Development and assessment of protein loop modeling methods: Application to CDR loops in antibodies

Amélie Barozet

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ FÉDÉRALE TOULOUSE MIDI-PYRÉNÉES

Délivré par :

l'Institut National des Sciences Appliquées de Toulouse (INSA de Toulouse)

Présentée et soutenue le *18/10/2019* **par :**

AMÉLIE BAROZET

Development and assessment of protein loop modeling methods: Application to CDR loops in antibodies

---

JURY

LIONEL MOUREY	Directeur de Recherche	Président du jury
CHARLOTTE DEANE	Professeur	Rapporteur
PABLO CHACÓN	Docteur	Rapporteur
CHANTAL PREVOST	Chargée de Recherche	Examinatrice
JUAN CORTÉS	Directeur de Recherche	Directeur de thèse
MARC BIANCIOTTO	Docteur	Directeur de thèse
THIERRY SIMÉON	Directeur de Recherche	Invité

École doctorale et spécialité :

EDMITT : Spécialité Informatique et Télécommunications

Unité de Recherche :

Laboratoire d'Analyse et d'Architecture des Systèmes

Directeur(s) de Thèse :

Juan CORTÉS et *Marc BIANCIOTTO*

Rapporteurs :

Charlotte DEANE et *Pablo CHACÓN*

Abstract

This thesis deals with antibody modeling, and in particular the modeling of hypervariable loops found at the interface with the antigen. These protein loops are responsible for the specific recognition of the antigen, as well as the formation of the antibody-antigen complex with a great affinity. The specificity and affinity of this interaction are possible thanks to a great variability of sequence, and also thanks to the plasticity of these protein fragments. Indeed, contrary to other more stable structural elements like alpha helices or beta sheets, protein coils exhibit a flexibility that plays a crucial role in many biological processes.

This manuscript starts with describing the analysis of structural changes in antibodies upon antigen binding, which constitutes the first contribution of this PhD research work. This study, based on the analysis of experimental structural data, shows that antibody conformational changes (occurring mainly in the loops), can be substantial and are not sufficiently accounted for. In particular, docking algorithms show poor results when dealing with excessively flexible hypervariable loops in the antigen binding site.

In this context, the PhD research work then focused more generally on protein loop modeling. These flexible protein regions represent a challenge for structural biology. Most experimental data related to protein structure are obtained through X-ray crystallography. Although this technique is able to accurately determine the structure of the most stable elements, it cannot correctly represent parts that are more flexible. Indeed, it provides a unique structure, which is inappropriate for protein loops, which adopt an ensemble of different conformations with various associated probabilities.

As shown by multiple recent works, and because of this representation bias due to the nature of the employed experimental methods and to the lack of data related to conformational flexibility, current protein methods cannot properly model protein loops. Protein loop modeling is usually performed in two steps. First, a conformational ensemble must be produced. This step, called *sampling*, must exhaustively generate all possible conformations of the loop, in order to globally represent this protein fragment. The next step is called *scoring*, and consists in attributing scores to each of these sampled conformations. This score is meant to represent the energy differences between the models generated during the first step. Sampling and scoring remain open problems. Indeed, methods developed so far in the field mostly focus on predicting a single stable conformation, that is not representative enough.

The two next contributions of the PhD research work logically follow from this observation. The first one presents a method for exhaustive sampling, with a reinforcement learning component to speed up the generation of loop models. This robotics-inspired method uses a geometric representation that forbids steric clashes between atoms and consists in concatenating protein fragments from a database

built specially for this application. The second contribution is an in-depth analysis of the performance of several scoring methods on a set of flexible loops for which experimental data exist. Combining sampling and scoring allows the visualization of energy landscapes implicitly modeled by these methods. The analysis of these energy landscapes enables to precisely identify both the flaws of sampling and the limits of scoring methods.

Finally, these methods were applied to an antibody with a hypervariable loop which changes conformations upon antigen binding. Results show that the methods previously studied and developed enable to model a consistent energy landscape for this flexible loop, identifying both conformations: the one adopted when the antibody is free, and the one adopted upon binding the antigen. This suggests that these methods could be successfully applied to antibody design. Visualizing the modeled energy landscapes would indeed allow to predict a loop's stability in a position or another, thus discarding loop sequences that are insufficiently stable or that adopt undesirable conformations.

Although applied to antibodies, the research contributions presented in this work can perfectly be generalized to the analysis of protein loops in other systems, since the developed methods are not antibody-specific.

Keywords: protein loops, antibodies, robotics-inspired methods, protein loop sampling, protein loop scoring, flexibility of protein loops.

Résumé

Cette thèse porte sur la modélisation d'anticorps, et en particulier des boucles hypervariables situées à l'interface avec l'antigène. Ces boucles protéiques assurent la reconnaissance spécifique de l'antigène ainsi que la formation du complexe anticorps-antigène avec une grande affinité. La spécificité et l'affinité de cette interaction sont rendues possibles par une grande variabilité de séquence, mais aussi grâce à la plasticité de ces fragments protéiques. En effet, contrairement à d'autres éléments de structures plus stables comme les hélices alpha et les feuillets beta, les boucles protéiques possèdent une flexibilité s'avérant cruciale dans un certain nombre de fonctions biologiques.

Ce manuscrit commence par décrire l'analyse des changements structurels d'anticorps survenant suite à la liaison avec l'antigène, qui constitue la première contribution du travail de thèse. Cette étude, s'appuyant sur l'analyse de données structurales expérimentales, établit que les changements conformationnels de l'anticorps (qui concernent principalement les boucles) peuvent être substantiels et sont insuffisamment pris en compte. Notamment, les algorithmes de prédiction de l'amarrage anticorps-antigène sont particulièrement mis en difficulté par une trop grande flexibilité des boucles hypervariables au niveau du site d'interaction.

Fort de ce constat, le travail de thèse s'est ensuite concentré sur la modélisation de boucles protéiques de manière plus générale. Ces régions flexibles des protéines représentent un défi pour la biologie structurale. La grande majorité des données expérimentales liées aux structures protéiques proviennent de cristallographie aux rayons X. Bien que cette technique permette de déterminer avec précision la structure des éléments les plus stables, elle ne permet pas de correctement représenter les parties plus flexibles. En effet, elle ne peut fournir qu'une structure unique, inadaptée à la réalité des boucles protéiques qui peuvent adopter un ensemble de conformations différentes, avec diverses probabilités associées.

En raison de ce biais de représentation dû à la nature des méthodes expérimentales employées et au manque de données liées à la flexibilité conformationnelle, divers travaux soulignent l'insuffisance des méthodes actuelles de modélisation de boucles protéiques. Cette modélisation s'effectue généralement en deux étapes. Tout d'abord, un ensemble de conformations doit être constitué. Cette étape, appelée échantillonnage, doit générer toutes les conformations possibles de la boucle de manière exhaustive, afin de représenter ce fragment protéique de manière globale. L'étape suivante est l'évaluation, et consiste à associer un score à chacune de ces conformations. Ce score est censé représenter les différences d'énergie entre les modèles proposés durant la première étape. L'échantillonnage et l'évaluation restent des problèmes ouverts. En effet, les méthodes développées jusqu'à présent dans ce domaine se concentrent en majorité sur la prédiction d'une unique conformation stable, insuffisamment représentative.

C'est dans ce contexte que se positionnent les deux contributions suivantes de la thèse. La première propose une méthode d'échantillonnage exhaustif, intégrant une composante d'apprentissage par renforcement pour accélérer la génération de modèles de boucles. Cette méthode, inspirée par la robotique, utilise une modélisation géométrique interdisant les collisions stériques entre atomes et consiste à concaténer des fragments protéiques issus d'une base de données construite spécialement pour cette application. La seconde contribution est une analyse en profondeur des performances de diverses méthodes d'évaluation sur plusieurs boucles flexibles pour lesquelles des données expérimentales sont disponibles. La combinaison de l'échantillonnage et de l'évaluation permet de reconstituer une visualisation des paysages énergétiques implicitement modélisés par ces méthodes. L'analyse de ces paysages énergétiques permet d'identifier de manière plus précise à la fois les insuffisances de l'échantillonnage, et les limites des méthodes d'évaluation.

