

# Contents

---

|   |           |
|---|-----------|
| <b>1 Introduction and Overview</b>  | <b>1</b>  |
| 1.1 Introduction . . . . .  | 1         |
| 1.2 Overview . . . . .  | 2         |
| <b>2 Protein Motions</b>  | <b>5</b>  |
| 2.1 Protein Flexibility . . . . .   | 5         |
| 2.2 Protein Motion Classification . . . . .                                   | 6         |
| 2.2.1 Size Classification . . . . .   | 7         |
| 2.2.2 Packing Classification . . . . .  | 8         |
| 2.3 Protein Motions and Ligand Binding . . . . .                              | 13        |
| 2.4 Protein Flexibility in Rational Drug Design . . . . .                     | 16        |
| 2.4.1 Computational Generation of Protein Conformations . . . . .             | 17        |
| 2.4.2 Methods Employing Experimentally Determined Structures . . . . .        | 20        |
| 2.5 Summary and Conclusion . . . . .  | 21        |
| <b>3 Side-Chain Flexibility Analysis</b>                                      | <b>24</b> |
| 3.1 Side-Chains and Protein Flexibility . . . . .                             | 24        |
| 3.2 Rotamer Libraries . . . . .   | 25        |
| 3.2.1 Rotamers and Rotamer Libraries . . . . .                                | 25        |
| 3.2.2 Limitations of the Rotamer Approximation . . . . .                      | 29        |
| 3.3 Analysis of Conformational Changes Occuring Upon Ligand Binding . . . . . | 33        |
| 3.3.1 Najmanovich <i>et al.</i> . . . . .                                     | 33        |
| 3.3.2 Fradera <i>et al.</i> . . . . .   | 35        |
| 3.4 Study of the Inherent Flexibility of Protein Side-Chains . . . . .        | 39        |
| 3.4.1 Zhao <i>et al.</i> . . . . .  | 39        |
| 3.5 Summary and Conclusion . . . . .  | 43        |
| 3.6 Aim of this Project . . . . .   | 47        |
| <b>4 Conformational Analysis: Data Set, Methods and Results Overview</b>      | <b>49</b> |
| 4.1 Choice of the Data Set . . . . .  | 49        |
| 4.2 Methods . . . . .   | 51        |
| 4.3 Results Overview . . . . .  | 60        |

---

|          |   |           |
|----------|---|-----------|
| 4.3.1    | Root Mean Square Deviation . . . . .  | 60        |
| 4.3.2    | Different Amino Acid Conformational Changes . . . . .   | 61        |
| 4.3.3    | All-Environment Flexibility Trends . . . . .  | 71        |
| 4.3.4    | Buried and Exposed Residues Flexibility Trends . . . . .  | 76        |
| 4.3.5    | Differences between apo-/holo- and holo-/holo- protein confor-<br>mational changes . . . . .  | 83        |
| 4.3.6    | Differences Between All Residues and Only Binding Site Residues<br>Results . . . . .  | 86        |
| 4.3.7    | Comparisons of the Results Obtained for Structures Solved at<br>2.0 Å or better and 2.5 Å or better . . . . .                               | 90        |
| 4.3.8    | Relationships Between Holo-Protein Side-Chain Conformational<br>Changes and Ligand Similarities . . . . .                                   | 92        |
| 4.4      | Conclusion . . . . .  | 93        |
| <b>5</b> | <b>HIV-1 Protease . . . . .</b>   | <b>98</b> |
| 5.1      | Introduction . . . . .  | 98        |
| 5.2      | HIV-1 Protease and Aspartic Proteases: Structure . . . . .  | 99        |
| 5.3      | Background to Protein: Past Work . . . . .  | 104       |
| 5.3.1    | Structure and Dynamic Behaviour of HIV-1 Protease: X-ray<br>Structure Comparisons, Molecular Dynamics and Normal Mode<br>Analyses . . . . . | 104       |
| 5.3.2    | Structure-Based Thermodynamic Study of HIV-1 Protease In-<br>hibitors . . . . .   | 109       |
| 5.4      | Results: HIV-1 Protease Side-Chain Conformational Changes . . . . .   | 112       |
| 5.4.1    | All Environments, and Environment Specific Conformational<br>Changes . . . . .  | 114       |
| 5.4.2    | Relationships Between Holo-Protein Side-Chain Conformational<br>Changes and Ligand Similarities . . . . .                                   | 116       |
| 5.4.3    | Investigation on Cross-Correlated Motions of Different Residues'<br>Side-Chains . . . . .   | 116       |
| 5.4.4    | Conformational Analysis of Holo-Proteins bound to the Same<br>Ligand and/or Solved by the Same Authors. . . . .                             | 120       |

---

|          |   |            |
|----------|---|------------|
| 5.4.5    | Percentages of Conformational Changes per Residue Sequence  |            |
|          | Number . . . . .  | 124        |
| 5.5      | Conclusions . . . . .   | 138        |
| <b>6</b> | <b>Endothiapepsin</b>   | <b>144</b> |
| 6.1      | Structure . . . . .   | 144        |
| 6.2      | Background to Protein: Past Work . . . . .  | 147        |
| 6.2.1    | Structure-Based Thermodynamic Analysis of Endothiapepsin/<br>Pepstatin Binding . . . . .                      | 147        |
| 6.3      | Results: Analysis of Endothiapepsin Conformational Changes . . . .  | 150        |
| 6.3.1    | All Environments, and Environment Specific Conformational<br>Changes . . . . .                                | 151        |
| 6.3.2    | Percentages of Conformational Changes per Residue Sequence<br>Number . . . . .                                | 154        |
| 6.4      | Conclusions . . . . .   | 163        |
| <b>7</b> | <b>Streptavidin</b>   | <b>166</b> |
| 7.1      | Structure . . . . .   | 166        |
| 7.2      | Background to Protein: Past Work . . . . .  | 170        |
| 7.2.1    | Structure-Based Thermodynamic Analysis of Streptavidin Bind-<br>ing to Structurally Diverse Ligands . . . . . | 170        |
| 7.3      | Results: Analysis of Streptavidin Conformational Changes . . . . .  | 177        |
| 7.3.1    | All Environments, and Environment Specific Conformational<br>Changes . . . . .                                | 180        |
| 7.3.2    | Relationships Between Holo-Protein Side-Chain Conformational<br>Changes and Ligand Similarities . . . . .     | 181        |
| 7.3.3    | Percentages of Conformational Changes per Residue Sequence<br>Number . . . . .                                | 182        |
| 7.4      | Conclusions . . . . .   | 190        |
| <b>8</b> | <b>Conclusions</b>  | <b>194</b> |
| 8.1      | Summary . . . . .   | 194        |
| 8.2      | Future Work . . . . .   | 199        |

|  |     |
|--|-----|
| References   | 201 |
| A Thermodynamic Analysis of Ligand-Protein Interactions                  | 209 |
| A.1 Isothermal Titration Calorimetry of Protein-Ligand Binding . . . . . | 209 |
| A.2 Structure-Based Thermodynamic Analysis . . . . .                     | 211 |
| A.2.1 Prediction of Binding Enthalpies . . . . .                         | 211 |
| A.2.2 Prediction of Entropy Changes . . . . .                            | 214 |
| B Residue Stability Constants  | 218 |
| B.1 Introduction . . . . .   | 218 |
| B.2 COREX . . . . .  | 219 |
| C Additional Figures   | 223 |

## Acknowledgments

First of all I would like to thank my supervisor Jonathan Essex for his enthusiasm, sense of humor and dedication to scientific research. Also, thanks to my industrial supervisors Andrew R. Leach and Paul A. Bamborough for having been so kind and helpful during these years of work, and to GlaxoSmithKline for making possible this research.

Many, many thanks to my 'ex-supervisor' Richard Lewis, who made me love the first six months I spent in the U.K. and greatly encouraged me to begin a PhD in England (something I would have done just to walk everyday on the land he lives on).

Than many thanks to the wonderful people I have met in this country, from my English landladies and colleagues (Chris, Rob, Rich T., Steve... I really would not have written this thesis without them), to the great French, Spanish, Thai, Indian, German, Malay, Japanese, Australian, Swedish, Italian, Luxemburger, Irish, Brazilian, Danish, Portuguese, Finnish, etc., etc., etc. incredibly nice, generous, funny, big-hearted and open-minded people, who made me enjoy life and weekends here.

Huge thanks to my parents for everything they have done for me. More recently, I would like to thank them for the Italian salame, mortadella, parmesan and coffee they sent me on a regular basis and for bearing the 'emigration' of their daughter without complaining about it. I only do not thank them for voting... who they vote!)

Also, many thanks to my Italian friends for remaining always very 'close' to me. Special thanks to Betta for such instructive and delightful daily conversations by email, but also to Silvia, Don Matteo, my brother, my niece Teresa, my crazy cousins Vally and Bendi, Claudia, Federica, her husband and daughters, Tonno and Betta,

Mora, Sbond (Annalisa, Bond), Bicci, etc., etc., etc...

Finally, of course, lots of thanks to Paolo, who greatly supported and helped me during these four years (and eventually became my husband in the course of the last one!). This long time surely cost to him at least as much as it cost to me (not only in phone bills and flight tickets!), but he always tried to not make me feel the weight of it.

## Introduction and Overview

The purpose of this book is to provide a comprehensive overview of the current state of the art in the field of artificial intelligence, with a particular focus on the applications of machine learning and deep learning.

### Introduction

The field of artificial intelligence (AI) has seen rapid growth and development in recent years, with significant advances in machine learning, deep learning, and natural language processing. This book aims to provide a comprehensive overview of the current state of the art in the field of AI, with a particular focus on the applications of machine learning and deep learning. The book is organized into several chapters, each covering a different aspect of the field. Chapter 1 provides an overview of the field of AI, including the history and current state of the art. Chapter 2 covers the fundamentals of machine learning, including supervised and unsupervised learning. Chapter 3 covers the fundamentals of deep learning, including convolutional neural networks and recurrent neural networks. Chapter 4 covers the applications of machine learning and deep learning in various domains, including computer vision, natural language processing, and robotics. Chapter 5 covers the ethical and social implications of AI, including issues related to privacy, security, and bias. Chapter 6 covers the future of AI, including emerging trends and challenges. The book is intended for a wide audience, including students, researchers, and practitioners in the field of AI. It provides a comprehensive overview of the current state of the art in the field of AI, with a particular focus on the applications of machine learning and deep learning. The book is organized into several chapters, each covering a different aspect of the field. Chapter 1 provides an overview of the field of AI, including the history and current state of the art. Chapter 2 covers the fundamentals of machine learning, including supervised and unsupervised learning. Chapter 3 covers the fundamentals of deep learning, including convolutional neural networks and recurrent neural networks. Chapter 4 covers the applications of machine learning and deep learning in various domains, including computer vision, natural language processing, and robotics. Chapter 5 covers the ethical and social implications of AI, including issues related to privacy, security, and bias. Chapter 6 covers the future of AI, including emerging trends and challenges. The book is intended for a wide audience, including students, researchers, and practitioners in the field of AI.

# Chapter 1

## Introduction and Overview

---

### 1.1 Introduction

Protein flexibility and rearrangement upon ligand binding is one of the main complicating factors in structure-based drug design. Since even small changes in the receptor conformation can dramatically change ligand binding affinities, considering a single and rigid receptor structure is a dangerous limitation; protein flexibility and/or different structures of the same protein should ideally be included in ligand docking.<sup>1</sup>

The energy landscape of most proteins can in fact be described as a folding funnel in which many highly unfavourable states collapse via multiple routes into possibly several favourable folded states.<sup>2</sup> A single structure cannot describe adequately these substates; protein structures deposited in the PDB, although representing a statistical average of many similar conformations existing in the crystal lattice, only provide one of the many possible conformations accessible to a protein. Moreover, crystal



structure conformations can also be influenced by crystallisation conditions, such as pH, temperature and crystal packing effects,<sup>3</sup> and by the methods employed to solve and refine the structures.

While the motions that take place upon ligand binding can be quite significant, involving substantial backbone reorganisation, side-chain conformational flexibility is the most common. Obtaining a better understanding of side-chain flexibility, would therefore be very useful in structure-based drug design. The approach to this problem that has been undertaken in this thesis is to analyse experimental protein-ligand structures, comparing multiple structures of the same protein in the presence and absence of different ligand molecules, to determine what side-chain rearrangements take place. This study will yield useful insights into the intrinsic flexibility of protein systems, and allow ligand-induced conformational changes to be resolved; motions that depend on genuine ligand binding effects and those which can be ascribed to spontaneous conformational changes and/or possible crystallographic artefacts will be distinguished. The ultimate objective is clearly to determine whether it is possible to model side-chain flexibility in protein-ligand binding studies through the prediction of likely side-chain conformations using knowledge derived from experimentally determined protein structures.

## 1.2 Overview

The overall aims of this project are the following:

1. Selection of a data-set of PDB structures belonging to different protein families comprising several apo- and holo- structures of the same protein solved at good resolution.
2. Definition of side-chain conformational changes occurring in each pair of apo-

/apo-, apo-/holo- and holo-/holo- proteins of the data set on the basis of different angular thresholds and rotamer libraries. Evaluation of average backbone Root Mean Square deviation within the same protein system.

3. Identification of possible cross-correlations in residues' conformational changes.
4. Identification of possible correlations between the observed percentages of conformational changes induced by ligand binding and the ligands' similarity coefficients.
5. Evaluation of side-chain flexibility trends to check whether protein structures solved by the same author(s) are more likely to show less conformational changes than structures solved by different groups of crystallographers.
6. A detailed analysis of side-chain flexibility across a range of protein systems, to determine the extent of side-chain motion as a function of side-chain solvent exposure, and amino acid location in the protein structure. This study will help to resolve the issue of whether a given conformational change is systematic and driven by ligand binding, or merely that which would be expected given random influences such as the effect of the crystallisation conditions.
7. More in depth analysis of specific protein systems. Comparison of the observed side-chain flexibility with other relevant literature data.

In the following chapters of this thesis, protein flexibility, together with existing methods for calculating side-chain flexibility, will be reviewed. Ten protein ligand systems will be identified, and representative protein structures obtained from the protein databank. Flexibility analyses will be reported on all systems, with the final chapters focusing in more detail on three of the more interesting proteins (HIV-1

protease, endothiapepsin, and streptavidin), for which other experimental thermodynamic data are available.

## Chapter 2

# Protein Motions

## Protein Flexibility

In many cases, conformational flexibility of a protein is an important factor in its function. For example, the flexibility of a protein can be crucial for its ability to bind a ligand or to catalyze a reaction. The flexibility of a protein can also be a factor in its stability and its resistance to denaturation.

There are many ways to study protein flexibility. One common method is to use X-ray crystallography to determine the structure of a protein in different states. Another method is to use NMR spectroscopy to study the dynamics of a protein in solution. A third method is to use molecular dynamics simulations to study the flexibility of a protein in silico. Each of these methods has its own strengths and weaknesses, and the choice of method depends on the specific question being asked.

## Chapter 2

# Protein Motions

---

### 2.1 Protein Flexibility

Proteins are dynamic structures that are intrinsically mobile; the accessibility of alternative conformational states is essential for their assembly, regulation, biological activity and catalysis.<sup>4</sup>

After polypeptide chains have been synthesised by ribosomes, proteins leave the ribosomes and begin to fold.<sup>5</sup> This is a complex process which is still little understood; however, it is thought that secondary structure elements come first then, because of their hydrophobic effect, these elements fold into the tertiary structure, excluding water molecules from the interior of the protein.<sup>6</sup> The general backbone topology develops first, then side-chains tightly pack; there is some evidence that atomic packing is driving the last step in the folding pathway.<sup>6</sup>

The flexibility of proteins is also essential for their function. For example, enzymes

must absorb their substrates, stabilise their transition state, and release their products. Also, membrane proteins might have to transmit a message from the outside to the inside of the cell; most proteins can bind a range of different ligands, and thus they must be able to adapt their structure to some extent. In many cases, substantial differences in conformation can result from relatively modest external stimuli, such as a pH change.<sup>1</sup> Growing experimental evidence also suggests that proteins consist of a myriad of different possible conformations that are in equilibrium; a protein can shift from one conformational state to another, in the presence or absence of external stimuli or ligands.<sup>2</sup>

The need to account for the dynamic behaviour of proteins is fundamental in rational drug design; the accurate prediction of binding modes between ligands and proteins relies on it. The choice of only one, rigid, protein structure, even if very efficient from a computational point of view, is in fact arbitrary and insufficient; computational methods that are able to study, predict and include different conformational states of a protein are becoming more and more common.<sup>7</sup>

## 2.2 Protein Motion Classification

Protein motions can be classified in one of two ways: on the basis of size and on the basis of packing.<sup>4,8</sup>

Applying the size classification, protein movements fall into three categories of decreasing size: motions of subunits, motions of domains and motions of fragments smaller than domains.<sup>4</sup> Of course, the motion (i.e. rotation) of individual side-chains is also possible, often occurring on the protein surface. However, this is on a much smaller scale than the motion of fragments or domains, it also occurs in all proteins and is normally observed in subunits, domains and fragment motions. Side-chain

motions can thus be considered a kind of background, intrinsic protein flexibility: they can also bring about movements of the backbone, and thus be the primary cause of large displacements of domains, fragments and any other structural parts.<sup>9</sup>

The packing classification divides motions into various categories depending on whether or not they involve sliding over a continuously maintained and tightly packed interface.<sup>4</sup>

### 2.2.1 Size Classification

Almost all large proteins are built from domains, and domain movements provide, excluding side-chains motions, the most common examples of protein flexibility. The motion of fragments smaller than domains usually involves the motion of surface loops, but it can also involve the motion of secondary structure elements.<sup>8</sup>

Often domain and fragment motions involve the closing of the protein around a binding site, with a bound substrate favouring a “closed” conformation. This closure around a binding site has been analysed in particular detail and its observation has probably been the basis of the ligand binding “induced-fit” theory.<sup>10</sup> Active site closure positions important chemical groups around the substrate, shields it from water, prevents the escape of reaction intermediates, and optimises favourable interactions between residues and ligands.

Subunit motion is distinctly different from fragment or domain motion; it affects two large sections of a protein that are not covalently connected and is often involved in allosteric transition and regulation. An example of subunit motion is provided by hemoglobin, in which the motions of the subunits change the affinity with which this protein binds to oxygen.<sup>11</sup>

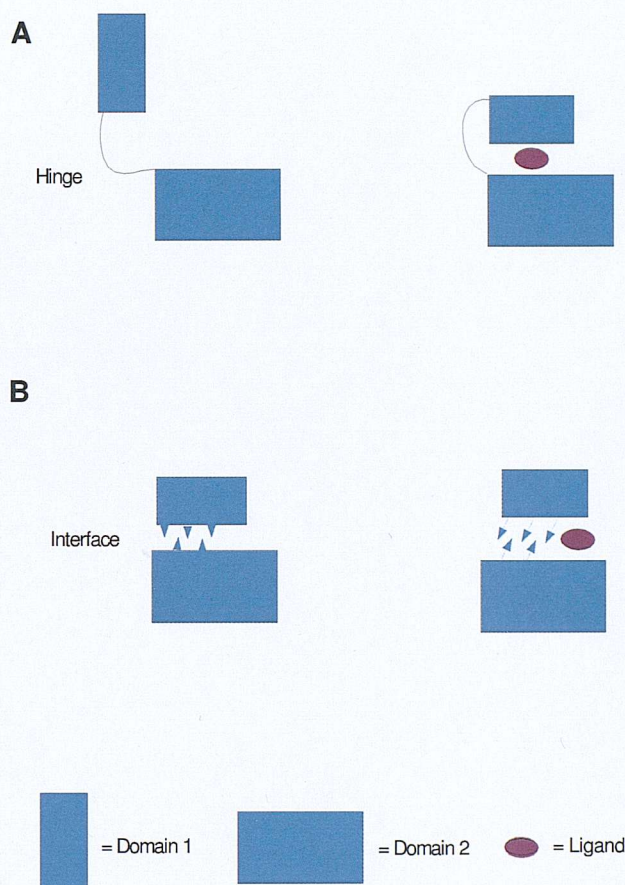
### 2.2.2 Packing Classification

Motions of protein domains and smaller units can be categorised on the basis of packing. The tight packing of atoms inside proteins provides a constraint on protein structure;<sup>6</sup> usually the atoms inside a protein cannot move much without colliding with neighbouring atoms, with the exception of cases where there is a cavity or packing defect. Also, internal interfaces between different parts of a protein are packed very tightly<sup>12,13</sup> and are formed from interdigitating side-chains; maintaining packing throughout a motion means that the side-chains at the interface maintain their relative orientation and inter side-chain contacts in both conformations (e.g. open and closed).

Individual movements within a protein can be described in terms of two basic mechanisms, shear and hinge,<sup>13</sup> depending on whether or not they involve sliding over a continuously maintained interface. These are illustrated in Figure 2.1.

Complete protein motions can be built up by many of these smaller movements and are classified as shear if they predominantly contain shear movements and as hinge if they predominantly contains hinge movements.

The shear mechanism is the sliding motion occurring in proteins that maintain well-packed interfaces. Because of the constraints on interface structure, individual shear motions have to be very small; side-chain torsion angles maintain the same rotamer configuration and there is no appreciable backbone deformation. The whole motion is parallel to the plane of the interface and is limited to total translations of about 2 Å and side chain rotations of 15°. Since individual shear motions are so small, a single one is not sufficient to produce a large overall motion; several shear motions must be summed to give a large effect, similar to each plate in a stack of



**Figure 2.1:** Domain closure represented for (A) hinge and for (B) shear mechanisms. While the overall motion is built up by many small local motions and is parallel to the interface between domains in (B), motion in (A) corresponds to a few large conformational changes occurring at the hinge and is perpendicular to the newly created interface between domains.

plates sliding slightly to determine an overall considerable bending of the whole stack.

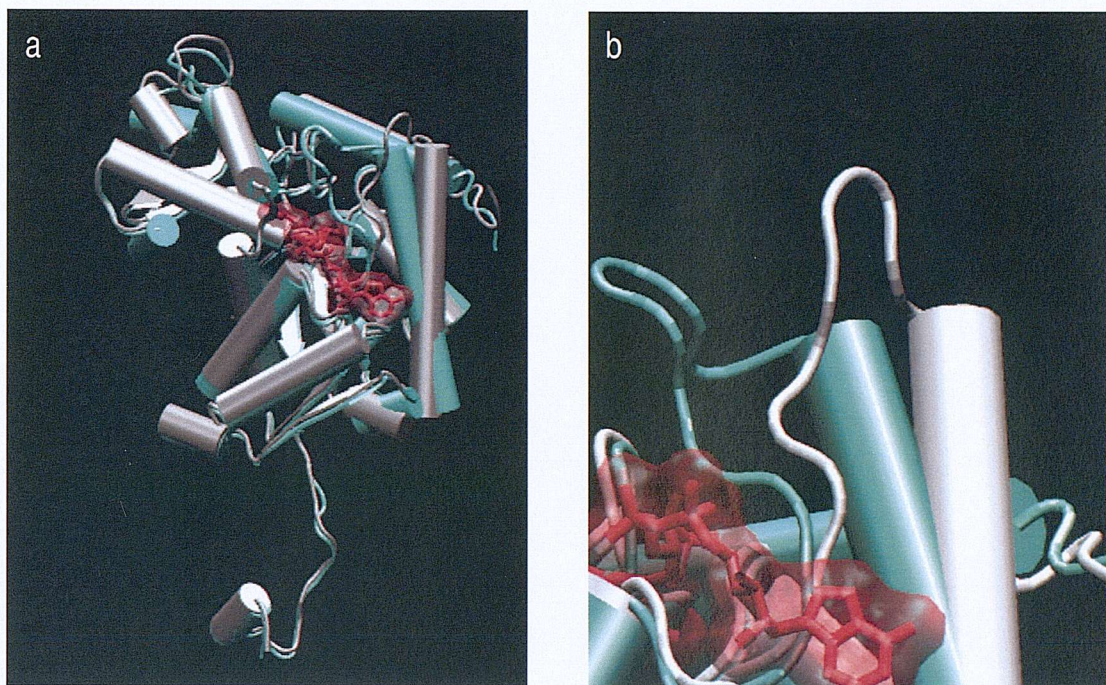
Hinge motions occur when there is no continuously maintained interface constraining the motion. They usually occur in proteins that have two domains or fragments connected by linkers, i.e. hinges, that are relatively unconstrained by packing. A few large torsion angle changes in the hinges are sufficient to produce almost the whole motion; the rest of the protein rotates essentially as a rigid body, with the axis of the overall rotation usually passing through the hinges. The overall motion is perpendicular to the plane of the interface; the interface exists in one conformation but not in



the other, similar to the opening and closing of a book.

Protein hinges are characterised by an exposed main chain and by few packing constraints.<sup>14,15</sup> Most main chain atoms are instead usually buried beneath layers of other atoms (mainly atoms of side-chains) that prevent them from undergoing large torsion angle changes. Hinges do not appear to be related to chain topology or secondary structure; they can be found in loops, sheets and helices.<sup>16</sup> It is important to remember that most shear motions do contain hinges that join the various sliding parts; the existence of a hinge is not the main difference between the two basic mechanisms. Rather, it is the existence of a continuously maintained interface.

An example of multiple-hinge protein motion is represented by lactate dehydrogenase (Figure 2.2). Upon binding lactate and NAD, this enzyme undergoes a large conformational change, with a surface loop (particular enlarged in Figure 2.2 (b)) moving about 10 Å to cover the active site.<sup>17</sup> This motion occurs at two hinges: while the first has few steric constraints and depends on large conformational changes in only two torsion angles, as in a classic hinge motion, the second has many more constraints and distributes its deformation over more torsion angles. The motion of the second hinge, which is part of a helix connected to the end of the loop, causes the helix to stretch and split into two distinct components, and side-chain repacking to occur at the interface with the end of a neighbouring helix. Through a network of contacts (most involving hydrophobic residues), the motion of the loop is propagated to parts of the protein that are not in direct contact with the ligands. These moving structures (five helices and three other loops move approximately 2 Å) are on the surface of the protein; the whole enzyme can therefore be subdivided into concentric shells of increasing mobility.<sup>17</sup>

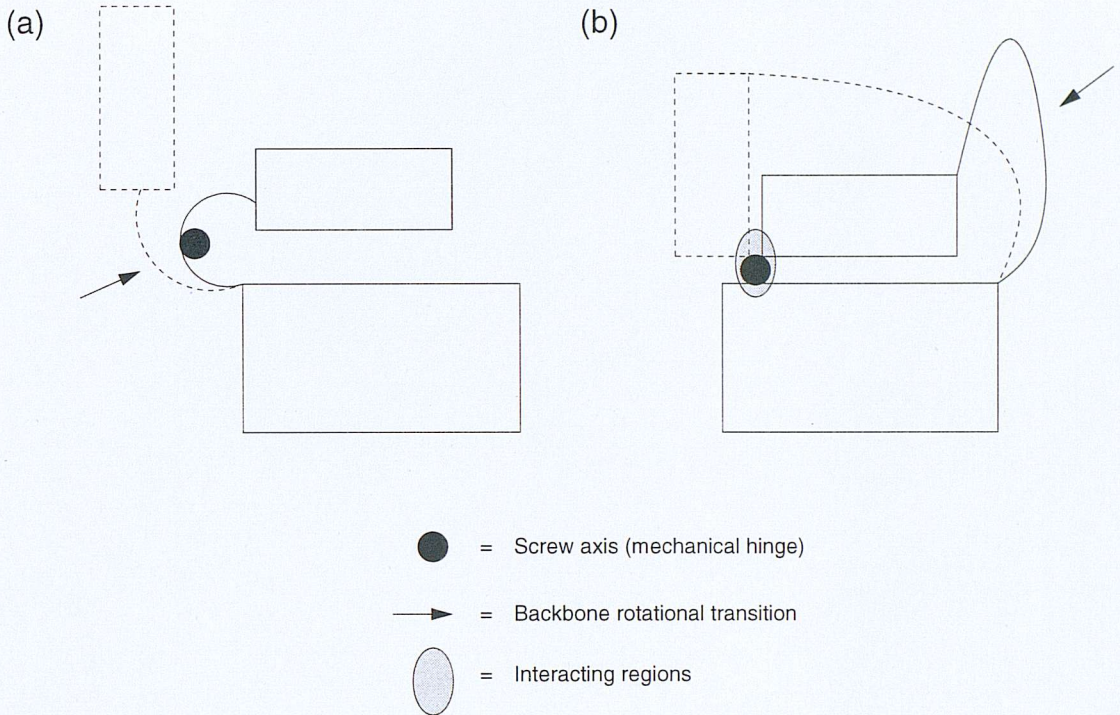


**Figure 2.2:** Motion of Lactate Dehydrogenase (LDH) upon pyruvate and NADH binding. (a) Superimposed structures of uncomplexed (grey) and complexed (cyan) LDH. The substrate's and coenzyme's accessible surface areas are shown. Large conformational changes occur: a surface loop moves roughly 10 Å to cover the active site (enlarged in (b)). Appreciable movements of five helices and three other loops are observed.

In most hinge motions the interdomain screw axis coincides with the backbone region where the rotational transition occurs. However, this is not always the case: proteins where the interdomain screw axis (i.e. the “mechanical hinge”) is distant from the region of the backbone where the rotational transition occur are known (e.g. endothiapepsin, tomato bushy stunt virus).<sup>16</sup> In these cases, the interdomain screw axis is found where interactions among residues from the different structures are preserved; the overall rotational transition can be located in the side-chain dihedrals of the residues, belonging to different domains or fragments, that establish noncovalent interactions between them.<sup>16</sup> The hinge is, in these cases, created by the intrinsic flexibility of these noncovalent interactions and of their side-chain dihedrals (Figure



2.3).



**Figure 2.3:** (a) Domain motion in which the interdomain screw axis (black circle) and the region where the rotational transition occurs (arrow) are the same. (b) Case in which the mechanical hinge is distant from the backbone region where the rotational transition occurs. The screw axis goes in fact through a part of the protein where interactions between residues that are not close in sequence (patterned region in the figure) are not preserved during the motion.

Most of fragment and domain motions fall within the hinge/shear classification. However, there are a number of exceptions: mechanisms that are clearly neither hinge nor shear, such as partial refolding of the protein that usually causes dramatic changes of the overall structure, order-to-disorder transitions, pro-enzymes that dramatically change structure after cleavage.

For the motions of subunits, different categories from hinge or shear may be applied:

1. allosteric motions (e.g. haemoglobin);
2. non-allosteric motions;

3. complex and large motions involving many subsidiary submotions that can be classified as subunit or domain motions themselves.

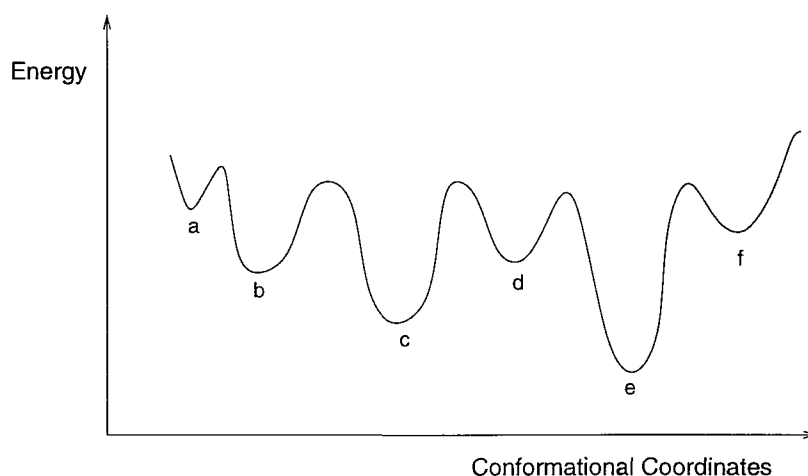
## 2.3 Protein Motions and Ligand Binding

It is now commonly accepted that proteins are molecules in constant motion.<sup>2</sup> The long-held rigid “lock and key theory”, according to which a protein exists in a single well-defined state with only one complementary ligand, was overtaken in the middle of the last century by the “induced-fit” theory,<sup>10,18</sup> which recognizes that flexibility in proteins is essential for protein and enzyme activity. According to this theory, a ligand binds to the lowest energy conformation of the protein, which is then distorted in order to accommodate it; this change in the protein structure causes a precise three-dimensional reorientation of the amino acid active site groups in a way that is necessary for protein function.

However, the theory behind protein-ligand binding is now moving away from this historical induced-fit theory. A new model describing proteins as molecules in equilibrium between different conformers that are possibly very similar in energy and constantly interchanging is now gaining large and growing experimental and theoretical evidence.<sup>2,19,20</sup> Alternative structures of the same protein exist for the unbound form that differ on many different size levels (e.g. the movement of one atom between two different stable position in a side-chain, the rotation of one or more side-chains, the movement of a loop, the shift of a helix, etc). Ligands might increase the proportion of one of the subpopulations of conformers simply by shifting the equilibrium to the conformer they preferably bind;<sup>9,21,22</sup> the “lock” (i.e. the receptor) is not described as rigid anymore but exists in a structural ensemble with some conformations fitting the “key” (i.e. the ligand). These fitting conformations are selected by the key

and not induced by it.

The evidence that the crystal structure of the unbound receptor is often different from the complexed form might find an explanation in the fact that, under crystallisation conditions, the unbound “open” conformation is the most populated, while ligands probably bind to alternative conformations, adjusting the equilibrium in their favour without the need for the ‘induced-fit’ mechanism. The crystal structure of the unbound protein represents a weighted average of the conformation ensemble and may, depending on the shape of the energy landscape, correspond to a state with a free energy higher than the minimum (Figure 2.4).<sup>7</sup>



**Figure 2.4:** Conformations (a) to (f) are accessible to a protein in solution. The two energetically most favourable, (c) and (e), might be resolved in high resolution crystal structures; their weighted, higher free-energy state average, somehow similar to (d), could instead be resolved in lower resolution structures.

The range of conformations that are accessible to a protein in solution can be described by the laws and distributions of statistical mechanics. The lower the barriers between the diverse conformers at and around the bottom of the folding energy funnel, the more flexible the molecule and the larger the number of conformations it can adopt. As stated by Teague in a recent review,<sup>20</sup> funnels characterised by an irregular bottom might determine broad range, non-specific binding, while funnels characterised by a few minima with high barrier between them or a smooth single

minima could imply rigid, more specific binding. The degree of substrate specificity of a receptor is thus likely to be correlated to its flexibility and function; for example, relatively nonspecific proteolytic enzymes or endonucleases possess a higher degree of flexibility, while enzymes with a narrow substrate specificities have fewer and more similar conformations.<sup>20</sup>

While alternative conformational states can be defined by significant differences in protein structure and large energy barriers, more modest conformational substates can involve smaller energetic barriers. In many cases significant conformational changes appear to be a consequence of modest external stimuli (such as ligand binding, pH variations, different crystallisation conditions) that act with two different mechanisms: kinetic regulation (when the barrier between two conformational states is reduced or eliminated by the stimulus) or thermodynamic regulation (when the free energy of an alternative state is lowered and becomes the new free energy minimum).<sup>2</sup> However, both multiple protein conformational states (characterised by significant differences in protein conformation and large energy barriers) and multiple protein conformational substates (differing for more modest coordinate changes and smaller energy barriers) are often observed in the absence of specific external condition variations. Multiple substates can sometimes be observed in a single crystal solved at high resolution.<sup>2</sup>

Protein rearrangements upon ligand binding in order to optimise ligand-protein interactions, something occurring in the backbone and particularly in protein side-chains, cannot be ruled out even if ligands largely bind to pre-existing protein conformations.<sup>23</sup> The capability of proteins to bind different ligands may fall beyond the simple consideration of ensembles and distributions; the conformational-selection model and the induced-fit model are equivalent from a thermodynamic point of view<sup>20</sup>



and do not necessarily rule each other out. In the process of redistributions of protein conformations, a certain extent of induced-fit may still operate locally.<sup>23</sup>

## 2.4 Protein Flexibility in Rational Drug Design

The evidence that proteins can accommodate many molecules in their binding sites, rearranging themselves with relatively little penalty either by small torsional changes in their side-chains or by large domain motions, has clear implications for rational drug design. The use of a single protein structure is only useful to find ligands that are able to bind that particular state of the subensemble or for proteins existing in a single well-defined conformational state (i.e. in the case of a system corresponding to the “lock and key” theory).

The use of a single protein structure has for a long time been the standard in rational drug design and often still is because of its speed and computational efficiency. It is in fact usually impractical to perform long time scale calculations on each possible conformer of a protein, as well as to screen every possible conformer of a ligand in a large database of compounds with each possible conformational state of a protein. However, advanced methods have been recently introduced to take into account protein flexibility in computational drug design.<sup>1,7,24,25</sup> In general, they can be divided into methods that employ experimentally determined structures and methods that use instead computer-generated conformations.

The use of multiple protein structures (MPS) is probably one of the best options to take into account the full flexibility of a protein. These structures can come from NMR studies, multiple crystal structures, multiple conformations generated by molecular-dynamics simulations, low-frequency normal modes, simulated annealing and other computational techniques.

This section only tries to present some examples of the many recently developed methods that include protein flexibility in drug discovery; full reviews about this topic are available in recent literature.<sup>7, 24, 26</sup>

## 2.4.1 Computational Generation of Protein Conformations

### Soft-Docking

The first attempt to accommodate small conformational changes in proteins was made by Jiang and Kim in 1991.<sup>27</sup> the protein was held fixed and a “soft” scoring function applied allowing for some overlap between the ligand and the protein. Soft scoring functions only include a small estimate of the flexibility of the receptors; the search for successful ligands is still limited to a restricted range of drug molecules’ sizes and conformations. Soft scoring functions are however computationally very efficient; there is no additional time required to calculate the fit of the ligand to the receptor, and the new function is relatively simple to implement in existing programs. This efficiency is the reason why there are still improvements being made to soft docking methods.<sup>28</sup>

### Sampling Side-Chain Conformations in the Receptor

One of the earliest incorporations of side-chain mobility in protein-ligand docking made use of a rotamer library (Leach, 1994).<sup>29</sup> Rotamer libraries include the lowest energy conformations of residue side-chains and provide a good estimate of accessible conformational states; the problem of energy barriers between different conformations is overcome (i.e. there is no risk of remaining trapped in a local minima as in other minimisation routines) and the search of the conformational space available to the receptor is relatively quick.



More recently, Schaffer and Verkhivker<sup>30</sup> employed a larger rotamer library to generate conformations of the side-chains, and in a second step minimised the docked ligand and the local side-chains in order to get the strongest possible contacts between the ligand and the receptor.

Docking optimisation routines are often applied only to a subset of atoms surrounding the ligand. They can use simulated annealing, steepest descent or conjugate gradient minimisations in order to optimise the energy, and Monte Carlo sampling and multicanonical molecular dynamics techniques to improve the conformational sampling. In multicanonical Monte Carlo, the energy surface is smoothed and unfavourable conformations allowed; after employing this technique Nakajima *et al.*<sup>31</sup> fully re-introduced the energy landscape and finally assessed the binding properties of the structures with a Boltzmann-weighting of the multicanonical sampling.

New interesting optimisation methods employ novel algorithms to avoid being trapped in local minima. For example, a method proposed by Apostolakis *et al.*<sup>32</sup> allows the ligand to initially adopt a conformation having some overlap with the receptor, i.e. a soft docking technique, then gradually switches on the van der Waals terms of the potential function to induce minor conformational changes in the ligand and in the receptor to avoid steric clashes.

## Generation of Subensembles

Molecular dynamics (MD) or Monte Carlo (MC) simulations are probably the most rigorous way to generate a subensemble of states. The free energy values calculated using these methods are very reliable and often comparable with experimental data;<sup>33</sup> however, the comparison of different ligands using these methods is probably too slow and inefficient for rational drug design.

The issue of conformational sampling in standard MD simulations has been addressed by Philippopoulos and Lim,<sup>34</sup> who compared the results of a 1.7 ns MD simulation with an ensemble of 15 structures obtained by NMR. Conformational sampling for the MD simulation and the NMR structures was found to be consistent, but the MD simulation appeared to be less exhaustive than NMR data in sampling the conformational space, with the experimental structures showing more flexibility both in the side-chains and in the backbone atoms of the protein. The results of the MD simulation were also compared to two X-ray structures: a correlation between the thermal factors of the atoms and the MD and NMR data was found.

Simplified versions of MD simulations with an implicit solvation model have been implemented to overcome the problem of slowness and inefficiency in conformational sampling. For the same reason, much of the protein structure has often been held fixed and only the binding site and the protein ligands fully sampled.<sup>35</sup>

A different method aimed at improving the efficiency and speed of MD simulations without penalising the level of conformational sampling of the system and with the inclusion of explicit solvent was developed by Mangoni *et al.*<sup>36</sup> This group applied the Molecular Dynamics Docking (MDD) method. In this technique, the centre of mass motion of the ligand is separated from its internal and rotational motions and a separate coupling to different thermal baths for both types of motion of the ligand and for the motion of the receptor is applied. By varying the coupling to the baths it is possible to increase the kinetic energy of the centre of mass of the ligand (i.e. its translational motion) without increasing the temperature of the internal motions of the receptor, the ligand and the solvent (which are kept at room temperature). This allows a fast exploration of the receptor surface, even in solution and in the presence of full flexibility of the binding site. The presence of explicit water prevents the system

from assuming improbable conformations; different weights for the ligand-receptor or ligand-solvent interactions were found to be necessary to avoid the shielding of the ligand-receptor interaction by the explicit molecules of water.

Masukawa *et al.* generated multiple protein conformational states by running an MD simulation; these states were then kept fixed in a docking procedure and the resulting docked structures overlaid to identify the complementary binding regions conserved over many protein structures.<sup>37</sup> The complementary regions were then used to describe a “dynamic pharmacophore” model that was exploited to look for successful inhibitors in the Available Chemicals Directory (ACD). The same principle was also employed on multiple crystal protein structures;<sup>38</sup> the so-obtained dynamic pharmacophore model was again found to be more efficient than a single protein structure in ranking known protein inhibitors.

### 2.4.2 Methods Employing Experimentally Determined Structures

Kuntz and coworkers employed the multiple protein structures (MPS) technique for the first time, in 1997.<sup>39</sup> They generated interaction grids for both NMR and X-ray protein structures and averaged them in order to create composite grids for the DOCK program; the accuracy and the speed of the calculations were dramatically improved.

In an extension of the FlexX software a different approach has been developed to take into account protein flexibility.<sup>40</sup> Multiple crystallographic structures of a protein were employed and backbone and side-chain regions in good agreement after MPS superimposition averaged; the conformational space of the more flexible, disordered regions was instead sampled using known structures.

In a method based on GRID and Consensus Principle Component Analysis (GRID/

CPCA), MPS interaction grids were analysed to find similarities among the different structures, and flexible regions (i.e. regions for which MPS interactions grids are significantly different) instead related to structural characteristics of the proteins that may determine their specificity.<sup>41</sup>

One of the most recent approaches using MPS was made in 2002 by Osterberg *et al.*:<sup>42</sup> they used the AutoDock software on 21 structures of HIV-1 protease complexed with peptidomimetic inhibitors and systematically compared four different methods to combine the resulting interaction grids. The two optimal methods to produce weight-averaged grids were found to perform well for all the 21 inhibitors.

The potential benefits of combining the information from an entire protein family in the search of small-molecule drugs has been discussed.<sup>43</sup> The design of combinatorial libraries based on related protein structures and the application of these libraries across a family of related enzymes was proposed by several groups of researchers.<sup>44,45</sup> Essential folds and general folds appear in fact conserved across families of proteins and the flexibility of the receptor seems to be based on observed evolutionary differences. Broad-spectrum ligands could thus be designed by employing a set of homologous structures; otherwise, the observed differences and peculiarities could be very useful in order to find specific ligands.

## 2.5 Summary and Conclusion

Proteins are intrinsically dynamic molecules. Their flexibility is essential for their assembly, regulation, biological activity and catalysis.

In this chapter, protein motions have been reviewed and classified. They can be distinguished on the basis of their size (motions of subunits, domains, fragments and

side-chains) or on the basis of packing, in terms of shear motions (if they maintain a continuously tightly packed interface) and hinge motions (if the interface between the moving protein parts exist in one conformation but not in the other one). Individual shear motions are very small; the tight packing of protein internal interfaces constrain backbone torsion to only small deformations, and side-chain torsion angles to the same rotamer configuration (rotations up to about  $15^\circ$ ). For this reason, several shear motions must be summed to produce an appreciable effect. The whole shear motion is parallel to the plane of the interface. Hinges are normally located in protein regions that are relatively unconstrained by packing. A few large torsional changes in them are generally sufficient to produce almost the whole motion. The overall motion is perpendicular to the plane of the interface, which exists in one conformation (“closed” conformation) but not in the other (“open” conformation). Most domain and fragment motions can be classified as prevalently hinge or shear motions. For the motions of subunits, different categories apply; we can distinguish allosteric motions, non-allosteric motions, and complex large motions involving many submotions that can be classified as subunit or domain motions themselves.

Protein motions are very often observed when ligand binding occurs. As illustrated in the second section of the present chapter, the long-held “induced-fit” theory is now being overtaken by the “conformational selection” and “populations switch” models. Alternative structures for the same protein exist and are constantly interchanging in equilibrium conditions; ligands might increase the proportion of one of the subpopulations of conformers simply by shifting the equilibrium to the conformers they preferably bind. The lower the barriers between the different conformers at and around the bottom of the conformational energy landscape, the more flexible the molecule and the larger the number of conformations it can adopt. Energy funnels

characterised by an irregular bottom might determine broad range, non-specific binding; funnels characterised by a single, smooth minima, or by a few minima with high barriers between them, could imply rigid, more specific binding.

The fact that proteins can accommodate many molecules in their binding sites, rearranging themselves with relatively little penalty by either small torsional changes in their side-chains or large domain motions, is fundamental for rational drug design. Treating the receptor as rigid is not a valid option; it can only help to identify ligands for a particular narrow state of the protein conformational subensemble. In the last section of this chapter, some examples of many recent methods that accommodate protein flexibility in computational drug design were given. In general, they can be divided into methods that employ multiple experimental determined structures, and methods that use computer-generated conformations. Flexible docking, MPS methods, side-chain conformations generation by using rotamer libraries or MD and MC simulations are all methods that have been developed towards the end of the last decade and that are the subject of ongoing improvements and modifications.

A full understanding of protein flexibility in ligand binding, something essential to improve its accommodation in rational drug design, still has to be achieved. In particular, it is often difficult to distinguish conformational changes actually due to the formation of a complex between the protein and a specific ligand, conformational changes that instead just depend on the intrinsically mobile nature of proteins, and conformational states that are observed because of poor resolution structures, biased refinement methods procedures, or crystallisation conditions that determine a protein conformation that does not actually exist in solution.<sup>2</sup> Many groups of researchers have addressed this problem; the study of conformational changes observed in multiple proteins X-ray structures is the subject of the next chapter.

## Chapter 3

# Side-Chain Flexibility Analysis

---

### 3.1 Side-Chains and Protein Flexibility

Ligand binding may involve a wide range of conformational changes in proteins, from hinge movements of entire domains to small side-chain rearrangements in the binding site residues. However, changes in the backbone of proteins are sometimes negligible, while only side-chain reorientation occurs upon ligand binding;<sup>46</sup> also, side-chain conformational changes generally accompany all larger scale motions of protein structure, thus representing a kind of intrinsic protein mobility.<sup>4</sup> The knowledge of the extent and the characteristics with which side-chain rearrangements occur is consequently very important to improve docking prediction algorithms, and an index of amino-acid side chain flexibility would be potentially useful in molecular biology and protein engineering studies.

Unfortunately, there are several issues that limit reliable side-chain modelling.

For example, a vast number of possible combinations of side-chain conformations is possible. Also, experimental or computational data are generally insufficient to give a true picture of protein flexibility and motions. This is because X-ray structures are either “snapshots” or an average of the many conformations available to a protein, and MD simulations take a long time and are dependent on the force field. Finally, when analysing X-ray structures, it is often difficult to distinguish genuine protein conformational states and/or changes of conformation, with motions that are instead dependent on crystallographic refinement artefacts and/or conditions. Several attempts to overcome these problems have been made by different groups of researchers; some examples are described in the following sections. The analysis of available protein X-ray structures is at the basis of all the methods and studies described in this chapter.

## 3.2 Rotamer Libraries

### 3.2.1 Rotamers and Rotamer Libraries

Side-chain conformer predictions have an essential role in the modelling of protein structures and molecular docking; several widely used approaches are based on rotamer libraries. These consist of a list of discrete side-chains torsion angles and their associated probabilities, as determined from their frequency of occurrence in the Protein Data Bank (PDB).

Early work based on structural surveys and energy calculations<sup>47,48</sup> indicated that side-chain dihedral angles in proteins generally correspond to the potential energy minima of the isolated amino-acids. In recent years, following the increase of high resolution crystal structures in the PDB, few side-chains have in fact been observed



to significantly deviate from one of the isolated amino acid minima.<sup>49</sup> On one hand, this could be due to the use of rotamer preferences in modern refinement programs;<sup>50</sup> on the other hand, the weighting factors normally employed by refinement programs are too weak to dominate the experimental data in high-resolution structures.

The observation that side-chain  $\chi_1$  and  $\chi_2$  torsions fall into clusters in the  $\chi$  space and that therefore a library of rotamers can usefully be defined, was confirmed in 1987 by Ponder and Richards.<sup>49</sup> By using 19 crystal structures with resolution equal or better than 2.0 Å, they derived a side-chain rotamer-library in which the centres of the clustering in the  $\chi_1$  and  $\chi_2$  conformational space corresponded to relaxed states of the side-chains, i.e. states where atomic contacts were not made with backbone or other side-chains atoms. They found that most side-chains are limited to a small number of the many possible  $\chi_1, \chi_2$  minima. For example, while Leucine has nine possible  $\chi_1, \chi_2$  conformers, only two account for 88% of the residue's observations in the survey.

As the database has grown, several groups have since compiled updated rotamer libraries;<sup>51-53</sup> their availability, together with the concept of rotamers, has changed the handling of side-chains in homology modelling, Monte Carlo and combinatorial calculations, and rational drug design.

It is common practice to compile rotamer libraries using only high-resolution structures (most often with a cutoff at 2.0 Å) and to use all residues of a chosen structure unless missing atoms mean the  $\chi$  angles are undefined. To build their “Penultimate Rotamer Library”, Lovell *et al.*<sup>54</sup> used a resolution cutoff at 1.7 Å and omitted side-chains having alternative conformations, missing atoms, steric clashes and with any B-factors greater or equal to 40. In their opinion, the use of these local quality in-

dicators is crucial to exclude inaccurate rotamers from libraries. In particular, the B-factor cutoff was found to be the most powerful and simplest filter to remove poor quality rotamers from their library. All of the rotamers listed by Lovell *et al.* correspond to local energy minima; their observations suggest that significantly strained side-chain conformations are rare in proteins, and only occur for good reasons.<sup>54</sup>

$\chi_1$  rotamer preferences show detectable patterns as a function of the  $\psi$  and  $\phi$  backbone dihedral angles.<sup>50</sup> These preferences can be explained by a simple steric conformational analysis: certain rotameric states will be higher in energy because of steric interactions that cause the side-chain to twist out of the way of neighbouring atoms, inflicting a high dihedral energy on the residue. The steric interactions can be “backbone-independent”, i.e. not dependent on the local backbone conformation of the residue, or they can be “backbone-dependent”, i.e. dependent on the local backbone conformation. Steric analysis and theoretical calculations<sup>55</sup> show that the most common rotamers usually correspond to the lowest energy minimum given the local backbone conformation.

Dunbrack and Karplus<sup>50</sup> compiled a “Backbone-Dependent Rotamer Library” that gives the side-chain  $\chi_1$  rotamer distribution and the  $\chi_1$  average angles for each amino acid type and each  $10^\circ$  by  $10^\circ$  region of the  $\psi$ ,  $\phi$  conformational space of the backbone. The library was calculated from 132 protein chains in 126 PDB entries refined at a resolution equal to or better than 2.0 Å. While proteins sharing the same sequence have been included in the list, several groups of homologous proteins were included to increase the size of the dataset. Rotamer populations for each  $\chi$  torsion angle were calculated. For all side-chains except Ala, Pro and Gly, the  $\chi_1$  angular ranges used to define the rotameric state corresponded to the rotamers of the tetrahedral carbon atom: bins were created for angular values ranging from  $-120^\circ$  to  $0^\circ$  (*gauche*<sup>+</sup>

conformer),  $0^\circ$  to  $120^\circ$  (*gauche<sup>-</sup>* conformer),  $120^\circ$  to  $240^\circ$  (*trans* conformer). The same limits were used for the  $\chi_2$  dihedral angle of all the amino acids that have  $\chi_2$  except for proline, the aromatics, asparagine and aspartic acid. For proline, both  $\chi_1$  and  $\chi_2$  were placed into two bins comprising values greater than  $0^\circ$  and values less than  $0^\circ$ , i.e. corresponding to the two possible proline conformations *C $\gamma$ -exo* and *C $\gamma$ -endo*.  $\chi_2$  values of phenylalanine, tyrosine and histidine were divided into bins of  $0^\circ$  to  $60^\circ$ ,  $60^\circ$  to  $120^\circ$  and  $120^\circ$  to  $180^\circ$ . If  $\chi_2$  was less than  $0^\circ$ , a value equal to  $\chi_2+180^\circ$  was used, since most crystal structures do not distinguish if Phe, Tyr and sometimes His residues have a value of  $\chi_2$  or  $\chi_2+180^\circ$ . This is true also for Asp and Asn, for which  $\chi_2$  or  $\chi_2+180^\circ$  were treated as equivalent and the following limits employed:  $-90^\circ$  to  $-30^\circ$  (*gauche<sup>+</sup>* conformer),  $-30^\circ$  to  $30^\circ$  (*trans* conformer) and  $30^\circ$  to  $90^\circ$  (*gauche<sup>-</sup>* conformer). Trp was treated as either  $0^\circ$  to  $180^\circ$  or  $-180^\circ$  to  $0^\circ$ . For the  $\chi_3$  and  $\chi_4$  angles, (Lys, Arg, Glu, Gln), ranges analogous to the ones generally employed for  $\chi_1$  were used, with the exception of glutamate and glutamine  $\chi_3$ . In their case, the limits employed for aspartate and asparagine  $\chi_2$  were adopted.

The Dunbrack and Karplus<sup>50</sup> analysis was later completed by a Bayesian statistical analysis<sup>55</sup> of the backbone-dependent rotamer library. In Bayesian analysis, the *prior distribution* of data (obtained either from previous data or by pooling some of the present data) is combined with the data to form the *a posteriori distribution*, i.e. a compromise between the prior distribution and the data. This statistical analysis provides a better estimate of a parameter of interest than the data alone provide. In fact, the *a posteriori* distribution is more than a point estimate for a parameter; it is instead a probability distribution over the full range of allowed values of the parameter.

For the  $\chi_2$ ,  $\chi_3$  and  $\chi_4$  rotamer prior distributions, the probability of each ro-

tamer was assumed to be only dependent on the previous  $\chi$  angle in the chain; for the backbone-dependence of the  $\chi_1$  rotamers, prior distributions were derived from the product of the  $\psi$ -dependent and  $\phi$ -dependent probabilities. The “Conditional Backbone-Independent Rotamer Library” by Dunbrack and Cohen consists of probability distributions of  $\chi_2$ ,  $\chi_3$  and  $\chi_4$  rotamers conditional on the  $\chi_1$  rotamer, in contrast to traditional backbone-independent rotamer libraries. For all side-chains except Asn and Asp, there is little or no dependence of  $\chi_2$  on  $\psi$  and  $\phi$ ; rotamers for all side-chain types except Asn and Asp are thus assumed to be backbone-independent, i.e. dependent on only the identity of the  $\chi_1$  rotamer. In the case of Asp and Asn, the  $\chi_2$  rotamer probabilities for each  $r_1$  rotamer type change instead dramatically with  $\phi$  and  $\psi$ ; these are thus included in the  $\chi_1$  and  $\chi_2$  population analysis.

### 3.2.2 Limitations of the Rotamer Approximation

The availability of rotamer libraries and the concept of rotamers have changed the handling of side-chains in homology modelling, Monte Carlo, combinatorial calculations, molecular docking and drug design.

However, despite the enormous advantages that derive from the rotamer approximation (most of all in the speed and efficiency of side-chain placement algorithms), one should not forget that it suffers from several drawbacks. On one hand, a rotamer represents a rigid conformation of the side chain whereas, in a real protein, side-chains are flexible, and their potential energy is a dynamic quantity that depends on the instantaneous conformation of a side-chain and on the instantaneous conformation of all the other side-chains.<sup>56</sup> On the other hand, even if the “right” rotamer for a given side-chain (i.e. a rotamer belonging to the same local potential energy minimum as the corresponding true average side-chain conformation) exists in a rotamer library,

this rotamer might deviate substantially from the true average side-chain conformation. Because of this, even if the side-chain is modelled with the correct rotamer, it will be affected by errors in its atomic coordinates.<sup>56</sup>

It was to overcome these problems that Mendes *et al.* developed a method based on a flexible rotamer model (FRM).<sup>56</sup> In FRM a rotamer is not a single, rigid conformation of a side-chain but a continuous ensemble of conformations of which the classic rigid rotamer is the average.

Another issue regarding the limitations of the rotamer model concerns the “non-rotameric” side-chains that are found in proteins, i.e. residues whose  $\chi_1$  or  $\chi_2$  deviate significantly from the nearest associated rotamer.