Enfin, ces méthodes ont été appliquées à un anticorps dont une boucle hypervariable subit un changement conformationnel lors de sa liaison avec l'antigène. Les résultats montrent que les méthodes précédemment étudiées et développées permettent de reconstituer un paysage énergétique cohérent pour cette boucle flexible, identifiant à la fois les conformations adoptées lorsque l'anticorps est libre et lorsqu'il se lie avec l'antigène. Cela souligne une application intéressante de ces méthodes pour le design d'anticorps. La visualisation des paysages énergétiques modélisés permettrait en effet de prédire la stabilité d'une boucle dans une position ou une autre, éliminant ainsi les séquences de boucles insuffisamment stables ou adoptant des conformations indésirables.

Bien qu'appliquées aux anticorps, les contributions du travail de thèse se généralisent parfaitement à l'analyse de boucles protéiques appartenant à d'autres systèmes, les méthodes développées n'étant pas spécifiques au cas des anticorps.

Mots-clés: boucles protéiques, anticorps, méthodes inspirées de la robotique, échantillonnage de boucles protéiques, évaluation de boucles protéiques, flexibilité des boucles protéiques.

Acknowledgments

Cette thèse a été pour moi une formidable expérience, autant sur un plan professionnel et scientifique que sur un plan humain. J'ai eu le privilège d'être très bien entourée tout au long de ces trois ans, et j'ai par conséquent un certain nombre de remerciements à adresser.

En tout premier lieu, je tiens à remercier Juan Cortés, qui a été un excellent directeur de thèse et qui a su m'accompagner à tous les niveaux. J'ai beaucoup aimé travailler avec toi et partager des discussions, qu'elles soient scientifiques ou non. Tes connaissances, ta rigueur scientifique et ta passion m'ont énormément apporté.

Je remercie également Marc Bianciotto, mon co-directeur de thèse. Malgré la distance, j'ai toujours pu compter sur ton aide et tu as toujours su te rendre disponible pour m'orienter ou m'apporter ton point de vue. Je remercie aussi tous les gens que j'ai croisés à Sanofi, et avec qui j'ai pu échanger voire travailler, en particulier Hervé Minoux et Jean-Philippe Rameau.

Mes remerciements vont également à Pablo Chacón et Charlotte Deane pour avoir accepté d'être rapporteurs de cette thèse, ainsi qu'aux autres membres du jury, Lionel Mourey, Chantal Prévost et Thierry Siméon.

Je remercie Marc Vaisset, avec qui travailler a été un plaisir. Merci pour ta bonne humeur, ainsi que pour toutes ces discussions, au bridge ou après les réunions MoMA ! Merci également à Maud Jusot, Laurent Dénarié et Alejandro Estaña, rares doctorants avec lesquels je pouvais parler tripeptides sans sourcils levés. Bien que travaillant chacun sur nos sujets, les discussions et les coups de mains échangés furent précieux, tout comme les bons moments partagés en dehors du travail. Enfin, toujours au sein de l'équipe MoMA, merci à Nicolas Sergent.

J'ai eu la chance de travailler dans un cadre très agréable, ce pour quoi je remercie l'équipe RIS du LAAS-CNRS dans son ensemble, mais également tous les doctorants qui ont contribué à rendre ces trois ans inoubliables. Un profond merci à Kathleen (je suis heureuse de pouvoir te compter parmi mes amis, et je te suis reconnaissante d'avoir mis l'ambiance dans ce sous open-space !), Guillaume (pour ton amitié, pour les discussions pendant les pauses, et pour ton soutien lors de ces week-ends d'écriture de thèse), Guilhem, Jules, Amandine, Alejandro, Rafa et David (pour ces discussions sur tout et surtout n'importe quoi, tous ces rires et votre soutien dans les moments qui n'ont grâce à vous pas été si difficiles), Ellon, Mamoun, Greg, Christophe et Raph (qui étaient là au début de ma thèse et le sont encore), Yoan, Arthur, et beaucoup d'autres.

Mes remerciements vont également à toutes les personnes, qui en dehors du travail, m'ont accompagnée pendant cette thèse. Je pense par exemple à Miki, Rémi, Aurélien et Clarisse, mais j'en oublie sûrement !

Enfin, merci à ma famille et mes amis d'être toujours là pour moi. Votre soutien est précieux.

Contents

Glossary

Ab	Antibody
ABR	Antigen Binding Region
Ag	Antigen
CDR	Complementarity-determining region
Fab	Antigen-binding fragment
Fc	Crystallizable fragment
FFT	Fast Fourier Transform
FR	Framework Region
Fv	Variable fragment
HCA	hierarchical cluster analysis
IK	Inverse Kinematics
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
PCA	Principal Component Analysis
PDB	Protein Data Bank
PEL	Potential Energy Landscape
REMD	Replica Exchange Molecular Dynamics
RL	Reinforcement Learning
RMSD	Root-mean-square deviation
SDR	Specificity-Determining Residue

Introduction

Antibodies are essential proteins of the immune system. They are capable of identifying specific molecules exposed by pathogens, called antigens, by forming complexes with them. Neutralization of the pathogen then either directly results from antibody binding, or from the immune response triggered by antibodies when they bind.

This thesis deals with antibody modeling, and in particular the modeling of hypervariable loops found at the interface with the antigen. These protein loops, called CDRs, are responsible for the specific recognition of the antigen, as well as the formation of the antibody-antigen complex with high affinity. The specificity and affinity of this interaction are possible thanks to a great variability of sequence, and also thanks to the plasticity of these protein fragments. Indeed, contrary to other more stable structural elements like α-helices or β-sheets, protein loops exhibit a flexibility that plays a crucial role in many biological processes.

These flexible protein regions represent a challenge for structural biology. Most experimental data related to protein structure are obtained through X-ray crystallography. Although this technique is able to accurately determine the structure of the most stable elements, it cannot correctly represent parts that are more flexible. Indeed, it provides a unique structure, which is inappropriate for protein loops, which adopt an ensemble of different conformations with various associated probabilities. As shown by multiple recent works, and because of this representation bias due to the nature of the employed experimental methods and to the lack of data related to conformational flexibility, current structural bioinformatics tools cannot properly model protein loops.

Protein loop modeling is usually performed in two steps. First, a conformational ensemble must be produced. This step, called *sampling*, must exhaustively generate all possible conformations of the loop, in order to globally represent this protein fragment. The next step is called *scoring*, and consists in attributing scores to each of these sampled conformations. This score is meant to represent the energy differences between the models generated during the first step. Sampling and scoring remain open problems. Indeed, methods developed so far in the field mostly focus on predicting a single stable conformation, that is not representative enough.

The main goal of this thesis is to develop new methods beyond the state of the art for flexible loop modeling, and to show the particular interest for antibody CDR modeling. This manuscript is structured in five chapters.

Chapter 1 provides background for the thesis work. It first gives general notions about protein function and structure, before focusing on antibodies, providing additional structural details and introducing the problems of structural prediction, antibody-antigen docking and antibody design. Protein loop modeling is then presented, together with a description of examples of state-of-the-art methods in that

field. Finally, this chapter describes energy landscapes, and various methods to explore them.

An analysis of structural changes in antibodies upon antigen binding is then presented in Chapter 2. This study, based on the analysis of experimental structural data, shows that antibody conformational changes (occurring mainly in the loops), can be substantial and are not sufficiently accounted for. In particular, docking algorithms show poor results when dealing with excessively flexible hypervariable loops in the antigen binding site.

Following these observations, the next two chapters are oriented towards general methods for protein loop modeling. Chapter 3 presents a method for exhaustive loop conformational sampling, including a reinforcement learning component to speed up the generation of loop models. This robotics-inspired method uses a geometric representation that forbids major steric clashes between atoms. The loop sampling process consists in concatenating protein fragments from a database built specially for this application.

Chapter 4 then presents an in-depth analysis of the performance of several scoring methods on a set of flexible loops for which experimental data exist. A process combining sampling, scoring and projection of loop samples is employed to visualize the energy landscapes implicitly modeled by these methods. The analysis of these energy landscapes enables to precisely identify both the flaws of sampling and the limits of scoring methods.

Finally, these methods were applied to an antibody with a hypervariable loop which changes conformations upon antigen binding. This study is described in Chapter 5. Results show that the methods previously studied and developed in Chapters 3 and 4 enable to model a consistent energy landscape for this flexible loop, identifying both conformations: the one adopted when the antibody is free, and the one adopted upon binding the antigen. This suggests that these methods could be successfully applied to antibody design, for instance. Visualizing the modeled energy landscapes would indeed allow to predict a loop's stability in a position or another, thus discarding loop sequences that are insufficiently stable or that adopt undesirable conformations.

At the end of this manuscript, an overall conclusion summarizes the work done in this thesis and provides some directions for future research.

Thesis contributions

With the aim of better modeling antibody structures, this thesis led to various contributions at several levels. These contributions range from a study to more precisely establish the methodological needs, to the development of techniques to fulfill these needs, ending with the validation of these methods on our system of interest, antibodies. Details are provided hereafter.

- *Analysis of conformational changes in antibodies upon antigen binding*: the thesis started with the measure and description of conformational changes in

27 antibody-antigen systems for which both free and bound structures exist. Breaking down the antibody structure into several subparts, a quantitative and qualitative analysis of the different types of motions was conducted. Combining these observations with reported results for four docking algorithms on the studied antibody-antigen systems led to designate large CDR loop movements as a probable source of failure for docking algorithms. This work has been published in *Immunology Letters* [Barozet 2018].

- *Development and extensive testing of a new loop sampling method*: this method constitutes the first part of the methodological contributions proposed to respond to the problem of bad modeling of loop flexibility. This method, called MoMA-LoopSampler, adopts a geometric representation with strict collision detection and concatenates tripeptide states from a dedicated database. A novel reinforcement learning approach was integrated to speed up sampling. Since this method was meant to be used in flexible loop modeling, particular care was taken to ensure a sufficient diversity in the sampled conformations, while maintaining high quality in the generated ensembles. Consistency of the methods was explored, and a thorough analysis of the sampled ensembles was conducted. This work has been accepted for publication in *Bioinformatics* [Barozet 2019b].

- *Comparison of state-of-the-art loop scoring methods*: this contribution encompasses the second part of the methodological developments proposed in this thesis. In this work, we developed a process to visualize and easily interpret energy landscapes of flexible loops, as modeled using diverse scoring methods. Conformational ensembles were sampled using MoMA-LoopSampler for multiple systems comprising a flexible loop. State-of-the-art methods were then used to score the different samples, and a 2D projection using meaningful descriptors was employed to plot the energy landscapes. This study identified the best methods to use for an accurate evaluation of conformations of flexible loops, and indicated potential sources of inaccuracy for the produced landscapes. A manuscript describing this work is currently under revision by *Proteins: Structure, Function and Bioinformatics* [Barozet 2019a].

- *Modeling of the energy landscape for a flexible CDR loop*: this work illustrates how the methods developed in the thesis can be applied to an antibody hypervariable loop. By successfully modeling the energy landscape of a CDR loop, we show how this thesis addressed the problem it initially identified: namely, properly modeling flexibility in antibody hypervariable loops. We are hoping to publish this work, along with the different structures, in a near future.

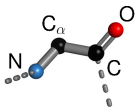
Although antibodies constituted the initial motivation behind the methodological developments performed in this thesis, the methods presented in this work are not antibody-specific and can perfectly be generalized to the analysis of protein loops in other systems.

Chapter 1

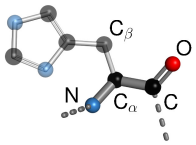
Background

Contents

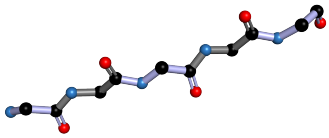
This chapter introduces basic concepts that are necessary to understand the work presented in this thesis. First, Section 1.1 briefly presents basic notions about protein function and structure. Antibodies, which are examples of protein systems, are introduced next (Section 1.2). The dedicated section describes their structure in detail, including the loops which are responsible for interacting with the antigen, and presents several applications requiring an accurate modeling of these loops. The next section (Section 1.3) introduces loop modeling methods, including loop sampling and scoring, which are at the core of Chapters 3 and 4, respectively. Finally, Section 1.4 presents notions of energy landscape exploration, which are



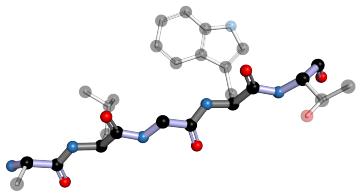
(a) Backbone (or main-chain) of an amino acid residue



(b) Amino acid residue with its side-chain



(c) Main-chain of a 5-residue-long chain



(d) Chain of 5 residues (with side-chains)

Figure 1.1. Chains of amino-acid residues. Hydrogen atoms were omitted for clarity.

necessary to understand proteins conformations and dynamics. Interest in these energy landscapes is twofold. First, being able to accurately model them would provide great insight into proteins and protein loops mode of action, which is why they constitute an inspiring goal for loop modeling methods. Second, they are used throughout the thesis to validate the consistency of the results returned by loop modeling methods.

1.1 Proteins

1.1.1 Definition and function

Proteins are macro-molecules produced by living organisms and involved in a multitude of biological processes [Kaiser 2007]. Enzymes, antibodies and hormones are all examples of proteins. From a structural point of view, a protein consists of one or several long chains of amino acid residues, themselves composed of a common backbone (or main-chain) and a variable side-chain attached to a carbon atom named C_α (Figure 1.1). There are 20 natural amino acids, differing only by the nature of their side-chains. A detailed explanation of the main concepts in structural biology can be found in textbooks such as references [Creighton 1993, Gu 2009].

The function of a protein is directly related to its three-dimensional structure but also to the flexibility of this structure, as it has been recently shown [Teilum 2009]. Understanding protein structure, flexibility and dynamics is therefore key to explaining the mode of action of a protein, predicting its function or explaining its malfunction. From another perspective, being able to accurately model such mechanisms is also crucial in order to design complex systems carrying out specific functions.

1.1.2 Structure and representation

Many proteins can fold to reach a relatively stable 3D structure determined solely by their amino-acid sequence. This section provides a brief introduction about protein structure. More details can be found in specialized textbooks such as [Brändén 1999]. This structure can be described at 4 different levels (Figure 1.2).

Primary level: the amino-acid sequence of the protein.

Secondary level: the local arrangement of amino acid fragments. The most frequent secondary structure elements are α-helices and β-sheets, considered relatively stable. Other secondary structure elements include π-helices, 3_{10}-helices, turns, β-bridges, bends and finally coils (or loops), which are the most unstructured fragments.

Tertiary level: organization of the secondary level elements into an autonomous three-dimensional structure, called domain.

Quarternary level: some proteins such as antibodies or hemoglobin consist of several domains or chains. This level describes their relative organization.

Despite the relative stability of these well-folded proteins, they may still include more flexible parts, including loops, that can be crucial to their functional role. Degrees of freedom in the structure include atom bond lengths, atom bond angles, dihedral angles (defined by four bonded atoms), and the relative orientation and position of the composing amino-acid chains.

One way to describe the conformation of a protein is to employ the 3D Cartesian coordinates of its atoms as in the Protein Data Bank (PDB) [Berman 2000] format. Another interesting alternative consists in using atom bond lengths and angles and dihedral angles, also called the *internal* coordinates of the protein, to describe the conformational state of a protein (Figure 1.3). Indeed, given that atom bond lengths and angles are determined by the nature of the involved atoms and scarcely vary, this representation allows to limit the number of variables in many applications, by focusing on dihedral angles and inter-chain poses to describe protein motion.

In the backbone, three dihedral angles can be defined per residue: ω, ϕ and ψ (Figure 1.4). ω is the dihedral angle corresponding to the peptide bond connecting two amino-acid residues. It usually takes values around π (this is called the *trans* conformation), but it may also take values close to 0 (*cis* conformation). The *cis* conformation is mainly observed when the residue after the peptide bond is a proline, although this conformation may also be adopted by other residue types. This limited range of values for the ω angles further motivates the use of internal coordinates to describe the conformations of proteins.