Schrauber *et al.* (1993)<sup>51</sup> compiled a rotamer library investigating how the number of side-chain conformations in a set of protein structures could be assigned to rotamers in this library. They found that their library, together with the one previously compiled by Ponder and Richards (1987),<sup>49</sup> covered only 90% of the observed  $\chi_1$  and 80% of the conformations when considering both  $\chi_1$  and  $\chi_2$ . Applying a resolution threshold equal to 2.0 Å, they observed an improvement in the “rotamericity”, i.e. in the percentage of residues that are found to be rotameric within the structures. Similarly, improvements in the rotamericity as resolution improves were observed by Carugo and Argos (1997)<sup>57</sup> and Heringa and Argos (1999).<sup>58,59</sup>

Heringa and Argos studied the relative and spatial positioning of “nonrotameric” side-chains, i.e. side-chains with atypical and strained dihedral angles in well-refined protein tertiary structures. They confined the analysis to buried protein cores, that are believed to be less subject to erroneous side-chain positioning, defining residues as buried if they had a solvent accessible surface area (ASA) equal or less than 5

Å<sup>2</sup>.<sup>58</sup> Their analysis revealed that 50% of the proteins with two or more nonrotameric residues (defined by them as residues whose  $\chi_1$  or  $\chi_2$  deviated more than 20° from the nearest associated rotamer) displayed clusters of two or more (up to five) nonrotameric residues. These clusters showed lower crystallographic temperature factors than isolated nonrotameric residues; according to Heringa and Argos,<sup>58</sup> this characteristic could suggest that spatially concentrated strain in protein folds could be minimized by lowered vibrational energy. Nonrotameric clusters also seemed to present tighter packing than rotameric ones, to prefer coils regions rather than helices and strands and to be preferentially composed by residues such as histidine, arginine, tyrosine and, to a lesser extent, leucine. The relation between the spatial positioning of nonrotameric residues and ligands was also studied:<sup>59</sup> clusters of nonrotameric side-chains were found to be associated with ligand binding sites. The authors compared 20 apo-protein structures (i.e. uncomplexed forms) to the corresponding holo-proteins (i.e. complexed forms); they suggested that ligand binding induced nonrotameric states, and thus strain, in the holo-protein forms.<sup>59</sup>

A study of the steric strain in the backbone of proteins, that is evident in the portions of the main-chain that are in disallowed regions of the Ramachandran map, had already been carried out in 1991 by Herzeberg and Moulton.<sup>60</sup> Later, Schrauber *et al.*<sup>51</sup> suggested that the interiors of proteins are not relaxed; strain exists, especially in certain regions of the protein, and they hypothesized that this could contribute to protein movements in solution. Similarly, Heringa and Argos<sup>59</sup> observed that the increased strain in ligand-protein complexes does not seem to be compensated by augmented hydrogen bonding or salt bridge formation involving the side-chains that become nonrotameric in the complexed form; in their opinion this internal energy

gain might help the formation and the ejection of enzymatic products, representing a kind of dynamic “rechargeable battery”<sup>59</sup> that could promote protein motion and activity. However, their survey was limited to 20 holo-/apo- protein pairs, for which some contradictory evidence was observed, and a totally different point of view was supported by Petrella and Karplus (2001).<sup>61</sup> Their study was aimed at comparing the energetics of rotameric and nonrotameric residues, to investigate whether non-rotameric side-chains commonly observed in crystal structures are real or artefacts. They predicted side-chain orientations for 24 proteins for which structures with a resolution equal or better than 2.0 Å existed, using the CHARMM energy function. They also calculated the probability of nonrotameric side-chains by performing an umbrella-sampling molecular dynamic simulation of isolated amino acids, in vacuum and in an average protein and aqueous dielectric environment. Their conclusion was that, while most of the nonrotameric conformations are real, many are likely to depend on artefacts of the X-ray refinement. This hypothesis is consistent with the statistical results obtained by Thornton and MacArthur,<sup>62</sup> who observed a systematic variation of the mean values of  $\chi_1$  rotamers with resolution. This correlation was more significant for some residue types (e.g. Ser, Thr, Leu, Lys) and absent for others (e.g. the aromatics). B-factor values appeared to be higher than average for the residues with the strongest correlation trend. Thornton and MacArthur suggested that this observation could depend on the existence of multiple rotameric states; the averaged electron density produced by a dual occupancy at low resolution could instead be resolved into its individual components at higher resolutions.<sup>62</sup>

## 3.3 Analysis of Conformational Changes Occuring Upon Ligand Binding

### 3.3.1 Najmanovich *et al.*

In a recent study by Najmanovich *et al.*,<sup>46</sup> two non-redundant databases of paired protein structures in complexed and uncomplexed forms from the PDB database were constructed (980 and 353 entries respectively). The aim of the work was to analyze side-chain rearrangements upon ligand binding; the number and identity of binding pocket residues that undergo side-chain conformational changes were determined. A PDB entry was considered the apo-protein of a given holo-protein if their amino acid sequences were identical and if none of the binding pocket residues of the apo-protein were in contact with another ligand not present in the holo-protein. A resolution equal or better than 2.5 Å was prerequisite. Side-chain dihedral angles for binding pocket residues in both holo- and apo-proteins were compared: a threshold of 60° of at least one torsional angle was chosen to define a conformational change. The group chose this angular value after performing the analyses at three different threshold values; the trends observed using 45°, 60° and 75° thresholds appeared to be all similar, and the differences in the amount of conformational changes not so pronounced. Also, the probability for a certain residue type to change side-chain conformation upon ligand binding was insensitive to the variation of the threshold.

In general, the study showed that only a small number of residues in the binding pocket undergo significant conformational changes; e.g. 85% of cases show changes in three residues or less.<sup>46</sup> The resulting flexibility scale had the following order: Lys > Arg, Gln, Met > Glu, Ile, Leu > Asn, Thr, Val, Tyr, Ser, His, Asp > Cys, Trp, Phe, which means that, for instance, Lys side chains in binding pockets change



conformation 25 times more often than the Phe side chains. Even though normalizing for the number of flexible dihedral angles in each amino acid attenuated this scale, the trend of large, polar amino acids being more flexible in the pocket than aromatic ones remained.

In general, 94% of  $\chi_1$  and 95% of  $\chi_2$  torsional angles did not undergo conformational change (a more detailed analysis for each residue showed larger differences in the averages). The parameter used to estimate the extent of backbone movements was the maximum displacement of the  $C\alpha$  atoms in the apo- and holo-protein entries:

$$\Delta d_{max} = \max_{\langle i,j \rangle} |d_{ij}^{apo} - d_{ij}^{holo}| \quad (3.1)$$

In the previous formula,  $\langle i,j \rangle$  denotes all pairwise combinations of  $C\alpha$  atoms from residues in contact with the ligand. 75% of the cases showed a  $\Delta d_{max}$  of less than 1 Å, and only 12% had backbone displacements larger than 2 Å; the authors concluded that backbone displacements in binding pockets are on average less important than side-chain flexibility. No correlation between backbone movement of a residue and the flexibility of its side chain was found in the study: in the few cases where very large backbone displacements occurred (i.e.  $\Delta d_{max} > 18$  Å), the fraction of residues undergoing side-chain conformational changes was not larger than average. This suggested that the flexibility of side-chains in pockets subject to very large motions does not differ from that of side-chains involved in smaller backbone displacement.<sup>46</sup>

Finally, the authors analysed pairs of apo-protein entries sharing the same amino acid sequence; changes of more than 60° were rarer among apo-/apo- compared to apo-/holo- protein pairs; the flexibility scale for the two surveys showed however the same trend, suggesting that this is probably an inherent property of amino acids.

### 3.3.2 Fradera *et al.*

More recently, Fradera *et al.*<sup>63</sup> analysed and compared the binding sites of 8 different proteins for which different high-resolution structures with different ligands exist. They calculated Root Mean Square deviation (RMSd), cavities' volumes and classical Molecular Interaction Potentials (cMIP)<sup>64</sup> of protein binding sites in order to quantify ligand-induced changes.

80 structures of 8 different proteins (lysozyme, dethiobiotin synthetase, cytochrome P-450 CAM, papain, trypsin, D-xylose isomerase, chymotrypsin, thymidine kinase) were chosen; for 6 of the 8 protein families, an apo-structure was included in the dataset. Only one monomer for dimeric proteins was randomly selected; a resolution cutoff "equal or around to 2.0 Å"<sup>63</sup> was applied. The structures belonging to the same protein family generally shared the same sequence; mutations were accepted only if far from the binding site. Binding sites were defined as the residues having at least one atom less than 4 Å from any atom of all ligands in the protein set. Complexes were oriented by superimposing the backbones of the protein binding sites.

RMSd calculations were performed on all heavy atoms, on only backbones and on only side-chains of the proteins, always using the backbones as the reference in fitting.

Cavities at the binding sites were computed using SURFNET<sup>65</sup> after removing the ligands; the shape of the binding site cavity was numerically compared using grids identically defined in all proteins of a family. Each grid point was assigned to 1 (point inside cavity) or 0 (point outside cavity or further than 4 Å from any binding site's atoms); the 3D matrices defined by the grids were then used to calculate the accessible volumes of each structure. The absolute and relative changes induced by

ligand binding in the accessible volume of the binding site cavity were elaborated and similarity indices obtained for pairs of compared proteins; the indices provided information regarding not only the volume but also the shape of the cavities.

cMIP were calculated using three chemical groups (an  $sp^3$  aliphatic carbon, a positive oxygen and a negative oxygen) placed in a grid that covered all the analysed binding sites. Spearman coefficients were then calculated for all the pairs of structures; each probe and cross-correlation matrix describing the degree of similarity between all possible pairs of proteins were finally examined with Principal Component Analysis (PCA).<sup>63</sup>

In all cases, more than 95% of the variance could be explained by the first and second principal components, that were then used to cluster the binding sites in terms of reactivity.

The RMSd of binding sites' backbone was found to be in general fairly small; significant deviations are represented by lysozyme and dethiobiotin synthase (whose backbone movements, even if less than 1 Å, are not negligible) and D-xylose isomerase, whose backbone is exceptionally rigid.

Regarding the changes in binding site cavities, a strong variability among proteins was found. For example, the relative change in the accessible volume of the binding site was less than 10% for D-xylose isomerase and greater than 60% for thymidine kinase. Significant variability was also found in the shape of binding sites. Great flexibility and capacity to adapt (i.e. a small similarity index of binding sites) was found for thymidine kinase, while great similarity of binding sites (i.e. rigidity) observed for D-xylose reductase. The authors concluded that small RMSd changes can still introduce important variations in the volume and shape of protein binding site

pockets.

The analysis of clusters produced by plotting the first and the second components obtained by PCA and cMIP analyses revealed two different kinds of situations:

1. families of structures not clearly clustered (lysozyme, thymidine kinase, papsin and chymotrypsin);
2. families where most structures are found in one cluster (dethiobiotin synthetase, trypsin, D-xylose isomerase, cytochrome P-450 CAM).

In the protein families belonging to the first case, different possible situations are observed: lysozyme is for example characterised by a total random distribution in the PCA plot (i.e. by a large flexibility of the binding site), while thymidine kinase can be solved into three different subfamilies that correspond to three different classes of ligands. However, the recognition differences of the first case did not seem to depend clearly on the presence or absence of the ligands. For example, lysozyme's similarity indexes were found to be totally similar for unbound/bound form comparisons and for bound/bound form comparisons.

Regarding protein families where most structures are found in one cluster and only a few outliers detected, the binding sites probably have a favourite conformation that is still subject to changes, perhaps in the presence of certain conditions. This, for example, is the case of trypsin: the only outlier present in its plot is generated by the rotation of the side-chain of Gln 192. Similarly, the different side-chain orientations of Tyr 193 and Phe 96 in cytochrome P-450 CAM determine the unique characteristics of one of its outlier structures, while in the case of D-xylose isomerase some small side-chain movements of several polar residues (His, Glu and Asp) determine the peculiarities of two isolated protein structures. In the case of dethiobiotin synthase,

an outlier corresponding to the unbound structure is determined both by side-chain rearrangements and backbone conformational changes.

Only in the case of dethiobiotin synthase does an outlier corresponds to an unbound protein. In the opinion of the authors, the fact that the largest differences in the recognition properties appears in the binding sites of complexed proteins suggests that the analysed proteins do not follow a two-step induced-fit mechanism but, rather, a conformation-selection model, in which the inherent flexibility of binding sites allow different molecules to fit. Similarly, the conformational changes of side-chains and backbones were not necessarily larger when unbound and bound structures were compared.

The authors of this survey<sup>63</sup> concluded that proteins are more conservative in terms of electrostatic distribution characteristics rather than in steric properties. In their opinion, the results of their analysis could support the fact that electrostatic properties are the main reason determining differential binding in proteins. In a recent review, Teague<sup>20</sup> observed that flexible receptors can often bind to structurally diverse ligands that occupy almost the same volume within a binding site, while a broad range of different residues' orientations is common.

Fradera *et al.*<sup>63</sup> also judged the cross-correlations of MIP found between holo-structures bound to different ligands to be poor, even when structures are solved by the same authors and in the same conditions; this highlights the potential dangers of using a single protein structure in molecular docking. In the authors' opinion, these risks could be overcome by choosing a representative protein structure with the help of cMIP analysis and PCA-Spearman's cross correlation matrices.<sup>63</sup>

## 3.4 Study of the Inherent Flexibility of Protein Side-Chains

### 3.4.1 Zhao *et al.*

Zhao and coworkers<sup>66</sup> compiled and analyzed a data set of 123 paired protein structures for which multiple high-quality uncomplexed atomic structures were available. Side-chain flexibility was quantified by evaluating the changes of torsional angles, yielding a set of residue- and environment- specific confidence levels which describe the range of motion around  $\chi_1$  and  $\chi_2$  angles. The data-set was restricted to high-quality structures, and the effects of the selection criteria on the results were tested: the least stringent resolution cutoff that yielded a consistent set of final confidence levels was 2.2 Å. Moreover, in order to choose structures with a relatively rigid backbone conformation, an iterative superimposition process was performed; its goal was to match the rigid core of the proteins and to omit their flexible loops or terminal extensions. All the backbone atoms for each pair of proteins were superimposed, and those with a difference greater than 1 Å excluded. A new backbone superimposition was then performed and the process repeated until convergence. Finally, a more stringent cutoff of 0.5 Å on the root-mean-square-deviation (RMSD) was applied, and protein pairs showing large flexible regions omitted from the analysis.

For each group of paired proteins, each amino acid position in the chain was compared to the identical position in the matched chains, and the results for each amino acid type represented in a diagonal plot. In this graph, the  $\chi_1$  angle of the given residue in two different structures was reported along the two perpendicular axis: if the amino acid adopted the same conformation in two structures, the corresponding point resulted along the diagonal. If the residue adopted two different conformations in the two structure analysis, it fell off the diagonal. Residues completely buried

within the protein (represented below the diagonal of the graphs) were distinguished from the solvent exposed ones (points above).

Most points were arrayed close to the diagonal in these graphs, showing that there is typically no change in the  $\chi_1$  torsion angle of a given residue from structure to structure. Also, they clustered into the three regions corresponding to staggered conformations, just as expected from previous rotamer studies.<sup>49,50</sup> The off-diagonal points, i.e. the points which change conformation from structure to structure, jumping to a different staggered conformation, also clustered in the regions of staggered conformations.

As one might expect, the large aromatics clustered closely along the diagonal, showing that they are packed into well-defined niches in proteins. Polar and charged residues however showed much variability, and while large buried aromatic residues appeared to be well locked within the packed core of the protein, the smaller aliphatic amino acids surprisingly showed large motions also when buried. For example, the Val plot is nearly symmetrical, reflecting a similar flexibility inside and outside the protein. Ser was instead markedly inflexible inside proteins, presumably because of its ability to form structural hydrogen bonds with the neighbouring portions of the chain. The flexibility of each amino acid type (taking into account also the smaller level of motion we observe when a side-chain adopts a slightly different position but stays within a given staggered conformation) was quantified as the angular thresholds that contained 90% of the observed structures.

In general, the obtained thresholds reflect the common-sense ideas about side-chain flexibility: buried residues are far less mobile than surface residues, and while Lys and Arg required a large angular tolerance to cover 90% of the observations, the large aromatic amino acids are less flexible, with most moving less than  $10^\circ$  between

the different structures of a given protein. When looking at all environment values, we can see the expected order of flexibility: Ser > Lys, Glu, Gln, Arg, Met > Leu, Asn, Asp, Val, Thr > Ile, His, Cys, Trp, Tyr, Phe. The large aromatics (Phe, Tyr and Trp) showed limited flexibility both when exposed and when buried. Surprisingly, Leu showed the largest amount of motion when buried ( $19.6^\circ$ ); we might expect the small aliphatic amino acids Ile, Val, and Leu to have the greatest mobility when buried since they lack of specific hydrogen-bonding or salt-bridge interactions with neighbouring parts of the protein. The fact that the largest of these three has the greatest flexibility might be a result of steric hindrance in the  $\beta$ -branched structure of Val and Ile.

The exposed residues were not seen as uniformly flexible. The large polar or charged residues (Arg, Lys, Glu and Gln) were found to be very flexible, but many of the smaller polar and charged residues appeared to be markedly inflexible; for example, Asn and Asp showed about  $15^\circ$  of flexibility, and His only  $11^\circ$ . Exposed Ser was the most flexible residue at  $102.7^\circ$ , while Thr ( $13.5^\circ$ ) could reflect the effect of  $\beta$ -branching.

A similar analysis was performed for  $\chi_2$  angles, and the differences in  $\chi_2$  in each paired residue were compared and plotted with the differences in  $\chi_1$ ; the distribution of points in these plots showed a random distribution, with no correlation between  $\chi_2$  and  $\chi_1$ .

The rationale of this research was that, while most analysis of side-chain conformations study the range of motion available to a given residue type, they do not analyze the flexibility of a given residue within a given protein environment. Steric contacts with the local peptide backbone and interactions with neighboring parts of the protein and solvent limit the intrinsic flexibility of each residue.



Since the proteins being compared were identical and did not contain bound ligands, the authors assumed that the differences between the paired structures are a result of the intrinsic flexibility under the influence of the different environments in each crystal determination. However, one must be aware that artefacts can arise from the process of crystallographic structure solution, such as the use of rotamers during the refinement of undetermined regions, and that the results could represent only a few snapshots of the range of flexibility that might be present in solution.

The authors suggest that the confidence levels and angular tolerances obtained by their analysis could provide better metrics to score any side-chain discrimination methods, since they depend both on the residue type and its local environment. The standard threshold of  $40^\circ$  for the difference in  $\chi_1$  angles, which is widely used to define correct and incorrect prediction of side chain position,<sup>67</sup> could be inappropriate in most circumstances. In fact, according to this study for most buried residues this tolerance is far too generous, while for some surface residues it is too restrictive. Rotamer analysis derived from surveys of databases that include a single structure for each protein possibly do not give an accurate estimate of the flexibility of residues within proteins, since they provide an estimate of the range of conformations that are available to a given side-chain when observed in all environments. The effects of the protein environment tend to average out, and the distribution of rotamers is dominated by the steric contacts between the side-chain and the protein main chain. The normal range of motion available to a given amino acid residue, when placed within a protein, is instead determined by a combination of short-range steric forces, along with the interactions with neighbouring portions of the chain and the surrounding environment. An analysis of the motions in individual residues from the paired-protein database can instead quantify the flexibility of a residue within its

normal protein environment.

The confidence levels are more restrictive than an estimate derived from the spread of points in a rotamer analysis, indicating that the motion of a given residue within a protein is more restricted than expected from a comparison of different residues in the chain. This is evident from the diagonal plots, presented in the article, which report on the x and y axis the  $\chi_1$  angle of a given residue found in two different structures of the same protein. The data points along the diagonal on the graphs cluster into elliptical rather than circular regions, with a greater spread along the diagonal (representative of the range of conformations across many proteins) than the spread perpendicular to the diagonal (representative of the range of conformation of a given residue position in a given protein).

## 3.5 Summary and Conclusion

Side-chain flexibility and motions play a crucial role in ligand binding. In this section, some studies and methods aimed at analysing and describing them have been presented.

The prediction of side-chain conformations in homology modelling, rational drug design and molecular docking is often based on rotamer libraries. Rotamers are approximate representations of the configuration of the torsion angles within an amino acid side-chain; they are usually derived from statistical analysis of experimental structures, and they generally correspond to local minima on the side-chains' potential energy maps. However, rotamer libraries derived from large datasets may often include very rare conformations, most of which are probably subject to a large dihedral energy strain. Some of these “nonrotameric” conformations could be real, but others are likely to depend on artefacts of the X-ray refinement. This hypothesis is

consistent with analyses that show a systematic variation of the mean values of  $\chi_1$  rotamers with resolution.<sup>62</sup>

Dunbrack and Karplus have compiled a “backbone-dependent” rotamer library where the distribution of the  $\chi_1$  rotamer is expressed as a function of the  $\psi$  and  $\phi$  torsion angles.<sup>50</sup> The  $\chi_1$  rotamer distribution shows in fact detectable patterns as a function of the  $\psi$  and  $\phi$  backbone dihedral angles; these preferences can be explained by a simple steric conformational analysis.

Despite some limitations that intrinsically affect the concept of rotamers (e.g. the representation of a side-chain as a single, rigid conformation, the dependency on the quality and accuracy of experimental data from which they are derived), rotamer libraries are still commonly employed to reduce the search space and successfully predict side-chain positions.<sup>68,68</sup>

To analyze side-chain rearrangements upon ligand binding, Najmanovich *et al.*<sup>46</sup> compared paired protein structures in complexed and uncomplexed forms from the PDB database. They determined the number and identity of binding pocket residues that undergo side-chain conformational changes using a threshold of 60° in at least one torsional angle to define a conformational change.

The study showed that only a small number of residues in the pocket undergo significant conformational changes. The flexibility scale had the following order: Lys > Arg, Gln, Met > Glu, Ile, Leu > Asn, Thr, Val, Tyr, Ser, His, Asp > Cys, Trp, Phe. Even when normalizing by the number of flexible dihedral bonds in each amino acid, the trend of large, polar amino acids having a greater flexibility than aromatics remained. The flexibility of side-chains in pockets subject to very large backbone motions did not differ significantly from that of side-chains involved in smaller backbone displacements.<sup>46</sup> The comparisons of pairs of apo-/apo- protein entries revealed rarer

conformational changes than apo-/holo- protein pairs; the flexibility scale for the different residue types was however the same, probably being an intrinsic property of amino acids.<sup>46</sup>

A more recent study from Fradera *et al.*<sup>63</sup> was also aimed at comparing and quantifying residues' conformational changes in the binding site of apo- and holo-proteins sharing the same sequences. They chose 80 protein structures belonging to 8 different protein families and calculated the RMSd, the volumes and the classical Molecular Interaction Potentials of their binding sites.

The RMSd of the binding sites' backbone was found to be in general fairly small, even if some significant variations among the different protein families were observed. Regarding binding-site cavities' volumes and shapes, a strong variability among the different protein families was found; the authors concluded that small RMSd changes can still introduce important variations in the volume and shape of protein binding site pockets.<sup>63</sup>

The analysis of clusters produced by plotting the first and the second components obtained by PCA and cMIP analyses revealed two different kinds of situations: families of structures that were not clearly clustered, and families where most structures were found in one cluster. The recognition differences revealed in the first case did not seem to necessarily depend on the presence or absence of the ligands; similar behaviour for unbound/bound comparisons and for bound/bound comparisons were found. Protein families where most structures are found on one cluster and only a few outliers are detected probably have binding sites with a favourite conformation that is subject to changes, perhaps in the presence of specific stimuli. The peculiarity of the outliers' structures in the graphs could generally be explained by some different side-chain orientations in the binding site.

The poor cross-correlations found between holo-structures highlights the potential risks of using a single protein structure in molecular docking.

Distinguishing their analysis from those of Najmanovich and co-workers<sup>46</sup> and Fradera *et al.*,<sup>63</sup> Zhao *et al.*<sup>66</sup> focused their attention on uncomplexed protein structures. They analysed a dataset of paired protein structures for which multiple high-quality uncomplexed atomic structures exist; side-chain flexibility was evaluated and a set of residue- and environment- specific confidence levels that contained 90% of the  $\chi_1$  and  $\chi_2$  observed flexibility obtained. In doing so, Zhao *et al.* hoped to determine the intrinsic flexibility of the amino acid side-chain in protein crystal structures since, in principle, the pairs of protein structures compared should be identical.

Compared to the large aromatics, polar and charged residues showed much variability, and while large buried aromatic residues appeared to be quite rigid in the core of the protein, the smaller aliphatic amino acids surprisingly showed large motions also when buried. Ser was markedly inflexible inside proteins, presumably because of its ability to form structural hydrogen bonds with the neighbouring portions of the chain, but was instead the most flexible residue when exposed. As we might expect, buried residues were less mobile than surface residues, and polar residues such as Lys and Arg required a large angular tolerance to cover 90% of the observations. The small aliphatic amino acids showed the greatest mobility when buried, probably because of the lack of specific hydrogen-bonding or salt-bridge interactions within the protein; many of the smaller polar and charged residues (e.g. Asp and Asn) appeared to be markedly inflexible.

Since the Zhao *et al.* study compared pairs of apo-proteins sharing the same sequence, i.e. the intrinsic flexibility of amino acid side-chains in the same environment, the angular thresholds obtained by this research should reflect steric contacts with

the peptide backbone and interactions with neighbouring parts of the protein. Rotamer analysis derived from databases that include a single structure for each protein estimate instead the range of conformations available to a certain residue type in all the environments; as highlighted by the diagonal plots presented in their survey, the confidence levels obtained by Zhao *et al.* are more restrictive, and probably more reliable, than an estimate derived from a rotamer analysis.<sup>66</sup>

### 3.6 Aim of this Project

As described in the last two chapters, several kinds of flexibility are observed in experimentally determined structures of protein-ligand complexes. In most cases only a few side-chains vary in the active site; however, rearrangements of single loops or other secondary structure elements and large motions involving entire domains are also commonly observed.

Side-chain flexibility is essential for molecular docking purposes. The rearrangement of a single side-chain can have considerable consequences in molecular recognition: for example, the shift of a tyrosine's side-chain from a *gauche*<sup>-</sup> to a *trans* conformation implies a motion of about 9 Å of its hydroxyl group.<sup>68</sup> Also, the inclusion of side-chain flexibility in rational drug design could compensate for the uncertainty of the available protein models. It has been estimated that the standard deviation of refined atomic positions at 2.0 Å resolution is in the range of 0.15 Å to 0.25 Å, implying a precision of dihedral angle determination of better than 10°.<sup>69</sup> However, the constraints and the method of refinement, and the connectivity of the involved atoms can largely affect the accuracy of atom positions.<sup>70</sup> Main chain atoms have lower positional errors than side-chain atoms bound to two non-hydrogens neighbours; these, in turn, have lower errors than side-chain atoms with one non-hydrogen neigh-

bour, and side-chain atoms with only one non-hydrogen neighbour have the greatest uncertainty of all.<sup>70</sup>

The aim of this project is to analyse and obtain a better understanding of side-chain conformational changes, particularly under the influence of a ligand. The ultimate goal is to determine whether it is possible, using knowledge derived from the PDB, to predict the conformation of a particular side-chain on binding by a given ligand. Multiple PDB structures of different proteins were analysed to identify and distinguish genuine ligand-binding induced conformational changes, motions that depend on the intrinsic flexibility of proteins, and possible artefacts arising from crystal structure refinement methods.

Najmanovich *et al.* 60° angular cutoff,<sup>46</sup> Zhao *et al.* residue type and environment specific thresholds,<sup>66</sup> and the Dunbrack and Karplus backbone dependent rotamer library<sup>50</sup> were employed to define conformational changes. The results obtained on applying these different methodologies were compared and assessed, and different conformational behaviours and trends in various protein families characterised.

## Chapter 4

# Conformational Analysis: Data Set, Methods and Results Overview

---

### 4.1 Choice of the Data Set

To analyse the characteristics and the extent of side-chain motion in proteins, a data set of ten proteins was chosen for which multiple PDB apo- and holo- protein structures at high resolution exist. The 10 selected proteins are glutathione S-transferase, HIV-1 protease, carbonic anhydrase II, thrombin, cytochrome P-450 CAM, streptavidin, trypsin, D-xylose isomerase, ribonuclease A and endothiapepsin.

These proteins were mainly chosen for the relatively high number of crystallographic structures at good resolution that are deposited in the PDB; other extensively studied protein systems, such as HIV-1 reverse transcriptase or neuraminidase, do not offer a similarly large number of PDB entries at low resolution. Also, the 10 protein systems were chosen since they are representative of a quite heterogeneous ensemble



of different protein families. HIV-1 protease and endothiapepsin are in fact aspartic proteases, thrombin and trypsin serine proteases; carbonic anhydrase is a lyase, cytochrome P-450 CAM an oxidoreductase. D-xylose isomerase is an intramolecular oxidoreductase, ribonuclease A is a hydrolase and glutathione S-transferase a transferase. Streptavidin is a biotin binding protein, very well known for its exceptionally strong affinity for biotin.

FASTA searches<sup>71</sup> were performed to identify PDB entries sharing an identical amino acid sequence for each of the 10 protein systems; entries with a sequence identity less than 100% (minimum sequence identity 97.7%) were occasionally accepted if mutations occurred far from the binding site and did not appear to interfere with protein function and/or folding.

Residues including one or more PDB atoms with an occupancy less than 1 were discarded. A 2.5 Å resolution cutoff was generally applied; a subset of structures having a resolution equal to or better than 2.0 Å was obtained from the larger dataset and analysed separately.

All holo-protein structures were inspected to identify and consider as ligands only the molecules that are actually bound within the protein active site. Water molecules and salt ions were excluded, with the only exception of anions that bind the catalytic zinc atom of carbonic anhydrase. Peptidic ligands, especially found in the case of proteases, were included in the survey.

If multiple entries sharing the same ligand were found within the same protein system, only that at best resolution was included in the main datasets. If several entries sharing the same ligand at the same resolution were found, only the first in alphanumerical order was arbitrarily considered. The holo-proteins discarded in this way were put set-aside and kept for eventual analyses at a second stage, to characterise

side-chain conformational changes occurring in the presence of the same ligand and to get a feel for the extent of “random”, non systematic conformational changes.

The selected PDB entries are listed in Table 4.1. Structures chosen for glutathione S-transferase, HIV-1 protease, carbonic anhydrase II and thrombin are human. The species of origin for trypsin and ribonuclease PDB entries is bovine, while cytochrome P-450 CAM is bacterial (*pseudomonas putida*), along with those of streptavidin (*streptomyces avidinii*) and D-xylose isomerase (*streptomyces rubiginosus*). The selected structures of endothiapepsin are instead derived from the chestnut blight fungus (*endothia parasitica*).

In this thesis, only data relative to structures solved at 2.0 Å or better will be generally presented. The comparison of these results with those obtained with a different resolution cutoff (2.5 Å) is described at the end of the present chapter (section 4.3.7).

## 4.2 Methods

Within a protein system, all possible apo-/apo-, apo-/holo- and holo-/holo- protein pairs were compared to characterise side-chain flexibility and backbone atoms' Root Mean Square Deviation in:

1. different PDB structures of the same uncomplexed protein (apo-/apo- protein comparisons);
2. complexed and uncomplexed forms of the same protein (apo-/holo- protein comparisons);
3. the same protein complexed with the same ligand and/or with different ligands (holo-/holo- protein comparisons).

|                              |                        | Resolution $\leq 2.0$ Å |      |               | Resolution $\in [2.0, 2.5)$ Å |              |      |               |  |
|------------------------------|------------------------|-------------------------|------|---------------|-------------------------------|--------------|------|---------------|--|
|                              |                        | Apo-Proteins            |      | Holo-Proteins |                               | Apo-Proteins |      | Holo-Proteins |  |
| Glutathione<br>S-Transferase | 16GS                   |                         | 13GS | 18GS          |                               |              | 10GS | 12GS          |  |
|                              |                        |                         | 1AQV | 1AQW          |                               |              | 20GS | 3PGT          |  |
|                              |                        |                         | 21GS | 2GSS          |                               |              |      |               |  |
|                              |                        |                         | 3GSS | 6GSS          |                               |              |      |               |  |
|                              |                        |                         | 9GSS |               |                               |              |      |               |  |
| Carbonic<br>Anhydrase II     | 1CA2 2CBA<br>2CBB 2CBE | 1AVN                    | 1BCD | 1CA3          | 1HCA                          | 1A42         | 1AM6 |               |  |
|                              |                        | 1BV3                    | 1CAH | 4CA2          | 4CAC                          | 1BN1         | 1BN3 |               |  |
|                              |                        | 1CAN                    | 1CIL |               |                               | 1BN4         | 1BNN |               |  |
|                              |                        | 1CNW                    | 1CNX |               |                               | 1BNQ         | 1BNT |               |  |
|                              |                        | 1CRA                    | 1F2W |               |                               | 1BNU         | 1BNV |               |  |
|                              |                        | 1G1D                    | 1G52 |               |                               | 1BNW         | 1CAH |               |  |
|                              |                        | 1G53                    | 1G54 |               |                               | 1CAY         | 1CIM |               |  |
|                              |                        | 1I8Z                    | 1I90 |               |                               | 1CIN         | 1CNY |               |  |
|                              |                        | 1I91                    | 1IF4 |               |                               | 1EOU         | 1IF6 |               |  |
|                              |                        | 1IF5                    | 1IF7 |               |                               | 1KWR         | 1OKL |               |  |
|                              |                        | 1IF8                    | 1IF9 |               |                               | 1OKM         | 1OKN |               |  |
|                              |                        | 1RAY                    | 1RAZ |               |                               | 5CA2         |      |               |  |
|                              |                        | 2CA2                    | 2CBC |               |                               |              |      |               |  |
|                              |                        | 2CBD                    | 3CA2 |               |                               |              |      |               |  |
| HIV-1<br>Protease            | 1G6L                   | 1AJV                    | 1AJX |               |                               | 1C70         | 1GNO |               |  |
|                              |                        | 1DIF                    | 1G2K |               |                               | 1HBV         | 1HIH |               |  |
|                              |                        | 1G35                    | 1HPS |               |                               | 1HOS         | 1HPS |               |  |
|                              |                        | 1HPV                    | 1HSG |               |                               | 1HTR         | 1OHR |               |  |
|                              |                        | 1HGT                    | 1HVI |               |                               |              |      |               |  |
|                              |                        | 1HVJ                    | 1HVK |               |                               |              |      |               |  |
|                              |                        | 1HVL                    | 1IIQ |               |                               |              |      |               |  |
|                              |                        | 2BPV                    | 2BPY |               |                               |              |      |               |  |
|                              | 7UPJ                   |                         |      |               |                               |              |      |               |  |
| Thrombin                     | 1C5L                   | 1AE8                    | 1AFE | 1HAH          | 1HGT                          | 1A2C         | 1ABI |               |  |
|                              |                        | 1AY6                    | 1BA8 | 1THR          | 1THS                          | 1AWF         | 1C4U |               |  |
|                              |                        | 1C1U                    | 1C1V |               |                               | 1C4V         | 1FPH |               |  |
|                              |                        | 1C1N                    | 1C5N |               |                               | 1IHT         | 1KTS |               |  |
|                              |                        | 1C5O                    | 1EBI |               |                               | 1KTT         | 1NRS |               |  |
|                              |                        | 1G30                    | 1G32 |               |                               | 1QHR         | 1QJ1 |               |  |
|                              |                        | 1H8I                    | 1IHS |               |                               | 1QJ6         | 1QJ7 |               |  |
|                              |                        | 1PPB                    | 1UMA |               |                               | 1TMB         | 1TMT |               |  |
|                              |                        |                         |      |               |                               | 2HGT         | 4THN |               |  |
|                              |                        |                         |      |               |                               | 5GDS         | 7KME |               |  |

|                       | Resolution $\leq 2.0$ Å |      |               |      | Resolution $\in [2.0, 2.5)$ Å |               |
|-----------------------|-------------------------|------|---------------|------|-------------------------------|---------------|
|                       | Apo-Proteins            |      | Holo-Proteins |      | Apo-Proteins                  | Holo-Proteins |
| Trypsin               | 1TLD                    | 1TPO | 1AZ8          | 1BJU |                               | 1AQ7          |
|                       | 2PTN                    | 3PTN | 1BJV          | 1BTW |                               | 1K1I          |
|                       |                         |      | 1BTX          | 1BTZ |                               | 1K1J          |
|                       |                         |      | 1C1N          | 1C1P |                               | 1K1L          |
|                       |                         |      | 1C1T          | 1C2G |                               | 1XUF          |
|                       |                         |      | 1C2H          | 1C2K |                               | 1XUK          |
|                       |                         |      | 1C5P          | 1C5Q |                               |               |
|                       |                         |      | 1C5S          | 1C5T |                               |               |
|                       |                         |      | 1EB2          | 1F0T |                               |               |
|                       |                         |      | 1F0U          | 1G34 |                               |               |
|                       |                         |      | 1G36          | 1G3B |                               |               |
|                       |                         |      | 1G3C          | 1GHZ |                               |               |
|                       |                         |      | 1GI0          | 1GI4 |                               |               |
|                       |                         |      | 1GI5          | 1GI6 |                               |               |
|                       |                         |      | 1GJ6          | 1JIR |                               |               |
|                       |                         |      | 1JRS          | 1K1N |                               |               |
|                       |                         |      | 1K1O          | 1K1P |                               |               |
|                       |                         |      | 1MAX          | 1MTS |                               |               |
|                       |                         |      | 1MTW          | 1QCP |                               |               |
|                       |                         |      | 1QL7          | 1TNG |                               |               |
|                       |                         |      | 1TNH          | 1TNI |                               |               |
|                       |                         |      | 1TNJ          | 1TNL |                               |               |
|                       |                         |      | 1TPP          | 1TPS |                               |               |
|                       |                         |      | 1TYN          | 1XUI |                               |               |
|                       |                         |      | 1XUJ          | 1ZZZ |                               |               |
|                       |                         |      | 2BZA          |      |                               |               |
| Streptavidin          | 1SLF                    | 2IZC | 1LCZ          | 1SLE |                               | 1I9H          |
|                       | 2IZD                    | 2IZE | 1SLG          | 1SRE |                               | 1LCW          |
|                       | 2RTB                    | 2RTC | 1SRF          | 1SRG |                               | 1SRH          |
|                       |                         |      | 1SRI          | 1SRJ |                               |               |
|                       |                         |      | 1STR          | 1STS |                               |               |
|                       |                         |      | 1VWL          | 2IZF |                               |               |
|                       |                         |      | 2RTH          |      |                               |               |
|                       |                         |      |               |      |                               |               |
| D-Xylose<br>Isomerase | 1XIB                    | 1XIS | 1GW9          | 1MNZ |                               |               |
|                       |                         |      | 1XIC          | 1XID |                               |               |
|                       |                         |      | 1XIE          | 1XIF |                               |               |
|                       |                         |      | 1XIG          | 1XIH |                               |               |
|                       |                         |      | 1XII          | 1XIJ |                               |               |
|                       |                         |      | 3XIS          | 4XIS |                               |               |
|                       |                         |      | 9XIA          |      |                               |               |

|                         |           | Resolution $\leq 2.0$ Å |               | Resolution $\in [2.0, 2.5)$ Å |               |
|-------------------------|-----------|-------------------------|---------------|-------------------------------|---------------|
|                         |           | Apo-Proteins            | Holo-Proteins | Apo-Proteins                  | Holo-Proteins |
| Cytochrome<br>P-450 CAM | 1PHC      |                         | 1CP4 1GEK     |                               | 1NOO 4CPP     |
|                         |           |                         | 1PHB 1PHD     |                               | 5CPP 8CPP     |
|                         |           |                         | 1PHG 2CPP     |                               |               |
|                         |           |                         | 6CPP 7CPP     |                               |               |
| Ribonuclease A          | 1AFU 1XPS |                         | 1AFK 1AFL     | 1A2W 1JVT                     |               |
|                         | 1XPT      |                         | 1EOS 1JVU     | 1JVV                          |               |
|                         |           |                         | 8RSA 9RSA     |                               |               |
| Endothiapepsin          | 4APE      |                         | 1E5O 1E80     |                               | 4ER4 2ER9     |
|                         |           |                         | 1E81 1E82     |                               | 1EPR          |
|                         |           |                         | 1EED 1ENT     |                               |               |
|                         |           |                         | 1EPL 1EPM     |                               |               |
|                         |           |                         | 1EPN 1EPO     |                               |               |
|                         |           |                         | 1EPP 1EPQ     |                               |               |
|                         |           |                         | 1ER8 2ER6     |                               |               |
|                         |           |                         | 2ER7 3ER3     |                               |               |
|                         |           |                         | 3ER5 4ER1     |                               |               |
|                         |           |                         | 4ER2 5ER1     |                               |               |

**Table 4.1:** PDB entries analysed in the present thesis. For each entry the PDB accession code is listed. Entries are divided into apo- and holo- proteins and according to their resolution.

In the case of dimeric and tetrameric proteins (HIV-1 protease, glutathione S-transferase, ribonuclease, xylose D-isomerase and streptavidin), only monomers sharing the same chain ID were compared. For example, in a pair of dimeric proteins comprising chain A and B, side chain conformational changes were obtained by comparing chain A and chain B of the first protein respectively with chain A and chain B of the second. In PDB entries having different chain IDs, the first chain of the first PDB structure was compared with the first chain of the second structure, the second chain with the second chain of the second structure and so on. This approach is however arbitrary; the differentiation between monomers in complex structures can be made only if the ligands belong to the same structural families or if the structures are frames of a molecular dynamics simulation of the same complex. To determine to what extent conformational changes depend on the choice of the chains that are compared, flexibility analyses were also carried out comparing all chains of a PDB entry with all chains of the other PDB entries (data not shown). The percentages of conformational changes obtained in this way are generally greater than those obtained by selectively comparing one chain against another chain in two protein structures. However, the differences between the percentages obtained in the two different ways are not systematic, and flexibility trends are not significantly affected by the “all chains against all chains” methodology of study; results obtained with the more straightforward and less data intensive methodology of study were therefore chosen and presented in this thesis.

Several methodologies were employed to define conformational changes in the compared protein structures. In accordance with the study by Najmanovich *et al.*,<sup>46</sup> an angular threshold of  $\pm 60^\circ$  was used to define side chain conformational changes. Also, the specific angular thresholds defined by Zhao *et al.*<sup>66</sup> were employed to iden-

| $\chi_1$ Angle Difference |         |        |
|---------------------------|---------|--------|
|                           | Exposed | Buried |
| Arg                       | 37.0    | 13.9   |
| Asn                       | 15.0    | 8.7    |
| Asp                       | 13.8    | 8.0    |
| Cys                       | 7.3     | 9.5    |
| Gln                       | 53.6    | 10.7   |
| Glu                       | 53.6    | 13.0   |
| His                       | 11.2    | 7.1    |
| Ile                       | 12.4    | 8.7    |
| Leu                       | 28.6    | 19.6   |
| Lys                       | 42.1    | 14.7   |
| Met                       | 47.5    | 17.7   |
| Phe                       | 8.0     | 6.4    |
| Ser                       | 102.7   | 16.1   |
| Thr                       | 13.5    | 9.3    |
| Trp                       | 9.0     | 6.7    |
| Tyr                       | 8.0     | 6.2    |
| Val                       | 21.0    | 10.0   |

**Table 4.2:** Angular thresholds ( $^{\circ}$ ) including 90% of  $\chi_1$  observations in Zhao *et al.* data set.<sup>66</sup>

tify  $\chi_1$  conformational changes; these thresholds, listed in Table 4.2, differ for different residue types and for buried and exposed residues (section 3.4). Finally,  $\chi_1$  and  $\chi_2$  rotameric states and their relative probabilities were defined employing the November 2000 release of “backbone dependent” and “conditional backbone independent” rotamer libraries developed by Dunbrack and Cohen.<sup>50,55</sup>

To make a distinction between the  $\chi$  angles and their corresponding rotamers,  $\chi_1$  rotamers were denoted as r1,  $\chi_2$  rotamers as r2, and so on.<sup>55</sup> Since each side-chain torsion angle can normally assume one of the three rotameric states corresponding to the classical conformations *gauche*<sup>+</sup>, *gauche*<sup>−</sup> and *trans*, there can be, for example, 81 possible rotameric states for residues with four rotatable torsions in their side-chains

(e.g. lysine); only two possible rotameric states are instead allowed for proline, whose  $\chi_1$  can only assume the *gauche*<sup>+</sup> or *gauche*<sup>-</sup> conformations. For residues with at least two rotatable bonds in their side-chains, the  $\chi_1, \chi_2$  rotamer was denoted as r12: r12 can assume 9 different states depending on the 9 different possible combinations of r1 and r2 conformations.

For each pair of compared structures, the number and the identity of residues whose r1, r2 and r12 rotamers and/or whose *rank* differ were identified. The “rank” of a residue is an integer, ranging from 1 to 9, that describes how far the actual r12 rotamer is from the most probable r12 rotamer, given the  $\phi$  and  $\psi$  values of the residue. For instance, rank = 1 means that the given r1 and r2 are the most probable for the  $\phi$  and  $\psi$  of that residue; if instead rank = 9, r1 and r2 are the least likely r1 and r2 rotamers that can be found in the PDB, given the residue’s  $\psi$  and  $\psi$  angles.<sup>72</sup> It must be noted that the rank of two corresponding residues can differ even if the relative  $\chi_1$  and  $\chi_2$  angles are almost the same; this can mainly happen if the corresponding  $\phi$  and  $\psi$  angles fall into a different bin of the backbone dependent rotamer library’s conformational space.

Several errors and/or ambiguities can occur in X-ray protein structures. For example, protein X-ray crystallography produces an electron density map that rarely allows carbon, nitrogen and oxygen atoms to be distinguished. Since their electron densities appear to be symmetrical,  $\chi_2$  torsions in asparagine and histidine residues and  $\chi_3$  torsions in glutamine residues can be assigned with difficulty; their orientation in a PDB entry is thus usually defined on the basis of the most favourable hydrogen bond network established with other protein residues. As the aim of this thesis is to analyse protein structures as deposited in the PDB without changing and/or correcting them, no measures were taken to modify them; the possible “flip” of Asn, His and



Gln residue side chains is thus part of the “noise” that inevitably affects our results. For side chain groups that are chemically symmetrical, a different action was applied. Because of symmetry, most PDB structures do not distinguish whether phenylalanine, tyrosine and aspartate have a  $\chi^2$  or  $\chi^2 + 180^\circ$  angular value, and if glutamate has a  $\chi^3$  or  $\chi^3 + 180^\circ$  value. To avoid redundancy in results, both angular values were considered possible for these symmetric residues when applying Najmanovich *et al.*<sup>46</sup> and Zhao *et al.*<sup>66</sup> methodologies of study.

In PDB structures there is often little information about  $\chi^3$  and  $\chi^4$ .<sup>51</sup> The position of the corresponding side-chain atoms is in fact usually poorly determined and, consequently, the convention of using a fully extended conformation is often applied. Further, the rotation around the bonds corresponding to  $\chi^3$  and  $\chi^4$  is less restricted, given the absence of bulky substituents in the long methylene chains of Glu, Gln and Lys that could raise the rotational barrier for these dihedral angles. For these reasons, and since the resolution of the structures in the larger dataset can be up to 2.5 Å, the results reported in this thesis will be mainly restricted to  $\chi^1$  and  $\chi^2$  rotamers and conformations.

Flexibility analyses were carried out on both all residues and on only binding site residues. The binding site of each protein system was defined as composed of the residues in contact with at least one atom of a ligand in the holo-protein data set. If any one residue was identified as being part of the binding site of a particular holo-protein, it was considered as being part of the binding site of all holo-protein structures. Atomic contacts were identified using the LPC and CSU servers (Sobolev *et al.*);<sup>73</sup> these are available through the PDB and define contacts on the basis of the van der Waals radii of the involved atoms and their interatomic distances.

All analyses were performed both by considering separately buried and exposed residues, and without any distinction based on residues' accessibility. Residues in the data set were defined as buried or exposed with the same procedure employed by Zhao *et al.*:<sup>66</sup> the accessible surface area (ASA) of each residue was calculated with the NACCESS program,<sup>74</sup> using a probe radius of 1.4 Å and including only heavy amino acid atoms in the calculation. The ASA of each residue was then compared to that of the same residue in the extended polypeptide Ala-X-Ala, where X is the particular amino acid type; if the exposed surface of the residue was 20% or less of the surface area of the same residue in the extended polypeptide, the residue was classified as buried. A residue was instead classified as exposed if its solvent-exposed surface area was greater than 20% the surface area of the extended polypeptide.

To quantify the geometrical changes occurring in the protein backbone, backbone atoms' Root-Mean Square deviation (RMSd) was computed for all possible pairs of protein structures. The program ProFit<sup>75</sup> was employed to perform the least squares fitting procedures and the RMSd calculations. The atoms chosen for the fitting procedure and the RMSd calculation were the same; residues and atoms that did not exist or had been discarded due to the occupancy criteria in one protein were not considered.

To investigate whether similar ligands induce similar conformational changes in the same apo-proteins, Tanimoto similarity coefficients<sup>76</sup> were computed for all possible pairs of ligands binding to the same protein and plotted against the percentages of side-chain conformational changes observed in the corresponding holo-/holo- protein pair. The Tanimoto coefficients were computed employing bi-dimensional 1024-bit Daylight fingerprints,<sup>77</sup> which calculate all pairwise distances for atoms in a molecule and represent them as a signature key, i.e. as a bit string where each position cap-

turing the presence or absence of a unique pattern. As a similarity value for pairwise comparison of fingerprints the Tanimoto coefficient ( $T_C$ ) is calculated as:

$$T_C = N_{AB}/(N_A + N_B - N_{AB}) \quad (4.1)$$

where  $N_{AB}$  is the number of common bits set on, and  $N_A$  and  $N_B$  are the bits set on in the fingerprints of the first (A) and the second (B) molecules, respectively.<sup>76</sup> Molecules with a high  $T_C$  are considered similar. Since the default path range of 1024-bit Daylight fingerprints is 0-7 bonds, they do not discriminate between molecules that differ only in bond paths longer than the maximum; multipeptidic ligands are thus treated as disconnected amino-acid molecules, and their sequence informations is lost. For this reason, peptidic ligands were excluded from this kind of study. In the case of endothiapepsin, for which only peptidic ligands exist, no ligand similarity analyses were performed.

While Tanimoto similarity coefficients were provided by another worker using the Daylight Fingerprint Toolkit,<sup>77</sup> most of the analyses described in this thesis were performed by code written by the author in Perl, a language particularly convenient for text processing. Perl scripts were also employed to invoke C or C++ programs.

## 4.3 Results Overview

### 4.3.1 Root Mean Square Deviation

Table 4.3 reports the average Root Mean Square deviation (RMSd) of backbone atoms for all possible apo-/apo-, apo-/holo- and holo-/holo- protein comparisons within the 10 different systems. The extent of backbone rearrangement appears to be generally small; HIV-1 protease, endothiapepsin, streptavidin, thrombin and ribonuclease are, in order, the systems for which backbone atoms show the greatest flexibility.

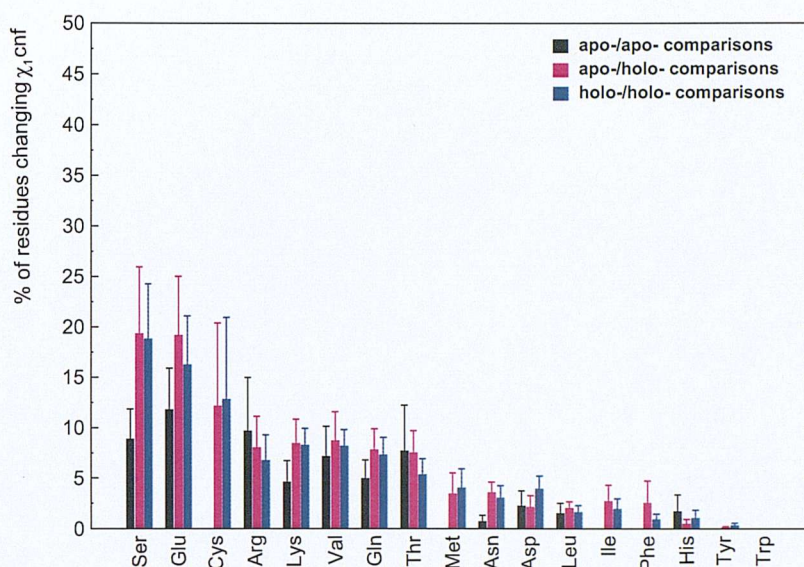
| Protein System            | RMSD (Å)            |            |             |
|---------------------------|---------------------|------------|-------------|
|                           | protein comparisons |            |             |
|                           | apo-/apo-           | apo-/holo- | holo-/holo- |
| HIV-1 Protease            |                     | 0.59       | 0.50        |
| Endothiapepsin            |                     | 0.46       | 0.37        |
| Ribonuclease              | 0.37                | 0.30       | 0.30        |
| Streptavidin              | 0.11                | 0.35       | 0.45        |
| Thrombin                  |                     | 0.35       | 0.34        |
| Glutathione S-Transferase |                     | 0.24       | 0.29        |
| Trypsin                   | 0.21                | 0.23       | 0.24        |
| Carbonic Anhydrase        | 0.12                | 0.19       | 0.26        |
| Cytochrome P-450 CAM      |                     | 0.18       | 0.18        |
| D-Xylose Isomerase        | 0.17                | 0.16       | 0.18        |

**Table 4.3:** RMSd in Å between the structures of the ten analysed protein families. The averaged RMSd for all possible apo-/apo-, apo-/holo- and holo-/holo comparisons within a protein system is reported; the same backbone atoms were considered in the fitting and in the RMSd calculation procedure.

The RMSd that is detected by comparing apo- against apo- proteins is potentially linked to the intrinsic flexibility of proteins, i.e. to conformational changes that do not depend on ligand binding and might inevitably bias our observations. As it appears from the apo-/apo- protein comparison results reported in Table 4.3, the analyses performed on ribonuclease and, to a lesser extent, D-xylose isomerase and trypsin, might be complicated by some intrinsic backbone flexibility. Nevertheless, the apo-/apo- protein comparison RMSDs appear small enough to justify our conformational analysis on all ten selected protein systems.

### 4.3.2 Different Amino Acid Conformational Changes

Figures 4.1-4.4 represent the percentages of different residue types for which  $\chi_1$  and  $\chi_2$  torsions change more than 60°; while Figures 4.1 and 4.3 refer to all protein residues, Figures 4.2 and 4.4 correspond to only binding site residues. The error bars



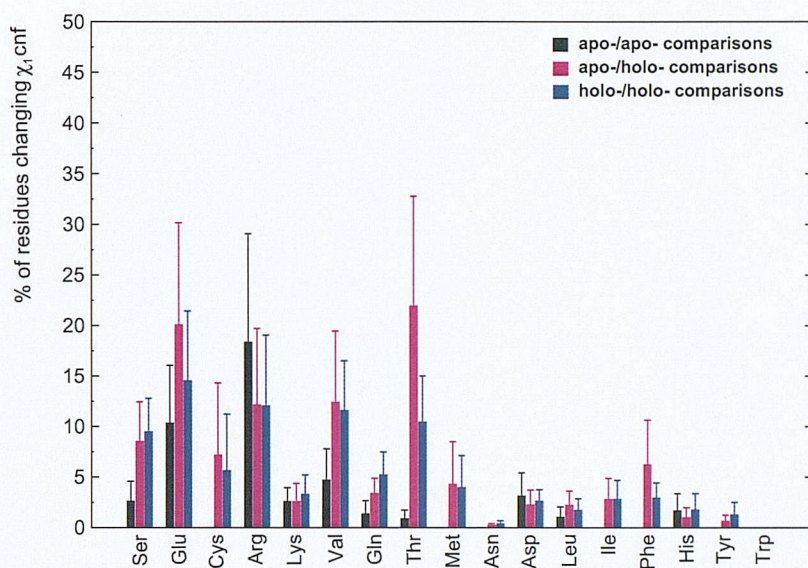
**Figure 4.1:** Percentages of  $\chi_1$  conformational changes revealed for different residue types when a  $60^\circ$  angular cutoff is applied; all residue results.

in these figures, as for all the figures in the present chapter, represent the standard error on the averages.

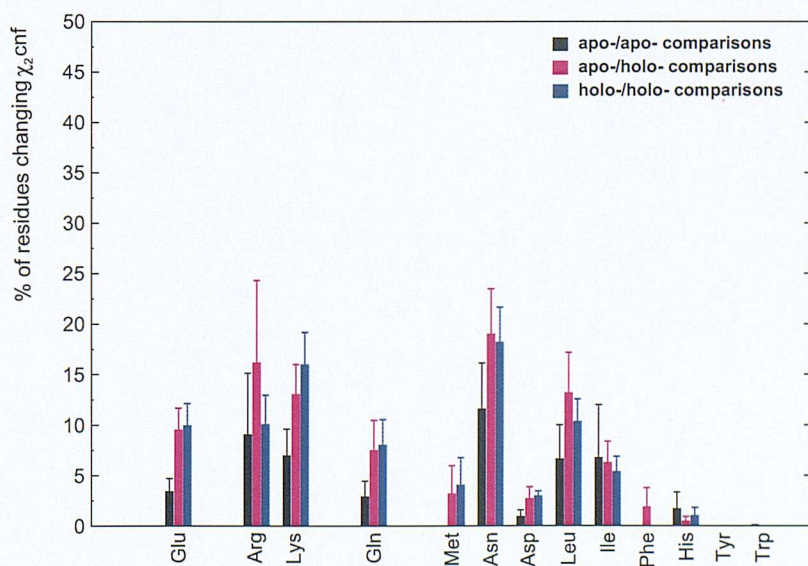
Figures 4.5 and 4.6 report the percentages of conformational changes observed when Zhao *et al.* residue- and environment- specific angular thresholds<sup>66</sup> are applied, while the percentages of  $\chi_1$  and  $\chi_2$  that change rotameric state according to Dunbrack and Cohen rotamer libraries are represented in Figures 4.7-4.10. Again, Figures 4.5, 4.7 and 4.9 refer to all protein residues, while Figures 4.6, 4.8 and 4.10 describe only binding site residues conformational changes.

Residues have been ordered on the x axis of this section's graphs in accordance with the  $\chi_1$  flexibility trends, as revealed for all protein residues with Najmanovich *et al.*<sup>46</sup> and Dunbrack *et al.*<sup>55</sup> methodologies of study. Since these methods employ undifferentiated angular thresholds to define conformational changes, they are likely to reveal the absolute propensity to move of different residue types. The  $\chi_1$  general flexibility scale that is obtained for all protein residues is: Ser, Glu  $\geq$  Cys  $\geq$  Arg > Lys, Val, Gln, Thr > Met, Asn, Leu, Ile  $\geq$  Phe, His > Tyr, Trp. If  $\chi_2$  percentages



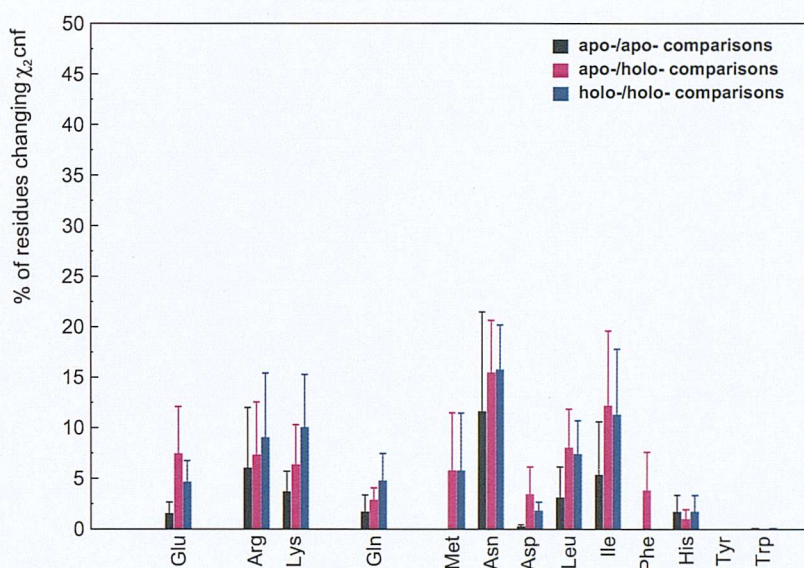


**Figure 4.2:** Percentages of  $\chi_1$  conformational changes revealed for different residue types when a  $60^\circ$  angular cutoff is applied; only binding site results.

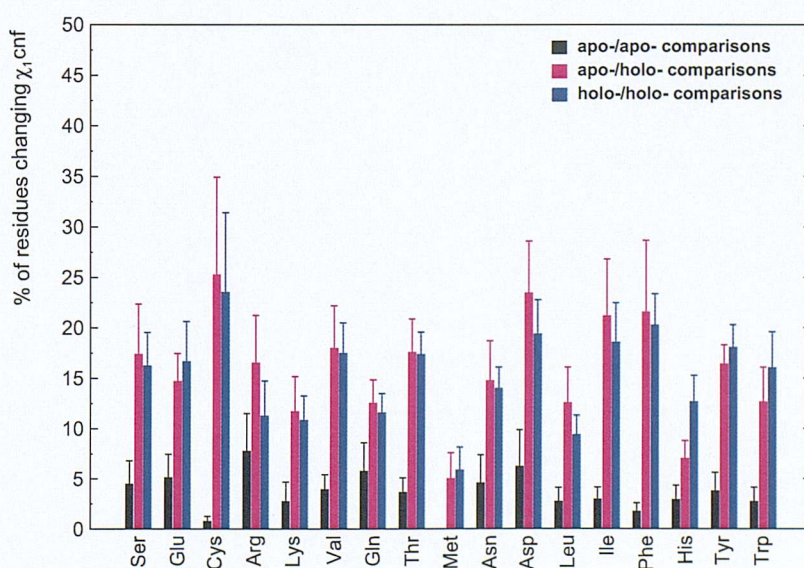


**Figure 4.3:** Percentages of  $\chi_2$  conformational changes revealed for different residue types when a  $60^\circ$  angular cutoff is applied; all residue results.