In side-chains, dihedral angles are usually referred to as χ angles (χ_1, χ_2, $\ldots$, χ_n, with n the number of dihedrals in the side-chain). Side-chains were shown to adopt a limited set of conformations given the state of the backbone they are attached to. This led to the development of so-called rotamer libraries containing the possible

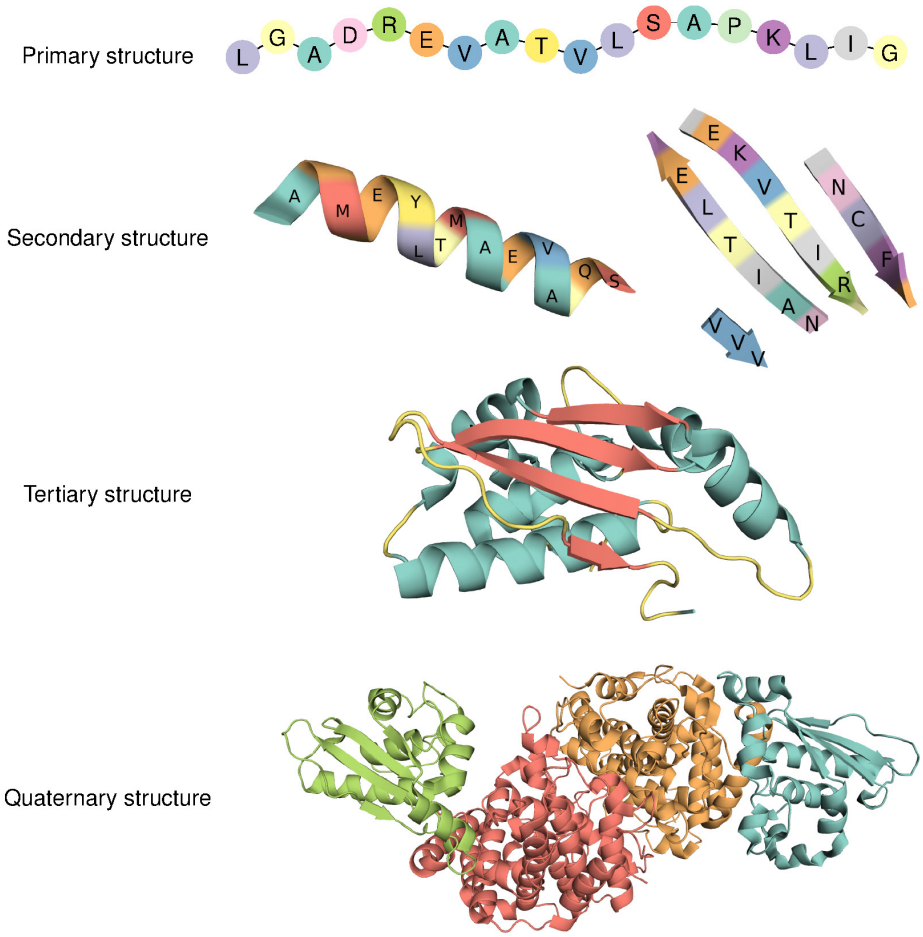


Figure 1.2. The four levels used to describe a protein structure.

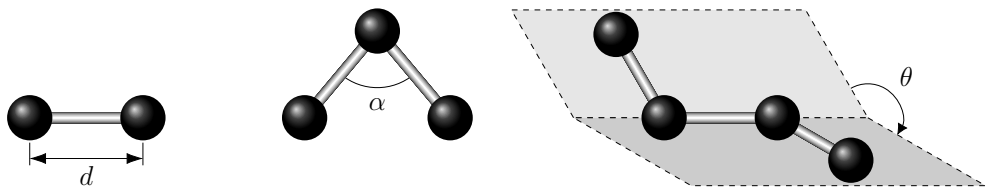


Figure 1.3. Internal protein coordinates.

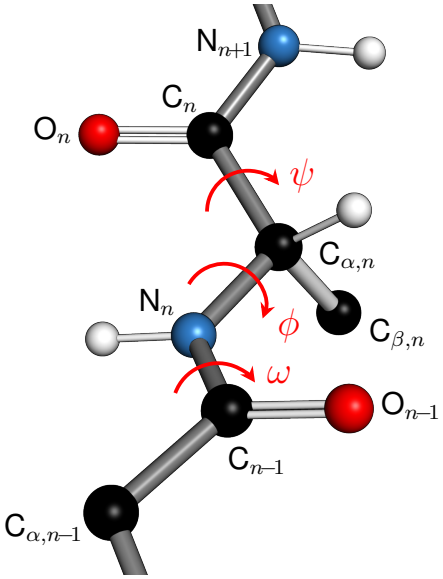


Figure 1.4. Definitions of dihedral angles ω, ϕ and ψ for the backbone of residue n in the chain.

values for χ angles. These rotamer libraries can then be used in methods for the optimization of side-chain placement [Krivov 2009]. While most rotamer libraries give a set of discrete possible conformations, BASILISK is an alternative providing continuous distribution functions for the different χ angles [Harder 2010].

1.2 Antibodies

1.2.1 Function

1.2.1.1 Role in the immune system

Antibodies (Abs) are essential molecules of the immune system. They are proteins of the immunoglobulin family, secreted by B-cells or expressed on the surface of their membrane. Their structure allows them to bind different molecular pathogens depending on their variable domain (see Section 1.2.2). The molecular entity targeted by an antibody is called its *antigen* (Ag). It may be a peptide (i.e. a short amino-acid chain), a protein, or other molecular substances attached to carriers such as proteins (e.g. haptens). The present chapter provides a limited and specific introduction to antibodies and their structure. A more detailed general presentation can be found in reference [Goding 1996].

There are several classes of antibodies, performing a range of different functions. These functions include: *opsonization* (the action of coating a pathogen, thereby triggering phagocytosis), *agglutination* (the action of forming precipitates by agglutinating, thereby facilitating phagocytosis), *antibody-dependant cell-mediated cytotoxicity* (the action of binding at the surface of a tumor cell, enabling the recognition

of the cell by natural killer cells, triggering cell death), *neutralization of toxins*, by attaching to their active site, *neutralization of viruses*, ...

1.2.1.2 Affinity and specificity

While high affinity is necessary to elicit an efficient immune response against a pathogen, an antibody also needs to be highly specific, so that it does not target self-proteins or other molecules involved in the functioning of the host organism. Thanks to their structure (See Section 1.2.2), antibodies can achieve both high specificity and affinity to their antigens, thus constituting a very interesting protein for therapeutic purposes [Chames 2009].

Other mechanisms such as polyspecificity (the ability to bind several potential antigens) and heterospecificity (a higher affinity to antigens other than the one that triggered the antibody response in the first place) have been observed in antibodies, suggesting that a more precise definition of specificity is necessary [Van Regenmortel 2014]. Specificity would thus be the ability of the antibody to recognize only one among very similar molecules, thus not excluding the possibility that the antibody can recognize other unrelated antigens.

1.2.2 Quaternary structure, domains and regions

An antibody is a protein formed by the association of two identical heavy chains and two identical light chains. A heavy chain is composed of a variable domain V_H and several constant domains (C_H1, C_H2...), while a light chain is composed of a variable domain V_L and only one constant domain C_L. Heavy and light chains are assembled so that the antibody adopts a symmetrical 'Y' topology (Figure 1.5). Each arm of this 'Y' consists of two variable domains (V_H and V_L) and two constant domains (C_H1 and C_L), while the base of the 'Y' contains the other constant domains from the two heavy chains.

The base of the 'Y' is the crystallizable fragment (Fc), responsible for the activation of the immune response. The arms of the 'Y' are the antigen-binding fragments (Fab fragments). Their tips, composed of the two variable domains, are called the variable fragments (Fv) and are the regions that directly interact with the antigen. V_H and V_L both contain three hypervariable loops (H1, H2, H3 for V_H and L1, L2, L3 for V_L) that play a crucial role in antigen binding. They are called the complementarity-determining regions (CDRs). The regions of Fv that are outside the CDRs are called Framework Regions (FRs).