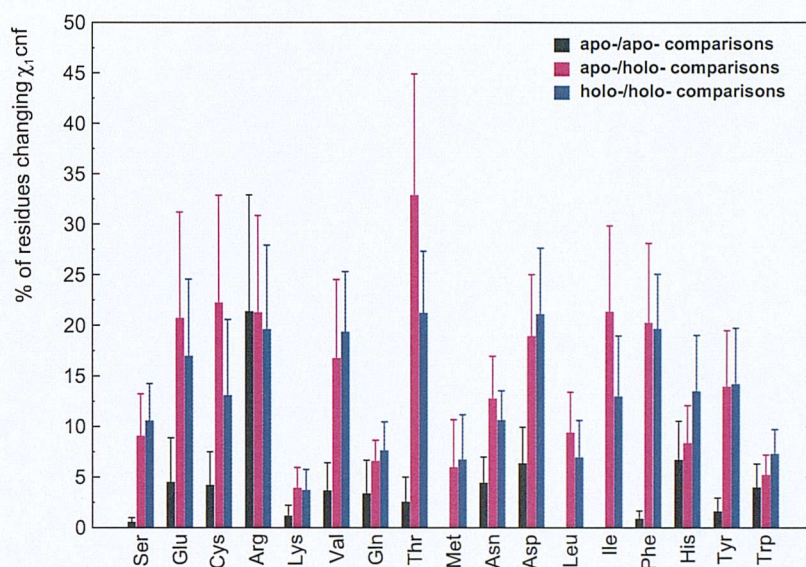


**Figure 4.4:** Percentages of  $\chi_2$  conformational changes revealed for different residue types when a  $60^\circ$  angular cutoff is applied; only binding site results.

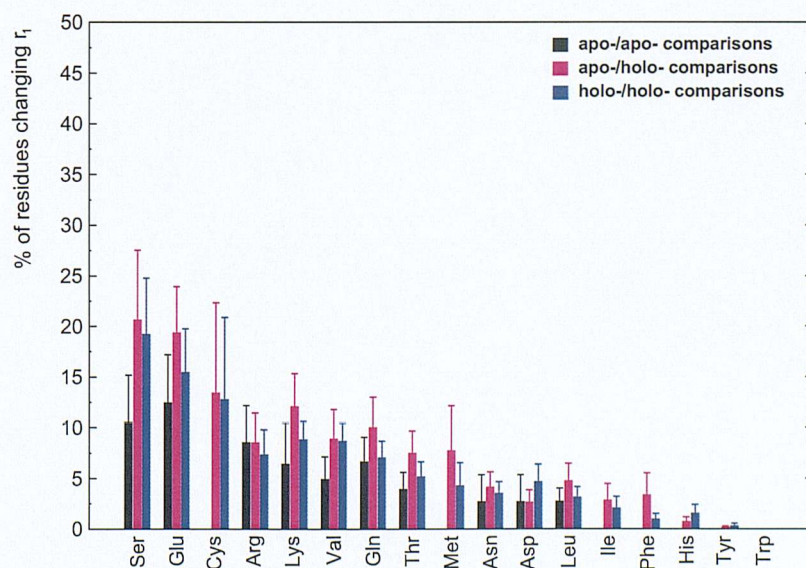


**Figure 4.5:** Percentages of  $\chi_1$  conformational changes revealed for different residue types by Zhao *et al.* methodology of study; all residue results.



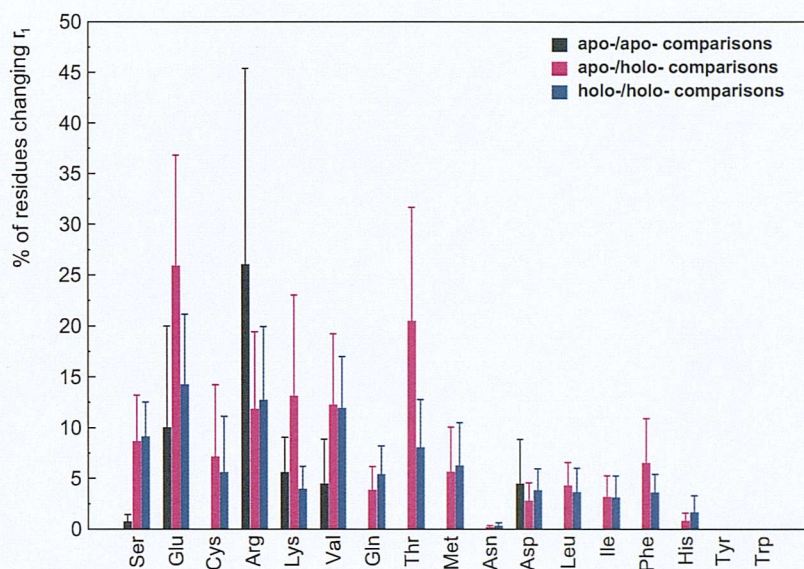


**Figure 4.6:** Percentages of  $\chi_1$  conformational changes revealed for different residue types by Zhao *et al.* methodology of study; only binding site results.

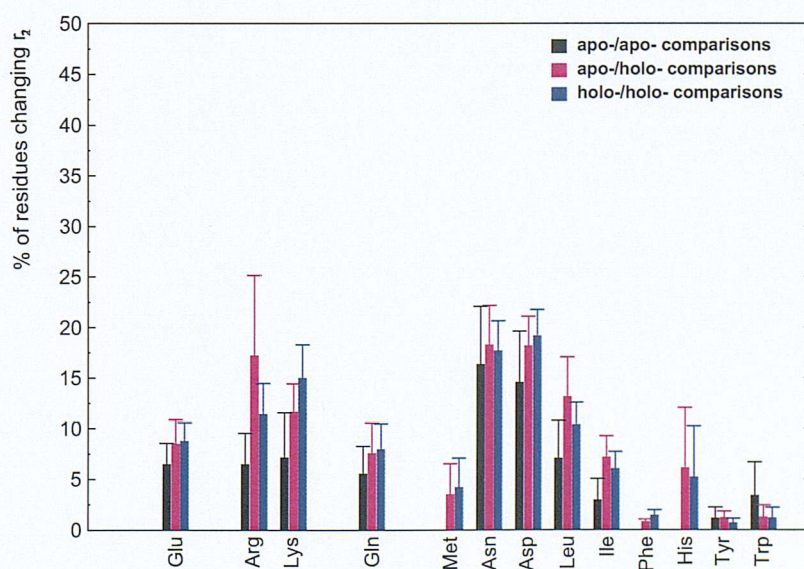


**Figure 4.7:** Percentages of  $r_1$  conformational changes revealed for different residue types when Dunbrack and Cohen rotamer libraries are employed; all residue results.



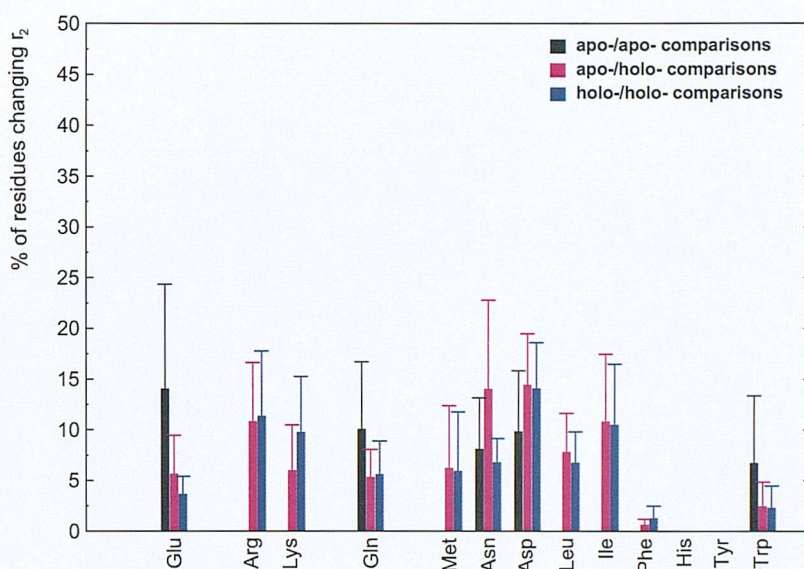


**Figure 4.8:** Percentages of  $r_1$  conformational changes revealed for different residue types when Dunbrack and Cohen rotamer libraries are employed; only binding site results.



**Figure 4.9:** Percentages of  $r_2$  conformational changes revealed for different residue types when Dunbrack and Cohen rotamer libraries are employed; all residue results.





**Figure 4.10:** Percentages of  $r_2$  conformational changes revealed for different residue types when Dunbrack and Cohen rotamer libraries are employed; only binding site results.

of conformational changes are considered, the flexibility scale obtained applying Najmanovich *et al.* methodology of study becomes: Asn > Arg, Lys  $\geq$  Leu > Glu > Ile, Gln > Met, Asp > His, Phe  $\geq$  Trp, Tyr. Apart from some residue types (e.g. Asp), the flexibility scale obtained for  $r_2$  with Dunbrack and Cohen rotamer libraries is not very different: Asn, Asp > Arg, Lys, Leu > Glu, Gln  $\geq$  Ile  $\geq$  Met, His  $\geq$  Trp  $\geq$  Tyr, Phe.

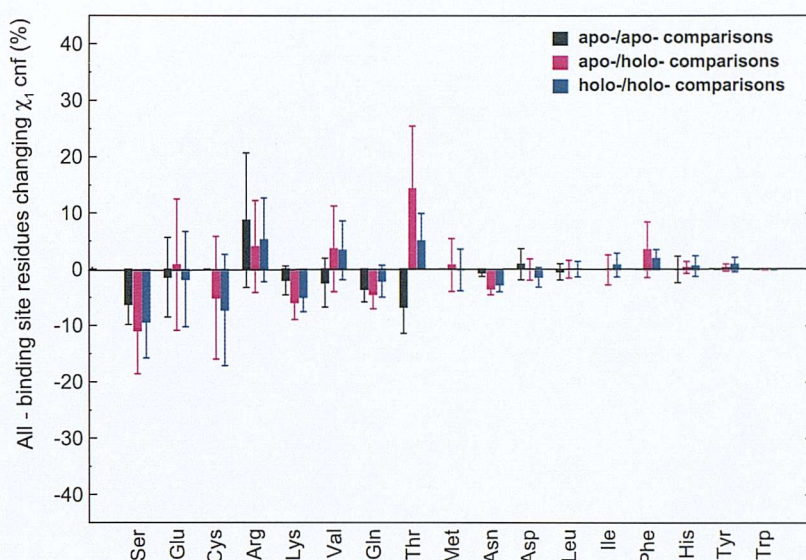
Broadly speaking, these flexibility trends are similar to those found by Najmanovich *et al.*: long, polar, sterically non bulky side-chains appear to be the most flexible, while aromatic residues seem to be the most rigid. However, in the present thesis data set residues such as serine, aspartate, asparagine and cysteine seem significantly more flexible than what Najmanovich and coworkers found.<sup>46</sup> At the same time, residues such as methionine change side-chain conformation significantly less often.

Different trends can be observed in the graphs obtained by applying Zhao *et al.* residue- and environment- specific angular thresholds<sup>66</sup> (Figures 4.5 and 4.6). Since

these thresholds were defined as the angular intervals comprising 90% of  $\chi_1$  side-chain torsions in a data set of apo-protein structures, peaks on these graphs correspond to residues whose  $\chi_1$  flexibility is significantly greater than average. Applying these thresholds to all protein residues, the scale of apo-/holo- and holo-/holo- protein conformational changes detected in the data set of the present thesis is: Cys > Asp, Phe, Ile > Val, Thr, Ser, Tyr, Thr, Glu.  $\geq$  Trp, Arg  $\geq$  Gln, Lys, Leu  $\geq$  His > Met. Clearly, the flexibility of Phe, Tyr, Trp and His appears to be unusually high for these aromatic residues, that seldom change their side-chain torsions by more than 60° and/or rarely change their rotameric state (Figures 4.1 and 4.7). Zhao *et al.* angular thresholds for these residues are in fact very small; only exposed His residues must change more than 10° for a conformational change to be detected. The residue for which the highest unexpected percentage of conformational changes is detected is Cys; this residue, whose  $\chi_1$  confidence levels determined by Zhao and co-workers are equal to 7.3° for buried residues and 9.5° for exposed residues, appears to be unusually flexible also when Najmanovich *et al.* and Dunbrack *et al.* angular thresholds are applied. This tendency, stronger in all protein residues rather than in binding site residues, might depend on the relatively small size of the present thesis's data set.

The aim of this thesis is to disentangle random, spontaneous motions of protein side-chains from motions that actually depend on ligand-binding effects. Ideally, the differences between the conformational changes that are observed in binding sites versus all protein residues should help to distinguish these different kinds of motions. However, the error bars on the graphs representing this differences are too large to allow the observation of any systematic effect (see for example Figure 4.11). To unambiguously identify ligand induced conformational changes of individual amino acids using this data, a larger dataset should be analysed and/or different protein





**Figure 4.11:** Differences between the  $\chi_1$  percentages of conformational changes occurring in the binding site and in all protein residues; negative values reveal greater flexibility of binding site residues. Conformational changes defined using a  $60^\circ$  angular cutoff.

systems considered.

It is interesting to compare the percentages of conformational changes that are detected in apo-/apo- versus apo-/holo- and holo-/holo- protein comparisons. With only few exceptions, fewer protein motions seem to occur in uncomplexed protein structures with all the methods of analysis. However, this trend is much more consistent when Zhao *et al.* angular thresholds are employed (Figures 4.5 and 4.6); this method of study only detects percentages of conformational changes in the uncomplexed PDB entries of greater than 10% in the case of arginine binding site residues. This fairly low  $\chi_1$  flexibility in apo-/apo- protein comparisons is something one would expect, since these angular thresholds were obtained comparing a data-set of apo-protein structures. However, it is somewhat striking, given the high flexibility that the same angular thresholds reveal for apo-/holo- and holo-/holo- protein comparisons. The standard errors associated with the apo-/apo- comparisons overlap with those of the apo-/holo- and holo-/holo- comparisons only in the case of Arg (all and only binding site residues) and Lys, Gln, His and Trp (only binding site residue data);

when Najmanovich *et al.* and Dunbrack and Cohen methods are applied (Figures 4.1, 4.2, 4.7 and 4.7), apo-/holo and holo-/holo- protein comparisons results are in fact significantly different from apo-/apo- protein comparisons ones only in very few cases.

These results suggest that Zhao *et al.* specific angular thresholds are probably more useful than generic, undifferentiated thresholds in identifying unusual, systematic conformational changes in proteins; unusually high peaks in these graphs are more likely to spot conformational changes that depend on ligand binding effects rather than the intrinsic propensities of different residues to move.

However, one should keep in mind that the selection criteria employed by Zhao *et al.* (resolution cutoff equal to 2.2 Å, side-chain flexibility analysis performed only on proteins and regions characterised by low backbone flexibility) are different from the ones employed in this thesis, and this might affect the results. Also, the apo-protein data set of the present thesis is significantly smaller than that employed by Zhao *et al.*; while Zhao and coworkers compared three apo-protein structures for a protein system, for a total of 123 pairs (i.e. 41 different protein systems),<sup>66</sup> only five out of ten protein systems analysed in this thesis have more than one apo-protein at resolution equal or better than 2.0 Å (4 uncomplexed protein structures for carbonic anhydrase II, trypsin and ribonuclease A, six for streptavidin and two for D-xylose isomerase).

The restricted size of this thesis's data set is likely to be the reason why the percentages of apo-/apo- protein conformational changes are always less than 10%, something surprising since Zhao *et al.* angular thresholds were obtained as the confidence levels that comprise all the analysed residues but 10%. In the next chapters, the issue of other possible reasons of bias (e.g. apo-protein structures prevalently

obtained by the same author or group) will be investigated.

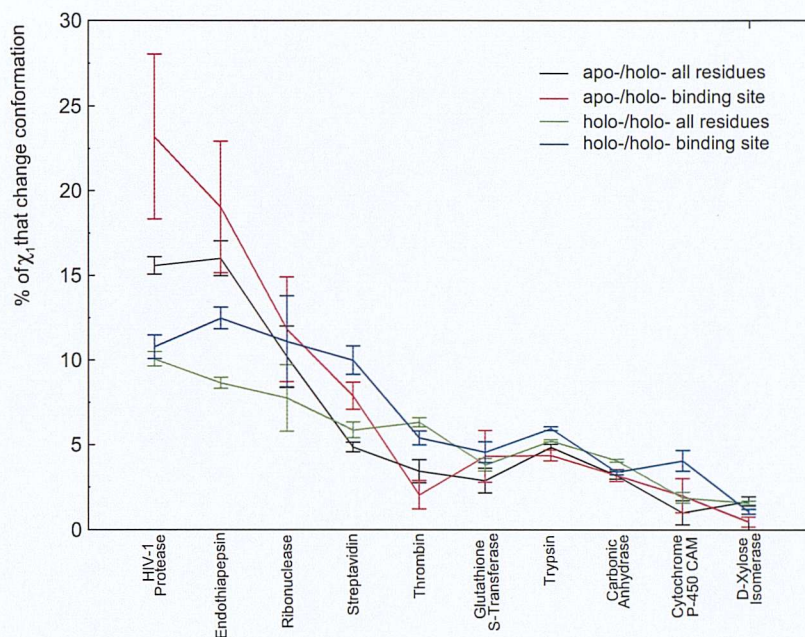
### 4.3.3 All-Environment Flexibility Trends

Applying Dunbrack and Cohen's rotamer definitions,<sup>55</sup> Najmanovich *et al.*  $\pm 60^\circ$  angular cutoff<sup>46</sup> and Zhao *et al.* specific angular thresholds<sup>66</sup> to define side-chain conformational changes, different flexibility trends emerge for the ten analysed protein systems. To evaluate a general protein flexibility order, the percentages of conformational changes observed in both apo-/holo- and holo-/holo- protein pair analyses were evaluated and compared.

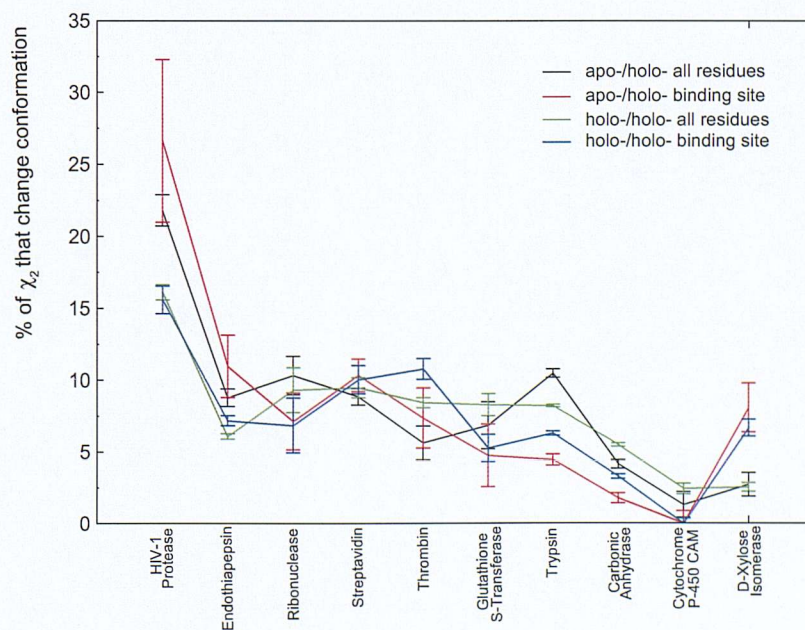
On the y axis of Figures 4.12, 4.13 and 4.14, side-chain propensities to move are represented by means of  $\chi_1$  and  $\chi_2$  percentages of observed conformational changes in the analysed protein systems (Najmanovich *et al.*, and Zhao *et al.* methodologies of study); Figures 4.15, 4.16 and 4.17 depict instead side-chain flexibility trends as the percentages of r1, r2 and rank (Dunbrack and Cohen's rotamer definitions) that change in the different protein systems.

Some general trends in terms of protein flexibility are broadly speaking conserved with all methodologies of study. HIV-1 protease, endothiapepsin and ribonuclease A are the most flexible protein systems. Streptavidin immediately follows, appearing to be slightly less flexible than thrombin only when Dunbrack and Cohen rotamers definitions are applied, while the other protein systems show a variable order of flexibility with different methods of analysis and for different  $\chi$  torsions. Roughly speaking, the flexibility order resulting from all methods of analysis is: HIV 1 protease > endothiapepsin > ribonuclease A > streptavidin  $\geq$  thrombin > trypsin  $\geq$  cytochrome P-450 CAM  $\geq$  carbonic anhydrase II  $\geq$  glutathione S-transferase  $\geq$  xylose D-isomerase (Figures 4.12, 4.13, 4.14, 4.15, 4.16 and 4.17).



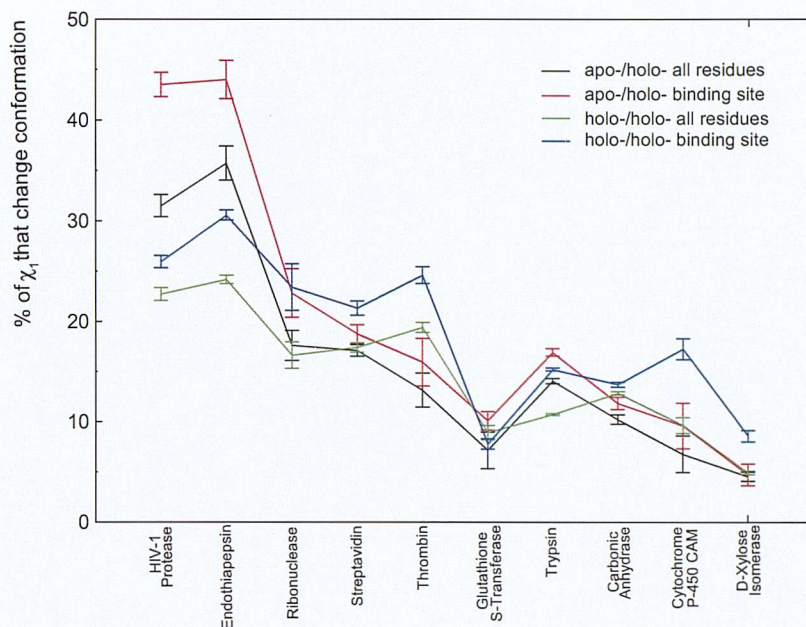


**Figure 4.12:** Percentages of  $\chi_1$  torsions that change more than  $\pm 60^\circ$  in apo-/holo- and holo-/holo- protein comparisons; data relative to only binding site residues and to all residues.

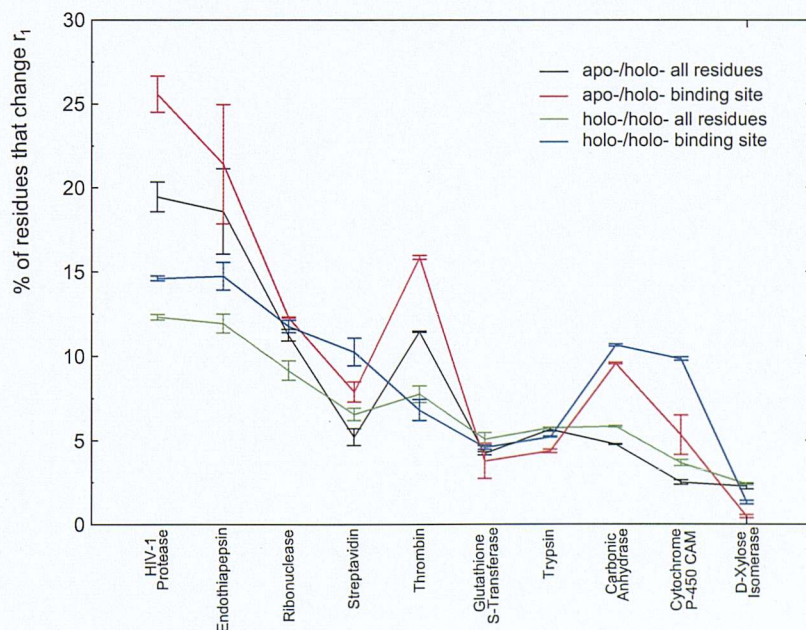


**Figure 4.13:** Percentages of  $\chi_2$  torsions that change more than  $\pm 60^\circ$  in apo-/holo- and holo-/holo- protein comparisons; data relative to only binding site residues and to all residues.



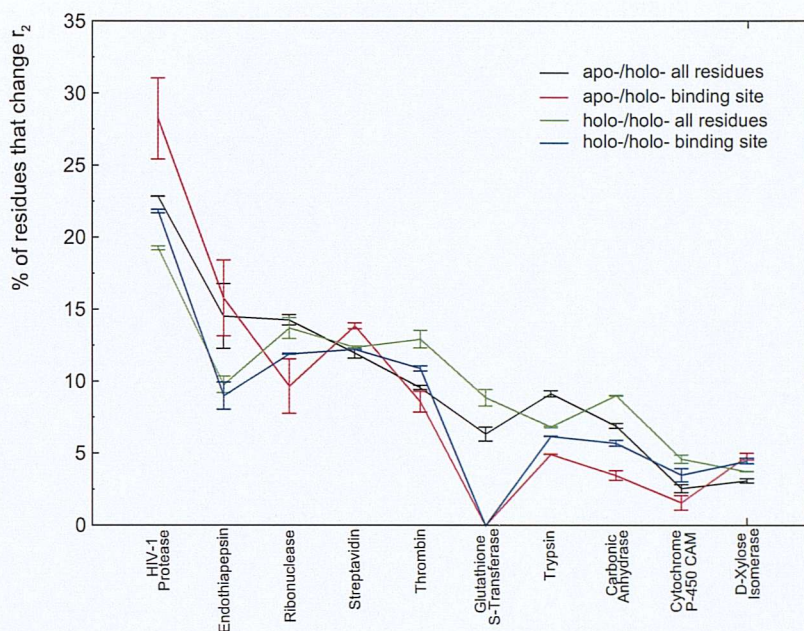


**Figure 4.14:** Percentages of  $\chi_1$  torsions that change more than Zhao *et al.* specific angular thresholds in apo-/holo- and holo-/holo- protein comparisons. Data relative to only binding site residues and to all residues.

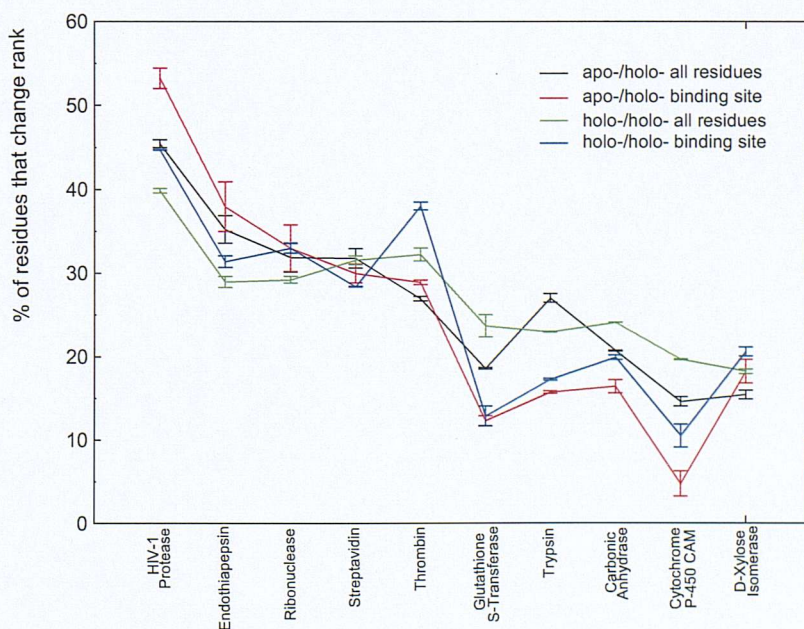


**Figure 4.15:** Percentages of  $\chi_1$  torsions that change their rotameric state ( $r_1$ ) in apo-/holo- and holo-/holo- protein comparisons; data relative to only binding site residues and to all residues.





**Figure 4.16:** Percentages of  $\chi^2$  torsions that change their rotameric state ( $r_2$ ) in apo-/holo- and holo-/holo- protein comparisons; data relative to only binding site residues and to all residues.



**Figure 4.17:** Percentages of residues that change  $r_1$  and/or  $r_2$  and/or rank. Data relative to apo-/holo- and holo-/holo- protein comparisons, binding site residues and all residues of the 10 analysed protein systems.

$\chi_1$  and  $\chi_2$  do not generally show the same flexibility patterns. For example,  $\chi_1$  in D-xylose isomerase appears to be significantly more rigid than  $\chi_2$ , especially in residues that are part of the binding site and when an angular threshold of  $\pm 60^\circ$  is applied (Figures 4.12 and 4.13). Also,  $\chi_2$  in HIV-1 protease is always remarkably more flexible than  $\chi_1$ , while in the binding site of cytochrome P-450 CAM,  $\chi_2$  is instead more rigid than  $\chi_1$  (Figures 4.12, 4.13, 4.15 and 4.16).

Apo-/holo- protein comparisons and holo-/holo- protein comparisons can show very different trends too. HIV-1 Protease and endothiapepsin show for example significantly greater conformational changes in apo-/holo- rather than holo-/holo- protein comparisons with all methods of conformational analyses (Figures 4.12, 4.13, 4.14, 4.15 and 4.16). Apo-/holo- and holo-/holo protein comparisons in the other systems generally give much more similar results.

Since data are inconsistent, trying to infer anything more than a broad flexibility trend from these graphs is problematic; apart from a few exceptions (HIV-1 protease and endothiapepsin), the proteins' flexibility trends are inconsistent when different methods of analysis are applied. If a single method had to be selected as most reliable, arguably the Zhao *et al.* methodology of study should be preferred, given the way it was parametrised.<sup>66</sup> However, the other two methodologies of study could possibly highlight specific protein characteristics too.

The unavailability of Zhao *et al.*  $\chi_2$  angular thresholds makes the comparisons of the results obtained for  $\chi_2$  with all methodologies of study impossible; in the next sections, only results for  $\chi_1$  side-chain torsions will be presented.

### 4.3.4 Buried and Exposed Residues Flexibility Trends

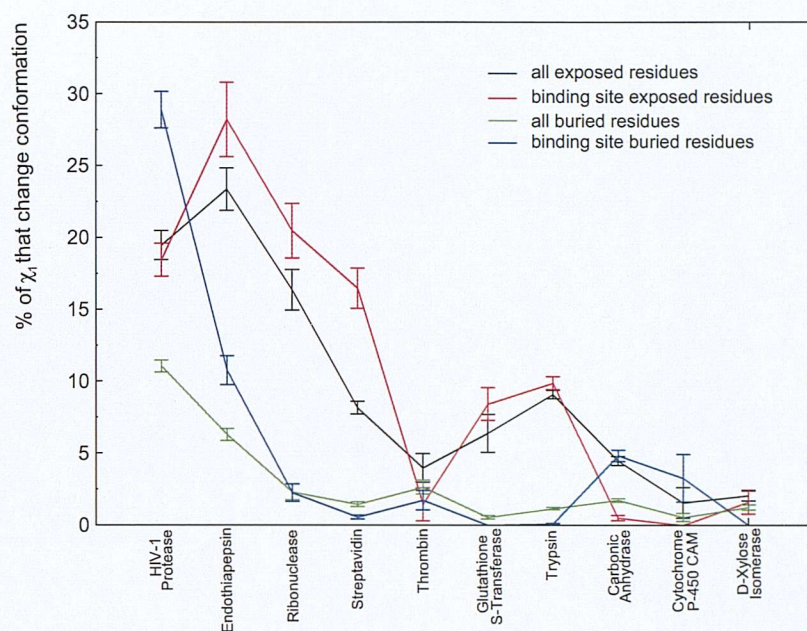
Figures 4.18 to 4.23 represent the percentages of  $\chi_1$  conformational changes that are observed when a distinction between buried and exposed residues is made.

The broad flexibility scale that is obtained from these graphs is similar to that obtained from the preceding section; again, different methods of analysis lead to different results.

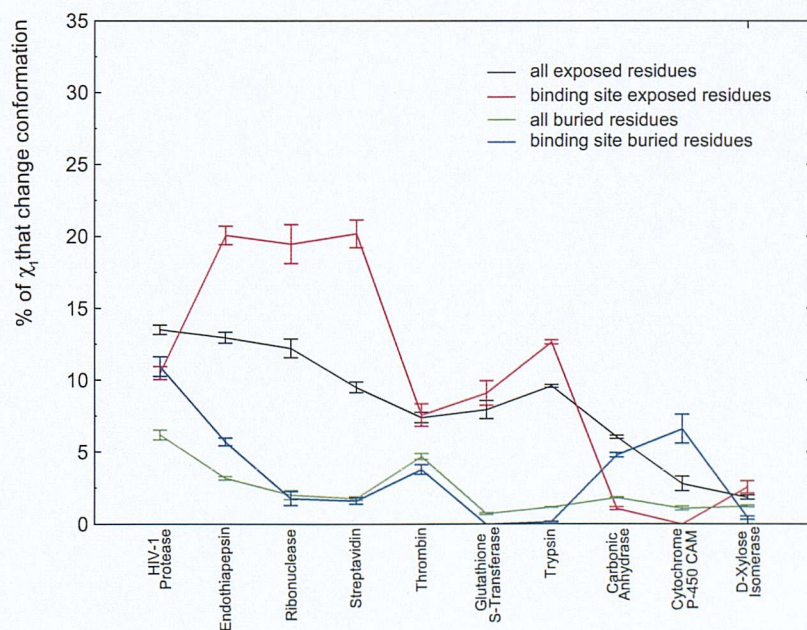
In the graphs obtained using undifferentiated angular thresholds<sup>46,55</sup> (Figures 4.18, 4.19, 4.22 and 4.23), the distance between the black and red lines and the blue and green lines is very large, i.e. a significant difference between exposed and buried residues' conformational changes is revealed. If Zhao *et al.*<sup>66</sup> methodology of study is applied (Figures 4.20 and 4.21), the spread of these data is much smaller; moreover, there is a significantly greater consistency between apo-/holo- and holo-/holo- protein comparisons results, especially with respect to the Dunbrack and Cohen method of study (Figures 4.22 and 4.23). Zhao *et al.* specific angular thresholds once more seem particularly helpful to disentangle ligand-binding dependent conformational changes from random motions dependent on residues intrinsic flexibility; peaks on these graphs potentially highlight important peculiarities of the analysed protein systems.

The characteristics of the 10 analysed protein systems and the differences and/or similarities among the results obtained with the three methods of analysis are much more easily depicted by plotting the differences between the percentages of conformational changes observed in buried and exposed residues (Figures 4.24-4.29). The shapes and the patterns on these graphs are very similar across all the employed methods of analysis; by plotting differences on the y axis some of the "noise" arising from how conformational changes are defined has been eliminated.



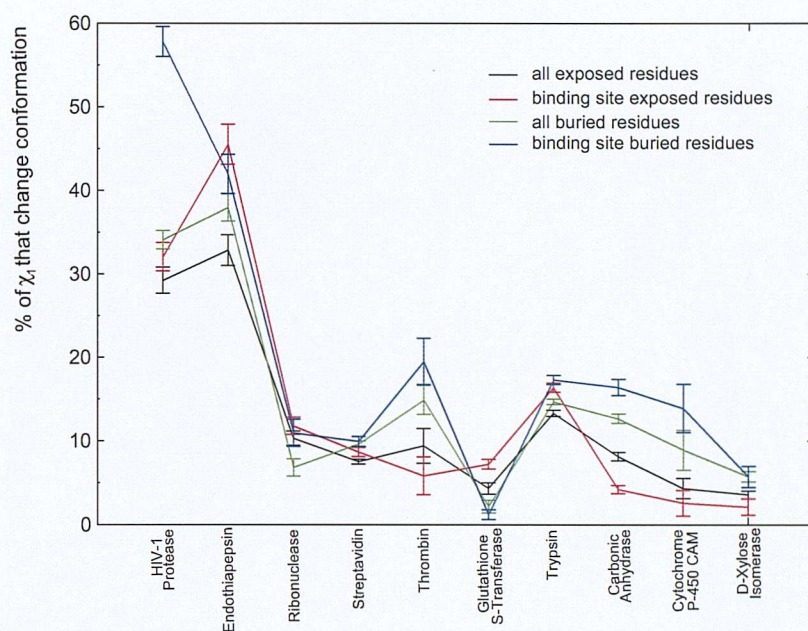


**Figure 4.18:** Percentages of buried and exposed  $\chi_1$  torsions that change more than  $60^\circ$  in apo-/holo- protein comparisons; data correspond to only binding site residues and to all residues.

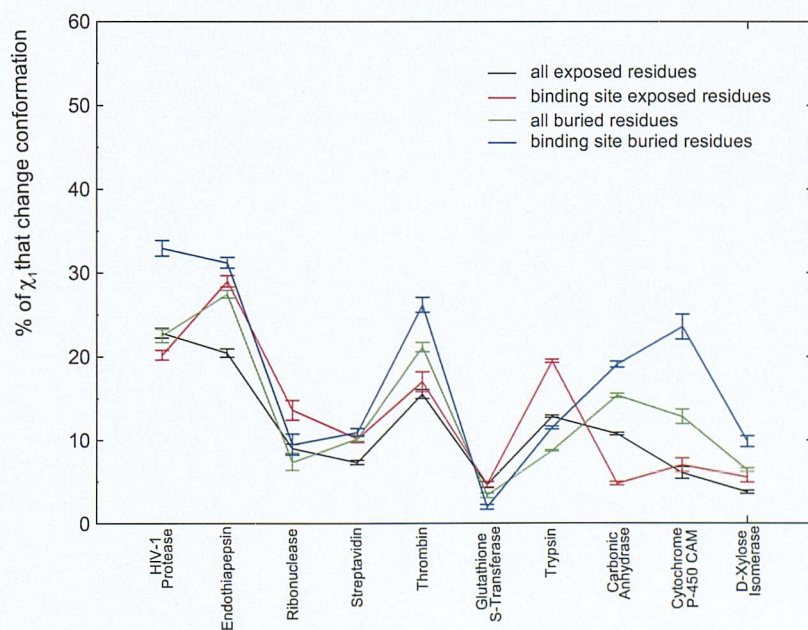


**Figure 4.19:** Percentages of buried and exposed  $\chi_1$  torsions that change more than  $60^\circ$  in holo-/holo- protein comparisons; data correspond to only binding site residues and to all residues.



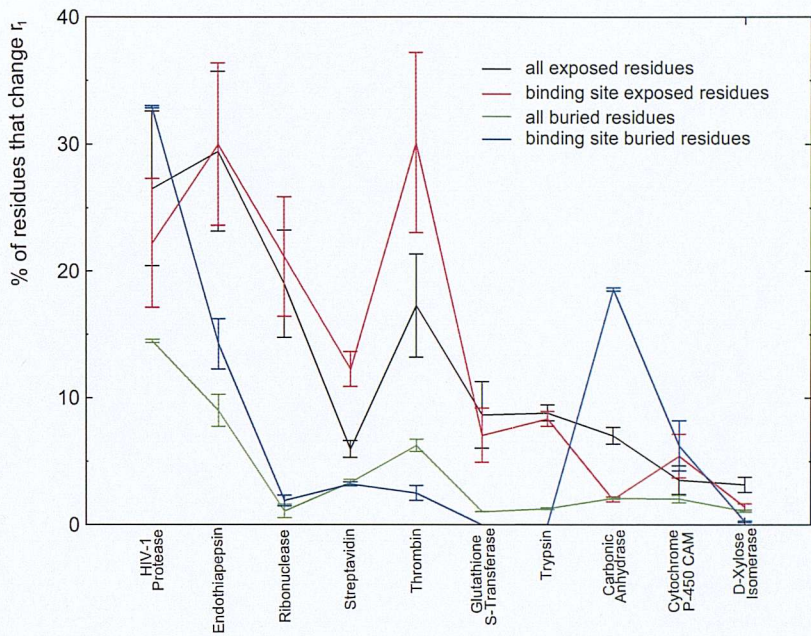


**Figure 4.20:** Percentages of buried and exposed  $\chi_1$  torsions that change more than Zhao *et al.* specific angular thresholds in apo-/holo- protein comparisons; data correspond to only binding site residues and to all residues.

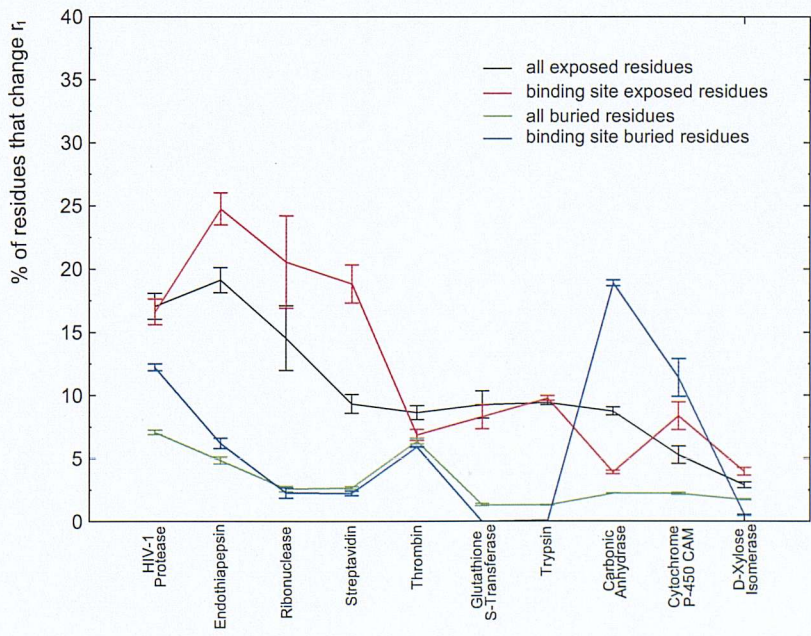


**Figure 4.21:** Percentages of buried and exposed  $\chi_1$  torsions that change more than Zhao *et al.* specific angular thresholds in holo-/holo- protein comparisons; data correspond to only binding site residues and to all residues.



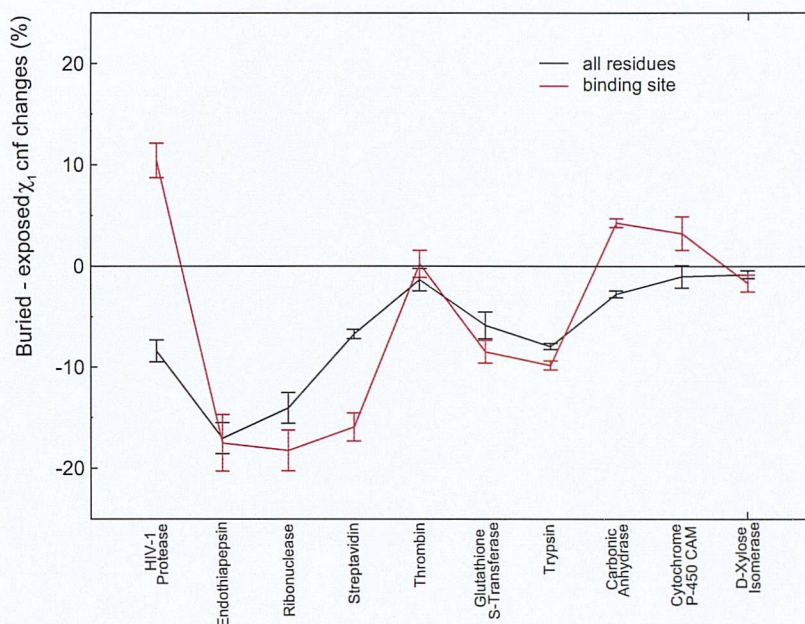


**Figure 4.22:** Percentages of buried and exposed  $\chi_1$  torsions that change their rotameric state ( $r_1$ ) in apo-/holo- protein comparisons; data correspond to only binding site residues and to all residues.

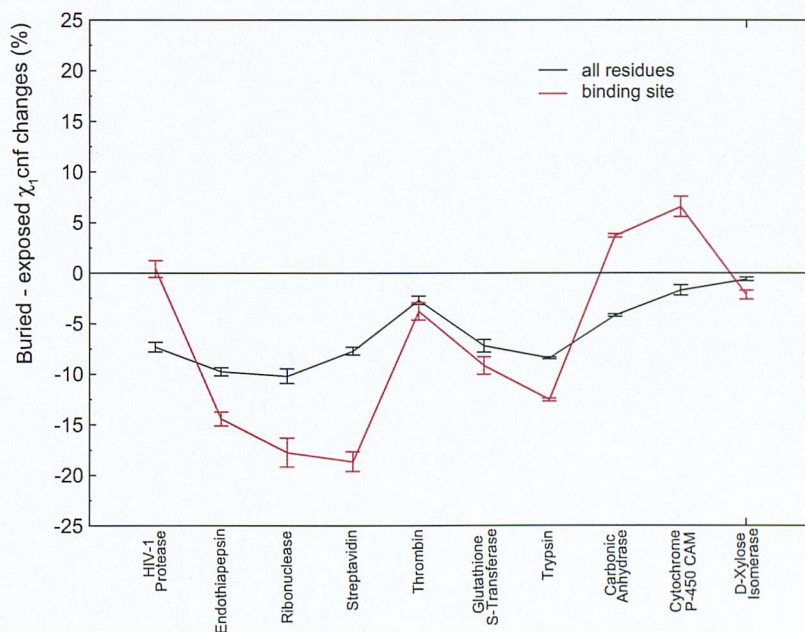


**Figure 4.23:** Percentages of buried and exposed  $\chi_1$  torsions that change their rotameric state ( $r_1$ ) in holo-/holo- comparisons; data correspond to only binding site residues and to all residues.



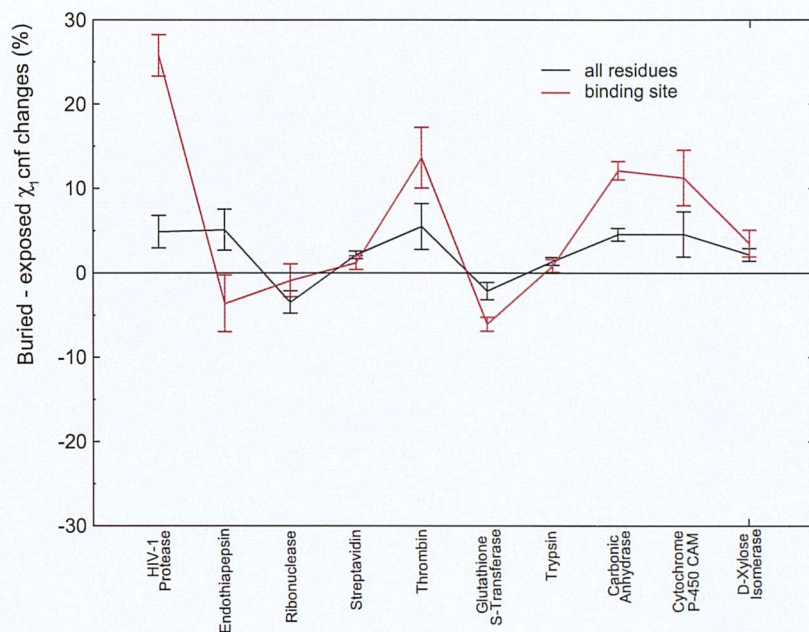


**Figure 4.24:** Differences between the  $\chi_1$  percentages of conformational changes occurring in buried and exposed residues (apo-/holo- comparisons); positive values reveal greater flexibility of buried residues. Conformational changes defined using a  $60^\circ$  angular cutoff.

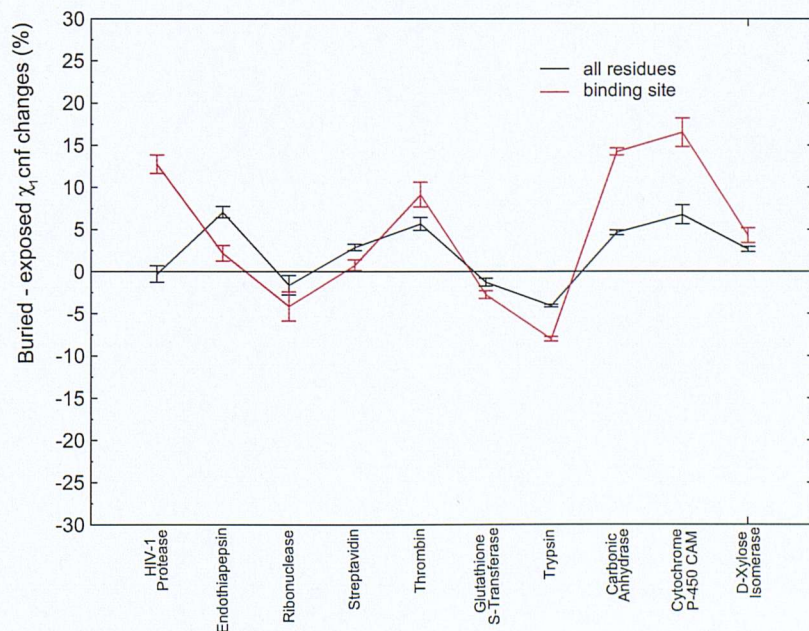


**Figure 4.25:** Differences between the  $\chi_1$  percentages of conformational changes occurring in buried and exposed residues (holo-/holo- comparisons); positive values reveal greater flexibility of buried residues. Conformational changes defined using a  $60^\circ$  angular cutoff.



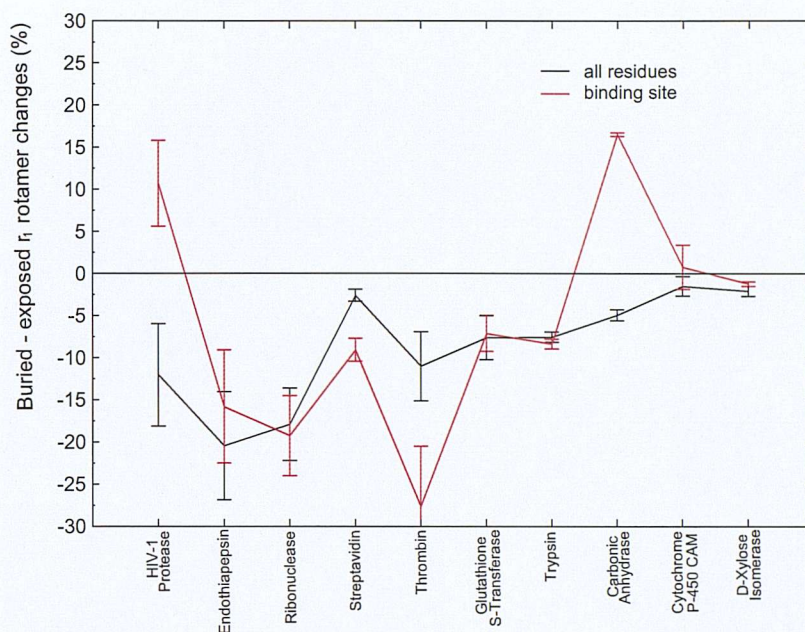


**Figure 4.26:** Differences between the  $\chi_1$  percentages of conformational changes occurring in buried and exposed residues (apo-/holo- comparisons); positive values reveal greater flexibility of buried residues. Conformational changes defined using Zhao *et al.* thresholds.

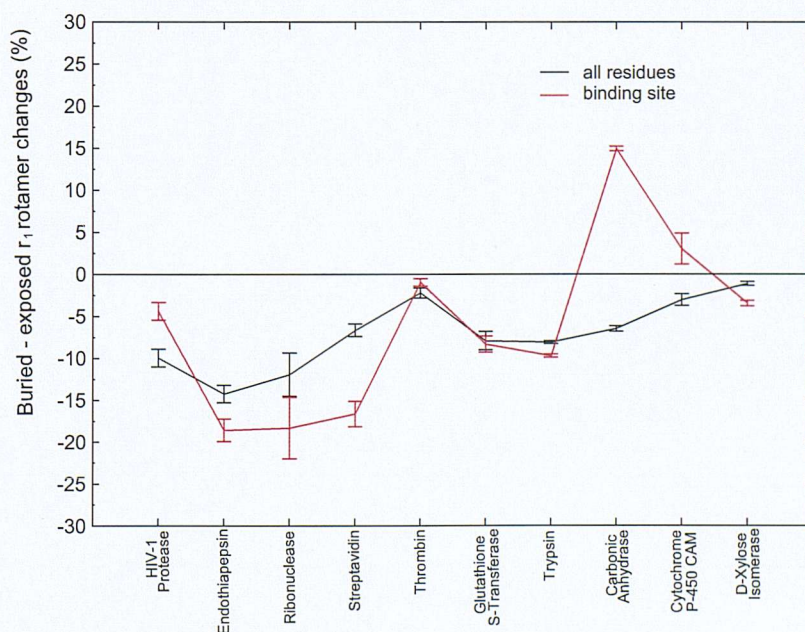


**Figure 4.27:** Differences between the  $\chi_1$  percentages of conformational changes occurring in buried and exposed residues (holo-/holo- comparisons); positive values reveal greater flexibility of buried residues. Conformational changes defined using Zhao *et al.* thresholds.





**Figure 4.28:** Differences between the  $\chi_1$  percentages of conformational changes occurring in buried and exposed residues (apo-/holo- comparisons); positive values reveal greater flexibility of buried residues. Conformational changes defined using Dunbrack and Cohen rotamer libraries.



**Figure 4.29:** Differences between the  $\chi_1$  percentages of conformational changes occurring in buried and exposed residues (holo-/holo- comparisons); positive values reveal greater flexibility of buried residues. Conformational changes defined using Dunbrack and Cohen rotamer libraries.

As expected, Najmanovich *et al.*<sup>46</sup> and Dunbrack *et al.*<sup>55</sup> methods show for most proteins greater conformational changes in the residues that are exposed to the solvent. An opposite trend is revealed by apo-/holo- comparisons in the binding site of cytochrome P-450 CAM, carbonic anhydrase and HIV-1 protease, in which buried residues show greater flexibility than exposed ones. Holo-/holo- comparisons confirm the same tendency for cytochrome P-450 CAM and carbonic anhydrase.

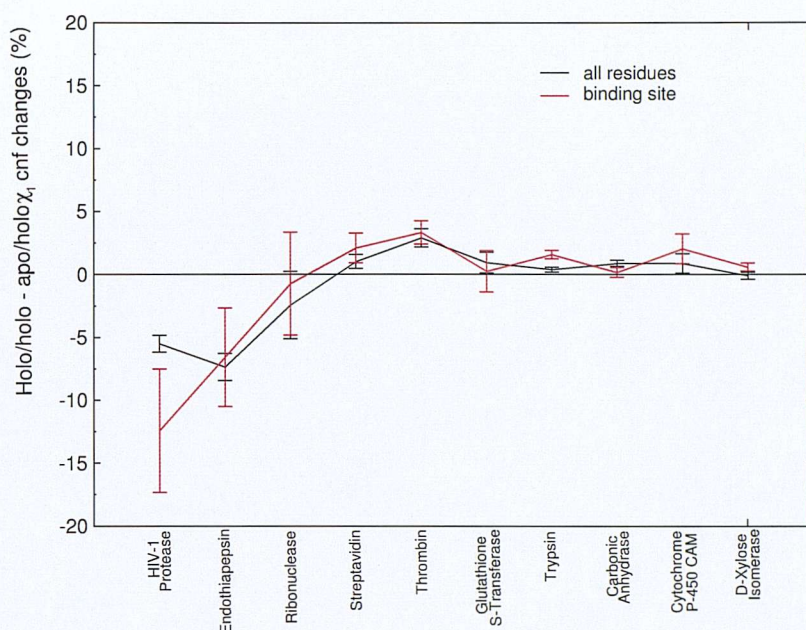
If the thresholds defined by Zhao and coworkers are instead applied, greater flexibility is in general detected for residues that are buried from the solvent. Again, buried residues in the binding site of HIV-1 protease, carbonic anhydrase and cytochrome P-450 CAM are significantly more flexible than exposed ones, both in apo-/holo- and holo-/holo- protein comparisons. In addition, the same trend is observed for thrombin and D-xylose isomerase (Figures 4.26 and 4.27).

In summary, the 10 analysed proteins can broadly be split in two classes: proteins in which binding site buried residues are the more flexible, and proteins in which binding site exposed residues move the most. HIV-1 protease, carbonic anhydrase and cytochrome P-450 CAM clearly belong to the first class; ribonuclease, glutathione S-transferase and trypsin consistently belong to the second one. Since Zhao *et al.* thresholds take into account residues' intrinsic flexibility, they are the most reliable; thrombin and D-xylose isomerase too can thus be considered part of the first class. Endothiapepsin, and streptavidin are borderline proteins.

### 4.3.5 Differences between apo-/holo- and holo-/holo- protein conformational changes

Figures 4.30, 4.31 and 4.32 report on the y axis the difference between the percentages of  $\chi_1$  conformational changes observed in holo-/holo- and apo-/holo- protein





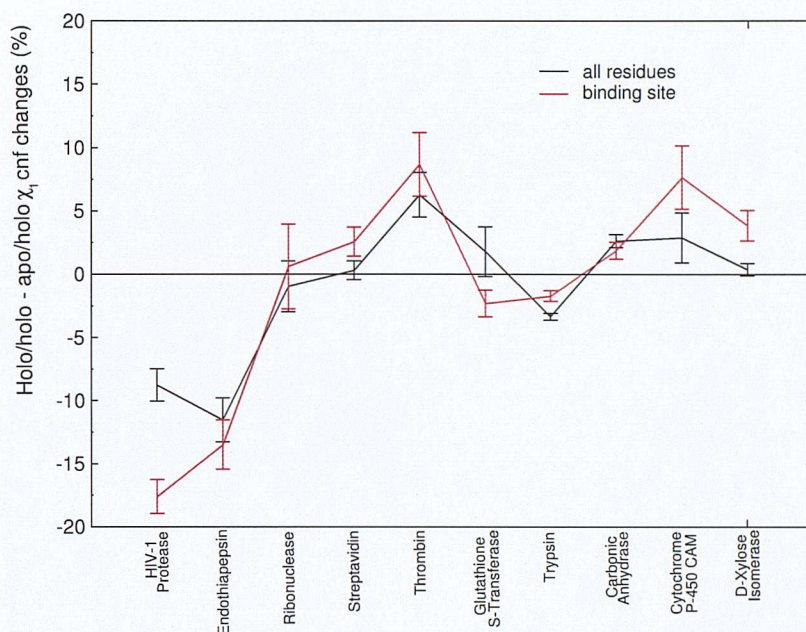
**Figure 4.30:** Differences between the  $\chi_1$  percentages of conformational changes occurring in holo-/holo- and apo-/holo- protein comparisons; positive values reveal greater conformational changes in pairs of holo-proteins. Conformational changes defined using Najmanovich *et al.* thresholds.

comparisons when the angular thresholds respectively defined by Najmanovich *et al.*, Zhao *et al.* and Dunbrack *et al.* are applied.

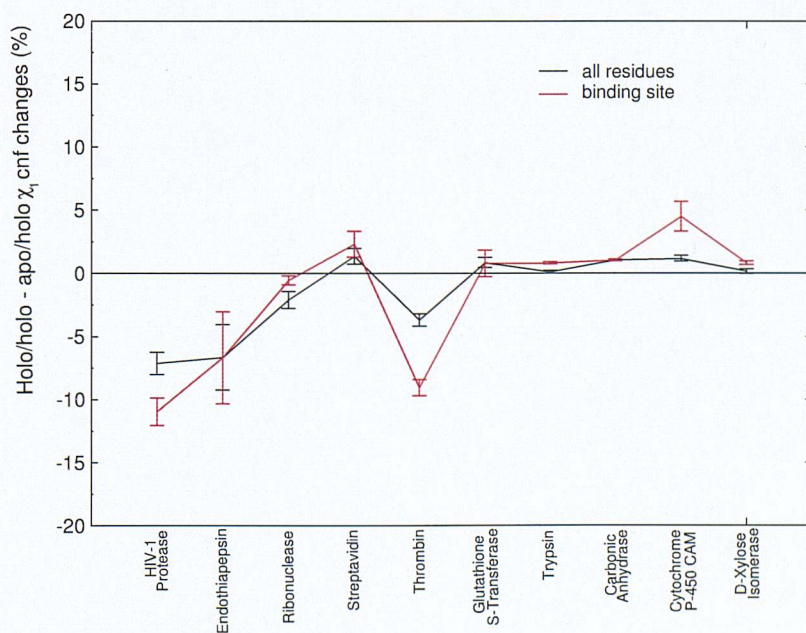
Again, by plotting differences rather than raw percentages on the y axis, consistently similar graphs are obtained for the three different methods of analysis, and trends that are more independent from arbitrary rotamer definitions are depicted.