Wilson and Stanfield reviewed the conformational changes that could be observed in Fab fragments upon antigen binding [Wilson 1993, Wilson 1994b], based on the limited data available at the time. At the Fab level, the main conformational changes include: side-chain repacking [Li 2000] (especially in CDRs), loop movements in hypervariable loops [Rini 1992], orientation change between Fv and the constant region formed by C_H1 [Guddat 1994] (the elbow angle), and repacking between heavy and light chains [Herron 1991].

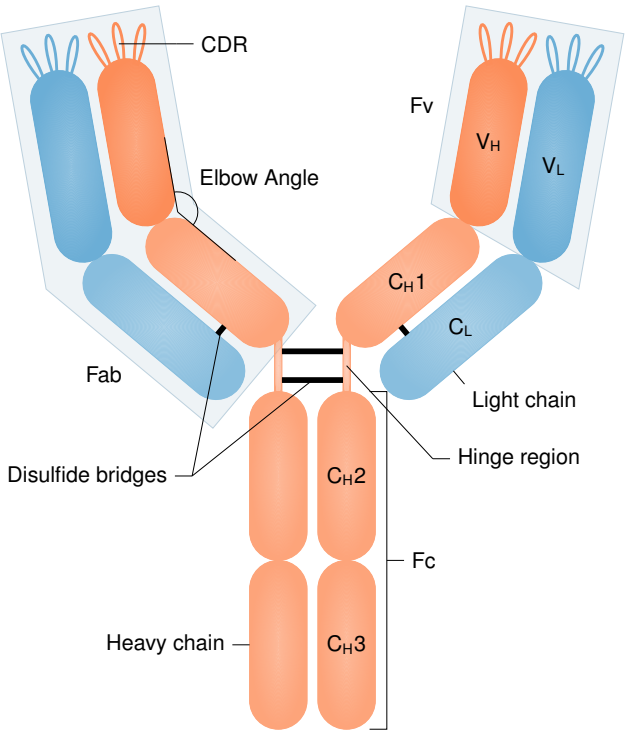


Figure 1.5. Structure of an antibody.

1.2.3 Complementarity Determining Regions (CDRs)

This section focuses on the regions that directly interact with the antigen. An in-depth analysis of loops in that region, along with their flexibility, is performed in Chapter 2.

1.2.3.1 Definitions

Definitions of the CDRs are not straightforward, and several have been proposed in the literature. Kabat and colleagues defined the CDRs as the most variable regions in the antibody sequences [Kabat 1977]. Chothia and colleagues analyzed 3D structures of antibodies and redefined the CDRs as the loop regions in the antigen combining site [Chothia 1987]. This definition differed slightly from that of Kabat *et al.* [Kabat 1977]. The IMGT numbering scheme [Lefranc 2003] is another definition of CDRs that can be used for all immunoglobulin variable domains (V_L, V_H and T-cell receptors). H3 is the most variable of CDR loops, with a length varying from a few residues to more than 30 residues.

Analysis of the relative positions of the CDR loops showed that H3 is at the center of the combining site and interacts with all the other CDR loops except H2 [Marillet 2015].

1.2.3.2 Canonical structures

Chothia and co-workers found that five out of the six hypervariable loops (L1, L2, L3, H1 and H2) displayed a limited set of conformations that they called canonical conformations [Chothia 1987, Chothia 1989, Al-Lazikani 1997]. The length of the loop and the presence of key residues at certain positions was found to determine the canonical conformation adopted by the CDR.

Attempts have been made to find rules for the structure of H3. These focus on the 'torso' region of the loop [Shirai 1996, Morea 1997] and try to differentiate between 'kinked' and 'extended' conformations from the nature of the residues at some key positions: it was suggested that the formation of the kink was conditioned by the formation of a salt bridge and/or several hydrogen bonds. An analysis of similar kinked conformation in the PDB [Weitzner 2015] showed that (i) this conformation is rare (ii) but it can be found in some isolated proteins and also in entire protein families (iii) formation of the kink can happen without the presence of the key residues found in antibodies. This suggests that the kink conformation has more to do with the loop environment than the presence of key residues at certain positions.

More recently, North and colleagues [North 2011] performed an automatic classification of the different CDR loop structures, in an attempt to determine whether the classification performed by Chothia and colleagues still held up with the increased PDB data. They found that despite the limited set of data available at the time of their definition of canonical structures, Chothia and colleagues captured most of the possible CDR conformations. They reported that the combination of a specific type of CDR (i.e. H1, H2, L1, L2 or L3) and a given length of loop corresponds most of the time to only one canonical structure. When it is not the case, sequence can easily discriminate between the different possible canonical structures.

Nikoloudis *et al.* report another recent classification [Nikoloudis 2014], which does not involve any arbitrary distance threshold. Three levels of classification are reported. For all loops except H3, classes defined by previous classification methods usually correspond to the first or second level of classification reported in this article. An interesting observation is that for most CDR type/loop length combination, the largest cluster at the first level contains a vast majority of the structures, suggesting that antibody-antigen recognition is very subtle, and specificity is ensured by only slight changes in CDR conformations.

Teplyakov and co-workers analyzed the structures of the 16 antibodies formed from 4 different light chains and 4 different heavy chains [Teplyakov 2016]. Results showed that when one of the heavy chain was paired with the 4 different light chains, different canonical structures were adopted by H1 in all 4 antibody structures, showing that the sequence of the loop (or even of the whole chain) does not fully determine the canonical structure, and that the pairing of the heavy and light chains may also have an influence.

Existing combinations of canonical structures have also been analyzed for patterns. Tomlinson and colleagues reported the sequencing of heavy chain V gene seg-

ments (see Section 1.2.4.1) from an individual [Tomlinson 1992]. Since the V gene segments from the heavy chain locus contain both H1 and H2 sequences, Chothia and co-workers used these results to investigate the number of different H1/H2 combinations of canonical structures [Chothia 1992]. Only seven such combinations are found. In all of these combinations, residue 33 from H1 interacts with residues 50 and 52 (as well as 52a when it exists) from H2. Interaction between the residues 33 from H1 ans 53 from H2 is also observed in most combinations. However, results here are limited to germline antibodies. The limited number of combinations of canonical structures was reported in another analysis [Vargas-Madrazo 1995], in which the authors explain this restriction by the necessity to adopt certain topographies of the binding site in order to accommodate certain types of antigens.

1.2.4 Variability, maturation and flexibility

1.2.4.1 Variability and germline antibodies

The germline DNA locus for the heavy chain contains a variety of V (variable) gene segments, separated by introns, followed by several D (diversity) segments, followed by several J (joining) segments, finally followed by the constant gene segments corresponding to the different isotypes.

In progenitor B-cells, VDJ recombination rearranges the heavy chain locus to select and place consecutively one V segment, one D segment and one J segment, followed by the C segments. Progenitor B-cells thus selected become precursor B-cells. The same process happens at the light chain loci, except that light chains do not contain D segments. The precursor B-cell then becomes an immature B-cell producing a specific class of antibodies called IgM. Subsequent class switching may then enable the production of other types of antibodies.

The multiplicity of V, D and J gene segments, as well as the variety of combinations resulting from V(D)J recombination is responsible for the variability of germline antibodies. During rearrangements, the ends of the different fragments may be extended or cut, producing even higher diversity. The position of H3 at the VDJ recombination site may thus explain its great variability.

1.2.4.2 Maturation and its effect on antibody flexibility

Maturation is a process involving somatic hypermutations affecting the immunoglobulin genes in B-cells following the recognition of an antigen by an antibody expressed on their membrane. Since this recognition causes the B-cell to proliferate, the competition between B-cells will tend to select those producing antibodies with higher affinity for the antigen.

The first contact of an antigen with the immune system elicits the *primary* immune response, involving germline antibodies. This triggers the maturation process described above, so that a later exposure to the same antigen triggers the *secondary* immune response, involving mature antibodies.

Manivel and colleagues [Manivel 2000] analyzed the thermodynamic properties of antibody/antigen association and dissociation, focusing on the enthalpic and entropic contributions. As will be later described (Section 1.4.2), enthalpic contributions result from the creation of (un)favorable contacts while entropic contributions are closely related to a system's flexibility. Although the thermodynamics of dissociation are very similar for primary and secondary response antibodies, association shows very important differences. The association between the primary response antibodies and their antigens is shown to be enthalpy-driven with a strongly unfavorable entropic contribution: the change in enthalpy is negative, and so is the change in entropy, suggesting that the flexibility of the antibody hinders the association between the antibody and the antigen. Conversely, the association between the secondary response antibodies and antigens shows unfavorable enthalpic effects, but neutral entropic effects. Overall, when taking into account the dissociation effects, the free energy of binding is lower and less temperature dependent for secondary response antibodies than for primary response antibodies, resulting in a higher affinity of mature antibodies to their antigen. While the reason for the difference in enthalpic contribution to binding is unclear, the difference in entropic contribution may come from a higher rigidity of the binding site of the secondary response antibodies, resulting in a lower loss in conformational entropy for more mature antibodies.

Indeed, higher rigidity of mature antibodies have been observed [Wong 2011], especially in H3, and was found to result from the formation of hydrogen bonds and salt bridges, as well as from an improved side-chains packing. Germline antibodies display very flexible H3 loops, in order to be able to recognize a broad range of antigens. Maturation was observed to lead to a preconfiguration of the H3 loop in the right position to bind the antigen [Schmidt 2013]. As a consequence, although both the germline and its corresponding mature antibodies exist in the right configuration to bind the antigen, the mature antibodies occupy this conformational state more often than their germline counterpart.

1.2.4.3 Flexibility, a key feature for antibodies

Antibody-antigen recognition is not necessarily a lock and key mechanism and can involve induced-fit, a mechanism in which the antibody undergoes conformational changes, in particular in the regions of hypervariable loops, to accommodate the antigen. These changes can involve large H3 rearrangement or a change in the relative position and orientation of the light and heavy chains [Wilson 1994a]. James and colleagues [James 2003] showed that induced fit was not enough to describe their observations on antibody multispecificity and that conformational diversity preceding the presence of the antigen was necessary to account for some of their observations. Induced fit would thus correspond to a displacement of the conformational ensemble equilibrium, with more antibodies adopting the conformation that is compatible with the antigen.

All these observations, along with the previous findings about the effects of affinity maturation, suggest that flexibility is a key feature of the antigen combining site and that it must be taken into account in order to correctly model the antibody structure, in particular in CDR loop portions of the structure. Flexibility is an important consideration when employing loop modeling (Section 1.2.5.2) to represent CDRs. It can also be handled via motion planning algorithms (Section 1.4.3.3).

1.2.5 Structure prediction

Due to the tight relationship between structure and function in proteins, structure prediction is an important problem that is not specific to antibodies. This section describes some tools to predict antibody structures, most of which originate from more general methods of protein structure prediction. More details about general methods for protein structure prediction can be found in reference [Zaki 2008].

1.2.5.1 Structural prediction of the variable regions of the antibody

Several solutions currently exist to predict the structure of Fv [Marcatili 2008, Pedotti 2011] using homology modeling and the knowledge about canonical structures. It is possible to choose the antibody with highest sequence identity as a template, or the antibody with the same canonical structures. Different templates can be used for the heavy and light chains, and/or for the loops and the framework regions. These different choices come with different potential problems. When sequence identity is low, taking the same antibody structure to model both the heavy and light chains or both the framework and the loops may lead to poor prediction. On the other hand, taking structures from different templates for the loops and the frameworks or the light and heavy chains requires an optimization phase to find the correct orientation of the different structural fragments relative to each other. Moreover, as mentioned previously, Teplyakov and co-workers showed that the canonical structure is determined not only by the loop's sequence, but also by the surrounding environment, in particular the structure of both the light and heavy chains [Teplyakov 2016].

However, canonical structures cannot be used to model H3. Yet, with its situation at the center of the binding site, it is of crucial importance that this loop be correctly described. H3 modeling still represents a major challenge in antibody structure prediction [Almagro 2014, Teplyakov 2014]. Efficient loop modeling methods are thus needed in order to solve this problem.

1.2.5.2 Loop modeling

Loop modeling aims at correctly describing the structural properties of loops. This is a problem whose difficulty mainly resides in the flexibility of these fragments, which, in many cases, do not adopt one single stable conformation. Numerous methods have been proposed to solve this problem, that is general to protein mod-

eling. Therefore, most of these methods are not antibody specific. The general problem of loop sampling, central to this thesis work, is the focus of Section 1.3.

1.2.6 The antibody-antigen complex

In this section, we present the bases of antibody-antigen binding. For a more complete description of the structural aspects of this interaction, we refer to a review by Sela-Culang and co-workers [Sela-Culang 2013].

1.2.6.1 Epitope and paratope

The *paratope* is defined as the region in the antibody that is responsible for binding the antigen. It is known to be overlapping with the CDR, at the extremity of the Fv region. Cross reactivity of the antibody suggests that the antibody does not have a single well-defined paratope, but several overlapping paratopes, depending on the antigen considered [Greally 1991, Van Regenmortel 2014].

The *epitope* is defined as the region in the antigen that binds the paratope. Contrary to the paratope, the epitope does not show any feature distinguishing it from the rest of the protein surface. In particular, it is not stickier than the rest of the protein surface. As the hypervariable loops differ greatly from one another, in length and in composition, it was suggested that the capacity of the antibodies' paratopes to bind virtually any protein surface patch was due to the combined use of the different binding propensities of the hypervariable loops [Kunik 2013].

1.2.6.2 Contacting residues

Several analyses have focused on the residues that actually interact with the antigen: the Specificity-Determining Residues (SDRs).

Padlan an co-authors [Padlan 1995] found that SDRs represent only between 25 and 37% of the CDRs in the complexes they studied and that antigen binding sometimes involves only five of the six hypervariable loops. Another interesting observation was that SDRs were mostly found in regions of high sequence variability.

MacCallum and co-workers [MacCallum 1996] found that neither Kabat's nor Chothia's definition of CDRs correctly identify contacting residues, even though Kabat's definition of the CDRs better correlated with contacting residues than Chothia's. A more recent study confirmed that CDR positions showing strong sequence conservation are usually not involved in antigen binding [Tsuchiya 2016]. The observation that previously defined CDRs did not exactly correspond to contacting residues led to another definition of CDRs called Antigen Binding Regions (ABRs). The multiple structural alignment of non redundant antibodies from the PDB led to the identification of 6 stretches of residues in which at each position, more than 10% of the antibodies bind the antigen [Kunik 2012].

Overall, the heavy chain was found to form more contacts with the antigen than the light chain [Almagro 2004, Raghunathan 2012]. Antigen binding was

found to happen at the center of the antigen-combining site or at 'very high' regions [MacCallum 1996, Tsuchiya 2016].

On the antibody's side, an interesting pattern of amino acid contributions to binding can be observed. Jackson [Jackson 1999] showed that tyrosine residues are responsible for more than a quarter of the interaction energy provided by the antibody whereas phenylalanine and lysine residues barely contribute to binding on the antibody's side. Specific contribution of the different amino acids in each hypervariable loop was analyzed in a study by Kunik and Ofran [Kunik 2013]. On the antigen's side, Jackson showed that arginines and lysines make a large contribution to binding. Jackson's analysis also underlined the fact that most of the interaction energy is contributed by side-chain/side-chain or side-chain (antibody)/ main-chain (antigen) interactions, which contrasts with the distribution of the interactions in protease-inhibitor complexes [Jackson 1999].

1.2.6.3 The influence of the antigen size

The size and class of the antigen targeted by an antibody influence several properties of the binding site. The combining site was found to adopt a rather concave or grooved topography in hapten-specific antibodies or antibodies targeting smaller antigens, while a flat topography was found for antibodies targeting proteins or larger antigens [MacCallum 1996, Almagro 2004]. The topography of the binding site was found to result from the length of the different CDR loops, in particular of L1 [Vargas-Madrazo 1995, Raghunathan 2012]. The difference in combining site topographies probably explains the differences that were observed in the propensities of CDR residues to bind the antigen for antibodies targeting haptens, peptides or proteins. Indeed, larger antigens were found to bind more apical residues than haptens.