In the aspartic proteases HIV-1 protease and endothiapepsin, significantly greater apo-/holo- conformational changes are detected with all methodologies of analysis, both in all residues and in only binding site residues. These two protein systems are the only ones for which backbone RMSd is significantly greater for apo-/holo- rather than holo-/holo- protein comparisons (section 4.3.1). Glutathione S-transferase (only binding site residues) and trypsin (all residues and binding site residues) show a similar trend with Zhao *et al.* angular thresholds; when Dunbrack and Cohen rotamer libraries are employed, this is also revealed for thrombin (all residues and only binding





**Figure 4.31:** Differences between the  $\chi_1$  percentages of conformational changes occurring in holo-/holo- and apo-/holo- protein comparisons; positive values reveal greater conformational changes in pairs of holo-proteins. Conformational changes defined using Zhao *et al.* thresholds.



**Figure 4.32:** Differences between the  $\chi_1$  percentages of conformational changes occurring in holo-/holo- and apo-/holo- protein comparisons; positive values reveal greater conformational changes in pairs of holo-proteins. Conformational changes defined using Dunbrack and Cohen rotamer libraries.

site residues).

Interestingly, the spread of the data is this time greater with the Zhao *et al.* methodology; the greater difference between all residues and only binding site residues' data might depend on the different effect that ligand binding exerts on different proteins and different parts of these proteins. In the next chapters, a deeper analysis of specific protein cases will be undertaken to confirm or exclude this hypothesis.

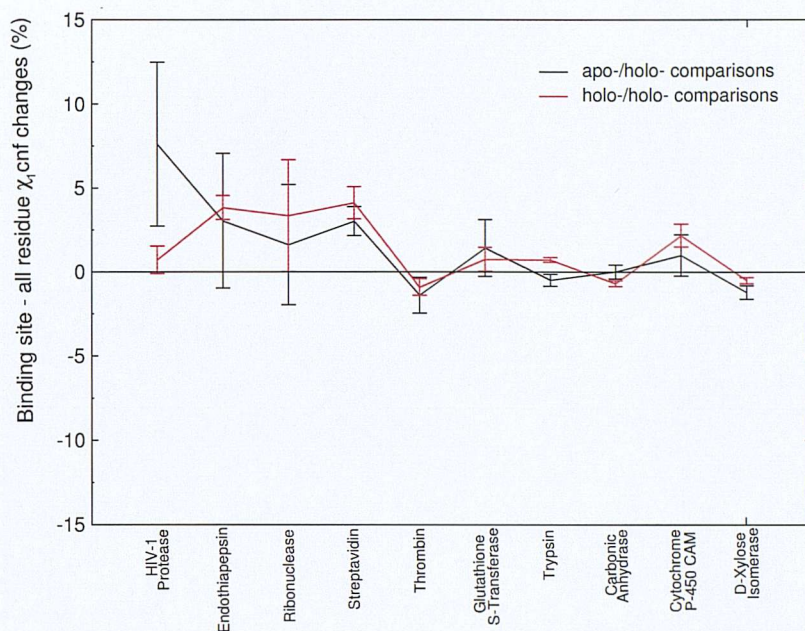
### 4.3.6 Differences Between All Residues and Only Binding Site Residues Results

It might be argued that the trends observed in the 10 analysed protein systems only depend on protein specific structures and/or characteristics, without any dependency on the effect and peculiarities of the ligand binding process. A possible approach to disentangle random, spontaneous conformational changes from those that actually depend on ligand binding is to compare the differences between the percentages of only binding site residues' and all residues' conformational changes.

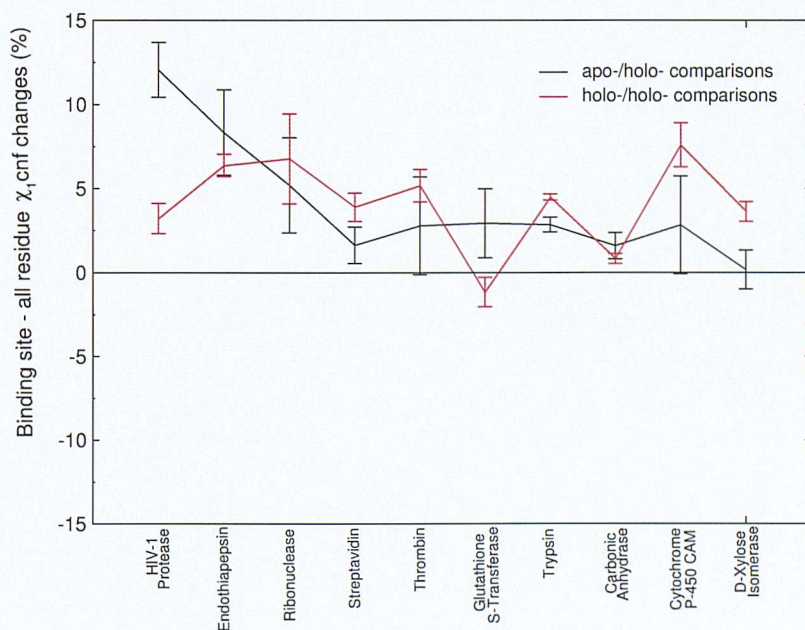
On the y axis of the graphs presented in Figures 4.33, 4.34 and 4.35, the differences between  $\chi_1$  conformational changes observed in the binding site and in all the residues of the 10 protein systems have been plotted.

When a 60° angular cutoff is applied (Figure 4.33), greater percentages of binding site residues conformational changes are observed for all proteins except thrombin (apo-/holo- and holo-/holo- comparisons), trypsin (apo-/holo- comparisons), carbonic anhydrase (holo-/holo- comparisons) and D-xylose isomerase (apo-/holo- and holo-/holo- comparisons). Binding site r1 rotamer changes are less than those observed in all residues only in the case of thrombin (holo-/holo- comparisons), glutathione S-transferase, trypsin and D-xylose isomerase (Figure 4.35), while the total percentages



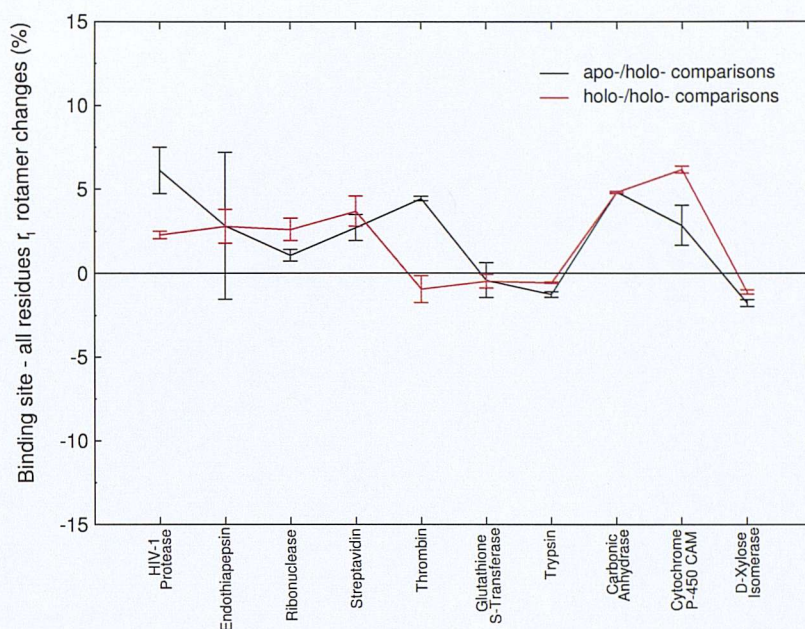


**Figure 4.33:** Differences between the  $\chi_1$  percentages of conformational changes occurring in the binding site and in all protein residues; positive values reveal greater flexibility of binding site residues. Conformational changes defined using a  $60^\circ$  angular cutoff.



**Figure 4.34:** Differences between the  $\chi_1$  percentages of conformational changes occurring in the binding site and in all protein residues; positive values reveal greater flexibility of binding site residues. Conformational changes defined using Zhao *et al.* thresholds





**Figure 4.35:** Differences between the  $\chi_1$  percentages of conformational changes occurring in the binding site and in all protein residues; positive values reveal greater flexibility of binding site residues. Conformational changes defined using Dunbrack and Cohen rotamer libraries.

of conformational changes defined using Zhao *et al.* specific angular thresholds are greater in the binding site of all protein systems but glutathione S-transferase (holo-/holo- comparisons) and D-xylose isomerase (apo-/holo- comparisons).

It is remarkable that there is a concordance between the results obtained applying the three different methodologies of analysis. HIV-1 protease, endothiapepsin, ribonuclease, streptavidin (i.e. the protein systems that generally show greater flexibility) and cytochrome P-450 CAM always show greater percentages of conformational changes in their binding site residues. Applying Zhao *et al.* specific angular thresholds, this trend is found in all protein systems but glutathione S-transferase.

Agreement between the different methods of analysis is found also when a distinction on the basis of the residues' accessibility to the solvent is made (Figures 4.24, 4.25, 4.26, 4.27, 4.28 and 4.29). When a  $60^\circ$  angular cutoff and Dunbrack and Cohen rotamer libraries are applied, only binding site residues generally show greater

conformational changes in their buried residues (HIV-1 protease, carbonic anhydrase and cytochrome P-450 CAM in Figures 4.24, 4.25, 4.28 and 4.29); the same proteins show the greatest differences between binding site and all residues employing Zhao *et al.* angular thresholds (Figures 4.26 and 4.27). Also, in those protein systems for which binding site exposed residues are more flexible than buried, the differences between the percentages of conformational changes that occur in exposed and in buried residues are with only few exceptions greater in the binding site of the proteins.

In summary, binding site residues are consistently more flexible than the whole protein, possibly reflecting an effect of ligand binding on protein side-chain conformations. This is generally more true for holo-/holo- rather than apo-/holo- comparisons; HIV-1 protease is the only protein system for which the difference between binding site and all residues conformational changes is consistently greater for apo-/holo- protein comparisons. When Zhao *et al.* angular thresholds are applied, the same trend is also revealed, to a lesser extent, for glutathione S-transferase and endothiapepsin. However, although consistent and possibly dependent on a systematic ligand-effect, these trends are arguably statistically insignificant; the error bars are quite large when compared to the absolute values of the plotted data. Deeper analyses of specific protein systems are necessary to infer correct conclusions from these graphs; their shapes and patterns could depend on genuine ligand binding effects, but also on specific characteristics and/or limitations of the studied data set (e.g. too few apo-structures).

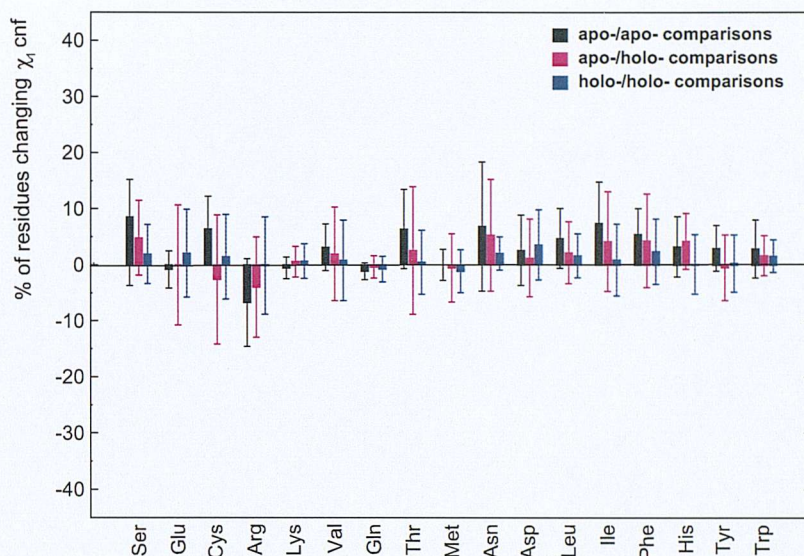


### 4.3.7 Comparisons of the Results Obtained for Structures Solved at 2.0 Å or better and 2.5 Å or better

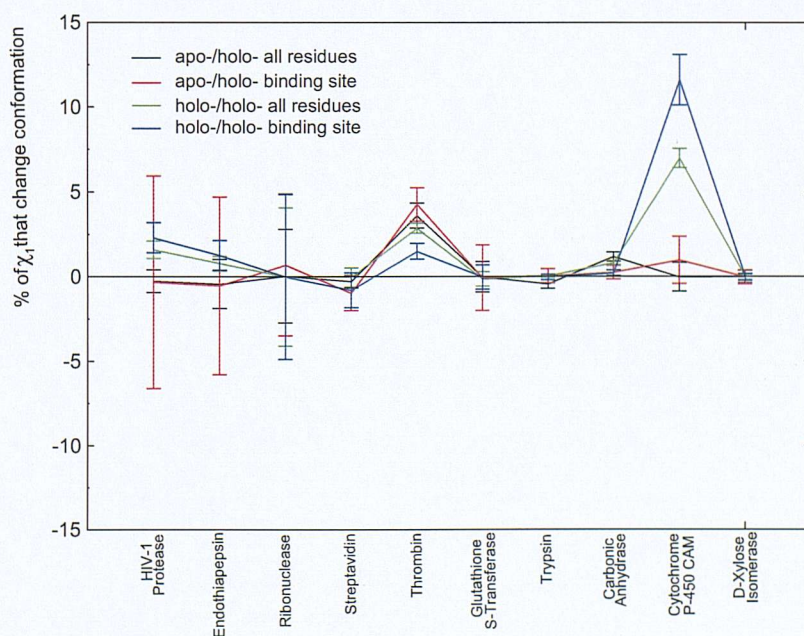
All the data reported so far are relative to structures solved at 2.0 Å or better. To analyse the dependency of the observed conformational changes on the PDB structures' resolution cutoff, a larger data set of structures solved at resolution up to 2.5 Å was also analysed (see Table 4.1). Graphs similar to those previously shown were obtained for this data set. Moreover, differential graphs plotting the difference between the percentages of conformational changes detected in the larger data set and the percentages of conformational changes observed in the subset of structures solved at resolution equal or better than 2.0 Å were obtained.

No significant differences were found between the flexibility trends of two data sets when the percentages of conformational changes of different residue types were compared across all protein systems. In fact, while structures solved at higher resolutions generally tend to show more flexibility in their side-chains, differential graphs such as that represented in Figure 4.36 are consistently characterised by error bars greater or equal to the corresponding difference values, not allowing any meaningful conclusions to be drawn.

If different protein families are distinguished, most protein systems show significant differences in the graphs that plot the percentages of conformational changes observed in the data set at the worse resolution minus the percentages of conformational changes observed in the data at the better resolution. These protein systems include thrombin and cytochrome P-450 CAM (see for example Figure 4.37), carbonic anhydrase and trypsin (for  $\chi^2$ ), and, to a minor extent, HIV-1 protease (Figure 4.37), endothiapepsin (for r1), streptavidin and glutathione S-transferase (for r2).



**Figure 4.36:** Differences between the percentages of different residue types that change  $\chi_1$  by more than Zhao *et al.* specific angular thresholds in structures solved at resolution  $\leq 2.5$  Å and  $\leq 2.0$  Å. Positive values reveal greater flexibility of PDB structures solved at higher resolution. Data relative to apo/apo-, apo/holo-, holo/holo- protein comparisons and only binding site residues are reported.



**Figure 4.37:** Differences between the percentages of  $\chi_1$  torsions that change more than  $\pm 60^\circ$  in structures solved at resolution  $\leq 2.5$  Å and  $\leq 2.0$  Å. Positive values reveal greater flexibility of PDB structures solved at higher resolution. Data relative to binding site residues and all residues in apo/holo- and holo/holo- protein comparisons.

Given the significant differences observed for the data sets at different resolution cutoffs, the use of protein structures at resolution greater than 2.0 Å is not recommended for the above mentioned protein systems, and especially for those for which the greatest and most consistent deviations are observed (i.e. thrombin, cytochrome P-450 CAM and trypsin). A significant influence of protein resolution on side-chain flexibility trends cannot be unequivocally ruled out in the case of the protein systems whose data sets at different resolution cutoffs show only little or no differences in their flexibility trends, but whose data sets' sizes are very different (see Table 4.1). For example, the differences in side-chain conformational changes detected in the case of streptavidin, for which only three holo-protein structures at resolution greater than 2.0 Å have been analysed, might well be more significant than those observed in the case of carbonic anhydrase, for which a similar number of structures at resolution equal or better than 2.0 and 2.5 Å have been studied.

In the case of xylose D-isomerase, no conclusion of any sort can be drawn, since all the analysed structures have resolution equal or better than 2.0 Å.

#### 4.3.8 Relationships Between Holo-Protein Side-Chain Conformational Changes and Ligand Similarities

To investigate the correlation between the similarity of ligands and the conformational changes they induce in the same apo-protein structure, Tanimoto similarity coefficients<sup>76</sup> were computed for all possible pairs of non-peptidic ligands. 1024-bit Daylight fingerprints and the default path range number of considered bonds (0-7) were employed.

In general, no significant correlations between the observed percentages of confor-

mational changes and the 2D-similarity of the corresponding ligands were found. The only trend which can be observed is that, in the case of HIV-1 protease, ribonuclease and cytochrome P-450 CAM, very similar ligands do not seem to generate large conformational changes in the corresponding holo-proteins. This observation should be tested however with a larger number of very similar ligands.

Some of the graphs describing the relationships between ligands' similarities and the percentages of conformational changes observed in the corresponding holo-protein pairs will be shown in the following chapters, which focus on specific protein systems. In the case of endothiapepsin, for which all ligands are peptidic, no reliable Tanimoto similarity coefficients can be calculated with 1024-bit Daylight fingerprints, and no graphs were produced.

## 4.4 Conclusion

Data sets of apo- and holo- protein structures at resolution equal to or better than 2.5 Å and a dataset of resolution equal to or better than 2 Å were chosen.

PDB entries were identified with FASTA searches; mutations in the protein sequences were occasionally accepted if they did not interfere with ligand binding and/or protein function. The ten protein systems belonging to the final data set are glutathione S-transferase, HIV-1 protease, carbonic anhydrase II, thrombin, cytochrome P-450 CAM, streptavidin, trypsin, D-xylose isomerase, ribonuclease A and endothiapepsin.

Most of the analyses described in this thesis were performed by code written by the author in Perl. Perl scripts were also employed to invoke C or C++ programs.

To study the dependence of side-chain flexibility on PDB structures' resolution, the results obtained for the two data sets at different resolution cut-offs (2.0 Å and 2.5

Å) were analysed. While the differences observed between the flexibilities of different residue types averaged across the ten protein systems are not significant, several systems show significant discrepancies when separate proteins are considered. This suggests that PDB structures solved at resolution greater than 2.0 Å should not be employed in the case of protein systems for which great differences between the two different data sets' results are observed (such as thrombin, cytochrome P-450 CAM and trypsin). Caution should also be applied when using structures solved at more than 2.0 Å for proteins that only show slight flexibility differences, and/or for which an insufficient number of structures solved at more than 2.0 Å were analysed (e.g. streptavidin). In the present thesis, only the results obtained for the data set at the best resolution are described throughout.

Side-chain conformational changes were defined on the basis of Dunbrack and Cohen rotamer libraries<sup>55</sup> and employing Najmanovich *et al.*<sup>46</sup> and Zhao *et al.*<sup>66</sup> angular thresholds. Conformational changes were evaluated for all possible apo-/holo-, holo-/holo- and apo-/apo- protein pairs. Buried and exposed residues were distinguished on the basis of their solvent accessible surface areas.<sup>74</sup>

Backbone Root Mean Square deviation was calculated using the program ProFit<sup>75</sup> for all possible pairs of structures within the 10 selected protein systems. Backbone rearrangements in the chosen data set seem to be generally small; the protein systems with the highest RMSd appear to be HIV-1 protease (apo-/holo- protein comparisons RMSd = 0.59 Å, holo-/holo- comparisons RMSd = 0.50 Å) and endothiapepsin (0.46 Å and 0.37 Å). The two aspartic proteases are also the only proteins for which apo-/holo- side-chain conformational changes are greater than holo-/holo- ones with all methods of analyses.

Generally speaking, no correlations were found between the similarity of ligands and the conformational changes they induce in the same apo-protein structures. In the case of HIV-1 protease, cytochrome P-450 CAM and trypsin, large side-chain conformational differences are not observed in holo-proteins bound to very similar ligands.

The flexibility order that is obtained for different residues types with Najmanovich *et al.* and Dunbrack and Cohen thresholds is broadly speaking in agreement with what one would expect; polar residues with long, non-bulky side-chains are the most flexible, while large, bulky aromatic residues seldom change conformation. Similar levels of flexibilities are detected in apo-/apo-, apo-/holo- and holo-/holo- protein comparisons.

Zhao *et al.* specific angular thresholds highlight unusual motions rather than the intrinsic propensities of residues to move; with this approach, residues such as Cys, Phe, Tyr and Trp are among the most flexible for the present data set. Also, the conformational changes identified by this methodology of study in apo-/apo- protein comparisons are significantly lower than those detected in complexed forms of the proteins.

HIV-1 protease, endothiapepsin and ribonuclease A are consistently the most flexible proteins in the data set. Streptavidin and thrombin follow, while the flexibility order of the remaining protein systems is subject to variations when different torsions are considered and different methods of study are employed.

In HIV-1 protease, endothiapepsin, streptavidin and cytochrome P-450 CAM, binding site residues are always more mobile than all protein residues. When a distinction between buried and exposed residues is made, trends are always more pronounced in binding site residues. If buried residues are always more mobile (HIV-1



protease, cytochrome P-450 CAM and carbonic anhydrase II), they are in the binding site. Similarly, for those protein systems in which exposed residues are instead always more flexible than buried, this tendency is always greater for binding site rather than all protein residues.

In summary, the three different methods of analysis here employed sometimes lead to consistent results, especially in the case of the most flexible proteins (HIV-1 protease, endothiapepsin, ribonuclease and streptavidin) and cytochrome P-450 CAM, for which all approaches of analysis generally detects similar trends. When the flexibility trends obtained with the three different methods of study appear to be totally inconsistent, different graphs, with differences rather than raw percentages plotted on the y axis, can often reveal similar trends. However, many discrepancies in the results are obtained, and significant “noise” in the observed conformational changes must be expected.

It is probable that, among the methods employed in this thesis, Zhao *et al.* specific angular thresholds are the most likely to detect systematic, ligand-binding dependent side-chain conformational changes, for a number of reasons. First, since they were obtained by comparing identical pairs of apo-proteins, they should reflect the intrinsic flexibility of residues under the influence of their given local environments, and not average out the effects of local backbone and side-chain interactions as do most rotamer libraries and side-chain flexibility studies do. Moreover, the results that are obtained by applying these angular thresholds to the data set of this thesis seem to be the most consistent. For example,  $\chi_1$  binding site conformational changes revealed by this method are greater than all residues conformational changes in 18 out of 20 of the apo-/holo- and holo-/holo- protein comparisons (Figure 4.34). Also, the significantly lower flexibility of residues detected by this methods in apo-/apo- protein compar-



isons, suggests the hypothesis that this approach could be the most reliable, helping to identify motions that are genuinely ligand-binding dependent rather than random, spontaneous ones.

It must however be remembered that the percentages of conformational changes obtained in this thesis with Zhao *et al.* angular thresholds might inevitably be an overestimation of genuine ligand binding induced fit. First, for the way in which Zhao *et al.* angular thresholds were defined, 10% of the different residues' side-chain torsions fell outside these thresholds even when a relatively restricted data set was analysed.<sup>66</sup> Moreover, the selection criteria and the size of Zhao *et al.* data set differ from the ones employed in this thesis. Also, Zhao *et al.* angular thresholds are available only for  $\chi_1$  torsions.

In the next chapter, side-chain conformational changes occurring in HIV-1 protease will be analysed in depth. Their characteristics and trends, and their comparisons with other workers' results, can provide useful insights on ligand binding induced fit in specific proteins.

## Chapter 5

# HIV-1 Protease

---

### 5.1 Introduction

The final aim of this thesis is to identify and characterise ligand binding induced fit in proteins, and to disentangle systematic ligand effects from random protein motions. Highly flexible proteins that undergo greater conformational changes in their binding sites are probably more likely to show genuine ligand-binding induced fit. Since HIV-1 protease is the most flexible protein of the present thesis data set and its binding site residues undergo greater side-chains rearrangements than all protein residues with all methodologies of study<sup>46,55,66</sup> (see chapter 4), this aspartic protease was chosen for the more in depth conformational analyses described in this chapter. Many previous studies of HIV-1 protease have been carried out to define the backbone flexibility and the nature of its ligand-binding process; the reliability of the methods employed in this thesis, and the extent to which they provide similar or novel data on protein

flexibility, will be assessed by comparing their results with previous work.

## 5.2 HIV-1 Protease and Aspartic Proteases: Structure

Aspartic proteases are a family of widely distributed enzymes found in vertebrates, fungi, plants and retrovirus. They are all characterised by specificity for extended peptides and for being inhibited by pepstatin, an amino acetylated peptide of sequence Iva-Val-Val-Sta-Ala-Sta (where Iva stands for isovaleric acid and Sta for the unusual amino acid statine), and by a low optimum pH. They all catalyse the cleavage of a peptidic bond, employing two catalytic aspartate residues in the protein binding sites.

Some members of this family, such as the aspartic protease encoded by the human immunodeficiency virus (HIV-1 protease), or renin (that plays a crucial role in the regulation of blood pressure in mammals), have recently become the object of intensive studies because of their medical relevance.

Aspartic proteases are characterised by two adjacent and coplanar aspartic side chains in their active site. They can be categorised in two sub-families: pepsin-like proteases, which consist of two similar but not identical lobes, and retroviral proteases (retropepsins), which are dimers consisting of two identical subunits.

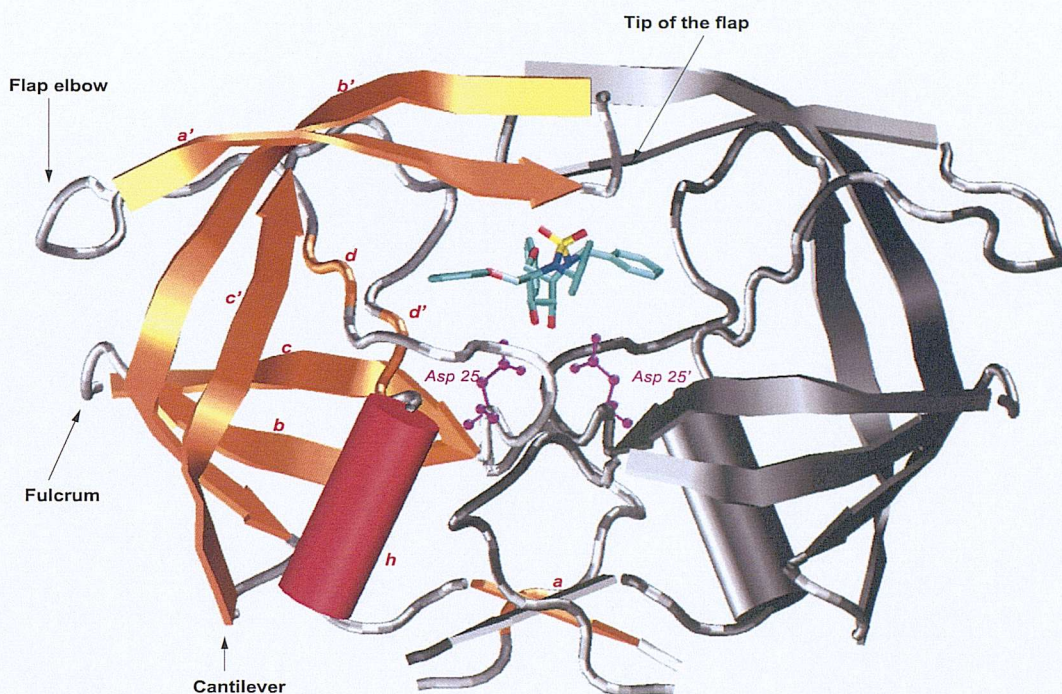
Structural changes occurring in aspartic proteases upon ligand binding are generally minor, with the exception of the pepsin-like enzymes' single  $\beta$ -hairpin loop known as the "flap" and the two equivalent flaps in retropepsins, which can move by up to 9 Å between the free and inhibited forms of an enzyme.<sup>78</sup> In all pepsin-like aspartic proteases, the conserved residue Tyr 75 (pepsin numbering), located near the tip of the flap, has been postulated to be involved in the capture and cleavage of the substrates.

As a member of the family of aspartic proteases, the catalytic activity of HIV-1 protease depends on a dyad of aspartic acid residues at the centre of the active site, Asp25 and Asp25', part of the highly conserved catalytic triad sequence Asp-Thr-Gly. As for the other members of its family, this enzyme shows a pH-dependent catalytic activity (optimum catalytic constant at pH 5-6) and the characteristic two-domain structure.

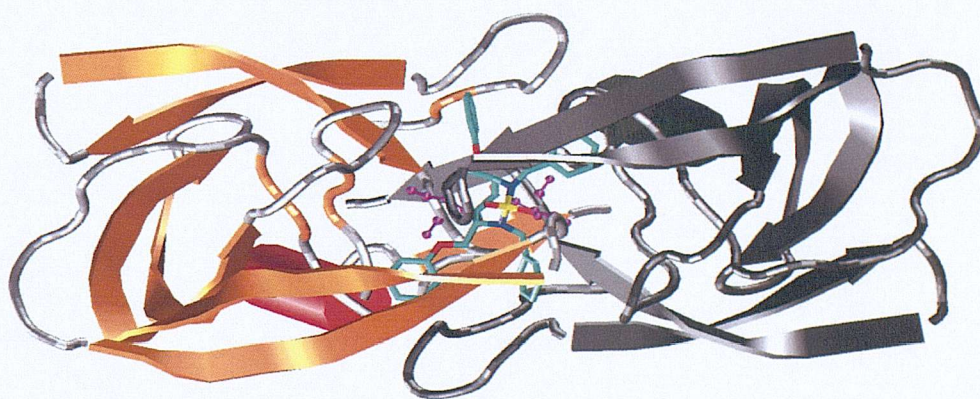
The function of HIV-1 aspartic protease is to hydrolyse viral polyproteins into functional protein products that are essential for viral assembly and activity. Since the inactivation of this enzyme results in the production of immature, non infectious viral particles, HIV-1 protease is an excellent target in anti-AIDS drug design, and has been the focus of intensive research. Several HIV-1 protease inhibitors are already employed as effective drugs which slow the growth of the virus. Drugs bind tightly to the protease, blocking its action, and the virus perishes because it is unable to mature into its infectious form. The first atomic structures of HIV-1 protease were reported in 1989; now, hundreds of structures are available in the PDB and in the proprietary databases of pharmaceutical companies, making HIV-1 protease one of the best-studied enzymes.

The protein is a symmetric homodimer; each monomer contains 99 amino acids and comprises one  $\alpha$ -helix and two antiparallel  $\beta$ -sheets.<sup>79</sup> Aliphatic residues stabilise each monomer in a hydrophobic core; moreover, the dimer is stabilised by noncovalent interactions, hydrophobic packing of side chains and interactions involving the catalytic residues. The active site of the protein is located in a deep cleft at the dimer interface, approximately at the centre of the molecule, and is covered by two extended turns of  $\beta$ -sheets (the so-called "flaps"), which are essential for HIV-1 protease flexibility and ligand binding (Figure 5.1).





(a) HIV-1 protease secondary structure. In each monomer, N-terminal  $\beta$ -strand *a* (residues 1-4) is followed by a turn and  $\beta$ -strand *b* (residues 9-15). After another turn and  $\beta$ -strand *c* (residues 18-23), the loop that contains the catalytic triad (Asp25-Thr26-Gly27) is found.  $\beta$ -strand *d* and the extensive loop comprising the so-called “elbow” (residues 38-42) follow; they form the “flap claws” together with  $\beta$ -strands *a'* and *b'* (residues 43-49 and 52-58) and residues 49-51 (the “tip of the flap”).  $\beta$ -strand *c'* comprises residues 69 to 78; a loop at residues 79-82 continues into strand *d'* (residues 83-85). Helix *h* (residues 86-94) and  $\beta$ -strand *e* (residues 95-99) are found C-terminal of the protein. Several studies<sup>79–81</sup> have identified a hinge between two  $\beta$ -strands of HIV-1 protease, in the region between residue 11 and 21 (“fulcrum”).



(b) Top view. The flaps are very flexible in solution; they cause regular exposure of the otherwise inaccessible active site and, together with the catalytic loops between  $\beta$ -strand *c* and  $\beta$ -strand *d*, form a sort of “claw” which is essential to ligand binding.

**Figure 5.1:** Cartoon representation of the homodimer HIV-1 protease complexed to cyclic sulfamide inhibitor AHA006 (1AJV PDB entry).  $\beta$ -sheets and helices are coloured in orange and red; catalytic aspartate residues Asp 25 and Asp 25' are represented in balls and stick.



The motion of the flaps allows or prevents the access of natural substrates and inhibitors into the active site pocket. Several studies classified this conformational change as a predominantly hinge motion of fragments smaller than domains; the hinge has been located in the “fulcrum” region, comprising residues from 11 to 21 (Figure 5.1). Also, interdomain correlated motions suggest that the flap motion occurs with a compensatory change in residues 59-75, that act as a sort of a “cantilever” (Figure 5.1); the flap is observed to close down as the cantiliver moves up, and some mutations in the cantiliver residues deactivate the protease and give non infectious virus particles, suggesting a key role of this region in the functional energetics of the protein.<sup>81</sup>

In the “open” conformation, the tips of the flaps are approximatively 7 Å far from each other, leaving enough space for the substrate polypeptide chains or inhibitors to fit into the active site. In the ligand-bound form of the enzyme, the so-called “closed” conformation, the flaps are instead folded over the inhibitor, held in this position by hydrogen bonds between Ile 50 and Ile 50' (NH groups) and a water molecule (hydrogen bonded to the inhibitor) or to the ligand itself.

The open form of HIV-1 protease was once believed to be the only conformation of the unbound form of HIV-1 protease. However, NMR studies have shown the flap tips to be highly mobile in solution, adopting a variety of conformations from fully closed to open.<sup>82</sup> Moreover, “closed-flap” X-ray structures of unliganded HIV-1 protease have been solved and deposited in the PDB. The PDB entries of these structures are 1G6L<sup>83</sup> (the apo-protein employed in the present thesis) and 1LV1, respectively solved at 1.90 Å and 2.1 Å resolution.<sup>84</sup> While the first is a single C95M mutant that conserves the native protein’s activity and backbone structure,<sup>83</sup> the second is

a double mutant (C95M/C1095A) in which the mutation of Cys 1095 causes the movement of catalytic residues 23-26/1023-1026 towards the flap. This motion might facilitate the access of substrates to the active residues in the binding site cleft; in fact, a greater autolysis rate of the double mutant protease was found, indicating that the subtle movement of the catalytic residues is of extreme importance for the enzymatic activity.<sup>84</sup>

The conformation and the dynamics of the flaps have been probed by theoretical<sup>85</sup> opening in HIV-1 protease. An initial impulse for flap opening was provided by applying harmonic restraints to non-flap residues; within 200 ps of simulation, the two flaps opened to a 25 Å gap, following backbone conformational changes at Lys 45, Met 46, Gly 52 and Phe 53. In contrast, similar molecular dynamics simulations on the M46I mutant, which is associated with drug resistance, indicates that this mutation stabilises the flaps in a closed conformation. Some theoretical calculations estimate that the open and the closed flap conformations have similar potential energies and are thus equally accessible for native HIV-1 protease.<sup>86</sup> while others estimate that the difference in potential energies between the two forms of the enzyme is 3.0 kcal/mole, the open conformation being favoured entropically and enthalpically.<sup>87</sup> In general, it is believed that the flaps occupy a shallow energy minimum, so that small perturbations and/or crystallographic conditions can easily affect their conformation. It is likely that crystal contacts are responsible of trapping the flaps of HIV-1 protease in the open conformation.<sup>2</sup>

In addition to large flap motions, HIV-1 protease undergoes substantial conformational changes in its binding site, as its cleft site tightens around a substrate. This allows the protease to change its shape and accommodate ligands with widely

different shapes and volumes, often assuming asymmetric conformations in its two monomers.

When closed- and open- flap structures are superimposed, the C $\alpha$  and backbone RMSd are significantly greater than 1 Å; the average backbone RMSd is instead less than 0.6 Å for apo-/holo- and holo-/holo- protein comparisons in the present dataset (see section 4.3.1). The comparison of ligand-bound structures of HIV-1 protease with the closed-flap apo-protein 1G6L is expected to be more rational and useful in identifying ligand-binding induced conformational changes.

## 5.3 Background to Protein: Past Work

### 5.3.1 Structure and Dynamic Behaviour of HIV-1 Protease: X-ray Structure Comparisons, Molecular Dynamics and Normal Mode Analyses

In a recent study by Zoete, Michielin and Karplus,<sup>79</sup> the flexibility of different regions of HIV-1 protease was examined as a model system for the analysis of protein flexibility. A database consisting of 73 X-ray structures of HIV-1 protease differing in terms of sequence, ligands or both was used, and the root-mean-square differences of their backbone calculated. These results were compared with those obtained by molecular dynamics simulations, normal mode analysis, and X-ray B-factors. Finally, the various approaches were used to examine the correlations between different parts of the structure.

HIV-1 protease complexes with identical sequences were collected into families. The family containing the largest number of members (25) is the one having the same sequence as the unliganded 3PHV structure, an apo-protein showing an open

flap conformation solved at 2.70 Å (hence not included in the present thesis data set). This sequence was considered the consensus sequence of structures, and 3PHV used as a template of the apo-protein structure. Among the three pairs of complexes sharing the same ligand, the structures with the worst resolution were excluded from the general analysis to eliminate redundancy. However, they were employed to calculate variations between X-ray structures of identical complexes. The complexes of this family are representative of the great diversity of HIV-1 protease ligands; they include symmetric and asymmetric molecules, cyclic urea and cyclic sulfonamide, peptide-like linear molecules, penicillin derived molecules, etc.

For all 73 structure, the global average RMSD for backbone atoms was found to be 0.56 ( $\pm 0.15$ ) Å, i.e. slightly larger than the average RMSd for pairs of HIV-1 protease holo-structures analysed in this thesis (0.50 Å). The backbone RMSd of each residue was found to range from 0.25 Å for the most stable residues of the active site (23-32) to 1.0-1.4 Å for the most variable regions of the protease, i.e. the flap elbows around residue 40, and residues 11-21, where a hinge between two  $\beta$ -strands can be located (“fulcrum”). Two other regions that show structural variations are residues 79-83, i.e. the outer part of the active site in contact with the solvent, and the loop between  $\beta$ -strands  $b'$  and  $c'$  in Figure 5.1 (the “cantilever”, residues 65-72,). The most rigid zone, i.e. the region containing the active site triplet, is localised in a loop stabilised by a network of hydrogen bonds. Residues in the binding site belong to both rigid and flexible regions; residues 8, 10, 23, 25, 27-30, 32 and 84 are located in regions that appear to undergo only small variations, whereas residues 47-50 and 81-82 are in more variable zones.

For the 22 consensus sequence HIV-1 protease complexes, the RMSd average and

distribution were essentially the same. Also, despite smaller RMS deviations, the mean global RMSd for backbone atoms of complexes with the same ligand and different sequences are similar to both those observed for the whole set of 73 complexes with different ligands and sequences and for the 22 complexes set with different ligands and identical sequence.

For PDB structures sharing identical ligands and sequences (i.e. 1HSG and 2BPX, 2BPV and 2BPW, 2BPY and 2BPZ PDB entries), the global average RMSd was 0.42 Å between 1HSG and 2BPX, and only 0.13 Å between 2BPY and 2BPZ (the authors did not mention data for the comparison of the two structures 2BPV and 2BPW). The deviations between the structures 1HSG and 2BPX are larger around residues 18, 40, 52, 68 and 82, similar to what was observed for the other sets of experimental structures. In the case of 2BPY and 2BPZ, the differences are essentially at a noise level, though somewhat larger around residues 18, 32, 40, 68 and 80. All these structures have been obtained by the same group; however, the crystal structure for 1HSG was obtained in 1994 by soaking a crystal of HIV-1 protease with a liquor that differed from that employed to obtain the others (solved in 1998). Zoete *et al.* suggest that this could explain the larger RMSd between 1HSG and 2BPX.

Since the trends of the RMS deviations found in the different crystal structures are similar, independent from the specific nature of the stimuli (such as the ligand and buffer condition), the authors concluded that the variation the RMSd as a function of residue number is an inherent property of the protein energy surface rather than a consequence of different ligand structures and/or small variations of the protein sequence or crystallisation conditions.

400 ps dynamics molecular dynamics simulations (MD) were also conducted on



the enzyme complexed with six ligands belonging to different structural families (PDB entries 1HVI, 1AJX, 1HOS, 1OHR, 1HPX, 1HSG) and forming tight complexes with HIV-1 protease. For all the six compounds except one, a water molecule is present in the active site, establishing hydrogen bonds with the ligand and residue Ile 50 of both chains; the exception (1AJX) is a complex formed by a cyclourea, whose structure was developed to replace this active site water molecule. One complex (1HPX) was chosen since both X-ray and NMR structures are available; the others were selected on the base of the symmetry (1HVI, 1AJX, 1HOS) or asymmetry (1OHR, 1HPX, 1HSG) of their structures.

RMS fluctuations for backbone atoms of each residue of the two monomers of the PDB complex 1HPX were calculated and compared to experimental atomic RMS fluctuations. These can be obtained from the PDB B-factor using the formula:

$$B = \frac{8}{3}\pi^2\langle\Delta r_i^2\rangle \quad (5.1)$$

where  $B$  is the B-factor and  $r_i$  is the RMS fluctuation of atom  $i$ .

Zoete *et al.* found an overall agreement in trend and magnitude between the fluctuations calculated with the MD for backbone atoms and the experimental B-factors. The experimental NMR parameters were in agreement too with the MD fluctuations; moreover, the RMS fluctuations per residue obtained with the six simulations essentially showed the same trends and amplitudes as those obtained from crystal structure comparisons.

Some differences were observed in the different complexes that can be explained on the basis of the different inhibitor/protease interactions; for example, symmetric ligands and ligands occupying the sub-sites of HIV-1 protease in a symmetric way result in a similar dynamic behaviour of the two monomers, while asymmetric ligands

appear to result in different kinds of motions in the two monomers. However, the trends for all the 6 complexes are similar, and the authors suggest that the dynamic behaviour of HIV-1 protease backbone in protease/ligand complexes is mainly influenced by the intrinsic flexibility of the protein, while specific interactions between the protease and ligands only play a secondary role. In fact, the structure variations observed in the different complexes remain small. In the opinion of the authors,<sup>79</sup> the regions of the protease that are in contact with the ligands generally show less RMSd fluctuations than the other parts of the protein.

In the opinion of Zoete and co workers,<sup>79</sup> the high similarity between the deviations found in different equilibrium crystallographic structures and the magnitude of the fluctuations obtained from the MD simulations is a confirmation of the hypothesis that the X-ray structures of the HIV-1 protease complexes are likely to correspond to different local minima on the potential energy surface of the protease. This is expected to consist of multiple local minima that differ little in energy and are separated by low energy barriers. Moreover, these results demonstrate that information on the dynamic properties of a given protein can be obtained by comparing different crystal structures; this a rational basis for the analyses performed in this thesis.

Normal mode analysis of the native closed HIV-1 protease, both in the presence (PDB structure 1HVI) and absence of a ligand (3PHV), was also carried out by Zoete *et al.* The resulting RMS fluctuations per residue number were very similar to those obtained by the other approaches. As expected, their amplitudes were however smaller; larger-scale fluctuations in proteins involve significant anharmonic contributions, related to the multiminima character of the potential surface. The results obtained for the protease in the absence of the ligand and of the active site water molecule show RMS fluctuation per residues that are essentially similar to that ob-

served in the presence of the ligand. The fluctuation around residues 41 and 53 are a bit larger without the ligand, indicating a higher mobility of flaps and flaps' elbows in the unliganded protease. Large localized asymmetries were found in the absence of the ligand between the two monomers in a loop, the tip of the flap, and proline 80 (end of the active site, in contact with the solvent).

Zoete *et al.* concluded that ligands have little systematic effect on backbone RMSd of HIV-1 protease, and that the backbone dynamic behaviour of this protein is mainly influenced by the intrinsic flexibility of the protein.

### 5.3.2 Structure-Based Thermodynamic Study of HIV-1 Protease Inhibitors

Bardi, Luque and Freire applied structure-based thermodynamic analyses (see appendix A) to quantitatively parametrize and predict the energetics of binding of 13 inhibitors to HIV-1 protease.<sup>88</sup> Residue stability constants<sup>89–91</sup> were obtained for the aspartic protease applying the algorithm COREX (see appendix B) on the structure of the ligand-free protein (PDB entry 1HHP) and of the holo-protein structures after the ligands were removed (PDB entries 1HVI, 1HVJ, 1HVK, 9HVP, 1HVL, 1HPV, 1SBG, 1HBV, 1HPS, 2UPJ, 1GNO, 1PRO and 1HIH). The states employed to calculate the stability constants were generated with a sliding block of windows of 16 amino acids each.

While the generic portion of the Gibbs energy  $\Delta G_{gen}$ , that arises from the formation of secondary and tertiary structure (van der Waals interactions, hydrogen bonding, hydration and conformational entropy) was calculated by separating its enthalpic and entropic components, the additional contributions to the Gibbs energy of

binding ( $\Delta G_{ion}$ , originating from ionization effects, and  $\Delta G_{tr}$ , due to the change in translational degrees of freedom) were not.

The authors succeeded in predicting their free energies of binding with a standard deviation of  $\pm 1.1$  kcal/mol and an uncertainty of  $\pm 10\%$ ; the correlation between the predicted and the experimental free energies of binding was very good, yielding a slope of 0.982 and a correlation coefficient of 0.85.

In the present thesis, the apo-protein 1HHP was discarded since its resolution is 2.70 Å, i.e. well above the cutoffs applied to the data set. Of the 13 holo-proteins studied by Bardi *et al.*,<sup>88</sup> only 9 of the holo-proteins were included in the present data set; 9HVP and 2UPJ were excluded as their resolution is higher than 2.5 Å. Also, 1SBG and 1PRO were not included in the study since their sequences differ from that of the other holo-proteins, and the number of holo-protein structures analysed in the present thesis was considered sufficient.

According to Bardi *et al.*, the binding of the 13 inhibitors to the enzyme is dominated by the hydrophobic effect. In fact, both the inhibitor and the protease bury a significant non-polar surface upon ligand binding; the average fraction of non-polar area buried from the solvent in the protein is  $0.737 \pm 0.02$ , i.e. much more than the fraction that is normally buried by a globular protein upon folding (0.55-0.60). In agreement with these observations, the major contribution to  $\Delta G_{binding}$  is provided by the favorable entropy resulting from the release of water molecules associated with the desolvation of those surfaces.

Given the highly hydrophobic nature of the inhibitors and their lack of strongly polar groups, the electrostatic interactions contribute very little to the intrinsic enthalpy of binding, and hence to the free energy of binding. The only significant electrostatic contributions are established by Asp 25, 29 and 30, which can contribute up to 0.7

kcal/mol to the ligand binding process, depending on the given inhibitor.<sup>88</sup>

The binding pocket was mapped according to the energetics of binding. Broadly speaking, the same residues contribute to the binding energetics, although with different contributions with the different inhibitors. This reflects the fact that all the inhibitors in the set target the same site on the protease. If the different energetic contributions of single residues are considered, the binding site is defined by residues belonging to four non contiguous regions in the protein. First, the amino acids in the zone comprising the catalytic aspartate, i.e. Asp 25, Gly 27, Ala 28, Asp 29 and Asp 30, are the major contributors to the binding energetics. Also, the flap region (Met 46, Ile 47, Gly 48, Gly 49, Ile 50), the strand between residues 80-86 (especially Pro 81, Val 82 and Ile 84) and Arg 8 significantly contribute to the favourable free energy of binding. Bardi *et al.* found that, probably because of the roughly symmetrical nature of the inhibitors, the two chains of the protease contribute in a symmetrical way to the energetics of binding.<sup>88</sup> The catalytic zone always gives the most important contribution to the energy of binding.

The residue stability constants calculated with the COREX algorithm<sup>89,91</sup> indicate the two regions comprising residues 23-32 (i.e. the ones surrounding the catalytic triad) and 82-92 (i.e. residues that are part of the *h*  $\alpha$ -helix in Figure 5.1) as the most stable regions in the protease. These regions, close to each other in three-dimensional space, are the main components of the hydrophobic core of the protein and are part of its dimerization interface; the active site triad is in particular located in the most stable part of the molecule.<sup>88</sup> These results are in agreement with those obtained by Zoete *et al.*<sup>79</sup> and by crystallographic analyses in general; the catalytic triad is involved in an intense network of hydrogen bonds, and the loop on which it is located strongly interlocks with the symmetric counterpart by extensive hydrophobic inter-



actions. The region of the protease showing the lowest stability constants appeared instead to correspond to the flaps (residues 40-60), which appear to be unstructured even under native conditions. Again, this result is in accordance with all studies carried out on HIV-1 protease.<sup>2,79,81</sup> The stability constants of this region were similar both when the unbound structure of the enzyme and the bound protein structures after the ligands had been removed were employed, proving that in the bound complexes the flap is stabilized by interactions with the inhibitor, and not with the protein.

The residues that contribute to the energetics of binding appeared to belong to both the most stable (residues Asp 25, Gly 27, Ala 28, Asp 29, Asp 30, Pro 81, Val 82, Ile 84) and to the least stable regions of the protease (residues Met 46, Ile 47, Gly 48, Gly 49, Ile 50).<sup>88</sup> The former exists in the active, ligand-bound conformation even in the absence of the ligand; the latter is unstructured before binding and is forced to assume a precise conformation only by the interaction with the inhibitor. This dual character of the binding site from a flexibility point of view<sup>92</sup> appears to be essential to the efficacy and to the dynamics of binding; residues that can energetically influence the ease of motion of the flaps, and that are not even in direct contact with the inhibitor, can influence the overall binding process.<sup>92</sup>

## 5.4 Results: HIV-1 Protease Side-Chain Conformational Changes

The holo-proteins analysed in this thesis (see Table 4.1) share the sequence that can be defined as the consensus sequence for HIV-1 protease; each residue number of their chain is occupied by the amino acid with the highest probability of presence at that position.<sup>79</sup> 1G6L, the PDB entry employed as reference apo-protein structure in this thesis, is a single C95M mutant of HIV-1 protease that conserves the activity and



**Figure 5.2:** Ribbon representation of an apo-protein structure of HIV-1 protease (PDB entry 16GL, grey) superimposed to a holo-protein structure of the protease (PDB entry 1AJV, purple; the structure of the sulfamidic inhibitor AHA006 is shown). The RMSd of all backbone atoms in the two protein structures is 0.68 Å.

backbone structure of the native protein<sup>83,84</sup> (see section 5.2). While all the other solved apo-protein structures of HIV-1 protease are characterised by an open-flap conformation and, more importantly, a resolution greater than 2.5 Å, 1G6L structure presents a closed-flap conformation of HIV-1 protease, and was solved at 1.90 Å.

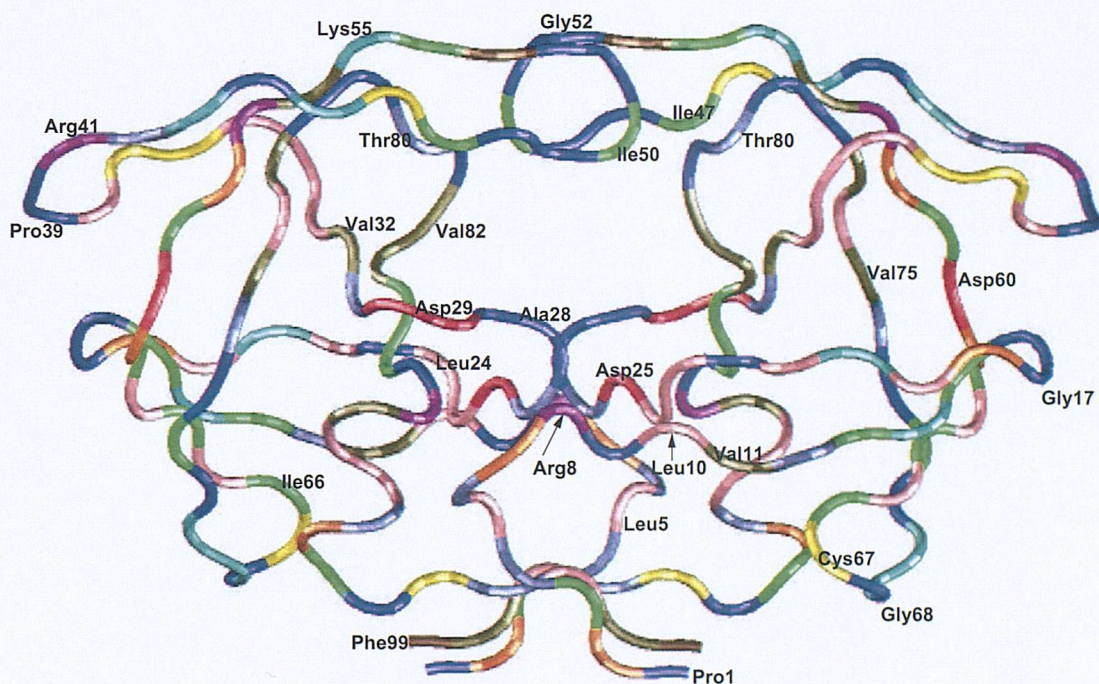
Figure 5.2 shows the superimposed ribbon representations of the apo- and holo-proteins 1G6L and 1AJV PDB entries respectively coloured in grey and purple. When a closed-flap conformation of the protein is considered, backbone conformational changes occurring upon ligand binding are not dramatic, even if somewhat larger than the average conformational changes generally detected in identical pairs of proteins at good resolution.<sup>79</sup>

Previous studies have described the ligand-induced conformational shift of HIV-1 protease as a predominantly hinge motion of the flaps, whose hinge is located



in the “fulcrum” region, comprising residues 11-21 (Figure 5.1). Also, many other interdomain correlated motions have been described for this enzyme;<sup>79–81,93</sup> however, while previous studies have focused their attention on backbone RMS deviation of HIV-1 protease, the aim of this section is to investigate side-chain conformational changes, and, eventually, their similarity and/or discrepancy with backbone RMS deviations of the protein backbone.

Figure 5.3 indicates some residue sequence numbers and types on a tube representation of HIV-1 protease (PDB entry 1AJV), to help identify some of the protein regions discussed in this chapter.



**Figure 5.3:** Tube representation of residue-name coloured HIV-1 protease. Some residue names and sequence numbers are indicated to help identify regions discussed in this section.

#### 5.4.1 All Environments, and Environment Specific Conformational Changes

Figures 4.12-4.17 in chapter 4 report on the y axis the percentages of  $\chi_1$  and  $\chi_2$  side-chain conformational changes observed in the protease when no distinction between

exposed and buried residues is made, and a 60° angular cutoff,<sup>46</sup> Zhao *et al.* angular specific thresholds<sup>66</sup> or Dunbrack and Cohen rotamer libraries<sup>50,55</sup> are applied.

It is evident that some common trends are detected in HIV-1 protease with all methods of analysis. First, in apo- and holo- protein comparisons, binding site residues always appear to be significantly more flexible than for all residues, both in  $\chi_1$  and  $\chi_2$  torsions, and in the r1, r2 and rank rotamer parameters. When holo-/holo- comparisons are instead considered, the difference between all residues' and only binding site residues' conformational changes is in general smaller, and, in the case of  $\chi_2$ , Najmanovich *et al.* methodology of study reveals the same amount of conformational changes in all and only binding site protein residues. In Figures 4.33, 4.34 and 4.35, the differences between conformational changes observed in the binding site and in all the residues of the 10 protein systems have been plotted.

If residue environments are considered and a distinction on the basis of residue exposure to the solvent is made, some consistent trends are again found with all the methodologies of study (Figures 4.18-4.23 in chapter 4). In apo-/holo- protein comparisons, buried residues' conformational changes are always greater than exposed residues' conformational changes in the binding site of HIV-1 protease. In the case of holo-/holo- protein comparisons, buried binding site residues are more flexible than exposed ones only when Zhao *et al.* residue- and environment- thresholds and, to a minor extent, a 60° threshold are applied. Zhao *et al.* specific thresholds are the only ones that detect greater buried residues conformational changes also in all protein residues (apo-/holo- protein comparisons, Figure 4.20).

In general, apo-/holo- comparisons reveal greater percentages of conformational change than those detected by holo-/holo- protein comparisons with all methodologies (Figures 4.30, 4.32 and 4.31). Since the Zhao *et al.* approach more clearly allows

systematic, non random, conformational changes to be identified, this is the preferred methodology for the aims of the present thesis.

### 5.4.2 Relationships Between Holo-Protein Side-Chain Conformational Changes and Ligand Similarities

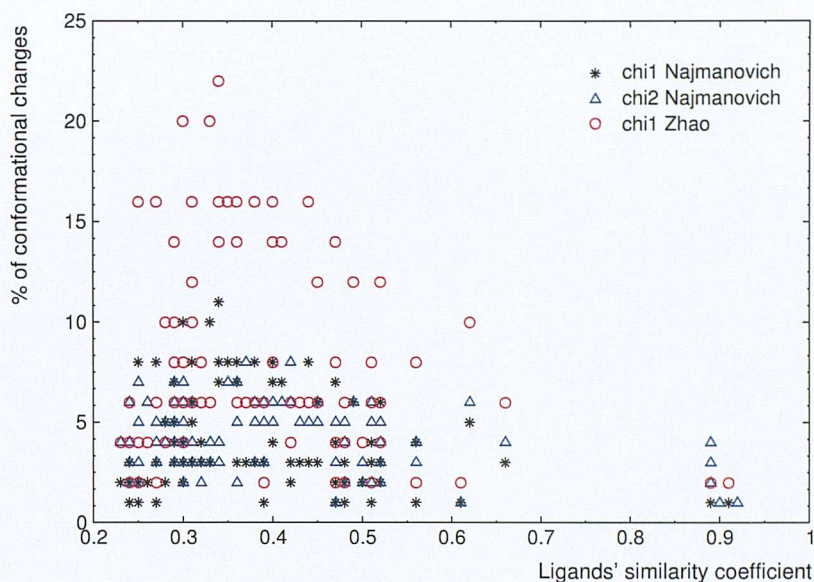
To investigate whether similar ligands induce similar conformational changes in the same apo-protein structure, Tanimoto similarity coefficients were computed for all possible pairs of HIV-1 protease ligands with 1024-bit Daylight fingerprints and the default path range number of considered bonds (0-7).

In Figure 5.4, the percentages of conformational changes observed in the binding site of HIV-1 protease holo-protein pairs have been plotted on the y axis; on the x axis, the Tanimoto similarity scores of the corresponding pairs of ligands are reported. As the graph shows, no correlation between the observed percentages of conformational changes and the 2D-similarity of the corresponding ligands was found. The only trend which can be observed in this graph is that very similar ligands do not appear to be related to large conformational changes. This is in contrast with the majority of the protein systems analysed in this thesis, which do not show any sort of correlations between the ligands' similarity values and holo-proteins conformational changes.

### 5.4.3 Investigation on Cross-Correlated Motions of Different Residues' Side-Chains

One of the aims of the present thesis is to investigate the possible cooperativity of conformational changes. To this end, the conformational changes of different residues in the same PDB entries were analysed, looking for correlated conformational changes induced by residue-residue contacts.





**Figure 5.4:** Percentages of conformational changes observed in the binding site of HIV-1 protease holo-/holo- protein pairs (y axis) plotted against the Tanimoto similarity scores of the corresponding pair of ligands (x axis). Only Najmanovich *et al.* and Zhao *et al.* data are shown.

The correlated conformational behaviour of each pair of amino acids was measured as the percentage of conformational changes of two residue in the same PDB entry, compared to the apo-structure, and reported a in table. For example, if  $c$  is the number of times residue A changes conformation at the same time as residue B, and  $a$  and  $b$  are respectively the total number of times residue A and residue B changed conformation in all pairs of apo-holo structures, the value found at the intersection of column A with row B was  $(c/b) \cdot 100$ . The resulting table was clearly not symmetrical: at the intersection of column B with row A was  $(c/a) \cdot 100$ , i.e. the number of times residues A and B changed together, versus the total number of times A changed. Ratios were expressed as a percentage.

If residue A always changes with respect to residue B,  $c$  will be always equal to  $b$ , thus column A will contain only values equal to 100% or 0%, depending on whether B changes or not. This can clearly happen even if residues A and B are not conformationally correlated: to highlight meaningful cross-correlated conformational

behaviour, and distinguish it from noise (possibly due to very mobile residues), other kinds of cross-correlation measures were derived, and conformational similarity indices analogous to those used to handle chemical information were defined.<sup>76,77</sup>

If we define  $n$  the number of compared pairs of structures ( $N_{AB}$  in equation 4.1), we can assume that this value represents the total number of bits two objects (in this case residues) can have "on" or "off", i.e. the number of times they can or cannot conformationally change. In this way, we will end up with 42 strings (one for each HIV-1 protease binding site residue)  $n$  bits long: each bit corresponds to a compared pair of structures, and can be on or off depending if the given residue changes conformation in the two structures. Given this definition of  $n$ , and the previous definitions of  $a$ ,  $b$ , and  $c$ , the correlated conformational behaviour of two HIV-1 protease residues was quantified using an expression analogous to the definition of the Tanimoto coefficient (see equation 4.1).<sup>76</sup>

The cross-correlation plots obtained using this method were easier to interpret. However, the Tanimoto coefficient does not consider a common absence of chemical attributes from a chemical point of view (i.e. the common absence of side-chain motion in two residues) as an evidence of similarity. To address this issue, a new index of conformational similarity which takes into account both "on" and "off" bits was defined. If  $d$  is the number of times residues A and B do not change in the same crystal structure, a modified coefficient  $MT$  can be defined as:

$$MT = \alpha * T_C + (1 - \alpha) * T_0 \quad (5.2)$$

where:

$$T_0 = d/[a + b - 2c + d] \quad (5.3)$$

$$\alpha = (2 - p)/3 \quad (5.4)$$

$$p = (a + b)/n \quad (5.5)$$

In these equations,  $T0$  is defined analogously to the Tanimoto Coefficient Complement, which takes into accounts of "off" bits, while  $MT$  is defined similarly to the Modified Tanimoto coefficient,  $T_C$ .<sup>94</sup>

Distance matrices for the whole data set HIV-1 protease holo-proteins were then obtained to investigate possible relations between the correlated conformational behaviour of amino acids and their atomic contacts.

The shortest atom-atom distance for each pair of residues was stored in a matrix, giving one residue-residue distance map for each PDB structure. The average of the data points over all matrices was then evaluated, obtaining an average of the shortest distance for each residue-residue pair, producing a single average plot for all the structures.

By multiplying the cross-correlated conformational coefficient by the relative normalised shortest distance of each pair of residues, new, combined matrix values were obtained.

In general, few of the highly conformationally correlated residues also appeared to be in contact. Example of residues that are highly correlated in conformational behaviour but are not at contact distance within each other are residues Asp 29, that changes conformation only once in the analysed data set, and Ile 47 of chain A, which changes in just two of the 20 analysed PDB entries. The only residue that appears close to both Asp 29 and Ile 47 of chain A is residue Asp 30 of chain A; however, this residue never changes its conformation in the 20 holo-proteins of the data set. Lots

of similar examples could be identified.

By and large, the correlated motions do not appear to be caused by protein-protein interactions; significant or systematic correlation between residue contacts and cross-correlated conformational data were not found.

This observation is in agreement with a study by Leach and Lemon,<sup>95</sup> who found that protein side-chains change conformation in a largely independent way.

#### 5.4.4 Conformational Analysis of Holo-Proteins bound to the Same Ligand and/or Solved by the Same Authors.

Among the protein systems analysed in more depth in the present thesis (see following chapters), HIV-1 protease is probably the most appropriate to evaluate the amount of side-chain flexibility observed in holo-proteins bound to the same ligand.

Table 5.1 reports the percentages of side chains which are found to be in a different conformation in PDB entries 2BPY and 2BPZ, which are bound to the same ligand (3IN) and were respectively solved at 1.90 Å and 2.50 Å resolution. While the first column refers to the conformational changes observed in all protein residues, the second one reports the percentages relative to only binding site residues.

Clearly, the amount of side-chain flexibility which is found in the two holo-proteins is very small, especially in comparison to the average percentages of conformational changes detected in all HIV-1 protease structures analysed in this thesis. However, this exceptional side-chain rigidity might depend not only on the presence of the same ligand in the two structures, but also on the fact that both 2BPY and 2BPZ were solved by the same group of crystallographers.<sup>96</sup>

In a recent work by DePristo *et al.* which addresses the issue of X-ray crystal structure accuracy,<sup>97</sup> the authors state that large differences among alternate ex-

|                                  | <u>Holo-/Holo- Cnf Changes %</u> |              |
|----------------------------------|----------------------------------|--------------|
|                                  | all residues                     | binding site |
| <b>Najmanovich <i>et al.</i></b> |                                  |              |
| $\chi^1$                         | 0.60                             | 0.00         |
| $\chi^2$                         | 0.83                             | 0.00         |
| <b>Zhao <i>et al.</i></b>        |                                  |              |
| buried $\chi^1$                  | 5.16                             | 12.92        |
| exposed $\chi^1$                 | 4.52                             | 0.00         |
| <b>Dunbrack and Cohen</b>        |                                  |              |
| $r1$                             | 0.60                             | 0.00         |
| $r2$                             | 2.50                             | 0.00         |
| $rank$                           | 19.17                            | 12.00        |

**Table 5.1:** Percentages of conformational changes observed in all and only binding site residues of HIV-1 protease PDB entries 2BPY and 2BPZ (both bound to ligand “3IN”). While 2BPY is solved at 1.90 Å resolution, the resolution of 2PBZ is 2.5 Å; this entry was thus excluded from the main data set analysed in this thesis.

perimental models have been limited to the rare cases where several groups have crystallised and solved the same protein independently. This observation, which had already been made by Ohlendorf<sup>98</sup> and Zoete *et al.*,<sup>79</sup> is crucial for the purposes of the present thesis; protein structures which have been solved by several different groups of crystallographers are more likely to reflect and represent the heterogeneity and flexibility that characterise protein backbones and side-chains. Proteins are in fact dynamic molecules showing structural heterogeneity, individual atomic anisotropic motion and collective, large-scale motion over a range of time scales which also occur in their crystal forms (due to the high solvent content in most protein crystals).<sup>97</sup> Modelling anisotropic motion and structural heterogeneity has been limited to proteins that diffract to atomic resolution; the vast majority of proteins diffract to worse than 1.6 Å resolution, and are solved as a single, average conformation with Gaussian, isotropic thermal motion.<sup>99</sup> The artifacts that are caused by ignoring heterogeneity



during structure determination are still largely uncharacterised; however, it is generally believed that they can lead to an incomplete description of crystallographic data and to a considerable degree of inaccuracy.<sup>97</sup> The presence of uncharacterised inaccuracies in crystal structures is troubling; without estimates of the uncertainty in atomic positions, genuine features or differences among structure cannot be identified and unreliable protein conformations might be overestimated.

To obtain an estimate of how much side-chain conformational changes are biased by the way in which the same author(s) solve different holo-structures of the same protein, all HIV-1 protease holo-proteins analysed in the present thesis were divided into groups of structures solved by the same author or group, and their conformational changes separately evaluated.

In the first five columns of Table 5.2 (**a-e**), the percentages of side-chain motions detected in the 5 so-obtained groups of structures are shown; their average is reported in column **f**, while column **g** reports the average percentages of conformational changes obtained for all the 17 holo-proteins structures solved at 2.0 Å or better which are part of the present thesis data-set.

2BPY and 2BPZ, the pair of proteins which are bound to the same ligand (Table 5.1), were solved by Munshi *et al.*,<sup>96</sup> i.e. the same authors who solved PDB entries 1C70 and 2BPV (column **c** in Table 5.2). As it can be observed by comparing the average percentages reported in Tables 5.1 and 5.2 for these authors, the flexibility found in structures binding to the same ligand is less than that of structures bound to different ligands. However, the amount of side-chain conformational changes detected in all protein residues of 2BPY and 2BPZ PDB entries is significantly lower than that observed in the other structures solved by the same authors. Whether this is a

| Holo-/Holo- Cnf Changes %        |       |       |       |       |       |       |       |  |
|----------------------------------|-------|-------|-------|-------|-------|-------|-------|--|
| all residues                     |       |       |       |       |       |       |       |  |
|                                  | (a)   | (b)   | (c)   | (d)   | (e)   | (f)   | (g)   |  |
| <b>Najmanovich <i>et al.</i></b> |       |       |       |       |       |       |       |  |
| $\chi^1$                         | 4.22  | 0.60  | 9.64  | 1.60  | 15.09 | 6.23  | 10.07 |  |
| $\chi^2$                         | 7.50  | 0.00  | 25.00 | 2.08  | 17.39 | 12.39 | 17.73 |  |
| <b>Zhao <i>et al.</i></b>        |       |       |       |       |       |       |       |  |
| buried $\chi^1$                  | 7.74  | 2.65  | 21.32 | 4.69  | 21.71 | 11.62 | 22.55 |  |
| exposed $\chi^1$                 | 12.43 | 1.11  | 19.56 | 4.08  | 30.12 | 13.46 | 22.64 |  |
| <b>Dunbrack and Cohen</b>        |       |       |       |       |       |       |       |  |
| $r_1$                            | 20.25 | 0.60  | 9.94  | 2.81  | 18.07 | 16.50 | 12.44 |  |
| $r_2$                            | 21.61 | 0.00  | 22.00 | 4.60  | 18.33 | 13.43 | 19.34 |  |
| $rank$                           | 44.92 | 10.00 | 39.17 | 18.83 | 41.67 | 30.19 | 40.03 |  |
| binding site                     |       |       |       |       |       |       |       |  |
| <b>Najmanovich <i>et al.</i></b> |       |       |       |       |       |       |       |  |
| $\chi^1$                         | 0.00  | 12.89 | 3.03  | 1.51  | 15.62 | 6.42  | 10.68 |  |
| $\chi^2$                         | 8.00  | 20.82 | 0.00  | 2.08  | 17.39 | 9.50  | 16.69 |  |
| <b>Zhao <i>et al.</i></b>        |       |       |       |       |       |       |       |  |
| buried $\chi^1$                  | 20.00 | 29.60 | 13.33 | 4.69  | 21.74 | 19.33 | 33.04 |  |
| exposed $\chi^1$                 | 5.56  | 25.64 | 0.00  | 4.08  | 30.12 | 15.90 | 20.29 |  |
| <b>Dunbrack and Cohen</b>        |       |       |       |       |       |       |       |  |
| $r_1$                            | 27.27 | 13.43 | 3.03  | 3.54  | 21.21 | 13.64 | 14.71 |  |
| $r_2$                            | 24.00 | 21.57 | 0.00  | 4.60  | 18.33 | 14.14 | 21.91 |  |
| $rank$                           | 46.00 | 37.25 | 16.00 | 18.43 | 41.87 | 32.19 | 44.92 |  |

**Table 5.2: (a-e):** Percentages of conformational changes observed in all and in only binding site residues of HIV-1 protease holo-proteins structures solved by the same authors. **(a)** 1AJV and 1AJX PDB entries, solved by Backbrock *et al.*; **(b)** 1G2K and 1G35, solved by Schaal *et al.*, **(c)** 1C70, 2BPV and 2BPY, solved by Munshi *et al.*; **(d)** 1HVI, 1HVJ, 1HVK and 1HVL (solved by Bath *et al.*); and **(e)** 1HTF and 1HTG (solved by Jhoti *et al.*). All these structures except 1C70 and 1HTF were solved at resolution equal or better than 2.0 Å. The average percentages of conformational changes for all five groups of structures **(a-e)** are indicated in column f; the averages observed in all HIV-1 protease structures solved at resolution equal or better than 2.0 Å are reported in column g for comparison.

structure refinement artifact (e.g. one structure was solved using the other one) or the observation of a real event cannot be established.

It is evident that the percentages of conformational changes observed in protein structures solved by the same authors are significantly less than those observed in all holo-/holo- protein comparisons. This suggests the opportunity to “normalise” the percentages of side-chain conformational changes observed in different protein systems according to whether different workers solved their structures.

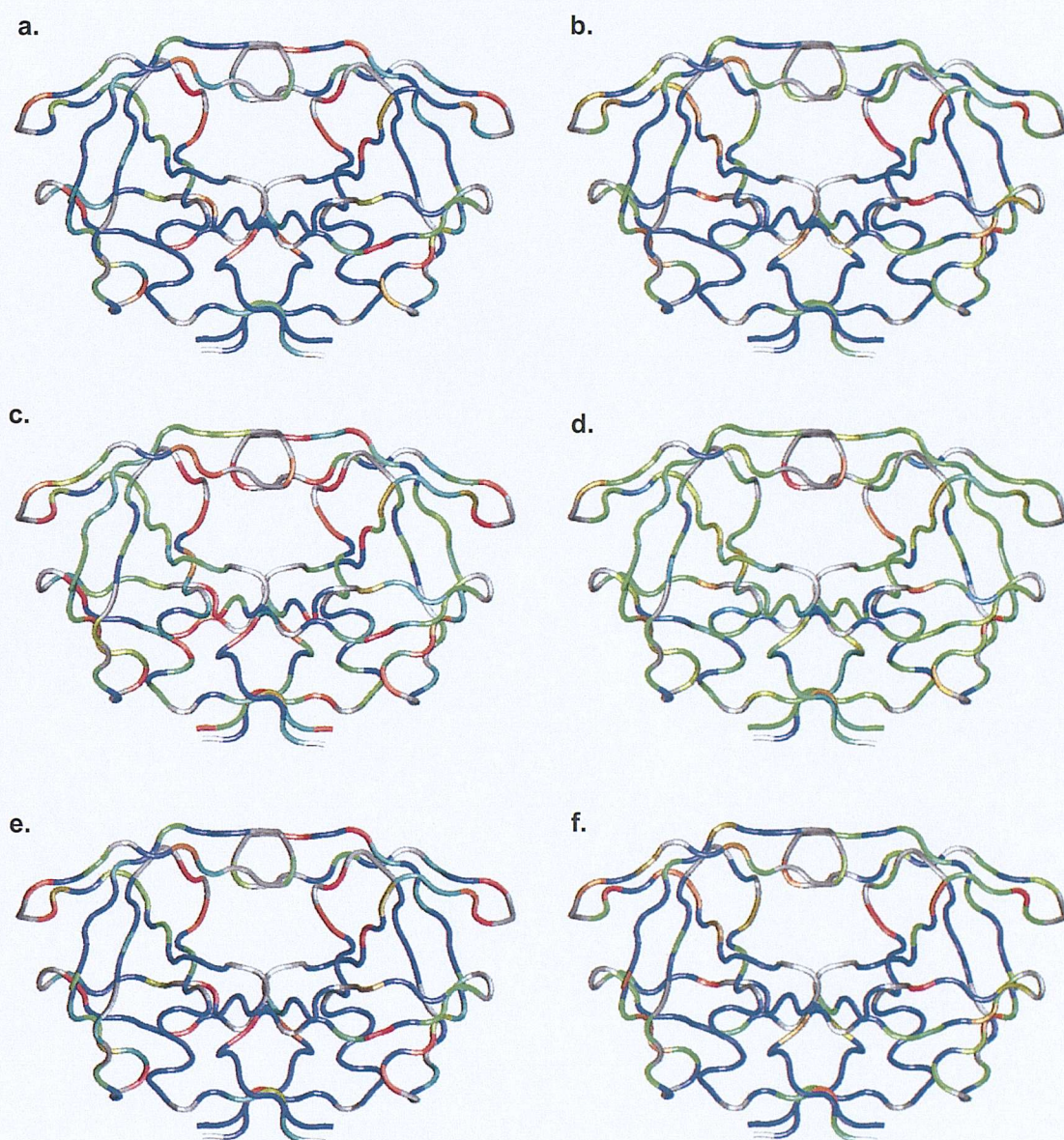
### 5.4.5 Percentages of Conformational Changes per Residue Sequence Number

Figure 5.5 shows the backbone structure of all HIV-1 protease residues coloured in accordance to the flexibility of their  $\chi_1$  side-chain torsions. The structures shown in the first row (indicated by letters *a* and *b*) represents the results obtained with Najmanovich *et al.* methodology of study, those in the second row (*c*, *d*) with Zhao *et al.* methodology, and the ones in the third row (*e*, *f*) with Dunbrack and Cohen rotamer libraries. Structures in the left column refer to apo-/holo- protein comparisons results, those in the right column to holo-/holo- protein comparisons data.

A quick visual inspection of Figure 5.5 reveals that the most flexible regions (red) are mainly concentrated at the top and in the side parts of the binding site, while its “lower” portion, that includes the catalytic triad, appears to be very rigid. Other consistently flexible zones appear to be the flap elbow, the cantiliver and the fulcrum. Broadly speaking, the same flexibility trends are observed with all methodologies of study.

Figures **a** and **e** (apo-/holo- comparisons), and **b** and **f** (holo-/holo- comparisons) show a very similar colour pattern; this is not unexpected, since the Najmanovich *et*





**Figure 5.5:** Tube trace of HIV-1 protease backbone structure; all residues are coloured in accordance with the average percentage of conformational change its  $\chi_1$  torsions undergo, as detected with Najmanovich *et al.* (a, b), Zhao *et al.* (c, d) and Dunbrack *et al.* (e, f) methodologies of study. The residue average percentage of conformational change increases from blue coloured residues (which never change conformation) to red coloured residues (that change conformation 100% of times), passing through cyan, green, yellow and orange. Proline, glycine and alanine residues have been coloured in grey. Figures on the left (a, c, e) refer to apo-/holo- protein comparisons, figures on the right (b, d, f) to holo-/holo- protein comparisons.

*al.* threshold and Dunbrack and Cohen rotamer libraries do not take into account residue types and/or residue specific environments, assigning residues' flexibility on the basis of an absolute scale. If residue type- and environment- specific thresholds are instead applied (Figures **c** and **d**), a similar distribution of flexible (red) and more rigid regions (blue) is found, but a greater overall flexibility is generally observed. Also, the differences between the most unstable and the most stable regions' of the protease are levelled out. This can again be expected; these thresholds are in fact very stringent for residues that are buried to the solvent and/or are generally rigid, and often very "permissive" for residues that are intrinsically more flexible and/or exposed to the solvent. They are thus likely to identify residues that are more mobile than expected, given their residue type and/or their exposure to the solvent, rather than residues that are very mobile on an absolute flexibility scale.

The distribution of residue side-chain flexibility appears to be roughly speaking symmetrical. However, some significant differences between the two protein chains appear with all methods of analysis. As other studies on protein flexibility,<sup>63,79</sup> homo-dimeric and homo-tetrameric proteins in this thesis (HIV-1 protease, glutathione S-transferase, ribonuclease, xylose D-isomerase and streptavidin) were treated by arbitrarily comparing chains with same PDB chain ID or, in the case of PDB structures with different chain names, by comparing the first chain on one protein with the first chain of the other protein and so on. However, unless protein structures are complexed with identical or with strongly asymmetric ligands belonging to the same structural family, it is not generally possible to distinguish identical monomers, and PDB chain IDs of homo-dimers and homo-tetramers are arbitrarily chosen by crystallographers. The results presented in this chapter for the two monomers of HIV-1 protease must thus be treated with caution, as chain definition relies on crystallographers' assign-



ments, i.e. are arbitrary and not necessarily meaningful.

The average percentages of  $\chi_1$  conformational changes have been plotted on the y axis of the graphs represented in Figures 5.6-5.9. In these graphs, residue numbers on the x axis have been distributed on two separate graphs (*a*, residues 1-50. and *b*, residues 50-99) for clarity.

In this chapter and in the following ones, only data relative to  $\chi_1$  will be presented. This choice is due to the overall similar trends found for  $\chi_1$  and  $\chi_2$  in this type of analysis, and to the lack of  $\chi_2$  thresholds in the case of Zhao *et al.* methodology; graphs reporting  $\chi_2$  data are omitted for purposes of greater clarity.

While the first two graphs (Figures 5.6 and 5.7) report data obtained with apo-/holo- protein comparisons, the last two (Figures 5.8 and 5.9) show holo-/holo- protein comparison data. Results obtained applying Najmanovich *et al.* angular thresholds are reported first (Figures 5.6 and 5.8), followed by those obtained with Zhao *et al.* angular thresholds (Figures 5.7 and 5.9). Since the graphs obtained with Najmanovich *et al.* methodology of study are strikingly similar to those obtained employing Dunbrack and Cohen rotamer libraries, the latter are not shown in the present chapter but reported in appendix 3 (Figures C.1 and C.2). To a lesser extent, the trends revealed by these two methods of analysis are also similar to those revealed by Zhao *et al.* angular thresholds (Figures 5.7 and 5.9). In the latter data it is, however, very rare to find areas of the protease in which residues never change conformation; a certain amount of conformational changes is also found for the catalytic aspartates, Asp 25 and 25'.

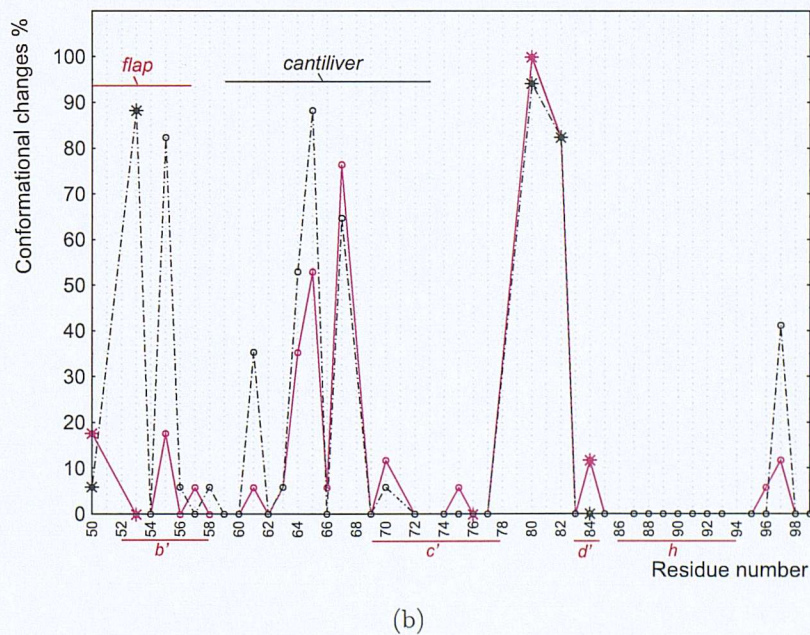
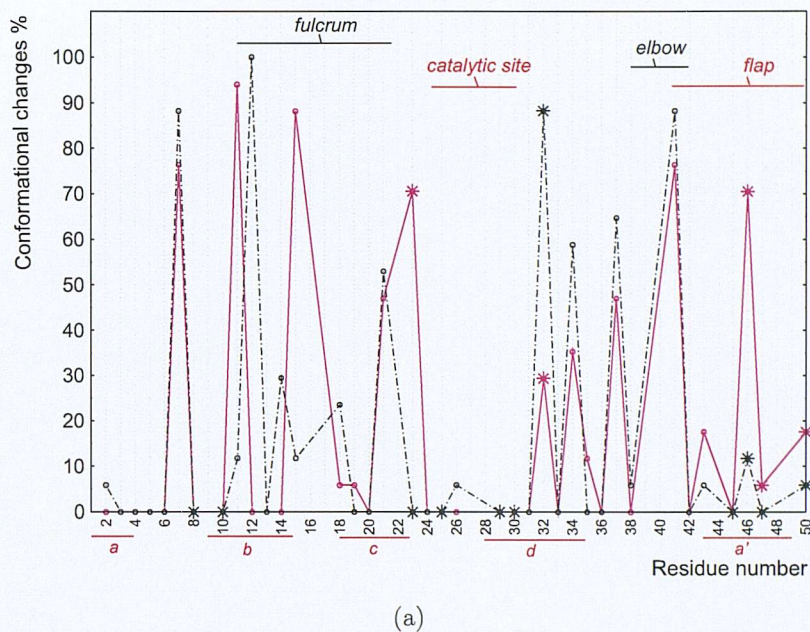
The most remarkable differences between the two protein monomers are found both in the binding site (e.g. the flap and the flap tip), and in parts of the protein

that are not in direct contact with the ligands, such as the fulcrum (residues 11, 12, 15) and the flap elbows (residue 40 and surroundings). The asymmetries are found in both apo-/holo- and holo-/holo- protein comparisons.

The regions with the highest side-chain flexibility consistently appear to be the fulcrum, the flaps and the flap elbows, the cantiliver, Thr 80 and Val 82 in apo-holo- comparisons and Val 82 in holo-/holo- comparisons. Also, some of the residues that immediately precede and follow the catalytic site are highly mobile: Glu 21, Leu 23 and Val 32 in apo-/holo- comparisons and Glu 21, Val32 and Glu 34 in holo-holo- protein comparisons. Interestingly, regions with highly flexible side chains belong to both  $\beta$ -strand secondary elements (fulcrum,  $\beta$ -strands *c* and *d*, flaps) and unstructured loops (flap elbows, cantiliver, residues 80 and 82). Luque *et al.* observed that binding site regions with very low stability constants<sup>91</sup> are often located in loops regions or turns, but that they can also be found in  $\alpha$ -helices and  $\beta$ -strands; making the assumption that flexible side-chains correspond to low stability residues, this HIV-1 protease side-chain conformational analysis confirms this finding.

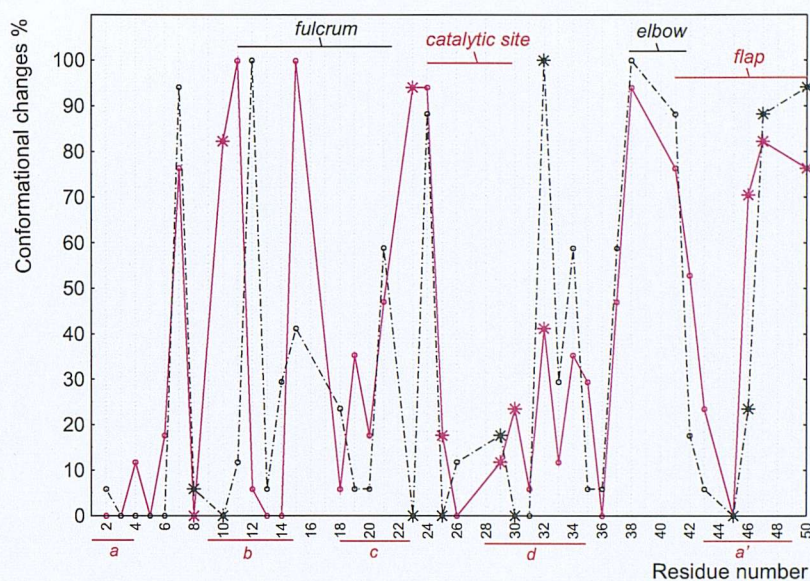
The most stable regions of the protease appear instead to be the catalytic site (residues 24-30), the N-terminal  $\beta$ -sheet *a* (residues 1-4),  $\alpha$ -helix *h* (residues 86-94) and residues belonging to  $\beta$ -strand *c'* (residues 69-78) with all methods of analysis.

These results are in strong agreement with previous MD studies on HIV-1 protease backbone flexibility,<sup>80,81,85,93</sup> and previous work by Zoete *et al.*<sup>79</sup> and Bardi *et al.*<sup>88</sup> (see previous section). However, a few differences are revealed between the side-chain flexibility results and the above mentioned studies. First, Leu 23, in the immediate proximity of the catalytic triad residues and part of the binding site, was found to be very stable both by Zoete and co-workers and Bardi *et al.*, but Najmanovich and Zhao

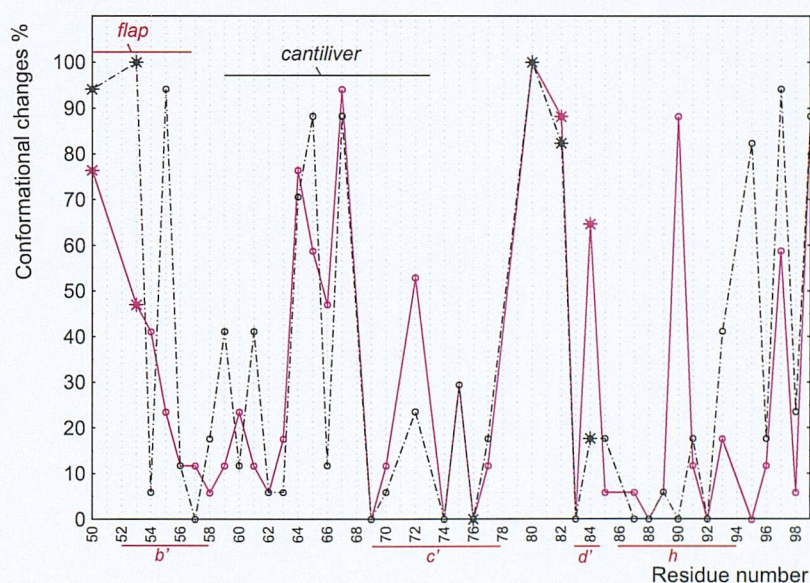


**Figure 5.6:** Percentages of times residues of HIV-1 protease change  $\chi_1$  by more than  $60^\circ$  in apo-/holo- protein comparisons; residues have been divided into Figure a (residues from 1 to 50) and b (residues from 50 to 99) for clarity. Data coloured in violet refer to the first monomer of HIV-1 protease, and black to the second chain of the enzyme. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data. Letters *a*, *b*, *c*, *d*, *a'*, *b'*, *c'* and *d'* on the x axis indicate HIV-1 protease  $\beta$ -strand regions, as described in Figure 5.1; letter *h* corresponds to the  $\alpha$ -helix comprising residue numbers 86 to 94 ( Figure 5.1). Some of the most common names employed to indicate regions of HIV-1 protease have been written in the top part of the graph; those that refer to regions that are in contact with the ligand are in red.





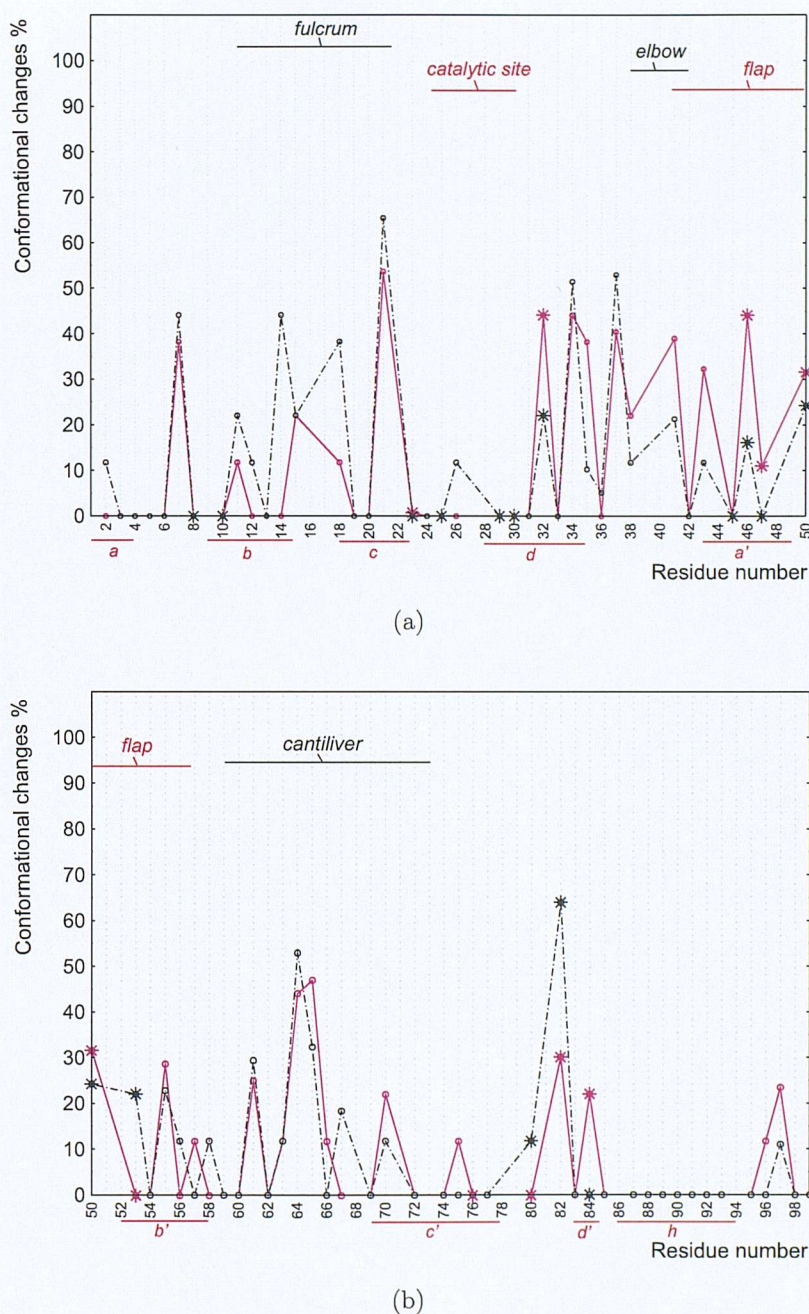
(a)



(b)

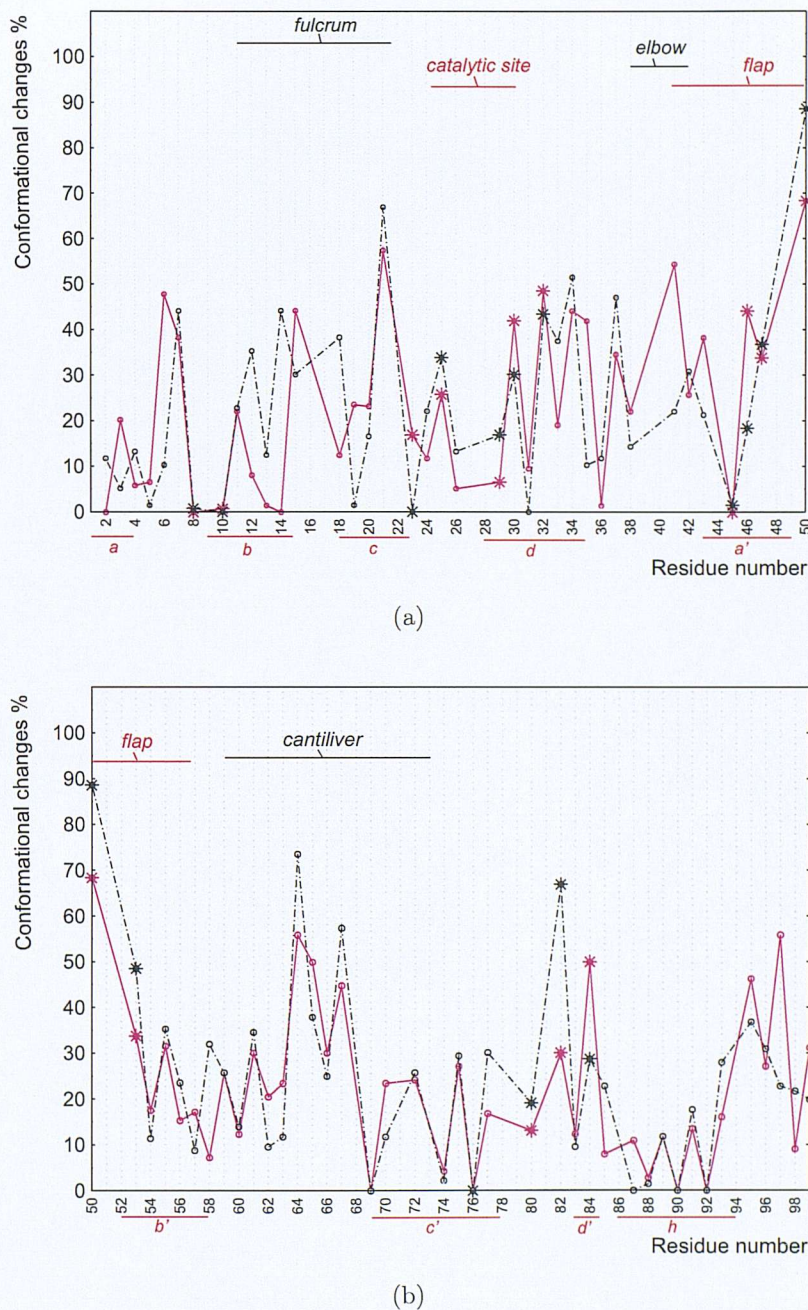
**Figure 5.7:** Percentages of times residues of HIV-1 protease change  $\chi_1$  by more than Zhao *et al.* specific angular thresholds in apo-/holo- protein comparisons; residues have been divided into Figure **a** (residues from 1 to 50) and **b** (residues from 50 to 99) for clarity. Data coloured in violet refer to the first monomer of HIV-1 protease, and black to the second chain of the enzyme. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data. Letters *a*, *b*, *c*, *d*, *a'*, *b'*, *c'* and *d'* on the x axis indicate HIV-1 protease  $\beta$ -strand regions, as described in Figure 5.1; letter *h* corresponds to the  $\alpha$ -helix comprising residue numbers 86 to 94 ( Figure 5.1). Some of the most common names employed to indicate regions of HIV-1 protease have been written in the top part of the graph; those that refer to regions that are in contact with the ligand are in red.





**Figure 5.8:** Percentages of times residues of HIV-1 protease change  $\chi_1$  by more than  $60^\circ$  in holo-/holo- protein comparisons; residues have been divided into Figure a (residues from 1 to 50) and b (residues from 50 to 99) for clarity. Data coloured in violet refer to the first monomer of HIV-1 protease, and black to the second chain of the enzyme. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data. Letters *a*, *b*, *c*, *d*, *a'*, *b'*, *c'* and *d'* on the x axis indicate HIV-1 protease  $\beta$ -strand regions, as described in Figure 5.1; letter *h* corresponds to the  $\alpha$ -helix comprising residue numbers 86 to 94 (Figure 5.1). Some of the most common names employed to indicate regions of HIV-1 protease have been written in the top part of the graph; those that refer to regions that are in contact with the ligand are in red.





**Figure 5.9:** Percentages of times residues of HIV-1 protease change  $\chi_1$  by more than Zhao *et al.* specific angular thresholds in holo-/holo- protein comparisons; residues have been divided into Figure **a** (residues from 1 to 50) and **b** (residues from 50 to 99) for clarity. Data coloured in violet refer to the first monomer of HIV-1 protease, and black to the second chain of the enzyme. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data. Letters *a*, *b*, *c*, *d*, *a'*, *b'*, *c'* and *d'* on the x axis indicate HIV-1 protease  $\beta$ -strand regions, as described in Figure 5.1; letter *h* corresponds to the  $\alpha$ -helix comprising residue numbers 86 to 94 ( Figure 5.1). Some of the most common names employed to indicate regions of HIV-1 protease have been written in the top part of the graph; those that refer to regions that are in contact with the ligand are in red.

*et al.* angular thresholds and Dunbrack and Cohen rotamer libraries reveal a high flexibility of its side-chain in at least one monomer of HIV-1 protease for apo-/holo-protein comparisons (Figures 5.6, 5.7 and C.1). In holo-/holo- protein comparisons, all methodologies of study do not find significant conformational changes in this residue (Figures 5.8, 5.9 and C.2). Also, Val 32, which immediately follows the catalytic site and is in direct contact with the ligands, appears to be highly flexible with all methodologies of study in both apo-/holo- and holo-/holo- protein comparisons, in contrast to previous backbone and residue stability constant analyses.

Ile 50, the tip of the flap, was among the most flexible residues of HIV-1 protease in the opinion of Zoete *et al.* and Bardi *et al.* While Najmanovich *et al.* and Dunbrack and Cohen methodologies fail to detect an exceptional level of flexibility for this residue, Zhao *et al.* results show percentages of conformational changes that are about 95% in the case of both apo-/holo- and 90% in the case of holo-/holo- protein comparisons. In particular, when holo-/holo- protein comparisons are considered, Ile 50 appears to be the most flexible residue (Figure 5.9). Another discrepancy between the results obtained with the different methodologies of study employed in this thesis appears to be Ile 84, to which the algorithm COREX<sup>89,91</sup> assigned a very high stability constant (see appendix 2 and section 5.3).<sup>88</sup> This residue seems to undergo many conformational changes with Zhao *et al.* angular thresholds (Figures 5.7 and 5.9); the same does not apply to Najmanovich *et al.* and Dunbrack and Cohen methods' results.

Somewhat unexpectedly, Zhao *et al.* methodology of study detects some conformational changes in the side-chain of the catalytic aspartatic residues (Figures 5.7 and 5.9) Also, both Najmanovich *et al.* and Dunbrack and Cohen methods show a

small amount of conformational changes in the second chain of HIV-1 protease for the catalytic residue Thr 26. Conformational changes in this residue are also found when applying Zhao *et al.* angular thresholds.

From one point of view, this should not be regarded as something surprising. The conformational changes detected for Thr 26 with Najmanovich *et al.* and Dunbrack and Cohen definitions are in fact extremely small, and, for the catalytic aspartates, one must not forget that Zhao *et al.* angular thresholds are very stringent, especially in the case of buried residues. In the case of aspartic residues, for example,  $\chi_1$  is considered to have changed conformation if it differs by more than  $13.8^\circ$  in exposed and  $8.0^\circ$  in buried residues. The percentages of conformational changes revealed for Asp 25 by these angular thresholds could be thus regarded as a kind of 'noise' data, in the range of the lowest levels of conformational changes found with this methodology of study.

From another point of view, the conformational changes observed in the catalytic residues might instead reflect something which has a significant physical meaning. In a recent article on HIV-1 protease flexibility, Kumar and Hosur<sup>84</sup> distinguished protein "flexibility", i.e. the natural fluctuations of residues and/or regions of proteins around their average positions described by Zoete *et al.* by means of X-ray structures comparisons, B-factors deviations, and MD and NMA studies,<sup>79,84</sup> from protein "adaptability", i.e. the ability of residues to alter their mean position in response to environment chemical changes and/or stress. In the opinion of the authors, the region of HIV-1 protease that includes the catalytic aspartates (residues 23-26) is structurally very adaptable,<sup>84</sup> in spite of being at the same time the least flexible of the protease (as proved by low crystallographic B-factors, Zoete *et al.* results and NMR data).<sup>79,84</sup> For Kumar and Hosur, such adaptability is shown by the rearrange-

ments that the catalytic aspartates undergo in the HIV-1 protease double mutant C95M/C95'A, both in complexed and uncomplexed forms of the protease. In fact, if the unliganded structure of this double mutant (PDB entry 1LV1, a "closed-flap" structure solved at 2.1 Å)<sup>84</sup> is compared to the single site mutant C95M (PDB entry 1G6L, i.e. the apo-reference structure employed in this thesis), a shift in the backbone of the catalytic aspartates is observed; more precisely, the backbone of residues 23-26 in the double mutant protein move towards the flap. Apparently, the C95M/C95'A mutation does not cause significant changes the structure in the nearby residues, but affects the position of residues 23-26 which are above it. Also, by comparing the same double mutant apo-structure to eight HIV-1 protease holo-structures carrying the double mutation C95A and bound to eight different ligands (two peptidic and six cyclic-urea based molecules), some interesting observations can be made.<sup>84</sup> On the one hand, when the two holo-structures bound to peptidic inhibitors are compared with the unliganded protein, no significant alterations in the position of the catalytic aspartates are found. On the other hand, if the six holo-proteins bound to cyclic urea molecules are considered, a shift of Asp 25 and Asp 25' in the direction of the flaps is instead again observed; this probably occurs to relieve a bad contact that the cyclic urea ring would otherwise establish with Asp 25 side-chains. Kumar and Hosur concluded that the catalytic residues of HIV-1 protease, in spite of being very rigid and stable, are in fact highly "adaptable". They can in fact move and adjust their position in response to internal stress or external stimuli, such as mutations and/or ligands, even if their electron density is very well defined and their B-factors appear to be the lowest in the protease structure.<sup>84</sup> This differentiation between the adaptability and the flexibility of a protein<sup>84</sup> is analogous to the distinction of "systemic

flexibility” and “segmental flexibility” proposed by another group of researchers.<sup>100</sup> According to them, systemic flexibility refers to small-scale fluctuations in side-chain and main-chain atoms in the protein native state and is distributed throughout the protein, while segmental flexibility refers to the motion of one part of the protein in respect to the other, and often occurs in response to a particular event which is related to protein function. While the time-scale of systemic flexibility is fast, segmental flexibility has lower time scales, and is related to much larger movements of restricted parts of the protein (such as hinges).

In the opinion of Kumar and Hosur, the analysis of an ensemble of structures, such as the analysis performed by Zoete *et al.*,<sup>79</sup> can give information on the inherent flexibility of a protein (i.e. systemic flexibility), but fails to detect information about structures’ adaptability (i.e. segmental flexibility). In fact, the study by Zoete *et al.* revealed a pattern of variability of different residues that is very much similar to the B-factors distribution that can be found in any single structure. The comparison of structures on a one by one basis could instead provide useful insights of the segmental induced-fit provoked by ligands.<sup>84</sup>

The structural rearrangements of the catalytic residues noticed by Kumar and Hosur<sup>84</sup> are, however, overall small, and certainly do not affect the geometry that is necessary to bind the substrates and other ligands. Similarly, the catalytic residues’ small side chain rearrangements observed in this thesis do not affect their catalytic requirements and are anyway very small when compared to the extent of conformational changes the most flexible residues undergo.

In accordance with literature data,<sup>88,92</sup> side-chains conformational changes obtained with all methodologies of study indicate that binding site residues belong to both very rigid and very flexible regions. While the side-chains of residues 8, 10 and



25-30 seldom change conformation, residues belonging to the flaps (such as residues 46, 47, 50, 53 and 55) and residues located in the loop between  $\beta$ -strands  $c'$  and  $d'$  (residues 80 and 82, i.e. the outer part of the binding site) are among the most flexible from the side-chain point of view.

The dual character shown by the active site of the protease is not something new: similar characteristics have also been noticed by Luque and Freire in at least 16 other non homologous proteins.<sup>92</sup> In the opinion of Luque *et al.*, the partially flexible character of binding sites might determine a higher binding affinity for ligands.<sup>92</sup> In fact, the binding affinity of ligands is often considered proportional to the accessible surface area (ASA) that becomes buried from the solvent upon binding, since this is in turn correlated to the solvation enthalpy and entropy contributions to the free energy of binding.<sup>88,92</sup> The reason why ligands with a small molecular weight are often found to be buried in clefts and cavities, shielded from water by loops and/or other secondary structure elements, might be that this location maximises the contacts between the protein and the ligand, and significantly increases the buried ASA (hence the Gibbs energy of binding).<sup>92</sup> However, the burial of ligands within a protein generally requires conformational changes in the protein structure; if the region that must undergo protein conformational rearrangements is relatively unstructured and flexible, and occupies a shallow energy minimum on the protein energy surface, the unfavourable  $\Delta H$  necessary to change the protein conformation would be minimised. This could explain why amino acid mutations that increase the structural stability of regions that must undergo conformational change to bind ligands could lower inhibitors' binding affinity, even if they do not affect the structure of the bound state and/or are not in contact with the ligand.<sup>92</sup> On the other hand, the fact that binding site residues

which are directly involved in the catalytic mechanism are usually located in the most stable regions of the binding site can be easily explained by the necessity of precise stereochemical arrangements for catalytic efficiency.

Residues that contribute significantly to the energetics of binding (see section 5.3) do not show any particular correlation between the specific nature of their contributions and their observed side-chain flexibility. For example, residues that strongly contribute to a favourable  $\Delta S$  of binding are both found on very flexible regions (e.g.: residues of the flap, Val 82) and very stable regions (e.g.: catalytic residues and Arg 8). The amino acids that significantly contribute to the enthalpy of binding with electrostatic interactions (Asp 25, Asp 29 and Asp 30) are too few to infer any conclusion about the relationship between their flexibility and their contributions.

## 5.5 Conclusions

In this chapter, the conformational changes of HIV-1 protease, the most flexible protein of the present thesis data set, have been analysed. With all methodologies of study, the binding site residues of this aspartic protease appear more flexible than all protein residues (see chapter 4). This characteristic, and the high flexibility of the protease, justify its choice as the object of the deeper conformational analyses performed in this chapter.

Previous studies have shown that HIV-1 protease mainly undergoes a hinge motion upon ligand binding, changing from a “closed flap” to an “open flap” conformation.<sup>79</sup> In addition to this motion, many backbone rearrangements are observed in several regions of the protease (e.g. cantiliver, fulcrum, loop between  $\beta$ -strands  $b'$  and  $c'$ ), while the catalytic sites were in general found to be highly rigid.<sup>79,81,88</sup>

The unbound form of HIV-1 protease was believed to naturally exist only in the

open flap conformation; however, recent theoretical and experimental studies<sup>82,83</sup> have shown that the flaps are highly mobile in solution, allowing a variety of conformations from fully closed to open. Unbound forms of HIV-1 protease in the closed-flap conformation have been recently solved;<sup>83,84</sup> a 1.9 Å flap-closed structure of HIV-1 protease (PDB entry 1G6L<sup>83</sup>) has been employed in this thesis as the reference apo-structure.

Zoete *et al.* investigated the flexibility of HIV-1 protease by means of crystal structures comparisons, molecular dynamics simulations, normal mode analysis and crystallographic B-factors comparisons.<sup>79</sup> The flexibility trends that they obtained with the different methodologies were all in agreement with previous experimental and theoretical studies on the protease.<sup>81,93</sup> Since the RMSd trends they found in the presence of different *stimuli* (such as the presence of the same or different ligands, sites of residue mutations and/or different crystallisation conditions) did not significantly differ, they hypothesised that the conformational changes observed in HIV-1 protease depend on the residues' intrinsic flexibility, rather than on the specific nature of the *stimuli* that might induce them.

Bardi *et al.*<sup>88</sup> applied structure-based thermodynamic analyses (see appendix A) to quantitatively parametrise and predict the energetics of binding of 13 inhibitors to HIV-1 protease,<sup>88</sup> and calculated HIV-1 protease residue stability constants (see appendix B) with the algorithm COREX.<sup>89-91</sup> These authors observed that ligand-binding reactions for HIV-1 protease are generally endothermic;<sup>88</sup> in fact, given the highly hydrophobic nature of HIV-1 protease inhibitors and their lack of strongly polar groups, the electrostatic interactions contribute very little to the intrinsic enthalpy of binding, and the unfavourable enthalpic solvation term is the dominant component of the generic enthalpy of binding. On the other hand, an exceptionally large portion

of the protein apolar surface area is buried upon binding, causing a highly favourable entropic component (hydrophobic effect) that is the driving force of the ligand-binding reaction. The highly flexible character of HIV-1 protease, especially in the regions that must undergo major rearrangements, is at the same time the reason of the small unfavourable conformational entropy term.<sup>88,92</sup>

In apo-/holo- comparisons, greater percentages of conformational changes consistently occur in the buried binding site residues of HIV-1 protease (see chapter 4). This characteristic differentiates the protease from the majority of the protein systems analysed in this thesis (see chapter 4). In the view of Bardi *et al.*,<sup>88</sup> the reasons for this peculiarity could be found in the nature of the ligand-binding mechanism; the great rearrangements of apolar residues in the binding site of HIV-1 protease, which are necessary to determine a favourable entropy of binding, could be the reason for the exceptionally high level of structural rearrangements which are observed in buried binding site residues.

When the trends of residue flexibilities along HIV-1 protease chains are considered, the results obtained with Najmanovich *et al.* 60° angular threshold and Dunbrack and Cohen rotamer libraries are strikingly similar. To a lesser extent, they are also similar to those found with Zhao *et al.* methodology.

Broadly speaking, most stable regions of the protease appear to be the catalytic sites (residues 24-30), the N-terminal  $\beta$ -sheet that includes residues 1-4, the  $\alpha$ -helix 86-94 and residues belonging to the  $\beta$ -strand 69-78. The regions with the greatest side-chain flexibility consistently appear instead to be the fulcrum (11-22), the flaps and the flap elbows, the cantiliver, Thr 80 and Val 82.

All these findings are in agreement with literature data; however, some differences between side-chain flexibility results and the above mentioned studies are found, es-

pecially when Zhao *et al.* thresholds are applied. Most of all, all methodologies of analysis detect conformational changes in the side-chain catalytic aspartic residues, particularly in the case of Zhao *et al.* methodology of study. This finding provides an independent corroboration to the view of Kumar and Hosur,<sup>84</sup> who noticed that the backbone of residues 23-26 in HIV-1 protease can move in the direction of the flap in response to specific *stimuli* (such as a mutation in the proximity of the binding site, or the presence of specific ligands).<sup>84</sup> This observation, that in the opinion of Kumar and Hosur proves the “adaptability” of the catalytic residues of HIV-1 protease in spite of their low flexibility, is further supported by the extremely high level of conformational changes that the Zhao *et al.* methodology detects in the catalytic aspartates of endothiapepsin (see chapter 6). A study by Yuan *et al.* reports that active site residues predominantly occur in regions characterised by low temperature factors;<sup>101</sup> while this observation suggests that active site residues are generally less flexible than the non-active site residues, i.e. have less vibrational motion within an energy well (“systematic flexibility”),<sup>100</sup> this does not exclude the possibility that they are instead rather “plastic”, i.e. that they can transform between discrete energy wells by overcoming potential energy barriers (“segmental flexibility”).<sup>100</sup> The catalytic effectiveness of an enzyme might very critically depend on its capability to change from one equilibrium conformation to another one in response to a *stimulus* (“triggered” conformational changes), while not depend on the fast collective motions of  $\alpha$ -carbons (vibrational motions).

In HIV-1 protease, residues that significantly contribute to the energetics of binding do not show any particular correlation between the nature of their contributions and the observed side-chain flexibility. For example, residues that strongly contribute



to the enthalpy of binding with electrostatic interactions are both found on very stable regions and on very flexible parts of the proteins. The same is true for residues which strongly contribute to the entropy of binding.

No significant correlations with the Tanimoto similarity scores of the ligands and the percentages of conformational changes observed in the corresponding pairs of holo-/holo- protein structures was found. Very similar ligands do not seem to be associated to large conformational changes. It is possible that a detailed analysis of ligand binding based on, for example, molecular interaction fields may yields a correlation between the nature of the ligand and the observed protein conformational change.

When 1BPY and 2BPZ, a pair of PDB entries that are bound to the same ligand, are compared, the percentages of conformational changes detected in their side-chain are surprisingly small. However, this trend should not be overestimated, since their crystallographic structures were solved by the same authors. The analysis of all pairs of holo-protein structures solved by the same crystallographers' group reveals in fact significantly less flexibility than that obtained by comparing all HIV-1 PDB entries. This fact could be used in an attempt to normalise the flexibility data to take into account such effects. This observation also reinforces the view that the extent of variability observed between protein X-ray structures is complicated by variable amounts of noise, making the disentangling of random effects from systematic trends very difficult.

A deeper conformational analysis of the protein system which appears to be the most flexible after HIV-1 protease, endothiapepsin, will be presented in the following chapter. This protein, similarly to HIV-1 protease, is an aspartic protease; eventual analogies and/or differences between the two proteins might help to understand the

peculiarities and trends of each protein system.

apter 5

ndottrapezino

Structure

Structure of the HIV-1 protease is a dimeric enzyme composed of two identical subunits. Each subunit is a small, globular protein with a molecular weight of approximately 25 kDa. The dimer is formed by two subunits interacting at their C-termini. The active site is located at the interface between the two subunits. The structure is characterized by a high degree of symmetry and a compact, well-defined fold. The active site is a deep, narrow cleft formed by the interaction of the two subunits. The structure is highly conserved across different HIV-1 strains, which is why it is a major target for antiretroviral therapy. The structure is also highly stable, which is why it is a major target for antiretroviral therapy. The structure is also highly stable, which is why it is a major target for antiretroviral therapy.

## Chapter 6

# Endothiapepsin

---

### 6.1 Structure

Endothiapepsin, an aspartic protease from the chestnut blight fungus (*endothia parasitica*), has been thoroughly studied to understand better the interactions that are important for ligand binding to this class of enzyme; as several inhibitors of endothiapepsin also bind renin, the design of endothiapepsin inhibitors could help the discovery of new antihypertensive drugs.

Many high-resolution structures of endothiapepsin co-crystallised with pepstatin A and several other inhibitors are available.<sup>102–104</sup> This enzyme is a 330 residue structure with the bilobal fold characteristic of the aspartate proteases, formed by two predominantly  $\beta$ -sheet domains of approximately equal size, and an extended binding-site cleft between them. Each of the two domains is mainly composed by  $\beta$ -sheets and exposed short helices on the outside of the protein; the two lobes are

connected by a six-stranded sheet (Figure 6.1).

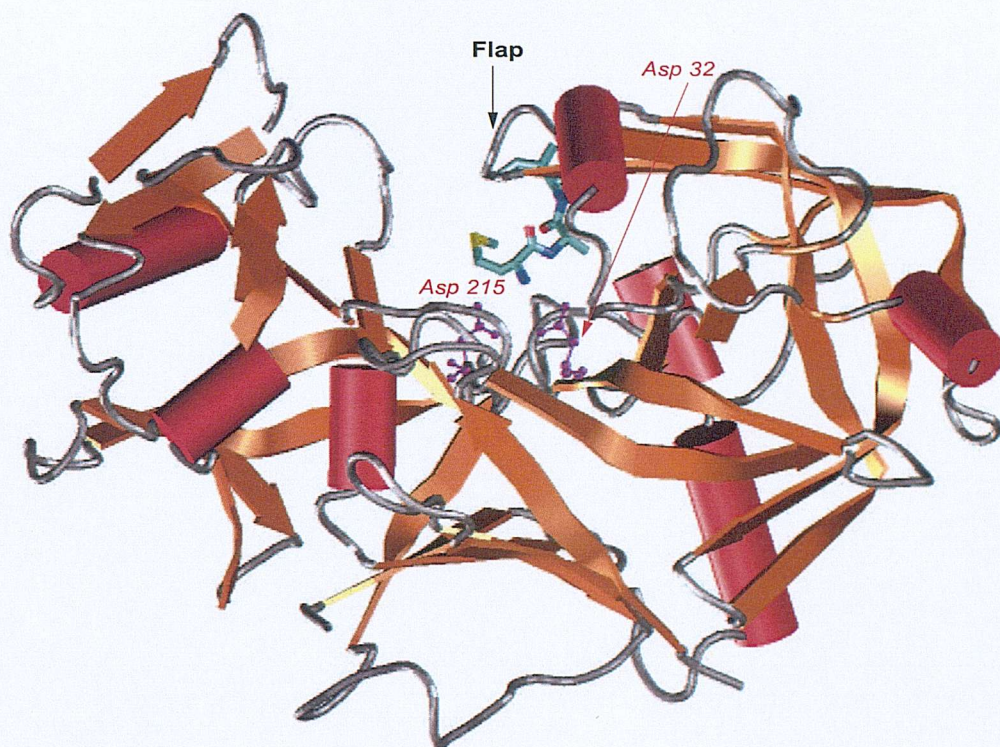
The active site of endothiapepsin is characterised by a symmetrical hydrogen-bond network that involves the sequences Asp32-Thr33-Gly34-Ser35 (from the N-terminal domain) and Asp215-Thr216-Gly217-Thr218 (from the C-terminal domain); the two catalytic sequences Asp-Thr-Gly are a conserved characteristic of the enzyme. Aspartic residues Asp 32 and Asp 215 are kept co-planar by the H-bonds established between the surrounding main-chain and side-chains atoms. Their carboxyl groups hold a water molecule, conserved in all aspartic protease native structures, which becomes deprotonated on substrate binding and initiate the general mechanism of catalysis.<sup>104</sup>

Residues 74 to 83 in the N-terminal domain lobe form a loop known as the *flap*: this highly flexible region in the native enzyme opens to allow access of the substrate or inhibitor to the active-site cleft. In the complexed form of the protein, the flap covers the catalytic groups of the enzyme and the central part of the inhibitor.<sup>105</sup>

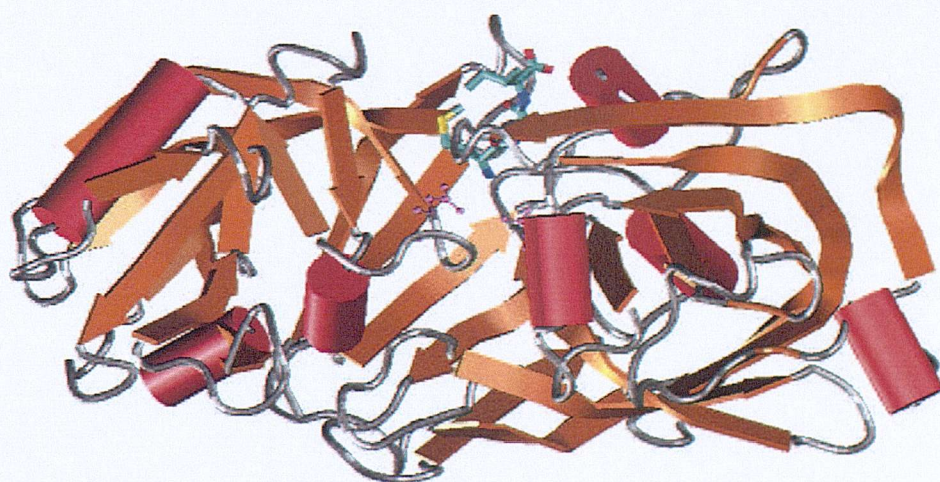
As for the majority of aspartic proteases, but different to HIV- protease, on binding endothiapepsin has been reported to undergo a predominantly shear motion (see section 2.2) at the interface between its two domains.<sup>102,106</sup> However, a more recent study<sup>16</sup> describes endothiapepsin's motion as a hinge motion, with the interdomain screw axis localised far from the backbone region where the rotational transition occur (see Figure 2.3). According to this theory,<sup>16</sup> endothiapepsin's overall rotational transition is located in the side-chain dihedrals of some residues that belong to the two different domains and establish noncovalent interactions between them; the hinge is created by the intrinsic flexibility of these noncovalent interactions and of their side-chain dihedrals.