1.2.7 Docking prediction

1.2.7.1 The general computational docking method

Docking prediction consists in identifying the relative pose of two molecular partners in a complex.

Computational docking methods usually involve two steps: a search phase during which the different poses of the two partners are investigated and a scoring phase during which the different docking structures obtained at the previous stage are ranked and the best-scoring predictions are selected. For a docking method to be successful, two major requirements have to be met [Halperin 2002] (i) the search algorithm must be fast enough while still covering enough of the conformational space to not miss near-native binding modes and (ii) the scoring function must be able to successfully discriminate between near-native poses and wrong docking predictions.

Studying results of computational docking on antibody/antigen complexes revealed that the search algorithm is quite successful at exploring the right bind-

ing modes, except in cases where large backbone conformational changes occur [Sotriffer 2000, Pedotti 2011]. However, scoring functions still struggle to distinguish between the right and wrong docking predictions [Pedotti 2011]: although the scoring function often places the docking prediction closest to the crystallographic structure among the best scoring structures, it is not precise enough to rank it as the first result.

1.2.7.2 Starting structures

The docking algorithm takes the structures of each of the two partners as input. For each partner, the starting structure may be either the structure of the unbound molecule, the structure of the molecule bound to another partner, a homology model, or even the structure of the molecule in complex with the partner with which docking is performed.

The last case is of course of limited interest from a modeling point of view, but may be used to assess how the algorithm performs with starting structures that are already in the right conformation.

The other cases represent more realistic situations. Indeed, complex formation is likely to involve conformational changes, which represents an additional difficulty for the docking procedure.

1.2.7.3 Representation of the system

The representation of the system may be the classical atomic model, allowing energy calculations with all-atom force fields, but a simplification, called *coarse-grained* representation, may also greatly speed up energy calculations and allow a more exhaustive search of configurations. For example, the ATTRACT [Zacharias 2005] docking algorithm uses a reduced representation where each amino acid is represented by 1 to 3 pseudo-atoms.

Shape complementarity between the partners is also very important in complex formation. Therefore, some docking algorithms choose to focus on the partners' surfaces, e.g. by representing the surfaces by a list of critical points representing holes and knobs [Connolly 1983]. Docking the two surfaces then consists in optimizing geometric complementarity between the two surfaces.

1.2.7.4 Rigid-body docking

Rigid-body methods assume that both complex partners are rigid, i.e. that their conformations within the complex are the same as -or sufficiently close to- their starting conformations. Using rigid-body docking considerably limits the search space, thus making a systematic grid search of the six-dimensional search space possible [Krumrine 2003]. Katchalski-Katzir *et al.* describe a grid method making use of the Fast Fourier Transform (FFT) to estimate the complementarity at the interface between the two partners [Katchalski-Katzir 1992].

1.2.7.5 Conformational flexibility in docking

When the starting structure is different from the structure that the partner will adopt upon binding, or in other words when conformational changes happen upon binding, flexibility may have to be taken into account for the docking algorithm to obtain satisfying results. This represents a considerable challenge for docking algorithms [Bonvin 2006] and advanced methods are needed to efficiently sample the conformational space in a reasonable time. The rigid docking problem considers 6 degrees of freedom (3 rotational and 3 translational) corresponding to the relative position and orientation of the two partners. Directly considering all the dihedral angles in the system as flexible would result in an explosion of the number of degrees of freedom, which would not be manageable. A choice thus has to be made about the components allowed to move.

Some methods consider the small ligand flexible and the receptor rigid, a larger molecule being less prone to conformational changes. This may be a reasonable assumption for antibody-hapten or maybe even to antibody-peptide complexes, but not reasonable for antibody-protein complexes. Other methods focus on side-chain flexibility, and consider the backbone rigid [Gray 2003]. This may constitute a good compromise [Pedotti 2011], although it is likely to fail in predicting complexes involving large conformational changes [Zacharias 2005]. Other methods integrate flexibility both in the backbone and in the side-chains [Wang 2007]. However, when backbone movements upon binding are actually limited, these methods are likely to disrupt the partners' structures and result in worse results than those neglecting backbone flexibility [Pedotti 2011]. Other rigid-body docking methods integrate flexibility either by allowing some degree of penetration between the two partners (soft docking) or by successively docking several starting structures for one or both of the partners, corresponding to structures representative of their conformational ensemble (ensemble docking).

1.2.8 Antibody design

The inverse problem of structural prediction is the problem of protein design. These two fields are closely related as they both make use of the knowledge on protein structure. Antibody design is seen as a promising field for therapeutic purposes.

Some general methods exist to predict sequences that may be compatible with a certain structure [Ponder 1987]. Some methods also offer to increase the affinity of a protein for another protein by investigating the sequence space around a given sequence [Babor 2011, Spassov 2013]. In the particular case of antibodies, OptCDR [Pantazes 2010] is a software that makes use of the knowledge about canonical structures to generate a library of antibodies that are likely to bind a given antigen.

It has been shown that rigidifying the binding site in the position favorable to binding the maturation increases the affinity of the antibody. Antibody design may thus benefit from restricting the flexibility of the CDR loops [Wong 2011]. Carefully

analyzing the energy landscape of a designed protein loop may greatly contribute to antibody design, by showing the rigidification of the loop, and by verifying that it is more likely to adopt the required conformation than any other.

Analysis of the residues that make contact with the antigen in an antibody may also be useful for antibody engineering. When humanizing antibodies for example, grafting only the residues that are responsible for contacting the antigen may limit the immunogenicity of the humanized antibody [Padlan 1995]. Nuclear Magnetic Resonance (NMR) [Cavanagh 2006, Jacobsen 2007] chemical shift has also been used to analyze which residues make contact with the antigen in an attempt to design an antibody against the dengue virus [Simonelli 2013].

1.3 Protein loop modeling

1.3.1 Motivation

Loops, together with linkers and terminal tails, are the least rigid fragments composing the secondary structure of a protein. Section 1.2.4.3 details how their flexibility is crucial for the mode of action of antibodies. Indeed, loop flexibility plays a key role in many protein-protein interaction processes, by allowing more complex modes of binding. It is also a key element in some enzymatic sites.

The vast majority of existing protein structures were obtained via X-ray crystallography [Woolfson 1997]. However, X-ray crystallography can only solve the most rigid parts of the structures, leaving out the most flexible portions. This is the reason why many "solved" protein structures within the PDB omit data for loop regions [Petoukhov 2002, Brandt 2008]. When loop information is included, it typically represents a single conformation, which does not adequately characterize the (local) structural diversity of the protein [Shehu 2006, Marks 2018].

The loop modeling process resembles the docking process, consisting of a sampling phase followed by a scoring phase. However, while docking focuses on finding *the* best pose, focusing on predicting a single loop conformation makes little sense if the loop exhibits flexibility. Nevertheless, many loop modeling methods are actually loop prediction methods, focusing on determining a single conformation. In this section, we present these sampling and scoring steps separately, although many methods have sampling and scoring phases intertwined, such as MODELLER [Fiser 2000] and DiSGro [Tang 2014]. However, intertwining these steps may create a bias towards a single stable conformation, preventing sampling of statistically likely alternative conformations.

1.3.2 The sampling phase

The addressed problem can be formulated as follows: given a protein structure with the loop region omitted, generate an ensemble of feasible loop configurations while leaving the remainder of the protein rigid. Figure 1.6 illustrates a cartoon representation of a protein. The end points of the loop, shown as red spheres, are

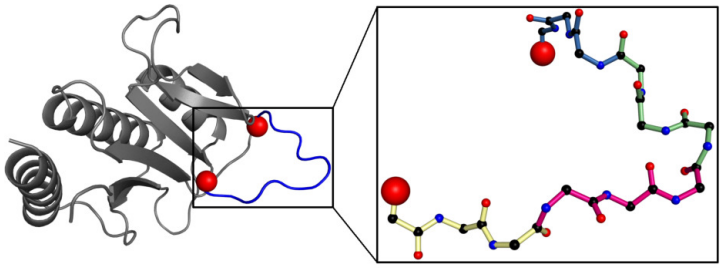


Figure 1.6. A loop region of a protein is illustrated between two stationary anchors (spheres). On the right, a more detailed picture highlights the tripeptides of the loop in alternating colors.

treated as anchors. A successfully generated loop connects these two anchors (this is known as the loop closure constraint) while satisfying a set of structural constraints (correct bond geometry, no atom overlaps, ...).