Since there are relatively few interactions between them, the two domains of en-





(a) Front view. The enzyme is composed by two distinct domains of roughly 170 amino acids each, both contributing a residue to the catalytic dyad and separated by the deep binding site cleft. The fold of the protein is typical of all aspartic proteases, comprising mainly  $\beta$ -sheets and small areas of  $\alpha$ -helix exposed to the solvent.



(b) Top view. The flap is a loop formed by residues 74-83 in the N-terminal domain; in the bound form of the enzyme, it covers the catalytic groups of the protein and the central part of the inhibitor.

**Figure 6.1:** Cartoon representation of aspartic protease endothiapepsin complexed to the peptide HPH-MET-ALA-ILE (1E5O PDB entry). Catalytic residues Asp 32 and Asp 215 represented in balls and sticks.



dothiapepsin can move as independent rigid bodies; comparing different structures of endothiapepsin, the RMSd values are reduced by up to 47% when the two parts of the structure are superimposed independently.<sup>106</sup> The overall displacement of the two domains is about 1 Å, and the subsequent total rotation approximately 17°. Even if considerable distortions also occur within the two domains, their shear motion is the most important, and might play a role in distorting the substrates towards the transition state for proteolytic cleavage.<sup>102, 106</sup>

## 6.2 Background to Protein: Past Work

### 6.2.1 Structure-Based Thermodynamic Analysis of Endothiapepsin/ Pepstatin Binding

Isothermal Titration Calorimetry<sup>107, 108</sup> experiments (ITC, appendix A) were carried out by Gomez and Freire to fully characterise the energetics of the binding of endothiapepsin to pepstatin A.<sup>105</sup> This high-sensitivity calorimetry technique allowed the temperature and the pH-dependence of the binding energetics to be fully characterised, and the heat capacity change ( $\Delta C_p$ ) and the enthalpy of ionisation ( $\Delta H_{ioniz}$ ) defined (appendix A). The association constant is maximised at low pH and decreases with increasing pH, indicating that the binding of pepstatin A involves the protonation of one of the aspartates of the catalytic dyad of the protein. The structure-based thermodynamic analysis by Freire and Gomez<sup>105</sup> allowed an estimation of the contribution of each residue to the total free energy of binding, and the identification of the protein's and inhibitor's regions that contribute the most to the total free energy of binding. The conformational entropy parameters for the different amino acids were obtained from the analysis performed by Lee *et al.*<sup>109</sup> (see appendix A).

The experimental results showed that the reaction of protein/ligand binding is characterised by a significantly exothermic heat effect, and that binding is favoured both enthalpically and entropically.

While the favourable enthalpic contribution is expected, the favourable entropy change can be explained on the basis of the large gain in solvent-related entropy,  $\Delta S_{solv}$  (see appendix A). This primarily depends on the burial of a large hydrophobic surface upon binding that compensates for the loss in conformational, translational and rotational degrees of freedom upon ligand binding.

The PDB files 4APE and 4ER2 (part of the data set analysed in this thesis), were used by Gomez and Freire as the reference structures for the free and the complexed protein structures. Previous studies had shown that no major conformational changes occur when the apo-structure of endothiapepsin is compared to the holo-protein structure complexed with pepstatin A.<sup>110</sup> However, the binding of pepstatin A to endothiapepsin results, in the case of the PDB entries 4APE and 4ER2,<sup>105</sup> in the burial of a total of  $732 \text{ \AA}^2$  of apolar and  $652 \text{ \AA}^2$  of polar area previously exposed to solvent. Most of the changes in the accessible surface area (computed using Lee and Richards algorithm<sup>111</sup> with a  $1.4 \text{ \AA}$  probe sphere and a  $0.25 \text{ \AA}$  slice width) can be attributed to the peptide residues. In endothiapepsin, the largest changes in the protein accessible surface area occur in residues that are located in the catalytic site and strongly interact with the ligand. These are the residues surrounding the catalytic dyad (Gly 34 in the N-terminal lobe and Asp 215 to Leu 220 in the C-terminal lobe), the flap region (Ile73-Asp77), and some hydrophobic residues that favourably interact with the aliphatic side-chains of the ligand (Leu 120, Phe 189, Ile 297, Ile 301). Moreover, some residues that are not in direct contact with the inhibitor also undergo

conformational rearrangements; this *domino* effect induces significant changes in their solvent accessible surface, and provides close to 40% of the protease contribution to the binding free energy

The enzyme also contributes to the free energy of binding from an enthalpic point of view. The enzyme regions that contribute the most are the residues surrounding the catalytic dyad Asp 32 and Asp 215 (Asp 30 to Ser 37 in the N-terminal domain, Ile 213 to Tyr 222 in the C-terminal domain) and the flap region, from Ile73 to Asp77 (especially Ser 74, Gly 76 and Asp 77). The flap shows a great flexibility in the unbound enzyme and shields the inhibitor from the solvent in the complex, contributing the largest changes in the solvent accessibility of the protein. Other residues that contribute to  $\Delta G$  are Asp 12 and Asp 13, which interact directly with residues Iva 1 and Val 2 of the inhibitor, and regions of the protein surface that stabilise by van der Waals contacts other complementary surfaces (mainly aliphatic side-chains) in the inhibitor. For example, Phe 275 and Phe 284 stabilise Iva 1, while the isopropyl side-chain of Val 2 interacts with the aromatic ring in Phe 111. The latter induces in turn an important change in the accessibility of Ser 110.

In the inhibitor, a very large negative heat capacity change and a favourable entropic component contribute to the free energy of binding; moreover, all pepstatin A residues strongly and favourably contribute to the free energy of binding with favourable enthalpic contribution, thanks to their strong interactions with the enzyme. In particular, the two statine residues appear to contribute the most; this is in agreement with the known essential role that these residues play in pepstatin A-endothiapepsin recognition.

Overall, the energetic components calculated for the binding of pepstatin A to endothiapepsin were -4.4 kcal/mole for the enthalpic change, and -4.8 kcal/mole for

the entropic change.

## 6.3 Results: Analysis of Endothiapepsin Conformational Changes

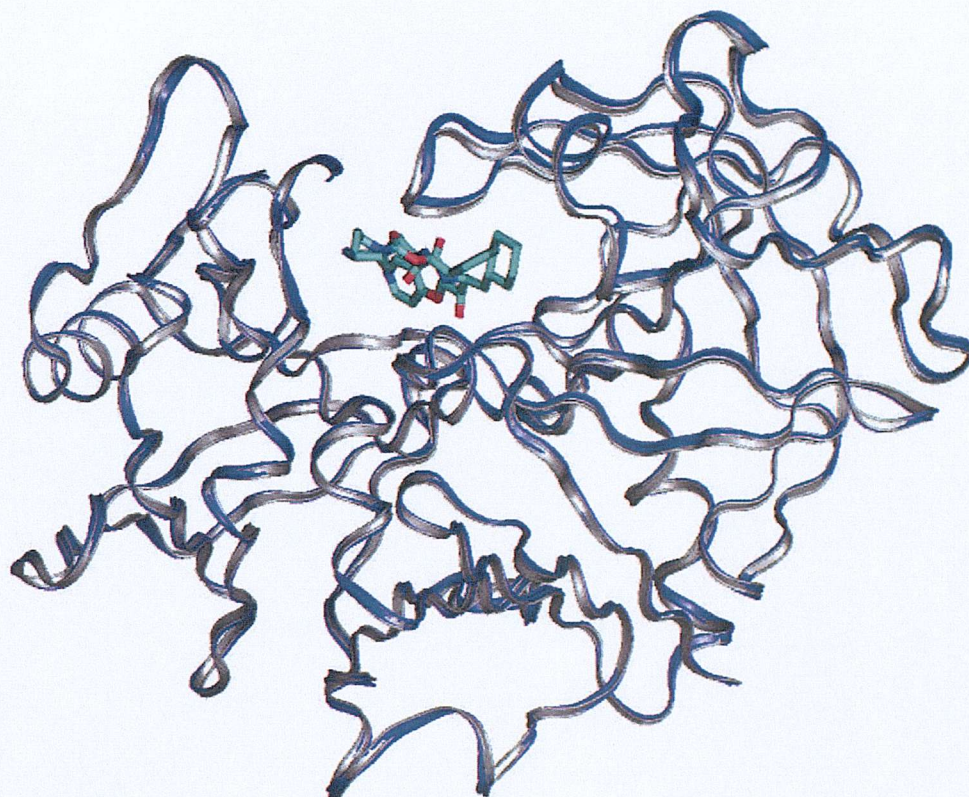
The data set of the present thesis includes 20 endothiapepsin holo-protein structures and 1 apo-protein structure (PDB entry 4APE) at resolution equal or better than 2.0 Å. The ligands in these PDB structures are all peptides of different sizes and often include unusual amino acids.

The binding of inhibitors to the enzyme, while not producing dramatic conformational changes in the protein (see Figure 6.2), cause small but significant changes in the orientation of the two domains, which move as independent rigid bodies with a small shearing motion at their interface.<sup>106</sup>

Since endothiapepsin's ligands are all peptidic, their Tanimoto similarity scores could not be evaluated employing 1024-bit Daylight fingerprints, which do not take into account paths bonds longer than 7, thereby loosing all information about the connectivity of the amino acids.

Although very similar peptides can be found among endothiapepsin ligands in the present thesis' data set, no 100% identical ones are found in different PDB entries; the analysis of side-chain flexibility in the presence of the same ligands was thus impossible in the case of this aspartic protease. Also, as all endothiapepsin structures share one or more PDB authors with all the other PDB structures in the data set (e.g. authors such as Blundell, Cooper and/or Veerapandian), it was not possible to divide endothiapepsin's PDB entries into groups of structures solved by the same authors which differed from the others in the data set. Because of these reasons, conformational analyses similar to those performed in the case of HIV-1 protease (see section 5.4.4) could not be carried out.





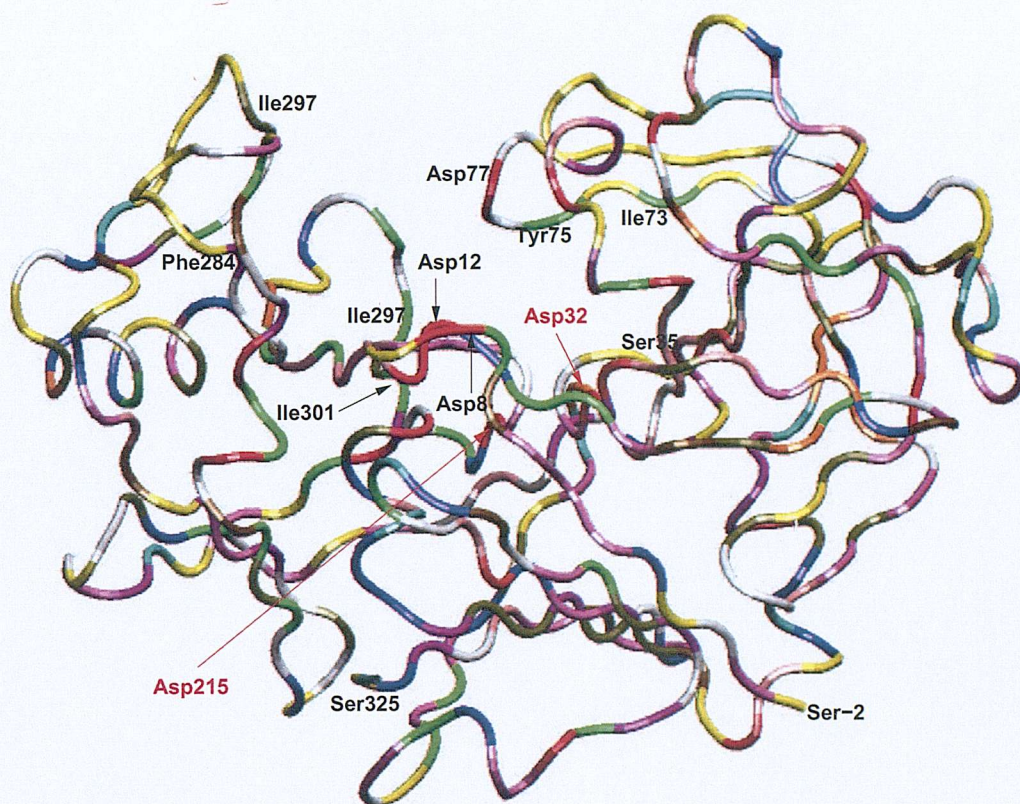
**Figure 6.2:** Ribbon representation of an apo-protein structure of endothiapepsin (PDB entry 4APE, coloured in grey) superimposed to a holo-protein structure of the protease (PDB entry 1E5O, blue; the structure of the 4 residues long inhibitor HPH-MET-ALA-ILE is shown). The backbone RMSd between the two superimposed structures is 0.30 Å.

Some residue sequence numbers and types have been indicated on a tube representation of endothiapepsin (PDB entry 1E5O) to help identifying some of the protein regions discussed in this chapter (Figure 6.3).

### 6.3.1 All Environments, and Environment Specific Conformational Changes

Figures 4.12-4.17 in chapter 4 report on the y axis the percentages of side-chain conformational changes observed in endothiapepsin when no distinction between exposed and buried residues is made; Najmanovich *et al.*,<sup>46</sup> Zhao *et al.*<sup>66</sup> methodologies and Dunbrack and Cohen rotamer libraries<sup>50,55</sup> were applied to obtain the data reported in these graphs.





**Figure 6.3:** Tube representation of residue-name coloured endothiapepsin. Some residue names and sequence numbers are indicated to help identify regions discussed in this section.

Some consistent trends are found in endothiapepsin with all methodologies of study. First, with the only exception of the rotameric parameter  $r2$  in holo-/holo-comparisons (Figure 4.16),  $\chi1$  and  $\chi2$  side-chain torsions are always more flexible in binding site residues rather than in all protein residues. These trends are probably better depicted in Figures 4.33, 4.34 and 4.35, where the differences between conformational changes observed in the binding site and in all the residues of the 10 protein systems have been plotted.

When residue environments are considered and residue are distinguished as exposed and buried, consistent trends are again found with all the methodologies of study (Figures 4.18-4.23 and 4.24-4.29 in chapter 4). Najmanovich *et al.* (Figures 4.24 and 4.25) and Dunbrack and Cohen methodologies (Figures 4.28 and 4.29) re-

veal that side-chain torsions of exposed residues are always more flexible than those of buried residues, both in apo-/holo- protein comparisons and in holo-/holo- protein comparisons. This is different from what was found for HIV-1 protease, in which apo-/holo- protein comparisons always detected greater buried residues' conformational changes in the binding site. If the more stringent cutoffs developed by Zhao and coworkers are applied, endothiapepsin's buried  $\chi_1$  torsions appear instead more flexible than the exposed in all protein residues (Figure 4.20 and 4.21). However, apo-/holo- comparisons fail to detect this trend in the binding site (Figure 4.20), and the difference between endothiapepsin's binding site buried and exposed residues' conformational changes revealed by holo-/holo- comparisons is almost negligible (Figure 4.21).

In summary, in contrast to HIV-1 protease, the uncomplexed form of endothiapepsin undergoes ligand-induced conformational changes that mainly involve exposed rather than buried binding site residues (apo-/holo- protein comparisons). In holo-/holo- comparisons, buried residues' conformational changes are slightly greater than those of exposed residues only with Zhao *et al.* angular thresholds.

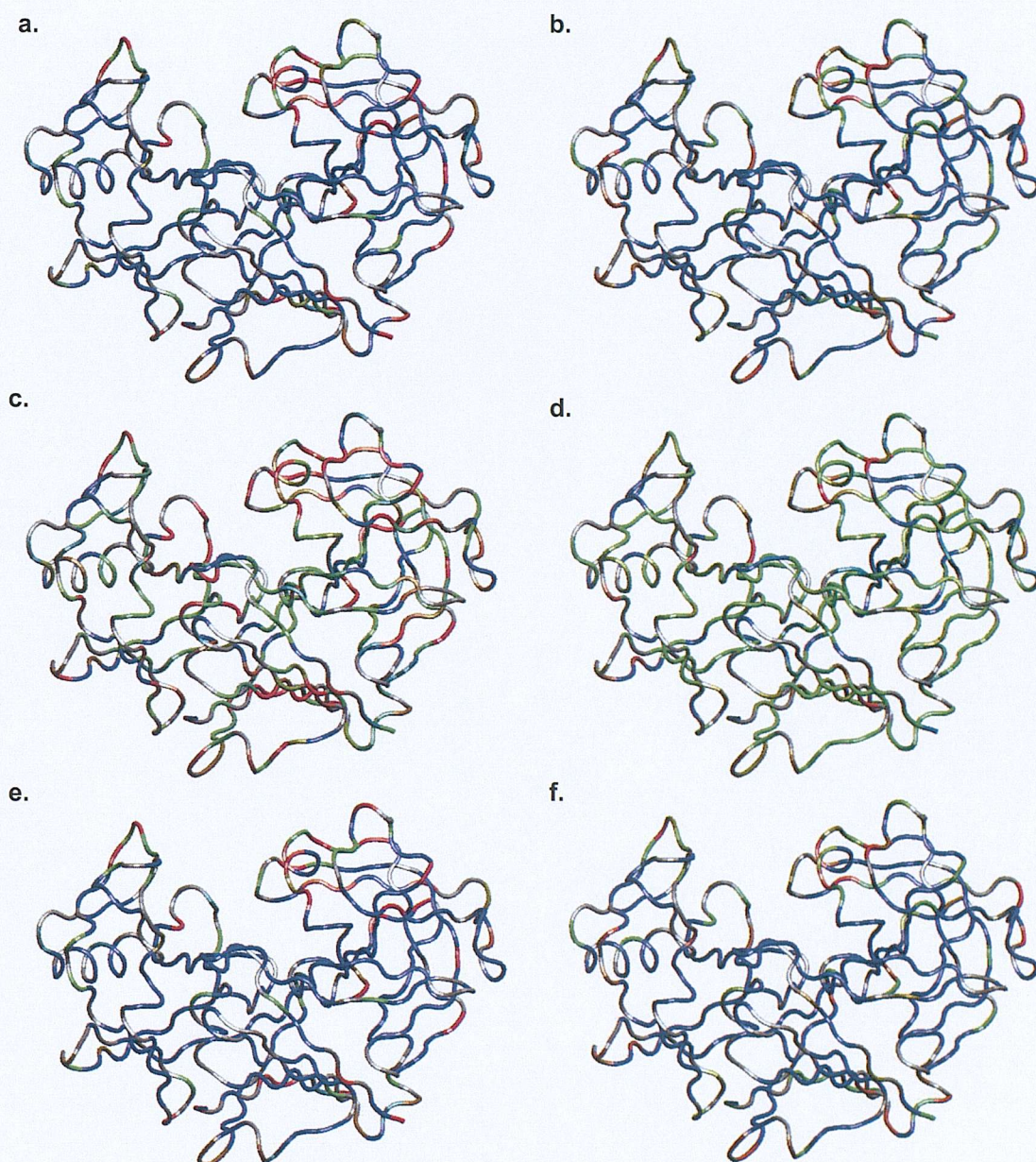
With all methodologies of analysis, similarly to HIV-1 protease, apo-/holo- comparisons reveal greater percentages of conformational changes than holo-/holo- comparisons. HIV-1 protease and endothiapepsin are the only two proteins in this thesis data set for which this peculiarity is found (See Figures 4.30, 4.31 and 4.32).

### 6.3.2 Percentages of Conformational Changes per Residue Sequence Number

Figure 6.4 shows the backbone structure of all endothiapepsin residues coloured in accordance to the flexibility of their side-chain torsions. The structures in the first row of the Figure (indicated by letters *a* and *b*), represent the results obtained with Najmanovich *et al.* methodology. The structures in the second row (*c* and *d*) are instead obtained with Zhao *et al.* methodology of study, and those in the third row (*e* and *f*) with Dunbrack and Cohen rotamer libraries. While structures in the left column refer to apo-/holo- protein comparisons results, those in the right column show holo-/holo- protein comparisons results.

A quick visual inspection of Figure 6.4 reveals that the most flexible regions (red) are mainly concentrated in the binding site of endothiapepsin, especially in the flap (residues 73-83). A few more regions, mainly located in loops, but that can also be found in more structured secondary structure elements such as  $\beta$ -strands and  $\alpha$ -helices, are also found to be very flexible. Broadly speaking, the same flexibility trends are observed with all methodologies. As observed in the case of HIV-1 protease, Figures **a** and **e** (apo-/holo- comparisons), and **b** and **f** (holo-/holo- comparisons) show a very similar colour pattern; this is expected, since Najmanovich *et al.* threshold and Dunbrack and Cohen rotamer libraries assign residues' flexibility on the basis of an absolute scale. When residue type- and environment- specific thresholds are instead applied (Figures **c** and **d**), a similar distribution of flexible (red) and more rigid regions (blue) is found, but a greater overall flexibility is generally observed. Also, the differences between the most unstable and the most stable regions' of the protease are levelled out. Again, this is not surprising; Zhao *et al.* thresholds are in





**Figure 6.4:** Tube trace of endothiapepsin backbone structure; all residues are coloured in accordance with the average percentage of conformational change its  $\chi_1$  torsions undergo, as detected with Najmanovich *et al.* (a, b), Zhao *et al.* (c, d) and Dunbrack *et al.* (e, f) methodologies of study. The residue average percentage of conformational change increases from blue coloured residues (which never change conformation) to red coloured residues (that change conformation 100% of times), passing through cyan, green, yellow and orange. Proline, glycine and alanine residues have been coloured in grey. Figures on the left (a, c, e) refer to apo-/holo- protein comparisons, figures on the right (b, d, f) to holo-/holo-protein comparisons.

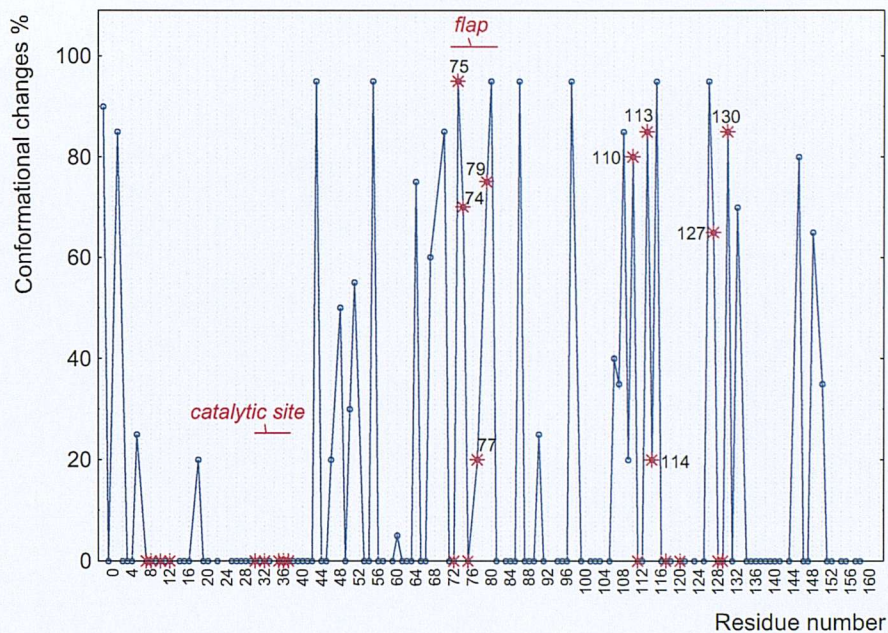
fact very stringent for residues that are buried to the solvent and/or are generally rigid, but are often very “permissive” with residues that are intrinsically more flexible and/or are exposed to the solvent. Thus, rather than residues that are very mobile on an absolute flexibility scale, they are likely to spot residues that are more mobile than expected for their residue type and/or their exposure to the solvent.

On the y axis of the graphs represented in Figures 6.5-6.8, the average percentages of residue conformational changes have been plotted. The residue on the x axis have been distributed on two separate graphs for clarity: *a* residues from -2 to 163 and *b* residues from 163 to 326. The first two graphs (Figures 6.5, and 6.6) were obtained with apo-/holo- protein comparisons and the last two (Figures 6.7, and 6.8) with holo-/holo- protein comparisons.

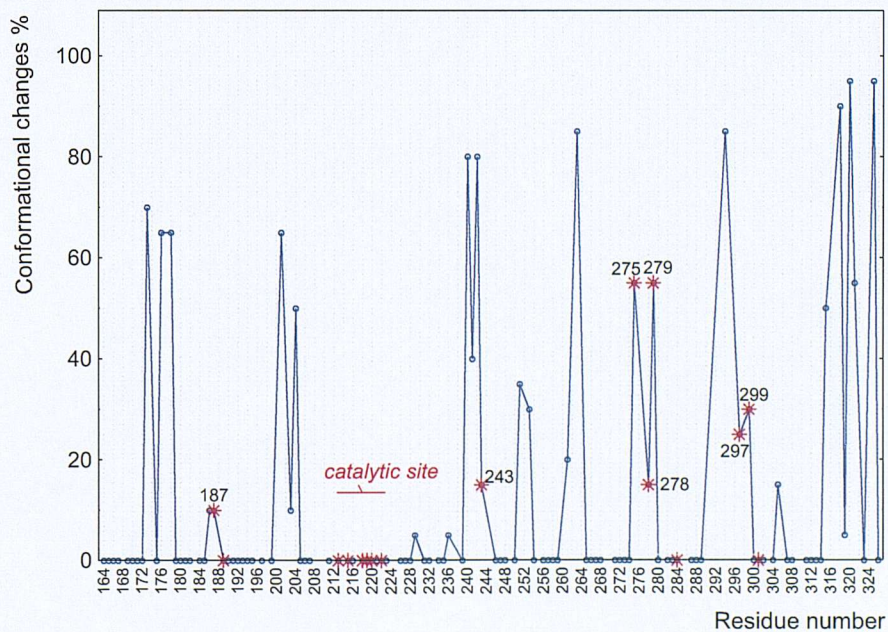
Results obtained applying Najmanovich *et al.* angular thresholds are reported first (Figures 6.5 and 6.7), followed by those obtained with Zhao *et al.* angular thresholds (Figures 6.6 and 6.8). Since the graphs obtained with Najmanovich *et al.* methodology of study are very similar to those obtained employing Dunbrack and Cohen rotamer libraries, the latter are not shown in the present chapter but reported in appendix 3 (Figures C.3 and C.4). To a lesser extent, the trends revealed by these two methods of analysis are also similar to those revealed by Zhao *et al.* angular thresholds (Figures 6.6 and 6.8). In the latter data, however, areas of the protease in which residues never change conformation are very rare. For example, conformational changes are also found for the catalytic aspartates: Asp 32 strikingly changing conformation 75% of times for apo-/holo- protein comparisons (Figure 6.6)

When Najmanovich *et al.* and Dunbrack and Cohen methodologies of study are applied, the binding site residues with the highest side-chain torsion flexibility that





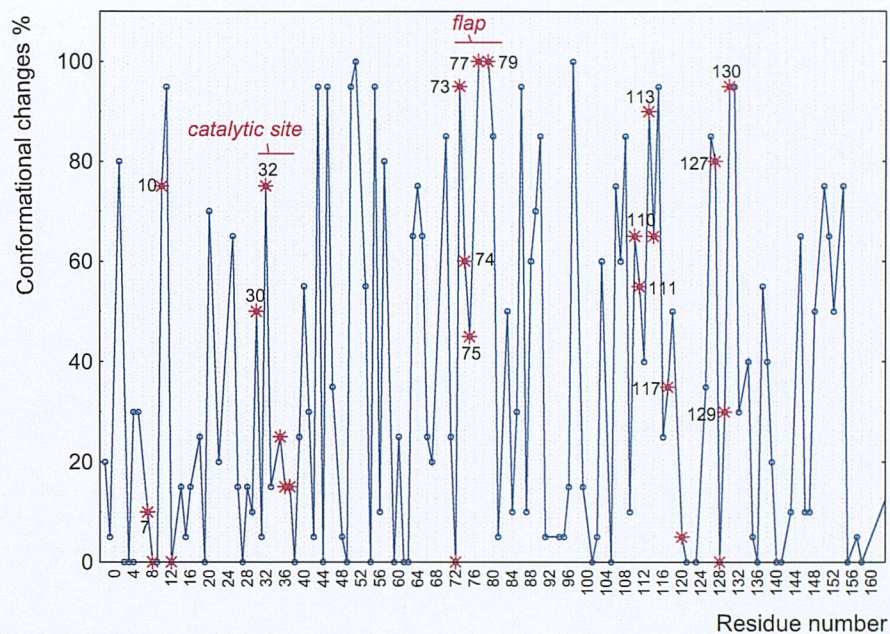
(a)



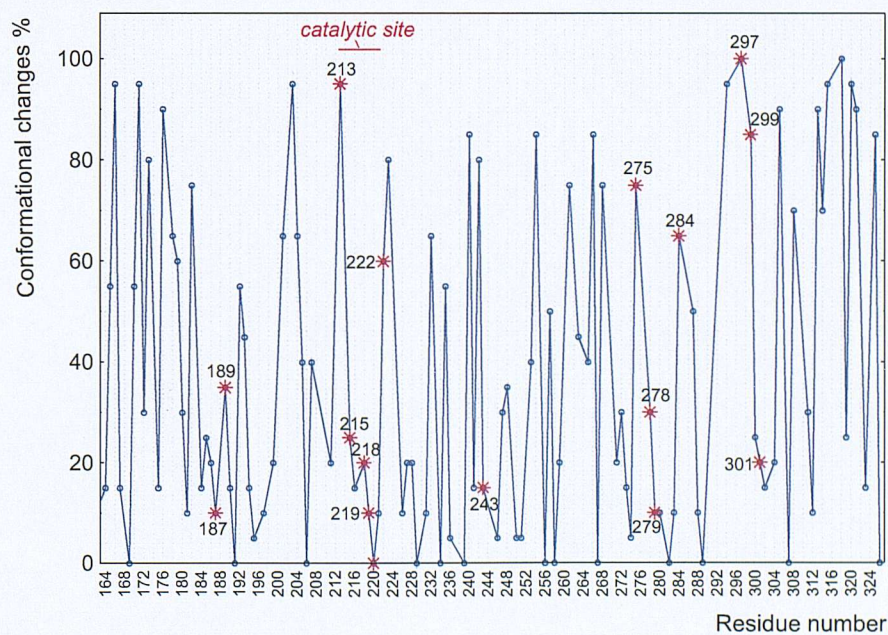
(b)

**Figure 6.5:** Percentages of times residues of endothiapepsin change  $\chi_1$  by more than  $60^\circ$  in apo-/holo- protein comparisons. Residues have been divided in Figure **a** (from residue -2 to 163) and **b** (residues from 163 to 326) for clarity. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data.





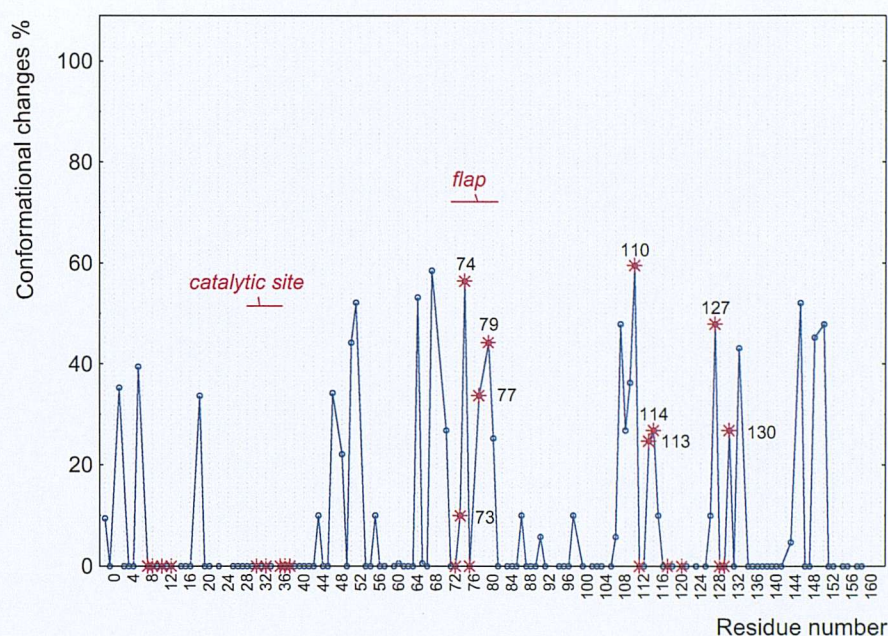
(a)



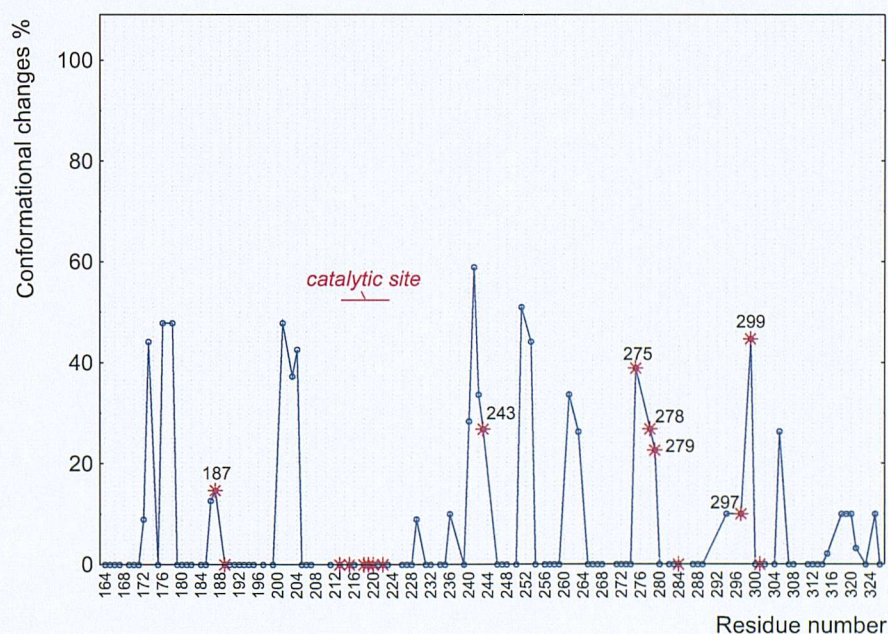
(b)

**Figure 6.6:** Percentages of times residues of endothiapepsin change  $\chi_1$  by more than Zhao *et al.* specific angular thresholds in apo-/holo- protein comparisons. Residues have been divided in Figure a (from residue -2 to 163) and b (residues from 163 to 326) for clarity. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data.





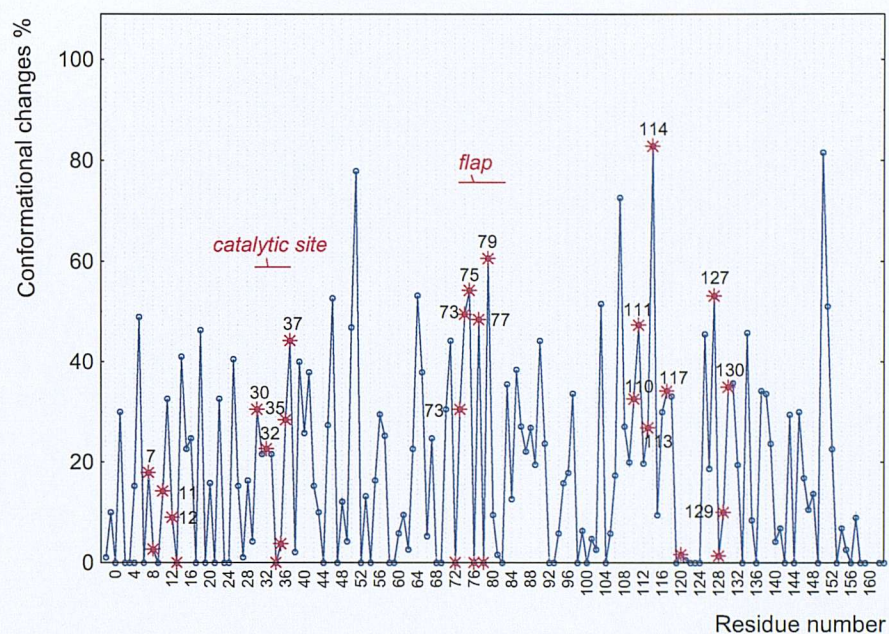
(a)



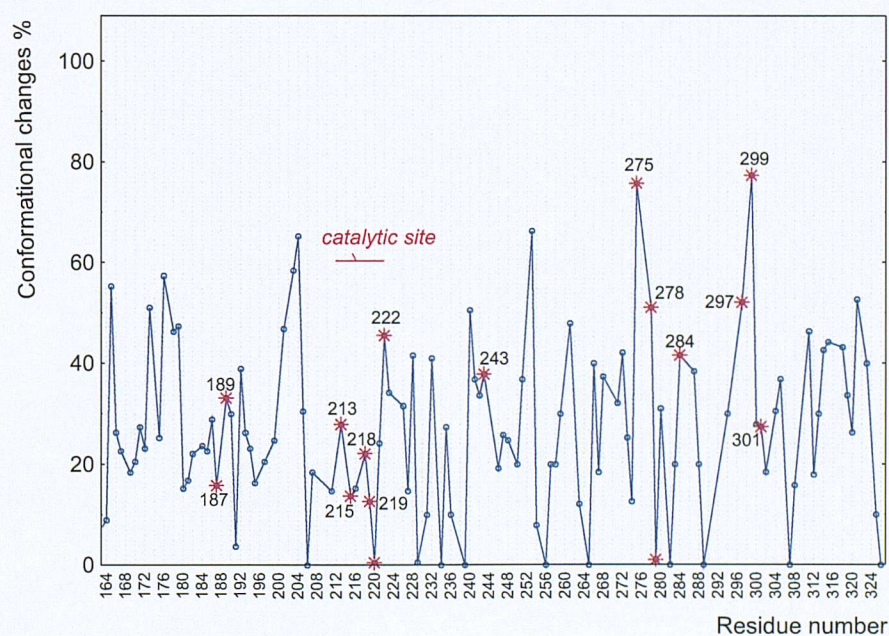
(b)

**Figure 6.7:** Percentages of times residues of endothiapepsin change  $\chi_1$  by more than  $60^\circ$  in holo-/holo- protein comparisons. Residues have been divided in Figure a (from residue -2 to 163) and b (residues from 163 to 326) for clarity. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data.





(a)



(b)

**Figure 6.8:** Percentages of times residues of endothiapepsin change  $\chi_1$  by more than Zhao *et al.* specific angular thresholds in holo-/holo- protein comparisons. Residues have been divided in Figure a (from residue -2 to 163) and b (residues from 163 to 326) for clarity. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any symbol.

are in contact with the ligands consistently appear to be located in the flap (Ile 73, Ser 74 and Ser 79 in apo-/holo- comparisons, and serines 74, 77 and 79 in holo-/holo-comparisons) and several residues in the region 110-130 (Ser 110, Glu 113, Thr 127 and Thr 130 in apo-/holo- comparisons, and Ser 110 and Thr 127 in holo-/holo-comparisons). Asp 114 shows a significant side-chain flexibility when the rotamer libraries by Dunbrack and Cohen are employed (Figures C.3 and C.4). Phe 275 and Ser 279 in apo-/holo- comparisons and Phe 275 and Ile 299 in holo-/holo- comparisons are also among the most flexible residues in the terminal part of the protein.

On the basis of Najmanovich *et al.* and Dunbrack and Cohen angular thresholds, the most stable regions of the protease in contact with the ligands comprise the catalytic site (residues 30-37 and 213-222), and residues 7-12; all are consistently characterised by percentages of conformational changes that equal to 0 (Figures 6.5, 6.7, C.3 and C.4). Further along the chain of endothiapepsin, Phe 111, Leu 120, Leu 128, Asn 129, Phe 189 and 284 and Ile 301 similarly never appear to undergo conformational changes in their  $\chi_1$ . The side-chains of Ser 72 (in the immediate proximity of the flap) and Tyr 75, on the flap, never change conformation; this is something unexpected, given the high flexibility of the flap structure.

The results obtained applying Zhao *et al.* angular threshold show significant differences with the above mentioned observations. First, the flap residue Tyr 75, for which no conformational changes are detected by the environment insensitive methods, shows with Zhao *et al.* methodology conformational changes which are greater than 40% in apo-/holo- comparisons, and 50% in holo-/holo-comparisons. On the contrary, the exceptional rigidity of the nearby residue Ser 72 is confirmed by this methodology of study. Asp 114, in support of the result found with Dunbrack



and Cohen methodology of study, is found to be the most flexible when holo-/holo-comparisons are considered (Figure 6.8). Ser 279, in disagreement with the results obtained with the other methods, is instead found to be rather inflexible. At the same time, several residues that were found to be highly rigid with Najmanovich *et al.* and Dunbrack and Cohen methodologies of study appear to be rather flexible applying Zhao *et al.*  $\chi^1$  thresholds. This is, for example, the case of Phe 111, Phe 189, Phe 284 and Ile 301, and, more interestingly, several catalytic residues of endothiapepsin.

Astonishingly, the catalytic residue Asp 32 appear to change conformation in approximately 75% of apo-/holo- protein comparisons. The same residue changes conformation in “only” 23% of holo-/holo- comparisons, a value not that dissimilar to the percentages of conformational changes that the second catalytic aspartate (Asp 215) shows in apo-/holo- and holo-/holo- protein comparisons (respectively 25% and 14%). The other residues in the catalytic site, including the catalytic residues Thr 33 and Thr 216 (which are not in contact with the ligands), show smaller flexibilities, with the significant exceptions of Asp 30 and Ile 213 in apo-/holo- comparisons (respectively 50% and 95% of conformational changes) and Asp 37 and Tyr 222 in holo-/holo- comparisons (44% and 46% of conformational changes). The finding of side-chain conformational changes in catalytic protein residues further supports the hypothesis that amino acids which are essential for enzymatic activity, although believed to be exceptionally rigid,<sup>105</sup> can effectively undergo side-chain motions that are larger than the average for the given residue type, in a similar environment.

Residues that significantly contribute to the energetics of binding with a strong entropic interaction (see section 6.2) do not show any particular correlation between the nature of their contributions and the observed side-chain flexibility. In fact, they

can be found both in very flexible regions (e.g.: Ile 73, in the flap) and very stable regions (e.g.: Leu 120, Asp 8). Similarly, amino acids that contributes to the free energy of binding also with a strong enthalpic contribution can also be found in very flexible (such as the flap) or very rigid (e.g. Asp 12) regions.

In agreement with previous studies,<sup>88,92</sup> binding site residues are found to belong to both highly rigid and highly flexible regions.

## 6.4 Conclusions

In this chapter, the conformational changes of endothiapepsin, an aspartic protease, have been analysed. With all methodologies of study, the binding site residues of this protein appear more flexible than all protein residues (see chapter 4). Also, apo-/holo- protein conformational changes are always greater than holo-/holo- protein comparisons conformational changes (see chapter 4); this trend is a peculiarity of endothiapepsin and HIV-1 protease, the only proteins among the ten protein systems of the present thesis' data set which show this behaviour.

Upon ligand binding, the two domains of endothiapepsin move as rigid bodies, with a motion that has been both described as predominantly shear<sup>102,106</sup> and as a hinge motion,<sup>16</sup> whose screw axis is located in a region of side-chain interactions rather than the backbone region where the rotational transition is found. However, many other residues rearrange themselves after ligand binding, both as a consequence of direct interactions with the ligand and as a sort of *domino* effect. The flap, comprising residues 74 to 83, is a highly flexible loop that opens to allow the access of inhibitors or substrates, and moves over the ligand shielding it from water.

Contrary to HIV-1 protease, endothiapepsin buried residues never appear to be more flexible than exposed residues when undifferentiated angular thresholds (i.e.

Najmanovich *et al.* and Dunbrack and Cohen methodologies of study) are applied (see chapter 4). Further, Zhao *et al.* angular thresholds also detect greater exposed  $\chi_1$  conformational changes in apo-/holo- comparisons and binding site residues (see chapter 4); this evidence suggests that the mechanisms at the basis of HIV-1 protease and endothiapepsin ligand-binding processes are different. As Gomez and Freire reported (see section 6.2), both strong favourable enthalpic and entropic contributions are responsible of the highly exothermic character of endothiapepsin/ligand binding reactions.<sup>105</sup> While strong favourable interactions between the protein and the ligands are the cause of the favourable enthalpic change, the greatest factor contributing to a highly positive entropic change is the burial of a large apolar surface area on the ligand molecules.<sup>105</sup>

The smaller flexibility of buried residues in the binding site of endothiapepsin (when compared to HIV-1 protease) could find a reason in the fact that endothiapepsin ligand-binding reactions are strongly exothermic, thus, in contrast to HIV-1 protease, a high, favourable  $\Delta S$  is not necessary for ligand binding to occur. Also, up to 40% of the protease  $\Delta S$  contribution to the binding free energy is provided by the burial of the apolar accessible surface area of residues which are not located in the binding site of the protein, but which change conformation upon ligand binding (*domino* effect).<sup>105</sup> This might explain the higher percentages of conformational changes which are found in all rather than in only binding site residues of endothiapepsin (apo-/holo- protein comparisons).

Similar trends in the distribution of residue side-chain flexibility are found with all methods of analysis. As expected, the results obtained with Najmanovich *et al.* and Dunbrack and Cohen methodologies of study are very much similar, analogously

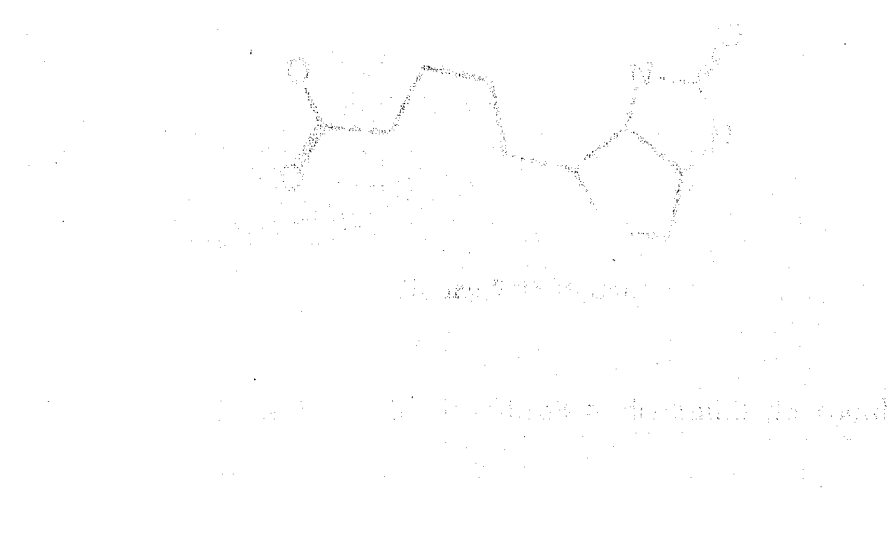
to those noticed in the case of HIV-1 protease.

The most stable regions of endothiapepsin that are in contact with the ligand include aspartates 8, 11 and 12, Ser 72, Leu 120 and Leu 128, Leu 220 and Ser 279. The most flexible regions of the protein consistently include the flap (residues 74-83) and many of the residues that favourably interact with the ligands (such as Ser 110, Glu 113, Thr 127, Phe 275 and Ile 299).

Similarly to what observed in the case of HIV-1 protease, the residues that significantly contribute to the energetics of binding of endothiapepsin do not show any particular correlation between the nature of their contributions and the observed side-chain flexibility. Residues that strongly contribute to the enthalpy and/or entropy of binding are both found on very stable regions and on very flexible parts of the protein.

In the next chapter, another highly flexible protein system, streptavidin, will be analysed. Its binding site, similarly to those of endothiapepsin and HIV-1 protease, appears more flexible than all protein residues with all methodologies. The existence of several streptavidin good resolution apo-structures in the PDB allows analysis which was not possible in the case of endothiapepsin and HIV-1 protease; also, since streptavidin binds non-peptidic ligands, the computation of Tanimoto similarity indices of its ligands with 1024-bit long fingerprints can be performed, and the correlation between the similarity scores of its ligands and the observed side-chain conformational changes will be sought.





The image shows the chemical structure of biotin, a small organic molecule. It consists of a central five-membered ring with a sulfur atom at the bottom. Attached to this ring are a methyl group, a propyl chain, and a side chain that includes a methylene group, a methoxy group, and a pyrimidine ring. The pyrimidine ring has a methyl group at the 6-position and a carboxylic acid group at the 2-position.

## Chapter 7

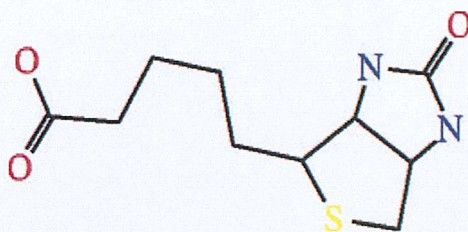
# Streptavidin

---

### 7.1 Structure

Streptavidin takes its name from the bacterial source of the protein, *Streptomyces avidinii*, and the hen egg-white protein avidin. Both avidin and streptavidin share an exceptionally large free energy of association with biotin ( $K_{\text{association}} = 10^{14} M^{-1}$ ). This high affinity is one of the largest of observed for noncovalent binding of a protein and small ligand in aqueous solution, and is the reason why, even if the biological function of streptavidin is not yet fully understood, many biochemical applications of the streptavidin/biotin system have been exploited, and the characteristics of this complex extensively studied as a model system of ligand/protein interactions.<sup>112–114</sup>

Streptavidin is a homo-tetrameric 159 residues protein in which each monomer binds one molecule of the vitamin biotin (represented in Figure 7.1) or other ligands (Figure 7.2). In many crystal structures, the visible electron density at both the N-



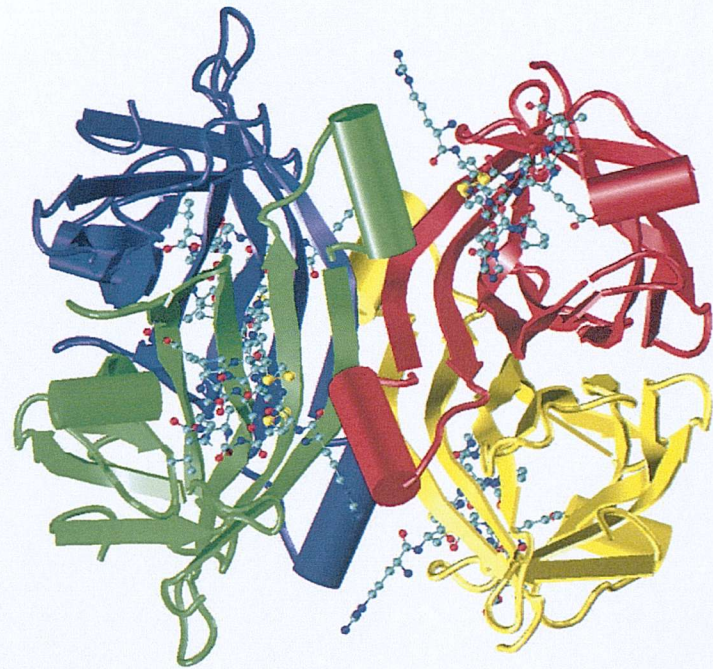
**Figure 7.1:** Biotin.

and the C-termini is weak, making it difficult to determine the coordinates of each protein chain up to its full length.

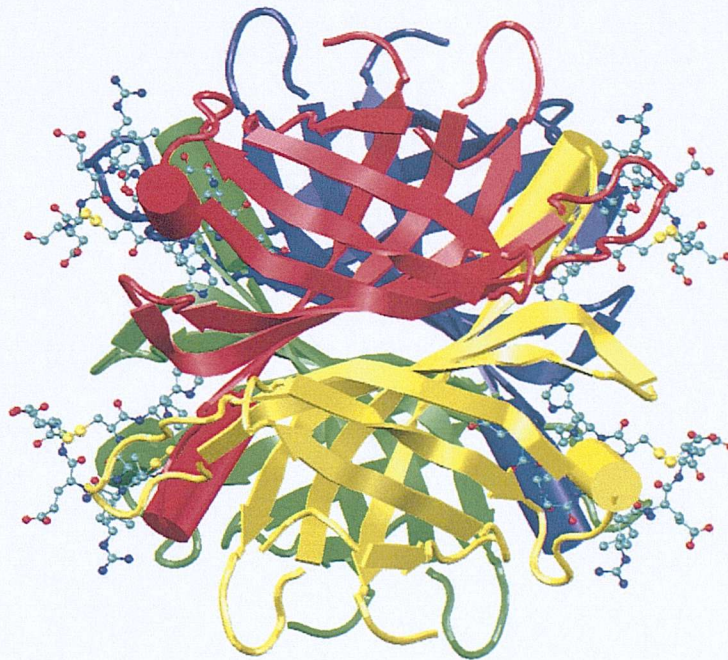
The streptavidin monomer is organised as an 8-stranded  $\beta$ -barrel; pairs of the barrels bind together to form symmetric dimers, pairs of which in turn interdigitate to form the naturally-occurring tetramer (Figure 7.2). Because of the intensive inter-subunit contacts that are shown by the pairs of subunits 1 + 2, and 3 + 4, streptavidin can be considered a *dimer of dimers*; the asymmetric unit of streptavidin generally contains two avidin polypeptide chains (Figure 7.3), which build up the functional tetramer through a crystallographic 2-fold axis. However, tetrameric asymmetric unit have been deposited in the PDB.

Several crystallographic forms of streptavidin have been described. In general, the flexible loop comprising residues 45 to 52 was found to be open in the unbound form of the protein and closed in the complexed form of the protein.<sup>115,116</sup> However, more recent crystallographic studies have found the loop closed even in the absence of biotin, supposedly as a consequence of crystal packing interactions.<sup>114</sup> The open and closed conformations of the mobile loop were found to be correlated to the crystallographic B-factor values, highlighting a greater disorder in the open rather than in the closed loop structure.<sup>114</sup> The loop is, however, not totally disordered in the open state, and a helical segment is observed in residues 49 to 53 in almost all unbound structures. If





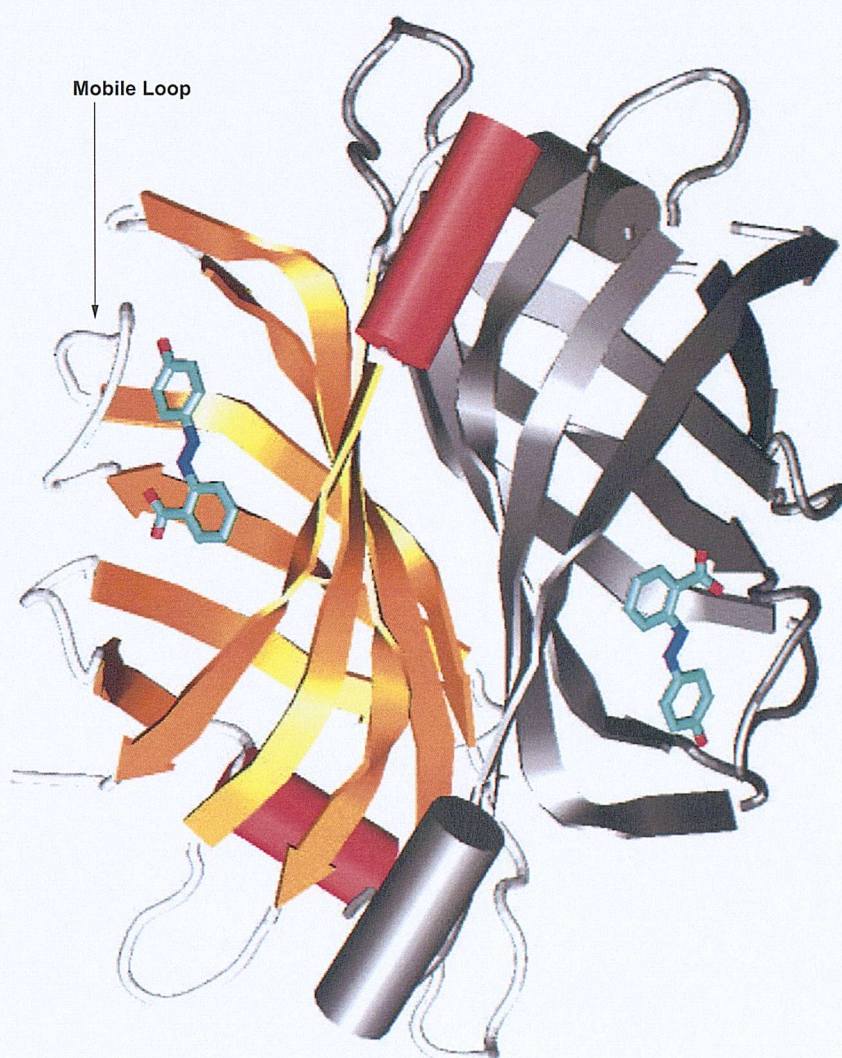
(a)



(b)

**Figure 7.2:** Cartoon representation of the streptavidin tetramer bound to the miniprotein mp-2 (PDB entry 1HQQ, 1.7 Å resolution; this structure is not included in the present thesis data set). Front (a) and side (b) view. Because of the intensive inter-subunits contacts between subunits 1 + 2 (red and yellow) and 3 + 4 (green and blue), the tetramer can be considered a *dimer of dimers*. The asymmetric unit of streptavidin generally contains two avidin monomers.<sup>114</sup>





**Figure 7.3:** Cartoon representation of the dimeric form of streptavidin complexed to HABA (1SRE PDB entry).  $\beta$ -sheets and helices are coloured in orange and red; each streptavidin's monomer is formed by a 8-stranded  $\beta$ -barrel, two  $\alpha$ -helices and extensive hairpin loops. The surface loop that folds over the ligand molecules (residues 45-52) is indicated.



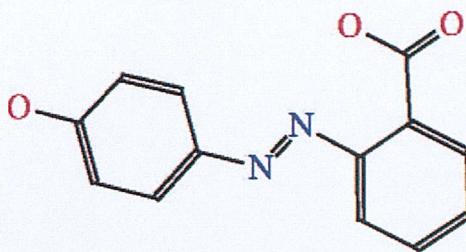
this secondary structure element is always present in solution, its “melting” when the loop goes from the open to the closed conformation could lessen the conformational unfavourable entropy contribution associated with loop ordering in the bound state. Conversely, the loss of the intra-helix hydrogen bond interactions may introduce an unfavourable enthalpic contribution; however, the formation of several hydrogen bond interactions with biotin can diminish this term.<sup>114</sup>

In the streptavidin tetramer, the Trp 120 side-chain is a key element for the connection of adjacent binding sites across the dimer-dimer interface. Crystal forms of streptavidin where not all four monomers are bound to biotin are observed; in these structures, only subunits 1 and 4 (numbering analogous to that employed in Figure 7.2) are complexed, and monomers 2 and 3 are unbound and characterised by the mobile loop open conformation.<sup>114</sup> This suggests the possibility of communication between the binding sites of different monomers, and in turn cooperative binding, as several studies on streptavidin have hypothesised.<sup>114,117,118</sup>

## 7.2 Background to Protein: Past Work

### 7.2.1 Structure-Based Thermodynamic Analysis of Streptavidin Binding to Structurally Diverse Ligands

Many studies have addressed the issue of the decomposition of the free energy of binding of streptavidin to biotin (Figure 7.1) and/or other ligands, and the identification of residues that are essential in the biotin binding process. The methods employed include molecular dynamics (MD) simulations,<sup>113</sup> free energy perturbations,<sup>119</sup> isothermal titration calorimetry and/or crystallographic structures studies,<sup>120</sup> site directed mutagenesis analysis<sup>121</sup> and “shotgun scanning”, a technique



**Figure 7.4:** 2-((4'-hydroxyphenyl)-azo)benzoate (HABA). This figure is derived from the ligand structure in the PDBsum, which omits hydrogen atoms.

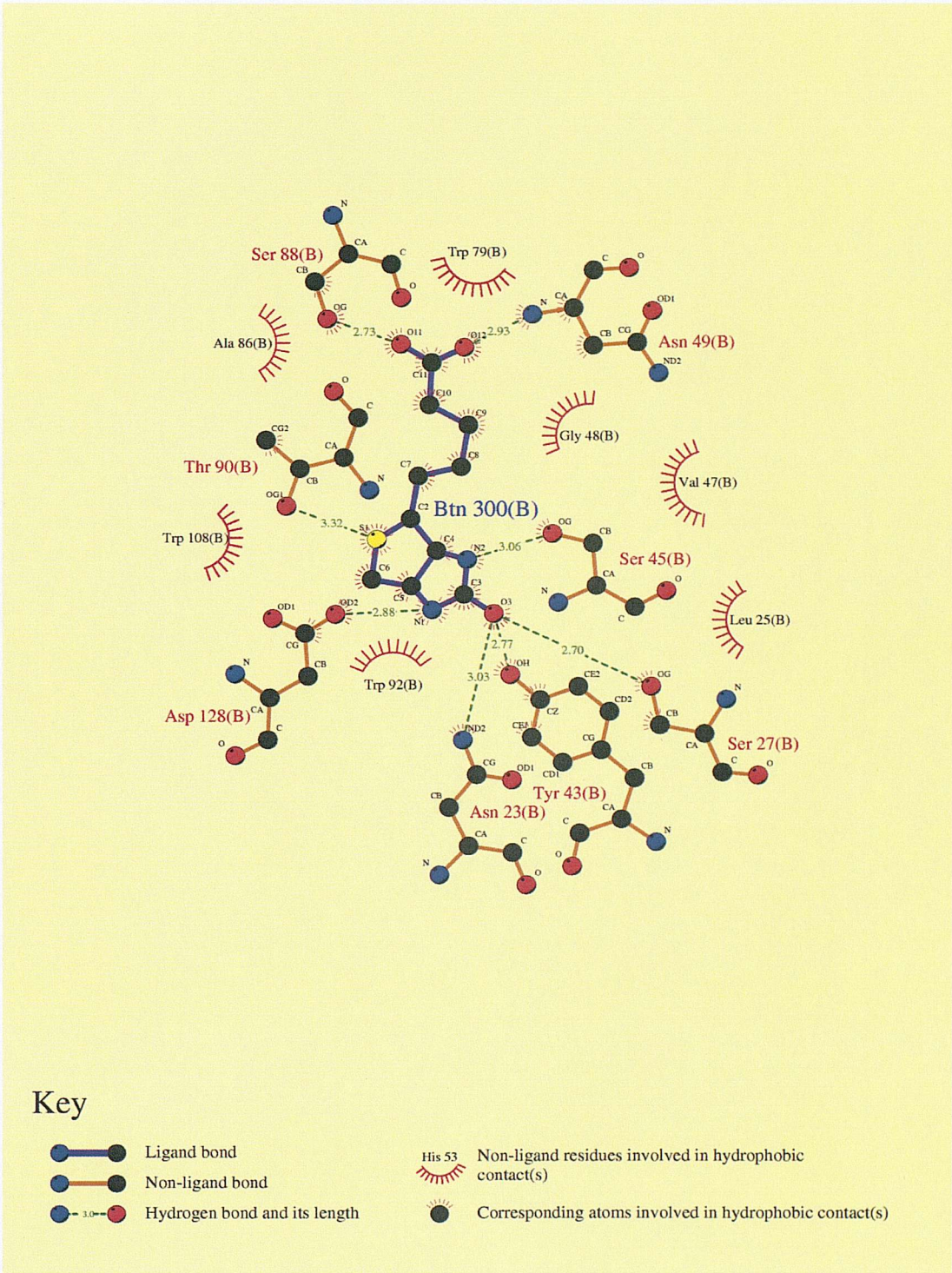
that employs a combinatorial library of proteins with the alanine side-chain substituted in every position.<sup>122</sup> In particular, the thermodynamic binding parameters and the crystallographic structures of three classes of streptavidin ligands (biotin, 2-[(4'-hydroxyphenyl)-azo]benzoate (HABA, represented in Figure 7.4) and derivatives, and peptidic ligands such as  $\text{NH}_2\text{-Phe-Ser-His-Pro-Gln-Asn-Thr-COOH}$ ) were compared.<sup>120</sup>

Thermodynamic measurements of biotin (Figure 7.1) binding to streptavidin (ITC, see appendix 3) show that the process is predominantly driven by an exceptionally large enthalpic contribution.<sup>116,120</sup> This is due to non-specific favourable packing interactions and, to a greater extent, to specific interactions with the ligand molecule (Figure 7.5) which are surprisingly strong, especially considering the small size of biotin.<sup>116,120</sup>

The binding site within each streptavidin monomer is located in a deep pocket at the centre of the  $\beta$ -barrel (see section 7.1), and includes both hydrophobic and polar residues involved in the recognition of biotin. In particular, three different biotin-binding motifs are observed:<sup>120</sup>

1. hydrophobic and van der Waals interactions, mainly involving four streptavidin tryptophan side chains (Trp 79, 92, 108, 120);
2. an effective hydrogen bonding network comprising Asn 23, Ser 27, Tyr 43, Ser





**Figure 7.5:** Interactions between biotin and streptavidin (PDB entry 2IZF) as obtained by the program LIGPLOT.<sup>123</sup> Hydrogen bonds and hydrophobic contacts are respectively represented by dashed lines and arcs with spokes radiating towards the ligand atoms they contact; the contacted atoms are shown with spokes radiating backwards. LIGPLOT calculates hydrogen bonds and non-bonded contacts with the program HBOND by McDonald and Thornton (1994).<sup>124</sup>

45, Asn 49, Ser 88, Thr 98, Asp 128;

3. the binding surface loop, which folds over the ligand, and whose Asn 49 forms a hydrogen bond with one of the vitamin's carboxyls (residues 45 to 52)<sup>120</sup>.

While calculations from Weber and co-workers have indicated the hydrogen bonding network as probably the most important factor contributing to the free energy of binding,<sup>120</sup> the results obtained by Kollman and co-workers have pointed at the importance of tryptophans 79, 92, 108 and 120.<sup>119</sup> A large part of the streptavidin/biotin interaction energy comes, in fact, from van der Waals interactions;<sup>113,119</sup> however, hydrogen bonds have also a key role in determining the very favourable streptavidin/biotin enthalpy of binding. In particular, the ureido group's of biotin is characterised by a resonance form which could be further stabilised by an additional hydrogen bond that the biotin's urea oxygen atom (Figure 7.1) can establish with streptavidin.<sup>120,122</sup> The possibility of biotin bound to streptavidin making three hydrogen bonds (*versus* the two hydrogen bonds established in water) allows a greater negative charge to be displaced on biotin's urea oxygen atom; this favourable contribution to binding might not have been taken into account in Kollman and co-workers,<sup>119</sup> as the polarisation of the ureido group induced by a charged aspartic residue of streptavidin might not have been included in classical force fields.<sup>113</sup>

Another strong factor contributing to the high affinity between streptavidin and biotin is the negligible ligand reorganisation energy; the conformational rearrangement of biotin in the binding site of the protein is in fact very small, and its conformational energy very similar to that in solution. Also, the small size of biotin contributes to a small ligand conformational entropy loss upon binding.

Most of the residues that bind biotin through hydrogen bonds are already or-



ganised in the correct binding geometry in the streptavidin apo-protein structure.<sup>122</sup> However, a recent MD simulation study by Lazaridis *et al.*<sup>113</sup> argues that the protein reorganisation energy makes a large unfavourable contribution to the free energy of binding, even if streptavidin's binding site is relatively well pre-organised; in the authors opinion, a significant reorganisation energy can arise even from subtle side-chains reorganisations, without significant main backbone conformational changes.<sup>113</sup>

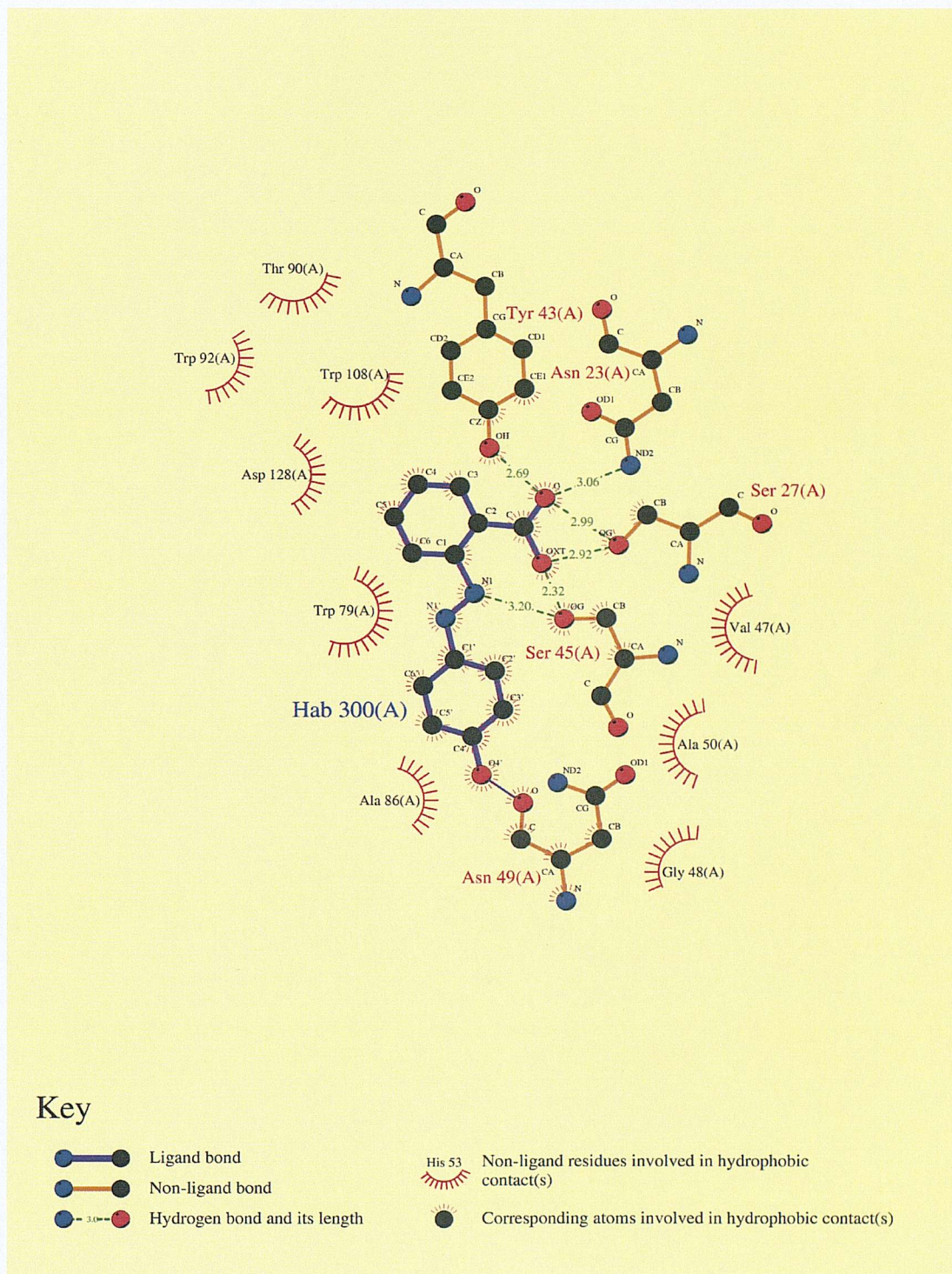
The protein reorganisation energy, principally due to the loss of intra-protein interactions, predictably increases with the increasing of ligand size, together with the ligand conformational entropy change and the ligand reorganisation energy.<sup>113</sup> Intuitively, the more rigid a ligand is, the higher its reorganisation energy necessary to adopt the binding conformation will be, and, in compensation, the smaller the conformational entropy cost associated with binding.

In addition to the contribution given by residues that are in direct contact with biotin, some studies point out the importance of previously unreported hydrophobic residues, which contribute with indirect contacts to the energetics of binding.<sup>121</sup> Residues that neighbour the active site, together with longer range networks of hydrophobic residues, could help to orient, in the optimal geometry, the side-chain of residues directly involved in biotin binding. Also, residues which are unimportant for biotin binding might be critical for the formation and the stabilisation of  $\beta$ -barrels, through hydrophobic collapse and inter-strands crossed interactions.<sup>121</sup>

HABA (Figure 7.4), an azobenzene employed to quantify the concentration of biotin binding sites in solutions of streptavidin and avidin, binds streptavidin approximately 10 orders of magnitude less tightly than biotin. Thermodynamic measurements show that the interaction energetics are dominated by favourable entropy components.<sup>120</sup>

The loop comprising residues 45-52, well ordered in the biotin-streptavidin complexes, is partially ordered in the HABA-streptavidin structure, its residues interacting with ligand atoms (key streptavidin/HABA interactions represented in Figure 7.6). The benzoic acid portion of this ligand binds at the bottom of the biotin binding pocket; its hydroxy-phenyl ring is partially buried inside the protein and partially exposed to the solvent.<sup>120</sup> Three hydrogen donor groups belonging to streptavidin residues Tyr 43, Asn 23 and Ser 27, the ones that stabilise the tetrahedral oxyanion formed by biotin's ureido group, establish a key ionic interaction with HABA. HABA also takes part to a hydrogen bond network that involves the hydroxyl group of Ser 45, while the side-chains of tryptophan 90, 92, 108 and 120 form a hydrophobic pocket that hosts the phenyl-benzoic portion of the ligand, Trp 108 forming a favourable inter-aromatic edge-to-face interaction (see Figure 7.6). The hydroxyphenyl ring of the ligand interacts with Trp 79 at the base of the pocket.<sup>120</sup> These interactions, represented in Figure 7.6, appear, however, as a kind of energetic compensation for the loss of hydrogen bonds that the ligand was establishing with solvent molecules prior to binding; ligand desolvation and/or other entropic factors appear to dominate the binding energetics of HABA.<sup>120</sup>

Several modified derivatives of HABA have been synthesised and their affinity for streptavidin tested. For example, X-ray structure studies of 3',5'-dimethyl-HABA show that this ligand is less ordered than HABA in the binding pocket of streptavidin; thermodynamic analyses show that it binds streptavidin approximately three times more tightly than HABA. The increase of the binding free energy appears to be totally attributable to a favourable entropy gain, probably due to a greater retention of conformational flexibility in the ligand molecule and to the displacement of an additional water molecule in the binding site pocket.<sup>120</sup>



**Figure 7.6:** Interactions between HABA and streptavidin (PDB entry 1SRE) as obtained by the program LIGPLOT.<sup>123</sup> Hydrogen bonds and hydrophobic contacts are respectively represented by dashed lines and arcs with spokes radiating towards the ligand atoms they contact; the contacted atoms are shown with spokes radiating backwards. LIGPLOT calculates hydrogen bonds and non-bonded contacts with the program HBOND by McDonald and Thornton (1994).<sup>124</sup>



ITC studies<sup>120</sup> show that the seven residue peptide  $\text{NH}_2\text{-Phe-Ser-His-Pro-Gln-Asn-Thr-COOH}$  binds streptavidin with a very similar affinity to HABA; however, the energetic components into which the energy of binding can be dissected are totally different. In the case of the peptide, the enthalpy of binding is in fact the driving force of the binding process, and is opposed by a large unfavourable entropy component. While the cause of the unfavourable entropic factor can be intuitively attributed to the necessity of freezing the numerous internal degrees of freedom that characterise peptidic binding, the reason for the highly favourable enthalpic contribution to binding must be probably sought in the extensive hydrogen-bonds network involving the heptapeptide molecule.<sup>120</sup> In the crystallographic structure analysed by Weber *et al.* (PDB entry 1PTS), only the three central residues of the peptide have a clearly visible electron-density map.<sup>120</sup>

### 7.3 Results: Analysis of Streptavidin Conformational Changes

The data set of the present thesis includes 13 streptavidin holo-protein structures and 6 apo-protein structures at resolution equal or better than 2.0 Å. All the 6 apo-proteins analysed in the present thesis data-set are characterised by a closed loop conformation, possibly as a consequence of crystal packing interactions (see section 7.1). The ligands in the complexed PDB structures comprise biotin and biotin derivatives, HABA and HABA derivatives, and peptidic ligands.

The binding of ligands to streptavidin does not cause dramatic conformational changes in the protein, even if some studies hypothesise that a significant reorganisation energy can arise even from subtle side-chain reorganisation.<sup>113</sup> Of course larger ligands, such as peptidic molecules, are likely to involve greater protein conforma-



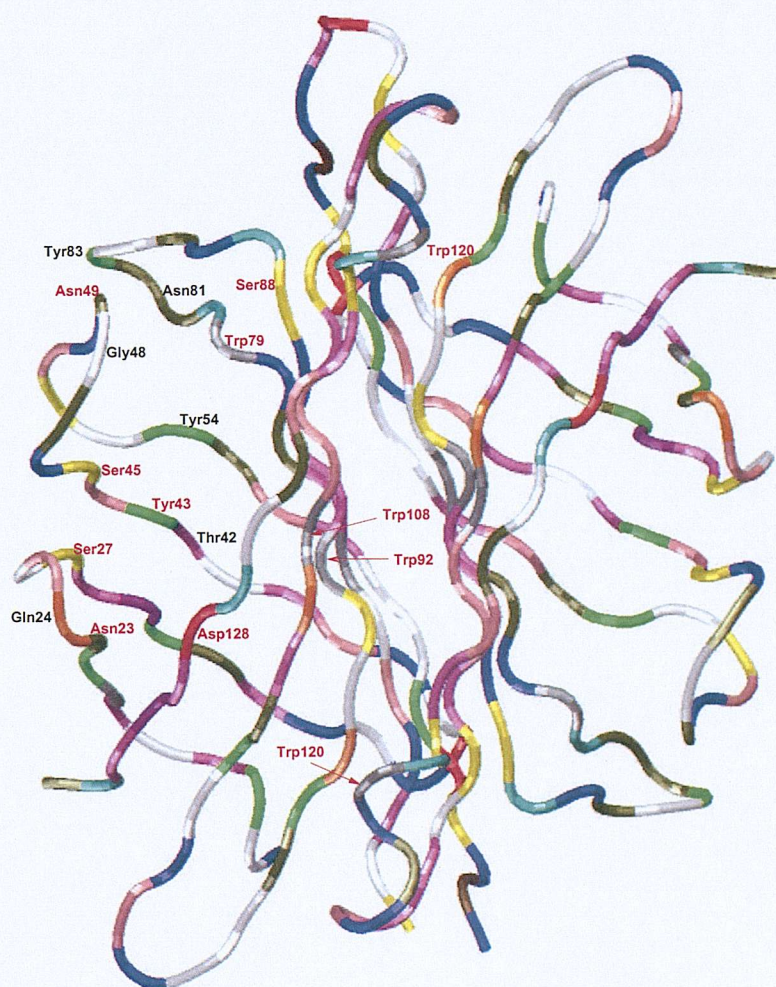
tional changes.

As in the case of endothiapepsin, side-chain flexibility analyses in pairs of holo-proteins bound to the same ligands were not performed. In the case of the biotin binding protein, the reason for this is not the lack of holo-protein structures binding to the same ligands, as several PDB structures solved at good resolution and bound to biotin and to 2-imino biotin exist. Rather, these protein structures were not analysed because:

1. they are all crystallised at different pH;
2. they are all from the same author.

Different crystallising conditions, such as a different pH, do not allow ligand-binding induced side-chain conformational changes to be distinguished from those determined by pH effects. Also, the fact that the author of all of the structures is unique does not allow exceptionally rigid side-chains that are a consequence of crystallographers' biases to be distinguished from those which are instead due to the presence of the same ligands. As all streptavidin structures of the present data set but two have been solved by Katz<sup>117</sup> or by Weber and coworkers,<sup>120</sup> the original data set of structures is biased by the restricted number of authors and groups who solved them; quantitative comparisons of side-chain conformational changes occurring in all PDB structures and in those solved by the same authors (see section 5.4.4) were thus not performed.

In Figure 7.7, some residue sequence numbers and types have been indicated on a tube representation of streptavidin (PDB entry 1SRE) to help identify some of the protein residues discussed in this chapter.



**Figure 7.7:** Tube representation of residue-name coloured streptavidin dimer structure. Some residue names and sequence numbers are indicated to help identifying regions discussed in this section; those referring to residues that are known to establish specific interactions with biotin have been indicated in red.

### 7.3.1 All Environments, and Environment Specific Conformational Changes

Figures 4.12-4.17 (chapter 4) show the percentages of side-chain conformational changes observed in the proteins of this thesis data set when no distinction between exposed and buried residues are made.

Some consistent trends are found in streptavidin with all methodologies of study. First,  $\chi_1$  and  $\chi_2$  side-chain torsions are in general more flexible in the binding site rather than in all protein residues; only in the case of  $\chi_2$  torsions, Najmanovich *et al.* and Dunbrack *et al.* methodologies of study, is the same amount of flexibility in the binding site and in all protein residues observed. These trends are probably better depicted in Figures 4.33, 4.34 and 4.35, where the differences between conformational changes observed in the binding site and in all the residues are plotted.

When residue environments are considered and residue are distinguished as exposed and buried (Figures 4.18-4.23 and 4.24-4.29 in chapter 4) Najmanovich *et al.* (Figures 4.24 and 4.25) and Dunbrack and Cohen methodologies of study (Figures 4.28 and 4.29) always reveal significant flexibility of exposed residues. When the more stringent specific cutoff developed by Zhao and coworkers are applied, buried  $\chi_1$  torsions are slightly more flexible than the exposed only in all protein residues (Figure 4.26 and 4.27).

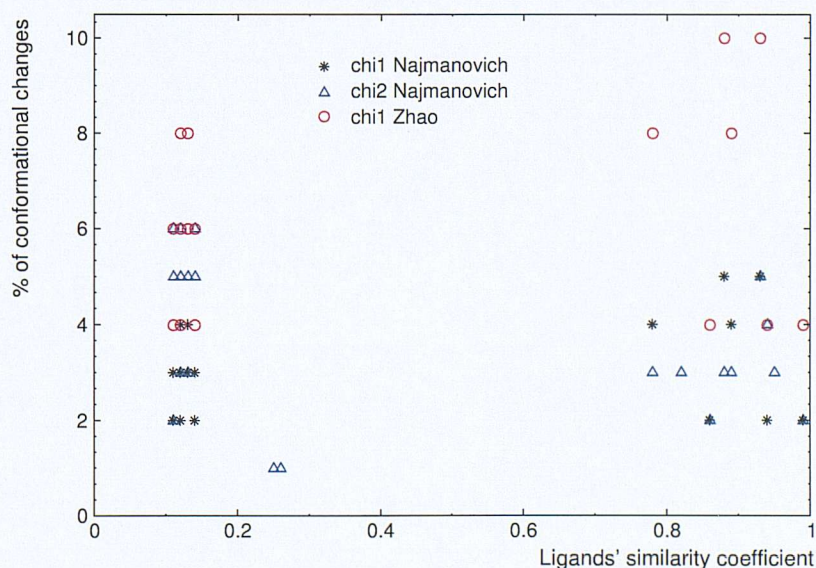
With all methodologies of analysis, streptavidin holo-/holo- comparisons reveal greater percentages of conformational change than apo-/holo- comparisons (Figures 4.30, 4.32 and 4.31). This trend is consistently stronger in the binding site residues of the protein.



### 7.3.2 Relationships Between Holo-Protein Side-Chain Conformational Changes and Ligand Similarities

To investigate whether similar ligands induce similar conformational changes in the same apo-protein structure, Tanimoto similarity coefficients were computed for all possible couples of non-peptidic streptavidin ligands with 1024-bit Daylight fingerprints and the default path range number of considered bonds (0-7).

In Figure 7.8, the percentages of conformational changes observed in the binding site of streptavidin holo-protein pairs have been plotted on the y axis. On the x axis, the Tanimoto similarity scores of the corresponding pairs of ligands have been plotted. This graph clearly shows that no correlations between the observed percentages of conformational changes and the 2D-similarity of the corresponding ligands are found; this is the case for the majority of protein systems analysed in this thesis.



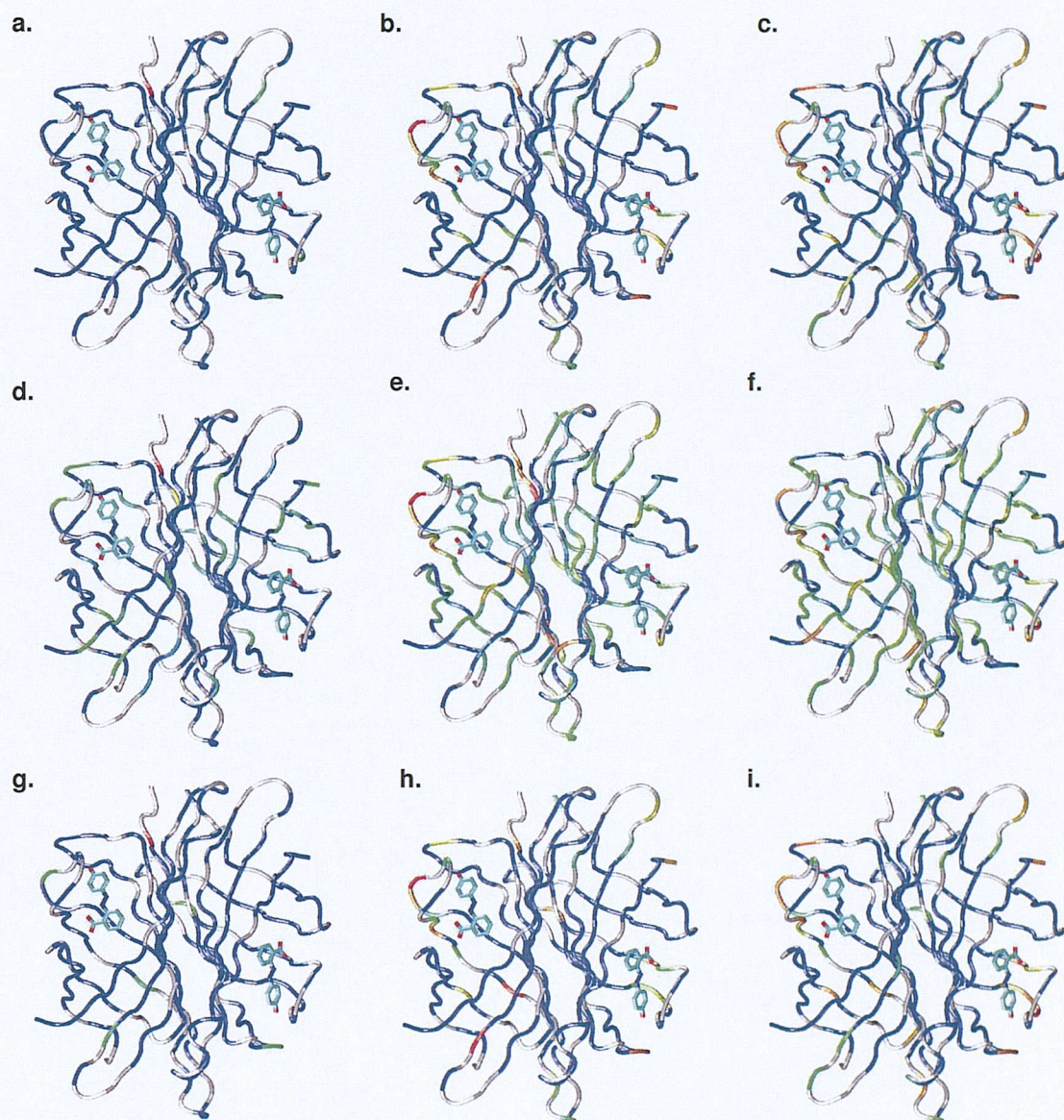
**Figure 7.8:** Percentages of conformational changes observed in the binding site of streptavidin holo-/holo- protein pairs (y axis) plotted against the Tanimoto similarity scores of the corresponding couple of ligands (x axis). Only Najmanovich *et al.* and Zhao *et al.* data are shown.



### 7.3.3 Percentages of Conformational Changes per Residue Sequence Number

Figure 7.9 shows the backbone structure of streptavidin dimer residues coloured in accordance to the flexibility of their side-chain torsions. The structures in the first row of the figure (indicated by letters *a*, *b* and *c*) represent the results obtained with Najmanovich *et al.* methodology of study. The structures in the second row (*d*, *e* and *f*) are instead obtained with Zhao *et al.* methodology of study, and those in the third row (*g*, *h* and *i*) with Dunbrack and Cohen rotamer libraries. While structures in the first column refer to apo-/apo- protein comparisons, those in the second column show apo-/holo- protein comparisons results, and those in the right column holo-/holo-comparisons results.

As previously observed in the case of endothiapepsin and HIV-1 protease, Najmanovich *et al.* and Dunbrack and Cohen methodologies produce very similar side-chain flexibility distribution patterns. Apo-/apo- protein comparisons clearly reveal less conformational changes than all other comparisons methods, especially when Zhao *et al.* angular thresholds are applied. Several parts of the protein which do not show significant flexibility in apo-/apo- protein comparisons (e.g. the flexible loop that folds and closes over the ligand in protein-ligand complexes) appear in fact rather flexible in apo-/holo- and holo-/holo- protein comparisons. Moreover, some side-chains which are found to be highly flexible in apo-/apo- protein comparisons clearly show less flexibility in apo-/holo- protein comparisons and are almost totally inflexible in holo-/holo- protein comparisons. This could be a random event or, more probably, a consequence of ligand-binding effects, which could 'freeze' the side chains of specific residues into one or a smaller range of conformations through



**Figure 7.9:** Tube trace of streptavidin backbone structure; all residues are coloured in accordance with the average percentage of conformational change its  $\chi_1$  torsions undergo, as detected with Najmanovich *et al.* (a, b, c), Zhao *et al.* (d, e, f) and Dunbrack *et al.* (g, h, i) methodologies of study. The residue average percentage of conformational change increases from blue coloured residues (which never change conformation) to red coloured residues (that change conformation 100% of times), passing through cyan, green, yellow and orange. Proline, glycine and alanine residues have been coloured in grey. Figures in the first column (a, d, g) refer to apo/apo- protein comparisons, figures in the central column (b, e, h) to apo/holo- protein comparisons and figures in the last column (g, h, i) to holo/holo- protein comparisons. Two bound molecules of biotin were left in all the protein structures to help identifying the binding site.

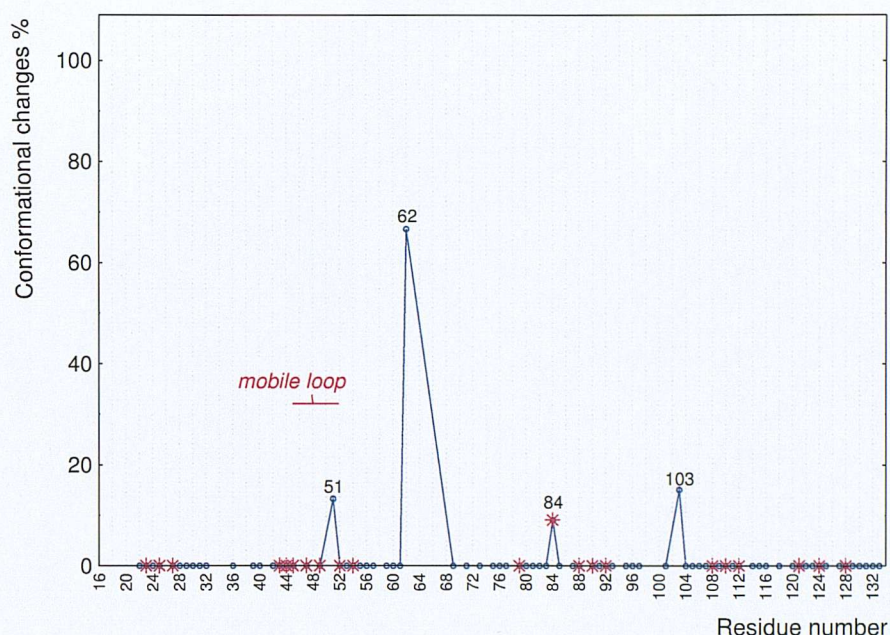
direct ligand/side-chains interactions and/or longer-range ligand-binding effects.

In the graphs represented in Figures 7.10-7.15, the average percentages of  $\chi_1$  side-chain conformational changes have been plotted on the y axis. Since each streptavidin monomer binds one ligand molecule, the percentages of conformational changes per residue sequence number have been averaged in these graphs over all four of streptavidin chains.

While the first three graphs (Figures 7.10, 7.11 and 7.12) were obtained with Najmanovich *et al.* methodology of study, the last three (Figures 7.13, 7.14 and 7.15) were obtained with Zhao *et al.* angular thresholds. Given the very high similarity between the graphs obtained with Najmanovich *et al.* angular thresholds (7.10-7.12) and those obtained with Dunbrack and Cohen rotamer libraries (Figures C.5-C.7), the latter are not shown in this chapter. To a lesser extent, their trends are also similar to those revealed by Zhao *et al.* methodology of study, which however generally detect greater side-chain conformational changes (especially in apo-/holo- and holo-/holo-proteins comparisons).

Since the present thesis data set includes six streptavidin's apo-proteins, i.e. enough structures to have statistically significant data regarding apo-/apo- protein binding, a new methodology of analysis to obtain side-chain flexibility values that are more likely to depend on genuine ligand-binding effects was allowed. The percentages of conformational changes observed in apo-/apo- protein comparisons were subtracted from those detected in apo-/holo- and holo-/holo- protein comparisons; in this way, spontaneous motions of intrinsically flexible residues are neglected, and apo-/holo- and holo-/holo- protein percentages of conformational changes refined. This sort of normalisation implies that the resulting flexibility values can be positive or negative;





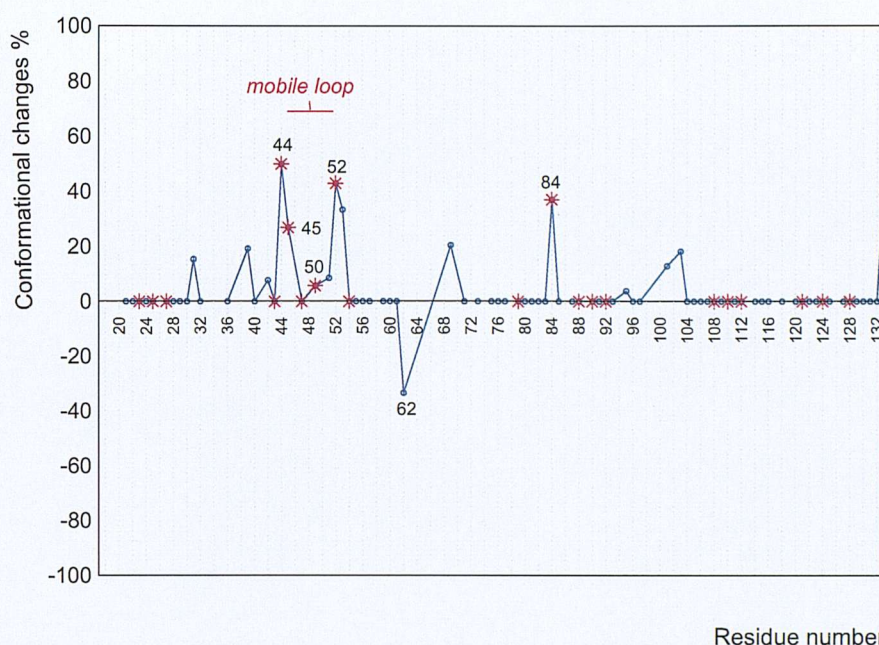
**Figure 7.10:** Percentages of times residues of streptavidin change  $\chi_1$  by more than  $60^\circ$  in apo-/apo- protein comparisons. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data.

while positive data correspond to residues which are more mobile in apo-/holo- and holo-/holo- comparisons, negative values might correspond to residues that have been 'frozen' in one or a few more preferred conformations as a consequence of long-range and/or direct contact ligand-binding effects.

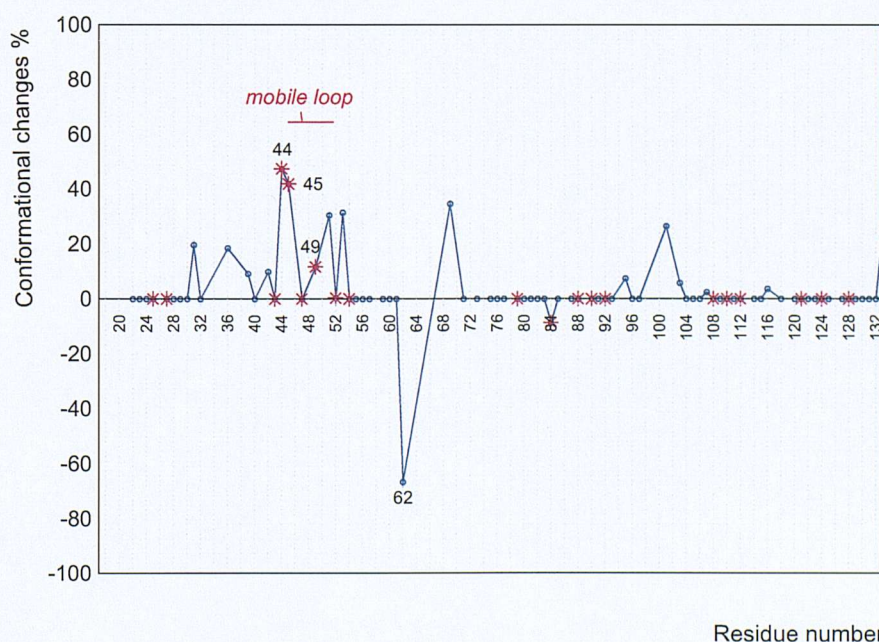
Few side-chains conformational changes of binding site residues are detected in apo-/apo- protein comparisons. In particular, no conformational changes at all are observed when  $60^\circ$  angular thresholds are employed (Figure 7.10) and only Trp 79, Asn 23 and Asp 128 (involved in the aromatic or hydrogen bonding network with biotin and other ligands) significantly move when environment- and residue type specific angular thresholds are considered (Figure 7.13).

Upon ligand binding, several binding site residues in the mobile loop appear to change conformation both with Najmanovich *et al.* (Ser 45 and 52, Asn 49) and Zhao *et al.* (Asn 49) methodologies. Also, binding site residues such as Arg 84 (Figure



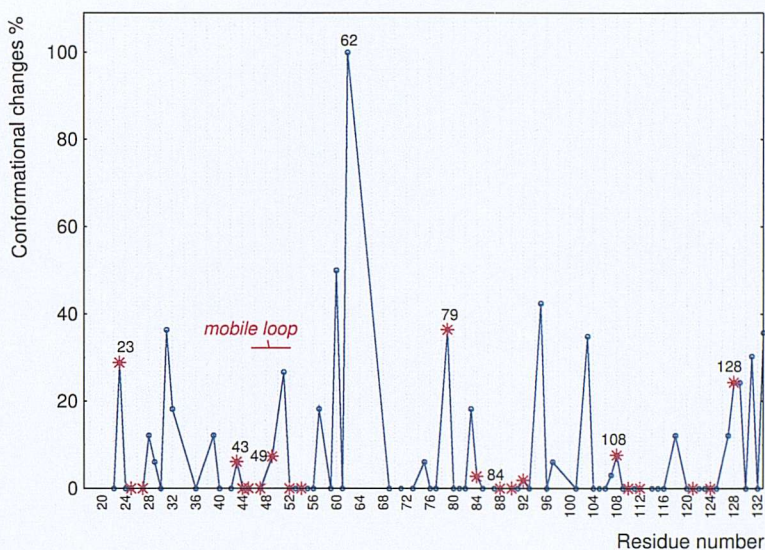


**Figure 7.11:** Percentages of times residues of streptavidin change  $\chi_1$  by more than  $60^\circ$  in apo-/apo- protein comparisons. The percentages of conformational changes detected in apo-/apo- protein comparisons have been subtracted; positive values correspond to residues whose flexibilities in apo-/apo- protein comparisons are greater than those observed in apo-/apo- protein pairs. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data.

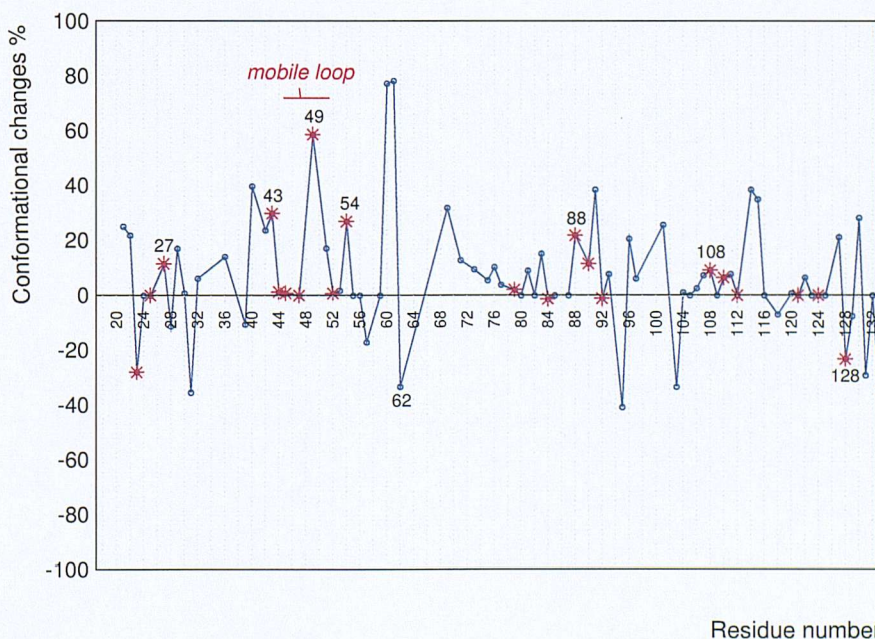


**Figure 7.12:** Percentages of times residues of streptavidin change  $\chi_1$  by more than  $60^\circ$  in holo-/holo- protein comparisons. The percentages of conformational changes detected in apo-/apo- protein comparisons have been subtracted; positive values correspond to residues whose flexibilities in holo-/holo- protein comparisons are greater than those observed in apo-/apo- protein pairs. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data.



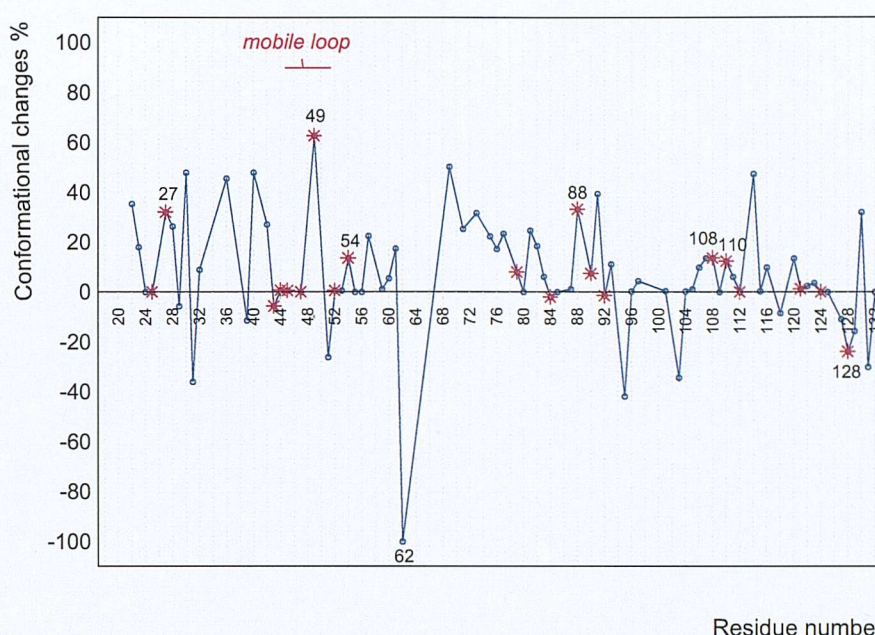


**Figure 7.13:** Percentages of times residues of streptavidin change  $\chi_1$  by more than Zhao *et al.* specific angular thresholds in apo/apo- protein comparisons. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data.



**Figure 7.14:** Percentages of times residues of streptavidin change  $\chi_1$  by more than Zhao *et al.* specific angular thresholds in apo/holo- protein comparisons. The percentages of conformational changes detected in apo/apo- protein comparisons have been subtracted; positive values correspond to residues whose flexibilities in apo/holo- protein comparisons are greater than those observed in apo/apo- protein pairs. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data.





**Figure 7.15:** Percentages of times residues of treptavidin change  $\chi_1$  by more than Zhao *et al.* specific angular thresholds in holo-/holo- protein comparisons. The percentages of conformational changes detected in apo-/apo- protein comparisons have been subtracted; positive values correspond to residues whose flexibilities in holo-/holo- protein comparisons are greater than those observed in apo-/apo- protein pairs. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data.

7.11), Ser 27, Tyr 43, Tyr 54, Ser 88, Thr 90, Trp 108 and Leu 110 (Figures 7.14 and 7.15) show significant side-chain flexibility after apo-/apo- protein conformational changes have been subtracted. Many of these residues are involved in H-bonds or hydrophobic binding networks with the ligands; however, several other residues which are critical for ligand binding do not show significant conformational changes with all methodologies and, generally speaking, binding site conformational changes appear to be significantly smaller than those observed in HIV-1 protease and endothiapepsin (see Figures 5.6-5.9 and 6.5-6.8). This could be because most of the residues that bind biotin through hydrogen bonds are already organised in the correct binding geometry in the apo-protein structure.<sup>122</sup>

In addition to residues that are in direct contact with biotin, Zhao *et al.* thresholds detect other flexible residues, which might indirectly contribute to the energetics

of binding<sup>121</sup> (Figure 7.14 and 7.15). Many of these neighbour the active site; others might be involved in longer range networks of hydrophobic binding, helping to optimally orient the side-chain of residues directly involved in ligand binding and/or be essential for the formation and the stabilisation of  $\beta$ -barrels.<sup>121</sup>

The highest peak which can be observed in Figures 7.10 and 7.13 corresponds to residue Ser 62. This amino acid, which is not part of streptavidin binding site, shows the highest flexibility observed in apo-/apo- proteins with all methodologies and few or no conformational changes when apo-/holo- and holo-/holo- proteins comparisons are considered. Rather than a random event, this peculiar behaviour can be realistically explained on the basis of a restriction of Ser 62 side-chain conformational freedom caused by ligand-binding effects.

Even if Ser 62 is not in direct contact with binding site residues, its side chain interacts with several residues which are part of the binding sites. These include Lys 80 (which is in contact with binding site residues Tyr 54, Trp 79 and Arg 84), Asn 85 (which interacts with Asn 49 and Trp 79), His 87 (which is also in contact with Ala 86 and Ser 88) and Asp 61 (that establishes contacts with binding site residues interacting with Ser 88). Further, it has been proved that biotin binding to streptavidin has a cooperative effect, and increases the tight association between the subunits of the tetramer by establishing and mediating many interactions on one side of the strong subunit interface.<sup>117</sup>

The residues which are mainly responsible for streptavidin inter-subunits interactions have been identified by crystallographic analyses and structure based engineering. They include Trp 120, which is often considered part of biotin binding site and is fundamental for inter-subunits interactions at the weak inter-subunits interface,



inter-loops hydrogen bonds between Arg 84, Glu 51 and Asn 49 at the strong inter-subunits interface, and the hydrogen-bonded salt-bridge between Asp 61 and His 87, which is fundamental for inter-subunits binding at the strong dimer interface.<sup>117</sup>

Biotin's closest contact with the strong intersubunit interface are van der Waals interactions with Ala 86 and a hydrogen bond with Ser 88. Bound biotin also induces the flexible flap into the closed conformation through van der Waals interactions and hydrogen bonds; since Asn 49 of the ordered flap in turn makes van der Waals interactions with Ala 86, ordering of the flexible loop by biotin and similar small molecules is also associated with ordering of the Arg 84 side-chain and of the Asp61-Ser69 loop.<sup>117</sup> Last but not least, ligand binding influences the hydrogen-bonded salt bridge between Asp 61 and His 87; since the strength of this bond determines different residue/residue interactions in Ser 62, the greater rigidity detected for this residue in protein/ligand complexes very likely depends on tighter inter-subunit and inter-residue contacts induced by ligand binding.

Other residues (such as binding site residue Asp 128) show negative peaks in Figures 7.14 and 7.15; the restriction of their side-chain conformational freedom imposed by ligand binding might be the reason of this observation. Residues which instead are detected as very flexible in apo-/holo- comparisons and as rigid in holo-/holo-comparisons probably adopt the same conformation in all protein/ligand complexes.

## 7.4 Conclusions

Streptavidin is a tetrameric protein, often referred as a *dimer of dimers*, which has been intensively studied for its exceptionally strong association with biotin. This peculiarity, together with the availability of several PDB apo-protein structures at good resolution, and the different conformational characteristics of this protein in

respect to HIV-1 protease and endothiapepsin are the main reasons for its choice as the object of the more in depth conformational analysis described in the present chapter. Further, streptavidin binding site residues, even if to a smaller extent than those observed in the aspartic proteases, appear more flexible than all protein residues; this suggests predominant genuine ligand-binding effects, rather than random events, at the basis of the observed conformational changes (see chapter 4).

Both strongly favourable enthalpic and entropic factors contribute to the highly negative biotin/streptavidin free energy of binding. HABA interaction energetics are instead dominated by favourable entropy contributions, which compensate for the loss of hydrogen bonds that this ligand can establish in solution; in the case of peptidic ligands, the enthalpy of binding is instead the driving force of the binding process, and is opposed by a large unfavourable entropy of binding (necessary to freeze the many degrees of freedom of peptides).

With all methodologies of study, holo-/holo- protein comparisons detect greater side-chain conformational changes than apo-/holo- protein comparisons. Zhao *et al.* environment-specific angular thresholds detect slightly greater flexibility of buried rather than exposed residues when all protein residues are considered, while undifferentiated angular thresholds reveal significantly greater flexibility of exposed rather than buried residues especially in the binding site (see chapter 4).

Apo-/apo- protein comparisons show significantly less conformational changes than all comparisons involving bound streptavidin structures. To obtain percentages of conformational changes which possibly depend on genuine ligand-binding effects rather than spontaneous movements of intrinsically flexible residues, the percentages of conformational changes obtained in apo-/apo- protein comparisons were subtracted

from those observed in apo-/holo- and holo-/holo- comparisons. As a result, several binding site residues still show significant side-chain flexibilities upon ligand binding. Once again residues that are essential in the process of ligand binding can be both flexible (e.g. Asn 49, ) or rigid (e.g. Asp 128) in apo-/holo- and holo-/holo- protein comparisons. The main side-chain responsible for the exceptionally favourable free energy of binding of biotin and derivatives have been identified in hydrogen bonding interactions (Asn 23, Ser 27, Tyr 43, Ser 45, Asn 49, Ser 88, Thr 98 and Asp 128), hydrophobic and van der Waals interactions (Trp 79, 92, 108 and 120) and the interactions established by a mobile loop which folds upon ligand binding (residues 45-52).<sup>113,119,120</sup> The majority of these residues, which are also involved in binding HABA and its derivatives, are in the correct binding geometry also in the apo-protein structures;<sup>122</sup> this explain the significantly smaller amount of flexibility of streptavidin when compared to HIV-1 protease and endothiapepsin. The greater flexibility observed in holo-/holo- rather than apo-/holo- protein comparisons probably also reflects the minor conformational changes which streptavidin must undergo to change from the unbound to the ligand-bound conformation.

While in apo-/holo- and holo-/holo- comparisons Najmanovich *et al.* and Dunbrack and Cohen methodologies hardly detect any conformational changes other than in the flexible loop residues and in Arg 84, Zhao *et al.* angular thresholds detect a significantly higher number of flexible residues, both in the binding site and in the rest of the protein (Figures 7.14 and 7.15).

At the interface between the strongly interacting pair of monomers, the interaction between residues 61-82 and the hydrogen bonds and van der Waals interaction network that is established between the mobile loop and Arg 84 are believed to change

nature and get stronger upon ligand binding.<sup>117</sup> The great peak that is found for Ser 62 in the apo-/apo- protein comparisons graphs (Figures 7.10 and 7.13) and changes sign in apo-/holo- and holo-/holo- protein comparisons graphs (Figures 7.11, 7.12, 7.14 and 7.15) is probably a consequence of tighter inter-subunit contacts and cooperative effects induced by biotin binding. Ligand induced effects might also be at the basis of the significant flexibility reduction observed for Asp 128 in protein/ligand complexes (Figures 7.14 and 7.15). The conformational behaviour of residues which, when Zhao *et al.* angular thresholds are employed, behave as Tyr 43 (which appears highly mobile when apo- and holo- proteins are compared but dramatically reduces its flexibility when holo-proteins are considered) could be typical of residues which rearrange themselves upon ligand binding but always adopt the same conformation in different holo-proteins.

No correlations at all between the percentages of conformational changes induced in pairs of holo-protein structures by non-peptidic ligands and the corresponding ligands' Tanimoto similarity scores were found.



## Chapter 8

# Conclusions

---

### 8.1 Summary

Proteins are intrinsically flexible molecules in constant equilibrium between a myriad of different conformations. The need to account for the dynamic properties of proteins is fundamental in rational drug design; computational methods able to evaluate and predict the conformations which are likely to be assumed by a protein in the presence of a ligand are needed. However, a full understanding of protein flexibility and, in particular, ligand-binding induced fit still have to be achieved. In fact, it is often very difficult to distinguish between genuine ligand-binding effects and random motions which proteins spontaneously undergo. Further, crystal structures deposited in the PDB could show conformations induced by crystallisation conditions and/or contacts, crystallographic artefacts and/or biased refinement methods.

The aim of this thesis is to analyse the patterns and trends of side-chain confor-

mational changes, trying to distinguish genuine ligand-induced effects from random protein motions.

A data set of 10 different protein systems for which multiple PDB apo- and holo-protein structures at high resolution exist was chosen (carbonic anhydrase II, cytochrome P-450 CAM, endothiapepsin, glutathione S-transferase, HIV-1 protease, ribonuclease A, streptavidin, trypsin, thrombin, D-xylose isomerase).

All possible pairs of structures within a protein system were compared to identify side-chain conformational changes occurring in apo-/apo-, apo-/holo- and holo-/holo-protein pairs. Side-chain conformational changes were defined on the basis of both constant<sup>46</sup> and environment- and residue- dependent<sup>66</sup> thresholds. Also, recently published rotamer libraries<sup>55</sup> were employed.

Some conserved trends were observed with all methodologies of study, especially in the case of the most flexible proteins in the data set (HIV-1 protease, endothiapepsin, ribonuclease and streptavidin) and cytochrome P-450 CAM. However, several discrepancies in the results obtained with the different methodologies of study are found; significant “noise” in the obtained flexibility trends is expected.

In HIV-1 protease, endothiapepsin, ribonuclease, streptavidin and cytochrome P-450 CAM, binding site residues are always more flexible than all protein residues. HIV-1 protease and endothiapepsin are the only two proteins for which apo-/holo-comparisons always show greater conformational changes than holo-/holo- protein comparisons. HIV-1 protease, cytochrome P-450 and carbonic anhydrase II are the only systems in which buried binding site residues appear more flexible than exposed binding site residues with all methodologies of study. In the protein systems where exposed residues are consistently more flexible than buried residues, this trend is

always more pronounced in the binding sites, suggesting that this observation is also genuinely related to the mechanism of ligand-induced fit.

Among the different approaches, the Zhao *et al.* methodology appears to be the most reliable. On the one hand, the other methodologies employed in this thesis do not make any distinction on residue environments and, in the case of Najmanovich *et al.*, residue types. As a consequence, they are likely to detect residues whose side-chain torsions are very flexible on an absolute scale, biasing the observations of ligand binding induced-fit towards residues that are intrinsically very flexible, and do not necessarily change conformation as a result of ligand binding induced fit. In fact, Najmanovich and co-workers reported that the pairwise comparisons of apo-protein structures with identical sequence, using a 60° cutoff, yielded the same flexibility trends obtained with apo-/holo- comparisons.<sup>46</sup>

On the other hand, the results obtained with Zhao *et al.* specific angular thresholds are often very consistent. For example, the conformational changes they detect in the 10 protein systems, in both apo-/holo- and holo-/holo- protein comparisons, are 18 times out of 20 greater in the binding site of the proteins, suggesting that they are indeed able to pick “unusual”, i.e. ligand-induced, rather than intrinsic residues’ flexibility. Moreover, Zhao *et al.* methodology gives the smallest percentages of apo-/apo- protein conformational changes, reinforcing the notion that specific angular thresholds are the best way to spot unexpected side-chain motions.

No correlations were found between the nature of a given residue’s contribution to the free energy of binding and the amount of their side-chain flexibility.

HIV-1 protease, endothiapepsin and streptavidin were chosen for a more detailed inspection. Their choice mainly depended on their high flexibility, their different flexibility trends, and on the observation that their binding sites are more flexible

than the whole protein with all methodologies. This suggests that the conformational changes observed in these proteins are more likely to depend on ligand-binding effects rather than random motions.

Some differences found in the flexibility results and trends of the two proteases might be explained on the basis of their different mechanisms of ligand binding; while the ligand-binding reaction is endothermic for HIV-1 protease, and depends on the burial of a large portion of the protein apolar surface for a favourable free energy of binding, the endothiapsin ligand binding reaction is favoured both enthalpically and entropically. Hence, buried residues' conformational changes in the latter protease might not be as essential as they are in the case of the viral protein.

The trends of side-chain flexibility that are found in these proteases are broadly in agreement with previous theoretical and experimental results (NMR data, X-ray B-factors, MD, NMA, backbone RMSd in crystallographic structures, ITC analysis and so on). Especially when Najmanovich *et al.* and Dunbrack and Cohen methodologies are applied, side-chain conformational changes seem to be larger in regions that also show significant backbone RMSd, and rigid side-chains appear to be located in protein zones that seldom undergo backbone movements.<sup>79</sup> However, significant deviations from these observations are found when Zhao *et al.* angular thresholds are employed to define conformational changes.

Surprisingly, not negligible side-chain motions are found in catalytic residues of both proteases, generally believed to be highly inflexible.<sup>101,105</sup> These findings provide an independent corroboration of the opinion of Kumar and Koshur,<sup>84</sup> according to whom the catalytic residues of HIV-1 protease, despite being very rigid and stable, are in fact highly “adaptable”, i.e. can move and adjust their position in response to internal or external stress (such as mutations and/or the presence of specific ligands),



even if their electron density is very well defined and their B-factors appear to be the lowest in the protease structure.<sup>84</sup> In fact, while low temperature factors suggests that active site residues have generally less “systematic flexibility” (i.e. less vibrational motion) than non-active site residues, this does not exclude the possibility that they are rather “adaptable”, i.e. able to reach discrete energy wells by overcoming potential energy barriers (“segmental flexibility”).<sup>100</sup> An enzyme’s capability to change from one equilibrium conformation to another one in response to a *stimulus* might be essential for its catalytic function, and not depend on the fast collective motions of  $\alpha$ -carbons (vibrational motions).

Streptavidin, a tetrameric protein well known for its exceptionally large free energy of binding with biotin, was also analysed in more detail. Najmanovich *et al.* and Dunbrack and Cohen methodologies of study detect more flexibility in exposed rather than in buried residues of this protein, while Zhao *et al.* angular thresholds do not spot significant differences between buried and exposed residues side-chain flexibilities. Holo-/holo- protein comparisons reveal greater conformational changes than apo-/holo- protein comparisons, while apo-/apo- protein comparisons show less flexibility than all comparisons involving bound structures.

To identify side-chain conformational changes which are more likely to depend on genuine ligand-binding effects rather than spontaneous motions of intrinsically flexible residues, the percentages of conformational changes detected in apo-structure comparisons of streptavidin were subtracted from those observed in apo-/holo- and holo-/holo- protein comparisons. The so-obtained positive peaks are often found to be involved in hydrogen bonds or hydrophobic interactions with the ligands. Large negative peaks are instead likely to correspond to side-chains which are highly flex-

ible in the unbound structures, but whose motions are instead 'frozen' in one or few conformations through ligand-binding effects. This is the case of Ser 62, whose side-chain motions are probably greatly restricted by the tighter inter-residues contacts and cooperative effects induced by biotin binding. Residues for which very high positive peaks are observed in apo-/holo- protein comparisons but very small flexibility is observed in holo-/holo- protein pairs might correspond to residues that move upon ligand binding, but which always assume the same side-chain conformation in different holo-structures.

## 8.2 Future Work

This thesis has identified a number of areas where further work would seem indicated. First, a deeper analysis of specific ligand/protein complexes should be performed, to avoid missing information that is easily lost when an ensemble of structures is studied. The comparison of structures on a one to one basis might provide useful insights on the induced-fit produced by similar or very different ligands; specific interactions and/or contacts they establish with protein residues might be investigated in the most interesting cases (e.g. protein/ligand complexes where catalytic residues are found to undergo conformational changes).

The relationships between side-chain conformational changes and the nature of the interactions that they establish with the ligands and neighbouring residues could also be quantitatively defined. The presence of specific interactions types (e.g. hydrogen bonds and/or hydrophobic interactions) might be directly or indirectly correlated with the amount of flexibility found in the associated side-chains.

The kind of analysis described in this thesis, and more in depth investigations of specific protein cases and ligands, is potentially very useful to predict which residues

are likely to move, and what conformations are likely to be assumed, in protein-ligand docking. Zhao *et al.* angular thresholds have proved to give the most consistent results when random motions and ligand-induced side-chain conformational changes have to be distinguished; however, one should not forget that the angular range defined by these thresholds is often very small. If broad conformational changes are desired, Najmaionovich *et al.* and Dunbrack *et al.* angular thresholds should also be taken into account.

P. A. Cornish, *Comp. Appl. Biotech. Med. Sci.*, 7 (1993).

M. Gervasio, A. M. L. de Almeida, C. Chaves, *Molecular Eng.*, 13, 673 (1993).

W. Ma, S. Kumar, Q. Li, L. Song and Q. Zhou, *Protein Eng.*, 12, 71 (1998).

C. E. Rasmussen, *Science*, 143, 173 (1962).

R. F. Drenth, J. C. Lumb, D. Moras, *Science*, 241, 1702 (1988).

M. Gervasio, R. C. Moras, D. Moras, *Proc. USA. Nat. Acad. Sci.*, 85, 1000 (1988).

M. Gervasio, A. M. L. de Almeida, R. C. Moras, *Protein Eng.*, 11, 673 (1997).

W. Gervasio, A. M. L. de Almeida, R. C. Moras, *Protein Eng.*, 11, 673 (1997).

W. Gervasio, R. C. Moras, *Protein Eng.*, 11, 673 (1997).

## References

---

- [1] H. A. Carlson and A. McCammon, *Mol. Pharmacol.*, **57**, 213, (2000).
- [2] P. A. Rejto and S. T. Freer, *Prog. Biophys. Molec. Biol.*, **66**, 167, (1996).
- [3] K. L. M. and H. A. Carlson, *J. Am. Chem. Soc.*, **126**, 13276.
- [4] M. Gerstein and W. Krebs, *Nucleic Acids Res.*, **26**, 4280, (1998).
- [5] B. Alberts, D. Bray, A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter, *Essential Cell Biology*, Garland Publishing Inc., New York, (1998).
- [6] M. Levitt, M. Gerstein, E. Huang, S. Subbiah and J. Tsai, *Annu. Rev. Biochem.*, **66**, 549, (1997).
- [7] H. A. Carlson, *Curr. Opin. Chem. Biol.*, **6**, 447, (2002).
- [8] M. Gerstein, A. M. Lesk and C. Clothia, *Biochemistry*, **33**, 6739, (1994).
- [9] B. Ma, S. Kumar, C. J. Tsai and R. Nussinov, *Protein Eng.*, **12**, 713, (1999).
- [10] D. E. Koshland, *Science*, **142**, 1533, (1963).
- [11] S. Franzen, J. C. Lambry, B. Bohn, C. Poyart and J. L. Martin, *Nat. Struct. Biol.*, **1**, 230, (1994).
- [12] M. Gerstein and C. Clothia, *Proc. Natl. Acad. Sci. USA*, **93**, 10160, (1996).
- [13] M. Gerstein, A. M. Lesk and C. Clothia, *Biochemistry*, **33**, 6739, (1994).
- [14] M. Gerstein, A. M. Lesk, E. N. Baker, B. Anderson, G. Norris and C. Chothia, *J. Mol. Biol.*, **234**, 357, (1993).
- [15] M. Gerstein, G. Schulz and C. Chothia, *J. Mol. Biol.*, **229**, 494, (1993).
- [16] S. Hayward, *Proteins*, **36**, 425, (1999).
- [17] M. Gerstein and C. Chothia, *J. Mol. Biol.*, **220**.



- 
- [18] J. A. Yankeelov and D. E. Koshland, *J. Biol. Chem.*, **240**, 1593, (1965).
- [19] D. W. Miller and K. A. Dill, *Protein Sci.*, **6**, 2166, (1997).
- [20] S. J. Teague, *Nat. Rev. Drug Discov.*, **2**, 527, (2003).
- [21] B. Ma, M. Shatsky, H. J. Wolfson and R. Nussinov, *Protein Sci.*, **11**, 184, (2002).
- [22] M. H. V. Van Regenmortel, *J. Mol. Recognit.*, **12**, 1, (1999).
- [23] S. Kumar, B. Ma, C. Tsai, N. Sinha and R. Nussinov, *Protein Sci.*, **9**, 10, (2000).
- [24] C. F. Wong and J. A. McCammon, *Annu. Rev. Pharmacol.*, **43**, 31, (2003).
- [25] R. D. Taylor, P. J. Jewsbury and J. W. Essex, *J. Comput. Aid. Mol. Des.*, **16**, 151, (2002).
- [26] M. L. Teodoro and L. E. Kavraki, *Curr. Pharm. Des.*, **9**, 1635, (2003).
- [27] F. Jiang and S. H. Kim, *J. Mol. Biol.*, **219**, 79, (1991).
- [28] R. D. Taylor, P. J. Jewsbury and J. W. Essex, *J. Comput. Chem.*, **24**, 1637, (2003).
- [29] A. Leach, *J. Mol. Biol.*, **235**, 345, (1994).
- [30] L. Schaffer and G. M. Verkhivker, *Proteins*, **33**, 295, (1998).
- [31] N. Nakajima, J. Higo, A. Kidera and H. Nakamura, *Chem. Phys. Lett.*, **278**, 297, (1997).
- [32] J. Apostolakis, A. Pluckthun and A. Caflish, *J. Comput. Chem.*, **19**, 21, (1998).
- [33] C. Chipot and D. A. Pearlman, *Molecular Simulation*, **28**, 1, (2002).
- [34] M. Philippopoulos and C. Lim, *Proteins*, **36**, 87, (1999).

- 
- [35] B. A. Luty, Z. R. Wassermann, P. F. W. Sotuten, C. N. Hodge, M. Zacharias and J. A. McCammon, *J. Comput. Chem.*, **16**, 454, (1995).
- [36] M. Mangoni, D. Roccatano and A. D. Nola, *Proteins*, **35**, 153, (1999).
- [37] K. M. Masukawa, A. H. Carlson and J. A. McCammon, *J. Phys. Chem. A.*, **103**, 10213, (1999).
- [38] H. A. Carlson, K. M. Masukawa, K. Rubins, F. D. Bushman, W. L. Jorgensen, R. D. Lins, J. M. Briggs and J. A. McCammon, *J. Med. Chem.*, **43**, 2100, (2000).
- [39] R. M. A. Knegtel, I. D. Kuntz and C. M. Oshiro, *J. Mol. Biol.*, **266**, 424, (1997).
- [40] H. ClauBen, C. Buning, M. Rarey and T. Lengauer, *J. Mol. Biol.*, **308**, 377, (2001).
- [41] M. A. Kastenholtz, M. Pastor, G. Cruciani, E. E. Haaksma and T. Fox, *J. Med. Chem.*, **43**, 3033, (2000).
- [42] F. Osterberg, G. M. Morris, M. F. Sanner, A. J. Olson and D. S. Goodsell, *Proteins*, **46**, 34, (2002).
- [43] P. R. Caron, M. D. Mullican, R. D. Mashal, K. P. Wilson, M. S. Su and M. A. Murcko, *Curr. Opin. struct. Biol.*, **5**, 464, (2001).
- [44] H. J. Bohm and M. Stahl, *Curr. Opin. Chem. Biol.*, **4**, 283, (2000).
- [45] M. L. Lamb, K. W. Burdick, S. Toba, M. M. Young, A. G. Skillman, X. Zou, J. R. Arnold and I. D. Kuntz, *Proteins*, **42**, 296, (2001).
- [46] R. Najmanovich, J. Kuttner, V. Sobolev and M. Edelman, *Proteins: Struct. Funct. Genet.*, **39**, 261, (2000).
- [47] J. Janin., S. Wodak, M. Levitt and B. Maigret, *J. Mol. Biol.*, **125**, 357, (1978).
- [48] B. R. Gelin and M. Karplus, *Proc. Nat. Acad. Sci. U.S.A.*, **72**, 2002, (1975).

- 
- [49] J. W. Ponder and F. M. Richards, *J. Mol. Biol.*, **193**, 775, (1987).
- [50] R. L. Dunbrack and M. Karplus, *J. Mol. Biol.*, **230**, 543, (1993).
- [51] H. Schrauber, F. Eisenhaber and P. Argos, *J. Mol. Biol.*, **230**, 592, (1993).
- [52] P. Tuffery, C. Etchebest and S. Hazout, *Protein Eng.*, **10**, 361, (1997).
- [53] M. De Maejser, J. Desmet and I. Lasters, *Fold Des.*, **2**, 53, (1997).
- [54] S. C. Lovell, M. Word, J. S. Richardson and D. C. Richardson, *Proteins: Struct. Funct. Genet.*, **40**, 389, (2000).
- [55] R. L. Dunbrack and F. E. Cohen, *Protein Sci.*, **6**, 1661, (1997).
- [56] J. Mendes, A. M. Baptista, M. A. Carrondo and C. M. Soares, *Proteins*, **37**, 530, (1999).
- [57] O. Carugo and P. Argos, *Prot. Eng.*, **10**, 777, (1997).
- [58] J. Heringa and P. Argos, *Proteins*, **37**, 30, (1999).
- [59] J. Heringa and P. Argos, *Proteins*, **37**, 44, (1999).
- [60] O. Herzberg and J. Moult, *Prot. Struct. Funct. Genet.*, **11**, 223, (1991).
- [61] R. J. Petrella and M. Karplus, *J. Mol. Biol.*, **312**, 1161, (2001).
- [62] M. W. MacArthur and J. M. Thornton, *Acta Crystallog. Sect. D*, **55**, 994, (1999).
- [63] X. Fradèra, X. De La Cruz, C. H. Silva, J. L. Gelpi, F. J. Luque and M. Orozco, *Bioinformatics*, **18**, 939, (2002).
- [64] J. L. Gelpi, S. Kalko, X. de la Cruz, X. Barril, J. Cirera, F. J. Luque and M. Orozco, *Proteins*, **45**, 428, (2001).
- [65] R. A. Laskowski, *J. Mol. Graph.*, **13**, 323, (1995).

- 
- [66] S. Zhao, D. S. Goodsell and A. J. Olson, *Proteins: Struct. Funct. Genet.*, **43**, 271, (2001).
- [67] M. J. Bower, F. E. Cohen and R. L. Dunbrack, *J. Mol. Biol.*, **267**, 1268, (1997).
- [68] P. Kallblad and P. M. Dean, *J. Mol. Biol.*, **326**, 1651, (2003).
- [69] J. Janin, *Biochemie*, **72**, 705, (1990).
- [70] R. M. Stroud and S. B. Fauman, *Protein Sci.*, **4**, 2392, (1995).
- [71] W. R. Pearson and D. J. Lipman, *PNAS*, **85**, 2444, (1988).
- [72] R. L. J. Dunbrack, BBDEP.C, University of California, San Francisco, (2000).
- [73] V. Sobolev, S. Sorokine, J. Prilusky, E. Abola and M. Eldeman, *Bioinformatics*, **15**, 327, (1999).
- [74] S. J. Hubbard and J. M. Thornton, NACCESS, Department of Biochemistry and Molecular Biology, University College, London, (1993).
- [75] ProFitV2.2, Martin, A. C. R., SciTech Software, (2002).
- [76] D. R. Flower, *J. Chem. Inf. Comput. Sci.*, **38**, 379, (1998).
- [77] C. A. James and D. Weininger, Daylight theory manual, Daylight Chemical Information Systems, Inc. (URL: [www.daylight.com](http://www.daylight.com)), Irvine, CA, (1995).
- [78] N. S. Andreeva and L. D. Rumsh, *Protein Sci.*, **10**, 2439, (2001).
- [79] V. Zoete, O. Michielin and M. Karplus, *J. Mol. Biol.*, **315**, 21, (2002).
- [80] S. Piana, P. Carloni and M. Parrinello, *J. Mol. Biol.*, **319**, 567, (2002).
- [81] W. E. Hartre, S. Swaminathan, M. M. Mansuri, M. J. C., I. E. Rosemberg and D. L. Beveridge, *Proc. Natl. Acad. Sci.*, **87**, 8864, (1990).
- [82] D. I. Freedberg, R. Ishima, J. Jacobs, W. Y. X., I. Kustanovich, L. J. M. and D. A. Torchia, *Protein Sci.*, **11**, 221, (2002).



- 
- [83] B. Pillai, K. K. Kannan and M. V. Hosur, *Proteins: Struct. Funct. Genet.*, **43**, 57, (2001).
- [84] M. Kumar and M. V. Hosur, *Eur. J. Biochem.*, **270**, 1231, (2003).
- [85] J. R. Collins, S. K. Burt and J. W. Erickson, *Struct. Biol.*, **2**, 334, (1995).
- [86] S. W. Rick, J. W. Erickson and S. K. Burt, *Proteins*, **32**, 7, (1998).
- [87] M. J. Todd, N. Semo and E. Freire, *J. Mol. Biol.*, **238**, 475, (1998).
- [88] J. S. Bardi, I. Luque and E. Freire, *Biochemistry*, **36**, 6588, (1997).
- [89] V. J. Hilser and E. Freire, *J. Mol. Biol.*, **262**, 756, (1996).
- [90] E. Freire, *Proc. Natl. Acad. Sci. USA*, **97**, 11680, (2000).
- [91] V. J. Hilser, D. Dowdy, T. G. Oas and E. Freire, *Proc. Natl. Acad. Sci. USA*, **95**, 9903, (1998).
- [92] I. Luque and E. Freire, *Proteins: Struct. Funct. Genet.*, **4**, 63, (2000).
- [93] W. E. Hartre, S. Swaminathan and D. L. Beveridge, *Proteins: Struct. Funct. Genet.*, **13**, 175, (1992).
- [94] M. Fligner, J. Verducci and P. Blower, in the Second Joint Sheffield Conference on Chemoinformatics, (2001).
- [95] A. R. Leach and A. P. Lemon, *Proteins: Struct. Funct. Genet.*, **33**, 227, (1998).
- [96] S. Munshi, Z. Chen, Y. Li, D. B. Olsen, M. E. Fraley, R. W. Hungate and L. C. Kuo, *Acta Crystallogr. D*, **54**, 1053, (1998).
- [97] M. A. DePristo, P. I. W. de Bakker and T. M. Blundell, *Structure*, **12**, 831, (2004).
- [98] D. H. Ohlendorf, *Acta Crystallogr. D Biol. Crystallogr.*, **50**, 808, (1994).
- [99] M. A. Wilson and A. T. Brunger, *J. Mol. Biol.*, **301**, 1237, (2000).

- 
- [100] S. Kumar, H. Wolfson and R. Nussinov, *IBM J. Res. Dev.*, **45**, 513, (2001).
- [101] Z. Yuan, J. Zhao and Z. Wang, *Protein Eng.*, **16**, 109, (2003).
- [102] A. Sali, B. Veerapandian, J. Cooper, S. I. Foundling, D. J. Hoover and T. L. Blundell, *EMBO J.*, **8**, 2179, (1989).
- [103] B. Veerapandian, J. Cooper, T. L. Blundell, R. L. Rosati, B. W. Domini, D. B. Damon and D. J. Hoover, *Protein Sci.*, **1**, 322, (1992).
- [104] L. Coates, M. P. Erskine, M. P. Crump, S. P. Wood and J. B. Cooper, *J. Mol. Biol.*, **318**, 1405, (2002).
- [105] J. Gomez and E. Freire, *J. Mol. Biol.*, **252**, 337, (1995).
- [106] A. Sali, B. Veerapandian, J. B. Cooper, D. S. Moss, T. Hofmann and T. L. Blundell, *Proteins: Struct. Funct. Genet.*, **12**, 158, (1992).
- [107] A. Schon and E. Freire, *Biochemistry*, **28**, 5019, (1989).
- [108] E. Freire, O. L. Mayorga and M. Straume, *Anal. Chem.*, **62**, 950A, (1990).
- [109] B. Lee, D. Xie., E. Freire and L. M. Amzel, *Proteins: Struct. Funct. Genet.*, **20**, 68, (1994).
- [110] D. Bailey, J. B. Cooper, I. J. Tickle, B. Veepandian, T. L. Blundell, B. Atrash, D. M. Jones and M. Szelke, *Biochem. J.*, **289**, 363, (1993).
- [111] B. Lee and F. M. Richards, *J. Mol. Biol.*, **55**, 379, (1971).
- [112] E. A. Merritt, *Acta Crystallogr. D*, **55**, 1997, (1999).
- [113] T. Lazaridis, A. Masunov and F. Gandolfo, *Proteins: Struct. Funct. Genet.*, **47**, 194, (2002).
- [114] S. Freitag, I. Le Trong, L. Klumb, P. Stayton and R. Stenkamp, *Protein Sci.*, **6**, 1157, (1997).

- 
- [115] P. C. Weber, D. H. Ohlendorf, J. J. Wendoloski and F. R. Salemme, *Science*, **243**, 85, (1989).
- [116] P. C. Weber, J. J. Wendoloski, M. W. Pantoliano and F. R. Salemme, *J. Am. Chem. Soc.*, **114**, 3197, (1992).
- [117] B. A. Katz, *J. Mol. Biol.*, **274**, 776, (1997).
- [118] H. W. Dudley, E. Stephens and M. Zhou, *Chem. Commun.*, **16**, 1973, (2003).
- [119] S. Miyamoto and P. A. Kollman, *Proteins*, **16**, 226, (1993).
- [120] P. C. Weber, *Acta Crystallogr. D*, **51**, 590, (1995).
- [121] S. K. Avrantis, R. L. Stafford, X. Tian and A. Weiss, *ChemBioChem*, **3**, 1229, (2002).
- [122] P. Stayton, S. Freitag, L. A. Klumb, A. Chilkoti, V. Chu, J. E. Penzotti, R. To, D. Hyre, I. Le Trong, T. P. Lybrand and R. E. Stenkamp, *Biomol Eng.*, **16**, 39, (1999).
- [123] A. C. Wallace, R. A. Laskowski and J. M. Thornton, *Prot. Eng.*, **8**, 127, (1995).
- [124] I. K. McDonald and J. M. Thornton, *J. Mol. Biol.*, **238**, 777, (1994).
- [125] A. Velazquez-Campoy, M. J. Todd and E. Freire, *Biochemistry*, **39**, 2201, (2001).
- [126] A. Velazquez-Campoy, Y. Kiso and E. Freire, *Arch. Biochim. Biophys.*, **390**, 169, (2001).
- [127] I. Luque and E. Freire, *Proteins: Struct. Funct. Genet.*, **49**, 181, (2002).
- [128] K. P. Murphy and E. Freire, *Advan. Protein Chem.*, **43**, 313, (1992).
- [129] I. Luque, S. A. Leavitt and E. Freire, *Annu. Rev. Biophys. Biomol. Struct.*, **31**, 235, (2002).

## Appendix A

# Thermodynamic Analysis of Ligand-Protein Interactions

---

### A.1 Isothermal Titration Calorimetry of Protein-Ligand Binding

The binding affinity of ligands is a key selection criteria in high-throughput screening and computational analysis. It is defined by the Gibbs energy of binding,  $\Delta G$ , that, in turn, is determined by entropy and enthalpy changes ( $\Delta G = \Delta H - T\Delta S$ ). In principle, many different possible combinations of  $\Delta H$  and  $\Delta S$  can yield the same  $\Delta G$ , i.e. the same binding affinity. In fact, since the magnitudes of the enthalpy and entropy changes reflect different underlying binding mechanisms, ligands that have been enthalpically or entropically optimised can present different responses to target or reaction condition changes even if they exhibited the same starting affinity.<sup>125,126</sup>

Isothermal titration calorimetry (ITC) is an universal tool for measuring the ener-



getics of binding reactions at constant temperature; it is performed at equilibrium, in solution phase and without any labelling or need for a fluorescent or other probe, and it provides direct information about the thermodynamics of binding. The enthalpy of binding and the equilibrium binding constant are directly measured; the temperature dependence of the enthalpy changes yields the heat capacity which, in turn, provides insight into the surface area buried during binding.

ITC experiments can be carried out to measure the energetics of protein-ligand binding,<sup>107,108,127</sup> under appropriate conditions (temperature, ionic strength, pH, etc.), a syringe containing a ligand is titrated into a cell containing a solution of the macromolecule. As the two elements interact, heat is released or absorbed, until the available binding sites in the cell becomes saturated; at that point, the heat signal diminishes until only the background heat of dilution is observed. Measuring the heat evolved in the stepwise addition of a ligand to the solution describes the association equilibrium in terms of the number of binding sites per ligand molecule (stoichiometry of the reaction), association constant ( $K_a$ ) and enthalpy of the reaction ( $\Delta H$ ). Since ITC provides a direct estimation of both enthalpy and Gibbs free energy changes ( $\Delta G = -RT \ln K_a$ ), the calculation of the entropy change upon ligand binding is straightforward ( $\Delta G = \Delta H - T\Delta S$ ).

However, the enthalpy, the entropy, and the Gibbs energy of binding are global properties of the system, which cannot be used to infer the precise nature of the interactions, nor the groups in the protein and in the ligand that are actually involved in them. To link the experimentally measured thermodynamic properties to the microscopic structure of the system, the structures of the complex and the free enzyme can be analysed; enthalpy and entropy changes can be decomposed into different predictable factors, and the contributions of specific regions of the protein to the total

free energy of binding distinguished (“structure-based thermodynamic analysis”).

## A.2 Structure-Based Thermodynamic Analysis

### A.2.1 Prediction of Binding Enthalpies

As a first approximation, the experimental binding enthalpy of small ligands,  $\Delta H_{exp}$ , can be expressed as the sum of three different factors:<sup>127</sup>

$$\Delta H_{exp} = n_{H+} \Delta H_{protonation} + \Delta H_{intrinsic} + \Delta H_{conformation} \quad (A.1)$$

The “protonation enthalpy”,  $\Delta H_{protonation}$ , is the contribution arising from the protonation/deprotonation of protein’s or ligand’s groups in the binding processes and  $n_{H+}$  the number of protons involved. The “intrinsic enthalpy”,  $\Delta H_{intrinsic}$ , reflects the interactions between the ligand and the protein, and the solvation changes upon ligand binding; the “conformational enthalpy”,  $\Delta H_{conformation}$ , is the enthalpic contribution arising from conformational changes.

### Protonation/Deprotonation Enthalpy

If a protonation/deprotonation process is associated to ligand binding,  $\Delta H_{protonation}$  has to be experimentally dissected by measuring the binding enthalpy at different pH values and using buffers characterised by different ionisation enthalpies. The contribution of this term can be, depending on the ionising groups, of the same order magnitude as the intrinsic binding enthalpy and must be explicitly considered. Once  $\Delta H_{protonation}$  has been evaluated, the “protonation-independent” enthalpy of binding can be calculated as the sum of  $\Delta H_{intrinsic}$  and  $\Delta H_{conformation}$ .

## Intrinsic Enthalpy

$\Delta H_{intrinsic}$  reflects solvation changes upon ligand binding, hydrogen bonds, van der Waals interactions and other interactions between the ligand and the protein; it is the most important term for lead optimisation. Several studies<sup>127,128</sup> have shown that most of the heat capacity and enthalpy change associated with the unfolding of the native state of a protein can be expressed as a linear combination of the differences in polar ( $\Delta ASA_{pol}$ ) and apolar ( $\Delta ASA_{ap}$ ) solvent-accessible surface areas between those states. This also applies to the binding of peptides to proteins or protein-protein interactions, since the atomic interactions are the same. Most binding processes involve dehydration of protein and ligand surfaces; a negative contribution is expected from the burial of apolar surfaces and a positive contribution from the burial of polar surfaces.

The “intrinsic enthalpy” corresponds to the enthalpy that would be observed if the ligand and, most of all, the protein had the same conformation in the unbound and in the complexed state, and if no changes in the protonation states of the two species occurred. The enthalpy at a given temperature  $T$  is expected to scale with the changes in the accessible surface area ( $\Delta ASA$ ) between the complexed and uncomplexed states according to the equation:<sup>127</sup>

$$\Delta H(T) = a(T) \times \Delta ASA_{ap} + b(T) \times \Delta ASA_{pol} \quad (\text{A.2})$$

In this formula,  $a(T)$  and  $b(T)$  are empirically determined scaling coefficients, temperature dependent, and  $\Delta ASA_{ap}$  and  $\Delta ASA_{pol}$  are the changes of accessible surface areas of respectively apolar and polar atoms.

The presence of buried molecules of water at the interface between the protein and

the ligand has to be taken into account. These buried molecules of water play in fact a crucial role in mediating the interactions between enzymes and ligands: they can fill non occupied volume in the complex, satisfy the hydrogen bonding potentials and help the dissipation of charges. While all these actions positively contribute to the binding enthalpy, the incomplete desolvation of the ligand-protein interface decreases the solvation entropy (enthalpy/entropy compensation); the effect of buried molecules of water on the Gibbs energy of binding is thus expected to be smaller than their enthalpic effect. Changes in solvent accessibilities must be calculated taking into account buried molecules of water within 5-7 Å of the ligand.<sup>127</sup>

## Conformational Enthalpy

The binding of small ligands is normally associated with a change in the protein conformation that can be global or, as it is often the case, involve only local rearrangements and/or stabilisation of unstructured regions near the binding site. Also, the ligand itself can be in enthalpically different states in the bound and unbound form.

Given the small size of binding enthalpies, the contributions of conformational changes to binding energies must be explicitly considered even if they are only local in nature. In theory, crystallographic structures of both free and bound conformations provide enough information; in practice, it is sometimes impossible to obtain a high resolution structure of the free protein in exactly the same conditions as those of the complex, and the conformation which is observed in crystals might not be representative of the native state ensemble. Also, small crystallographic differences that are not directly due to ligand binding (e.g. conformational changes in exposed side-chains that are far from the binding site) are often larger than the ligand binding



induced-fit. A possible way to address this problem is to use a minimum of two structures/thermodynamic datasets with different ligands for each protein, where ligands induce the same bound conformation in the protein. In this case, only the structure of the complex can be used and the conformational enthalpy considered as a parameter to be fitted in the parametrisation equation.

For cases in which different ligands determine the same bound protein conformation, if it is assumed that the enthalpy associated with ligand conformational changes is negligible in respect to that of the protein, the protonation-independent enthalpy of binding at a given temperature is equal to:

$$\Delta H_{binding}(T) = \Delta H_{intrinsic} + \Delta H_{conformation} \quad (A.3)$$

and

$$\Delta H_{binding}(T) = a(T) \times \Delta ASA_{ap} + b(T) \times \Delta ASA_{pol} + \Delta H_{conformation}(T) \quad (A.4)$$

In these previous formulae, the conformational enthalpy corresponds to a constant term and the intrinsic enthalpy of binding is a ligand-specific term.

## A.2.2 Prediction of Entropy Changes

The most important entropy contributions in protein-ligand binding are  $\Delta S_{solv}$ , the entropy change arising from the solvent and mainly reflecting the positive entropy that results from burial of apolar surfaces upon binding,  $\Delta S_{conf}$  and  $\Delta S_{rt}$ , which correspond to the reduction of conformational and rotational/translational degrees of freedom occurring in the association of two molecules, and  $\Delta S_{ion}$ , the entropy change associated with protonation/deprotonation processes:<sup>105</sup>

$$\Delta S = \Delta S_{conf} + \Delta S_{solv} + \Delta S_{rt} + \Delta S_{ion} \quad (A.5)$$

The translational/rotational entropy change has been correctly approximated for a 1:1 binding stoichiometry.<sup>105</sup> The entropy of protonation/deprotonation can be evaluated on the basis of the experimental  $\Delta G$  and the  $\Delta G$  associated with the ionisation process.

The other two contributions to the binding entropy change will be analysed in detail. While the entropy of solvation is temperature dependent, the conformational entropy is substantially constant at different temperatures.

### Entropy of Solvation

$\Delta S_{solv}$  at a reference temperature  $T$  can be expressed as a linear combination of the polar and apolar heat capacity changes,  $\Delta C_{p,ap}$  and  $\Delta C_{p,pol}$ :

$$\Delta S_{solv}(T) = \Delta S_{solv,ap}(T) + \Delta S_{solv,pol}(T) \quad (\text{A.6})$$

and:

$$\Delta S_{solv}(T) = \alpha(T)\Delta C_{p,ap} + \beta(T)\Delta C_{p,pol} \quad (\text{A.7})$$

where  $\alpha(T)$  and  $\beta(T)$  are equal to the natural logarithms of the ratio between the reference temperature and the temperatures at which respectively the apolar and the polar hydration entropies are zero (respectively 385.15 K and 335.15 K).<sup>88</sup>

The heat capacity change is weakly sensitive to temperature and has been parametrised in terms of changes in the solvent accessible surface area, as it mainly originates from changes in hydration:<sup>88</sup>

$$\Delta C_p = \Delta C_{p,ap} + \Delta C_{p,pol} \quad (\text{A.8})$$

and:

$$\Delta C_p = a(T)\Delta ASA_{ap} + b(T)\Delta ASA_{pol} + c(T)\Delta ASA_{OH} \quad (\text{A.9})$$

In the equation above,  $\Delta ASA_{ap}$ ,  $\Delta ASA_{pol}$  and  $\Delta ASA_{OH}$  are respectively the apolar, polar and aliphatic hydroxyl groups'  $\Delta ASA$  changes. The coefficients  $a(T)$ ,  $b(T)$  and  $c(T)$  depend on the reference temperature but, for low-temperature calculations ( $T < 80^\circ$ ) the temperature-independent coefficients are sufficient.<sup>88</sup>

## Conformational Entropy

The conformational entropy changes associated to protein unfolding and binding can be evaluated considering three contributions for each amino acid:<sup>109</sup>

1.  $\Delta S_{bu \rightarrow ex}$ , the entropy change associated with the transfer of a buried side-chain in the interior of a protein to its surface;
2.  $\Delta S_{ex \rightarrow u}$ , the entropy gained by a surface exposed side-chain when the backbone unfolds, i.e. changes from a unique conformation to many different ones;
3.  $\Delta S_{bb}$ , a contribution due to the immobilisation of the peptide backbone upon binding.

All these terms have been evaluated for each amino acid, taking into account the probability of different conformers as a function of the dihedral and torsional angles.<sup>88</sup>

The parametrisation of inhibitors' conformational entropy changes differ for peptidic and non-peptidic ligands. For peptidic ligands, the same terms described for protein residues apply; the conformational entropy changes in the protein are mainly restricted to the side-chains that become buried upon ligand binding, while  $\Delta S_{bb}$  is an important contribution for peptidic inhibitors that can be considered unstructured in solution and 'frozen' to a unique conformation in the complex. The conformational entropy of a free non-peptidic ligand is as a first approximation proportional to its number of rotatable bonds and to its total number of atoms. The coefficients of the





## Appendix B

# Residue Stability Constants

---

### B.1 Introduction

Experimental observations (especially NMR techniques detecting hydrogen/deuterium exchange) have shown that proteins undergo local unfolding reactions throughout their structure.<sup>90,129</sup> They therefore should be considered as complex statistical ensembles rather than structures in equilibrium between distinct conformational states. These unfolding reactions can occur independently of each other and can involve only a few amino acids; they determine a large number of states, each one defined by the presence of one or several locally unfolded regions. The native-state ensemble can be defined as the collections of all these states. The Gibbs energy of stabilisation of a protein is not uniformly distributed throughout its three-dimensional structure;

there are in fact regions whose folded conformation is very stable, and regions that are instead very likely to undergo local unfolding.

## B.2 COREX

The COREX algorithm developed by Hilser and coworkers can estimate the individual stability constants for all the residues in a protein.<sup>89–91</sup> The algorithm uses high resolution crystallographic or NMR structure of a protein as a template; the entire protein is considered as being composed of different folding units, and multiple states with some folded and some unfolded regions are computationally generated in all possible combinations. To maximise the number of distinct partially folded states, different “partitions”, i.e. different divisions of the protein into a given number of folding units, are employed. Each partition is defined by placing a block of windows that define the folding units over the entire sequence of the protein, irrespective of specific secondary structure elements. By sliding the whole block of windows, different partitions of the protein are obtained; two consecutive partitions have the first and last amino acids of each unit shifted by one residue. This procedure is repeated until the entire set of partitions have been exhausted.

If  $N$  is the number of the residues in a protein and  $w$  is the employed window size, each protein partitioning consists of  $n$  folding units, where  $n$  is equal to  $N/w$ ; if  $N/w$  is not an integer, the number of residues in the last unit is set equal to the remainder, and the number of folding unit  $n$  rounded up. To avoid partitions

into units composed of less than three residues, if any unit contains fewer than four residues, it is included as part of an adjacent unit. A partitioning results in  $(2^n - 2)$  partially folded intermediates, generated by folding and unfolding the units in all possible combinations. The total number of states generated by this methodology is  $2 + \sum (2^{n_i} - 2)$ , where the sum is performed on all partitions and  $n_i$  is the number of folding units in partition  $i$ . For example, a protein that is composed by 129 residues leads to 32,757 different states using a block of windows of 12 residues each. Different window sizes are generally chosen (ranging from 3 residues to the entire length of the protein) to check the consistency of the results; in many cases, the results are independent from the selected window size within a certain range (e.g. 3 to 20 residues).

For each microstate in the ensemble, the Gibbs free energy and thermodynamic quantities  $\Delta H$ ,  $\Delta C_p$  and  $\Delta S$  of each state are evaluated by using an empirical parameterisation of the energetics described in previous work<sup>105,127</sup> (see appendix A).

For the calculation of the changes of the accessible surface area ( $\Delta ASA$ ) when a given unit unfolds, this is computed as the difference between the *ASA* of the unit in the fully folded state, and the complementary *ASA* that is created when the unit is removed from the protein, plus the *ASA* of the unit in the fully unfolded state. The unfolded *ASA* of the unit is calculated based on each residue's exposed surface area in a structureless tripeptide conformation.

The probability that a given residue  $j$  is in the folded conformation,  $P_{f,j}$ , is equal to the sum of the probabilities of all the conformational states of the protein in which that particular state is folded. A descriptor for each residue in the protein can be evaluated as the ratio of the summed probabilities of all the states in the ensemble in which residue  $j$  is in a folded conformation ( $\sum P_{f,j}$ ) to the summed probabilities of all states in which residue  $j$  is in an unfolded conformation ( $\sum P_{nf,j}$ ):

$$k_{f,j} = \frac{\sum P_{f,j}}{\sum P_{nf,j}} \quad (\text{B.10})$$

Even if the stability constant is defined for each residue, its value does not represent the energy contribution of that residue; it is instead a property of the ensemble as a whole. Stability constants provide in fact the average thermodynamic environment of each residue, considering the energy difference between each partially unfolded microstate and the fully folded reference state, determined by the contributions of all amino acids in the folding units that are unfolded in the microstate, plus the energy contributions derived with the complementary exposed  $\Delta ASA$  on the protein.

The residue stability constants provided by COREX can be compared to hydrogen exchange protection factors.<sup>89</sup> Slow proton exchange of proteins with the solvent can occur only as a result of local partial or global unfolding; protons of residues that are unfolded in partially folded states become in fact exposed to the solvent. Moreover, the residues located in the complementary regions (i.e. the ones that



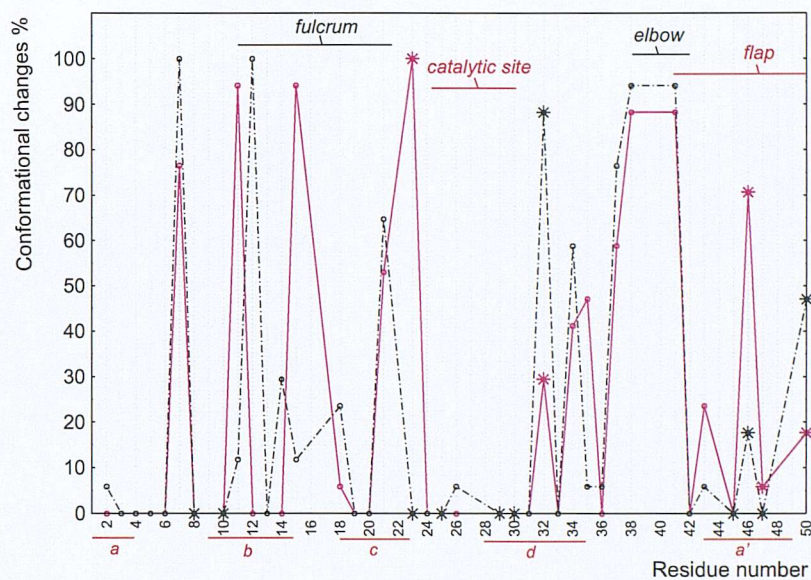
remain folded but are exposed to the solvent when the unfolded ones change their state) are also in contact with the solvent and can thus exchange their protons.<sup>89</sup> The good agreement between the calculated and experimental results suggests that the calculated native state ensemble provides a reasonable representation of the actual native state ensemble.

## Optional Figures

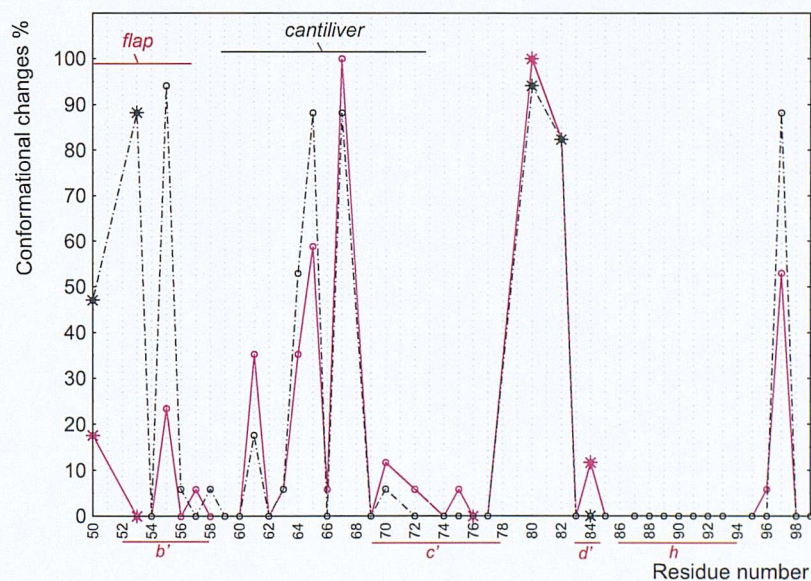
## Appendix C

# Additional Figures

---



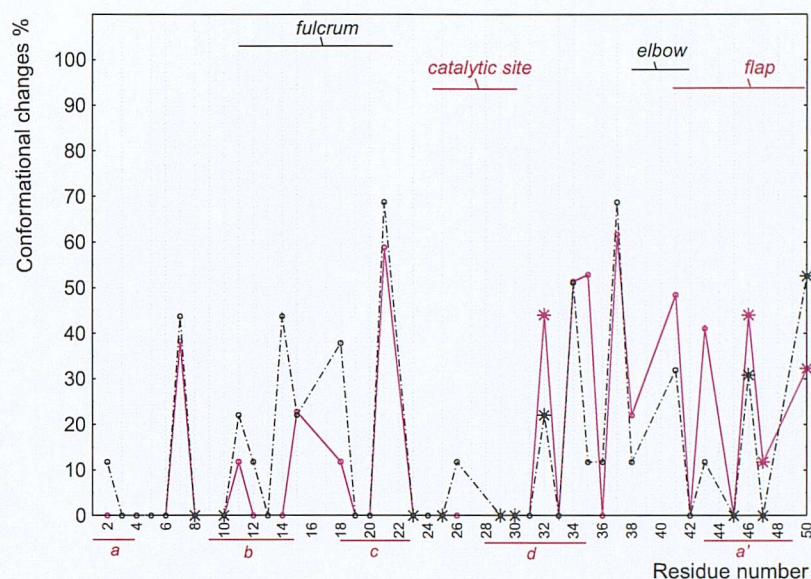
(a)



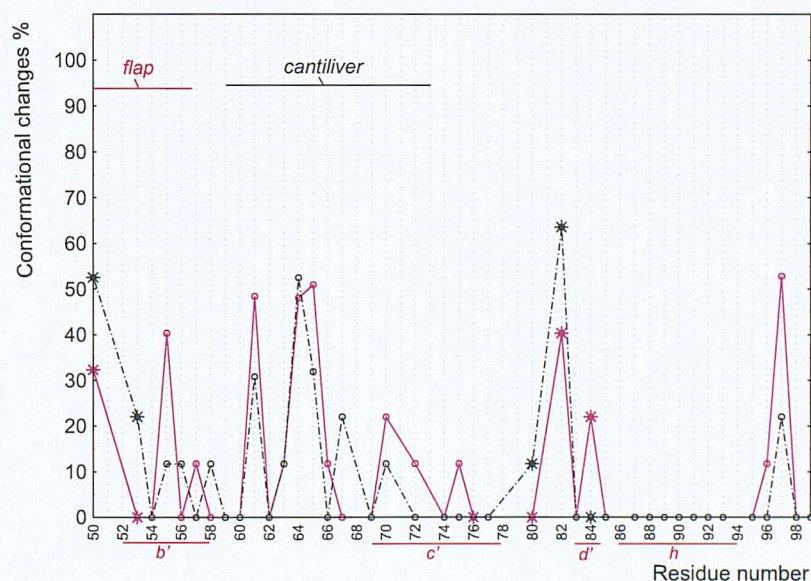
(b)

**Figure C.1:** Percentages of times residues of HIV-1 protease change their  $\chi_1$  rotameric state<sup>50,55</sup> in apo-/holo- protein comparisons; residues have been divided in Figure a (residues from 1 to 50) and b (residues from 50 to 99) for clarity. Data coloured in violet refer to the first monomer of HIV-1 protease, black ones to the second chain of the enzyme. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any symbol. Letters *a*, *b*, *c*, *d*, *a'*, *b'*, *c'* and *d'* on the x axis indicate HIV-1 protease  $\beta$ -strand regions, as described in Figure Alphfig:1AJV; the  $\alpha$ -helix *h* comprises residues from 86 to 94 (Figure Alphfig:1AJV). Some of the most common names employed to indicate regions of HIV-1 protease have been written in the top part of the graph; those that refer to regions that are in contact with the ligand have been written in red.





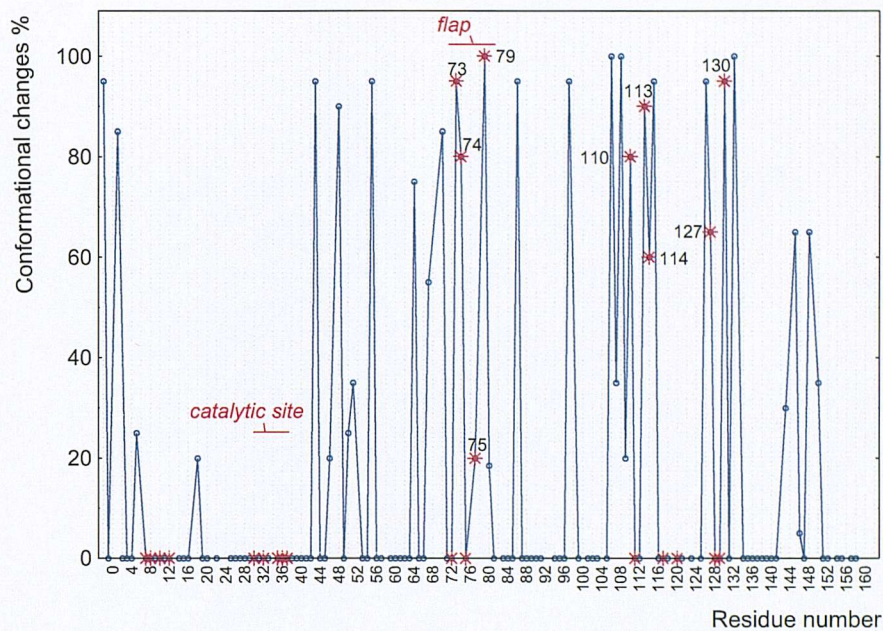
(a)



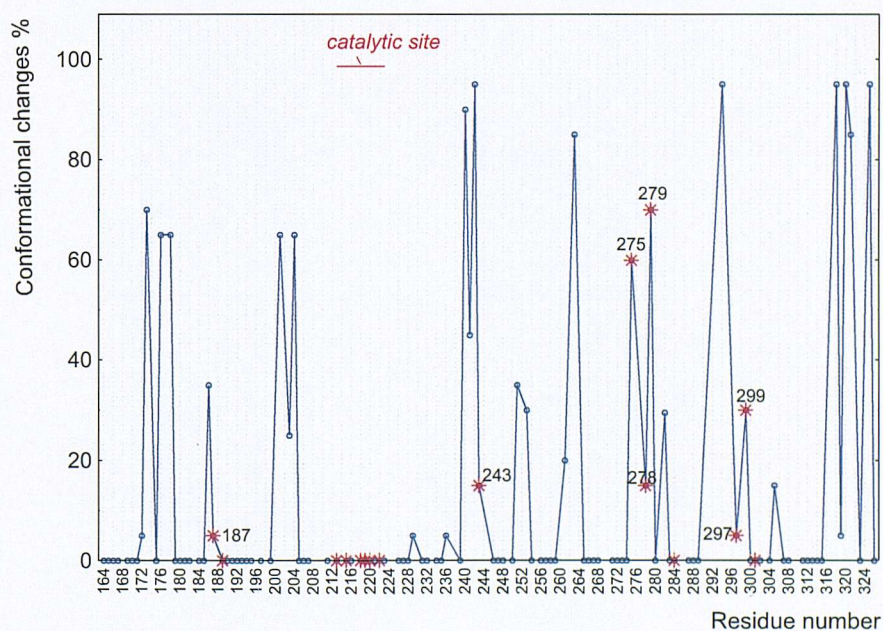
(b)

**Figure C.2:** Percentages of times residues of HIV-1 protease change their  $\chi_1$  rotameric state<sup>50,55</sup> in holo-/holo- protein comparisons; residues have been divided in Figure a (residues from 1 to 50) and b (residues from 50 to 99) for clarity. Data coloured in violet refer to the first monomer of HIV-1 protease, black ones to the second chain of the enzyme. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any symbol. Letters *a*, *b*, *c*, *d*, *a'*, *b'*, *c'* and *d'* on the x axis indicate HIV-1 protease  $\beta$ -strand regions, as described in Figure Alphfig:1AJV; the  $\alpha$ -helix *h* comprises residues from 86 to 94 (Figure Alphfig:1AJV). Some of the most common names employed to indicate regions of HIV-1 protease have been written in the top part of the graph; those that refer to regions that are in contact with the ligand have been written in red.





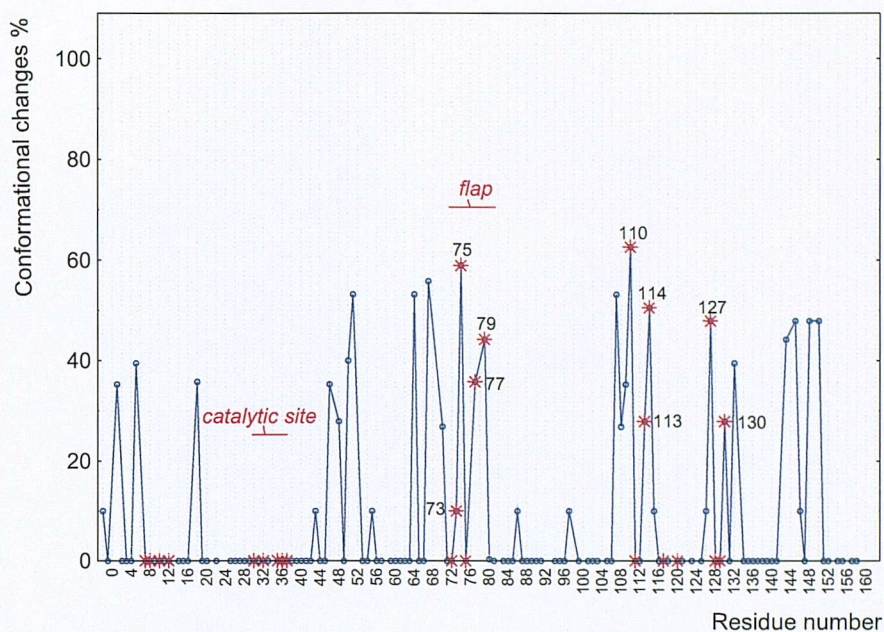
(a)



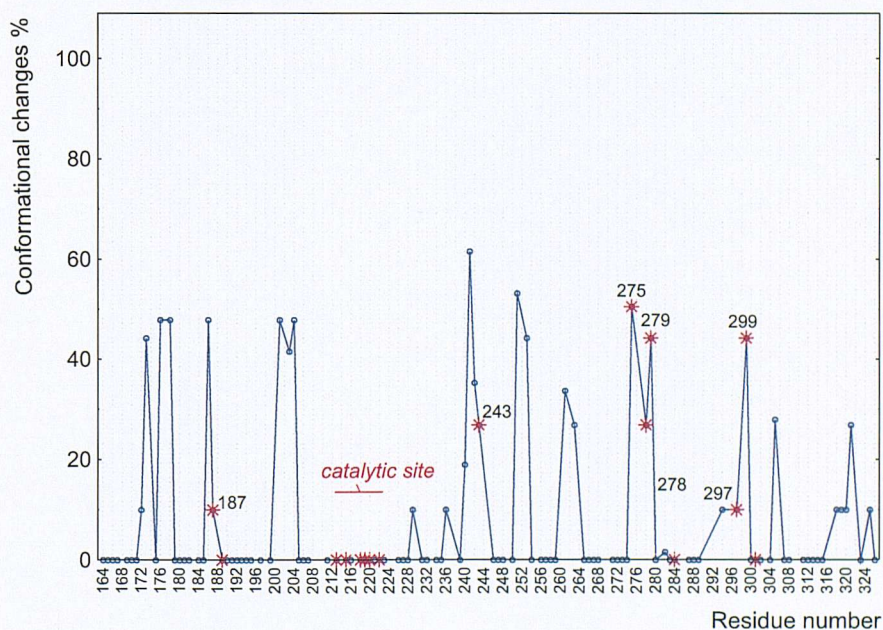
(b)

**Figure C.3:** Percentages of times residues of endothiapepsin change their  $\chi_1$  rotameric state<sup>50,55</sup> in apo-/holo- protein comparisons. Residues have been divided in Figure **a** (from residue -2 to 163) and **b** (residues from 163 to 326) for clarity. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any symbol.





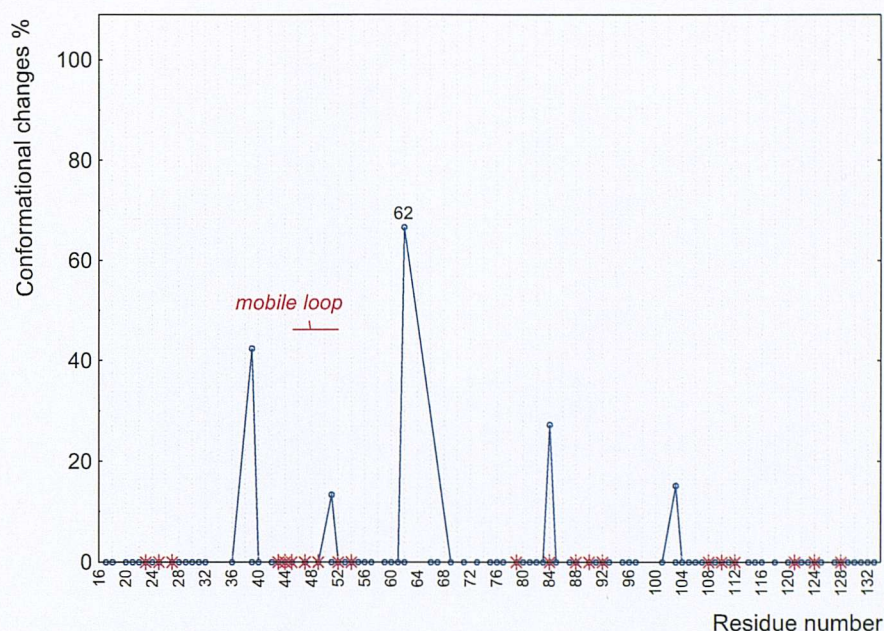
(a)



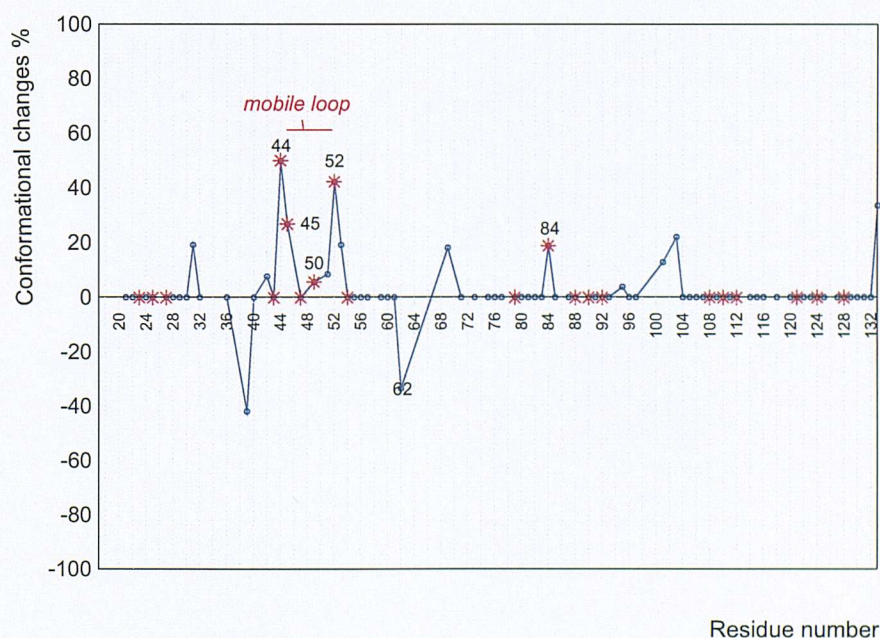
(b)

**Figure C.4:** Percentages of times residues of endothiapepsin change their  $\chi_1$  rotameric state<sup>50,55</sup> in holo-/holo- protein comparisons. Residues have been divided in Figure a (from residue -2 to 163) and b (residues from 163 to 326) for clarity. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any symbol.



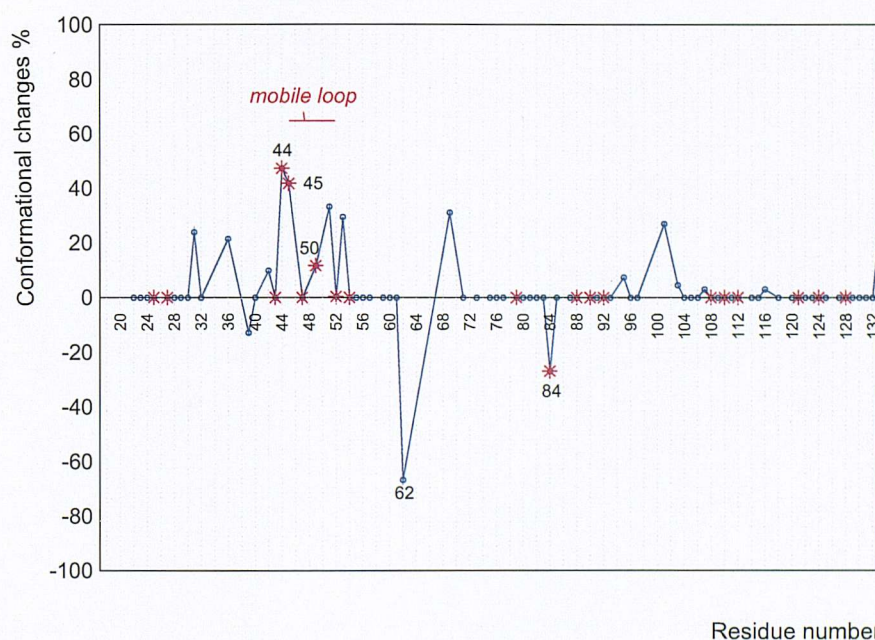


**Figure C.5:** Percentages of times residues of streptavidin change their  $\chi_1$  rotameric state<sup>50,55</sup> in apo/apo- protein comparisons. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data.



**Figure C.6:** Percentages of times residues of streptavidin change their  $\chi_1$  rotameric state<sup>50,55</sup> in apo/holo- protein comparisons. The percentages of conformational changes detected in apo/apo- protein comparisons have been subtracted; positive values correspond to residues whose flexibilities in apo/holo- protein comparisons are greater than those observed in apo/apo- protein pairs. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data.





**Figure C.7:** Percentages of times residues of streptavidin change their  $\chi_1$  rotameric state<sup>50,55</sup> in holo-/holo- protein comparisons. The percentages of conformational changes detected in apo-/apo- protein comparisons have been subtracted; positive values correspond to residues whose flexibilities in holo-/holo- protein comparisons are greater than those observed in apo-/apo- protein pairs. Data for residues that belong to the binding site are indicated with stars; other residues' data are indicated by small circles. Residue sequence numbers that correspond to prolines, alanines and glycines are not associated with any data.