Numerous methods have been proposed over the years to address this problem [Shehu 2012, Li 2013]. Classically, these methods are classified in three main groups: *ab initio* methods, that build loops without using any previous knowledge; homology methods, that propose loop structures based on known structural data of loops with similar length and sequence; and finally hybrid methods, that employ a combination of these approaches. These three types of methods are further commented below (Sections 1.3.2.2 to 1.3.2.4).

But first, Section 1.3.2.1 will introduce loop closure methods. They are indeed required by most loop sampling methods, to generate loop conformations that are attached to the loop anchors, while maintaining the correct bond lengths and angles.

Note that Sections 1.3.2.1 to 1.3.2.4 illustrate the different methods by briefly describing a few examples. The list is, however, far from exhaustive. The interested reader is referred to references [Shehu 2012, Li 2013] for more complete reviews about loop modeling.

1.3.2.1 Loop closure methods

This section presents loop closure methods through a few representative example. The list is not exhaustive but is meant to provide For example, Random Tweak [Shenkin 1987], whose performance have first been demonstrated on CDR loops, optimizes the base geometry of the loop using a linearized Lagrange multiplier with dihedral angles as variables. Cyclic Coordinate Descent (CCD) [Canutescu 2003] uses another minimization-based approach in which the loop is open at the C terminus, and the minimization of the distance between the C terminal residue and its theoretical static position is obtained through optimization of the ϕ and ψ dihedral angles.

Robotics-inspired methods constitute another category of loop closing methods. Indeed, a loop in which the bond angles and bond lengths are fixed can be seen as a robotic manipulator arm with revolute joints. In the general case, such a loop

satisfying closure constraints actually has $n - 6$ degrees of freedom, where n is the number of dihedral angles in the loop, meaning that once $n - 6$ of these angles are fixed, at most a finite number of solutions allowing the loop to close exist for the set of the 6 remaining dihedral angles. Methods that give the solutions to these 6 dependent variables are called 6R Inverse Kinematics (IK) methods. For example, Coutsias and colleagues described a method to solve closure equations for three non-consecutive residues in the loop [Coutsias 2004]. Other methods include the work by Manocha and Canny [Manocha 1994] or Dinner [Dinner 2000].

1.3.2.2 *Ab initio* sampling methods

Ab initio sampling methods include mostly geometric methods which attempt to close the loop while respecting the constraints of loop closure and of bond lengths and angles but also seek to avoid steric clashes, which are responsible for very unfavorable van der Waals interactions.

An example is the FEM method [Shehu 2006], that generates loop conformations by perturbing the backbone dihedral angles and closing the loop with CCD.

Random Loop Generator method (RLG) [Cortés 2004] is another geometric method. This method defines a 'passive' subchain containing 6 dihedral angles in the loop and considers the rest of the loop as 'active'. Once the 'active' part is sampled, IK exact analytical solutions are calculated for the remaining 'passive' subchain. Sampling of the 'active' part takes into account distance considerations to increase the probability of IK to find a solution to the loop closure equations. Once the loop is generated, collision detection is used to filter out loops that are in steric clash either with themselves or with the rest of the protein.

LoopTK [Yao 2008] is a 'seed sampling' method that divides the loop into three fragments: the front-end (F), mid-portion (M) and back-end (B). F and B are sampled separately. Then, for each pair of compatible F and B conformations, sampling of M is performed and the IK technique by Coustias *et al.* [Coutsias 2004] is used to close M.

Liu *et al.* described a method [Liu 2009] that starts by dividing the loop into several overlapping rigid fragments. The atoms of the loop are then randomly placed in space around the terminal atoms. Loop conformations are then generated by iterative rearrangement of the atoms so that two atoms from different rigid fragments are within acceptable distance and so that rigid fragments adopt the structure of their template.

Nilmeier and co-workers [Nilmeier 2011] described another loop sampling algorithm which combines the IK method by Coutsias *et al.* [Coutsias 2004] and the MCMC exploration method from reference [Nilmeier 2009]. Instead of generating first the backbone and then placing side chains, usually with a distinct method, as is often done in loop sampling algorithms, this method uses a hierarchical method to sample both the backbone and the side-chains within one MCMC algorithm but with different amplitudes of perturbation.

1.3.2.3 Homology methods

Typically, a method consisting in using canonical structures to model the CDR loops of an antibody is a homology method. It is a widespread method for antibody modeling.

Other homology methods that are not antibody-specific are the FREAD method [Choi 2010] and its variants. FREAD is a database search method that looks for chain fragments whose properties are similar to the ones of the loop that is to be sampled. These properties can be e.g. the distance between the two ends, the sequence profile or the contact profile. These methods have been successfully applied to CDR loops [Choi 2011]. The FREAD variants can yield even better results than methods based on canonical structures.

LoopIng [Messih 2015] uses a Random Forest approach that predicts the RMSD between a candidate loop model and the target loop, using a certain number of features including sequence similarity, distance between the loop anchors or the geometry of the loop ends. The Random Forest model is trained on a dedicated loop database, also used as template database to make predictions using the trained model.

DaReUS-Loop [Karami 2018] is another homology method, which consists in mining the PDB for fragments similar to the loop flanks (i.e. the regions before and after the loop to be modeled), and filtering the candidates based on sequence similarity and conformational profile. The method then performs a scoring step to identify the 10 best models.

1.3.2.4 Hybrid methods

The Protein Local Optimization Program (PLOP) [Jacobson 2004, Zhu 2006] generates loop halves that are steric-clash free and combines halves that (i) meet in the middle (ii) with correct bond angles (iii) with correct dihedral angles according to Ramachandran plots (iv) are steric-clash free and (v) leave enough space for the side-chain of the middle residue. The loops thus generated are ranked and for each loop l among the best scoring loops, additionnal runs are performed while restricting the loop around l. Improvements of the PLOP method have been made by Zhu and co-workers [Zhu 2006] to better model longer loops. Note that this method is mainly geometric, and as such could have been classified as a *ab initio* method. But available structural data is employed through the use of Ramachandran plots to validate dihedral angles, making this a hybrid method. Sellers and co-workers [Sellers 2010] used PLOP to predict H3 loops in modeled antibodies where other hypervariable loops were predicted using canonical structures. They tested improvements over PLOP for cases in which the surrounding environment may contain errors (e.g. side-chains in non native conformations), which is very likely to happen in such cases in which other CDR loops are grafted onto a homology-modeled framework. Another study showed PLOP's ability to model H3 loops [Zhu 2013].

The updated version of random coordinate descent (RCD) [Chys 2013, López-Blanco 2016], is another mainly geometric method using structural knowl-

edge for sampling or validating dihedral angles. The original method used a filter restraining ϕ and ψ angles to Ramachandran ranges. The updated method employs neighbor-dependent probability distributions derived from a data set of protein loops to sample torsion angles. The method performs loop closure through distance minimization of the closing distance, as in CCD. However, the torsion angles that are successively used as variables to minimize the distance to the target anchor are randomly selected. Geometric filters are used to prevent steric clashes.

The loop modeling method implemented within ROSETTA [Mandell 2009, Stein 2013] is a robotics-based method that uses KIC moves. These moves consist in sampling torsion degrees according to Ramachandran probabilities, before closing the chain using IK on six ϕ and ψ dihedral angles. The latest version of the method uses loop fragments.

DiSGro [Tang 2014] is a method in which the loop is built sequentially. Sampling of ϕ and ψ angles is done by careful placement of the C_i and N_{i+1} atoms of residues i and $i+1$ in the loop, respectively. For instance, when placing C_i once all previous atoms in the chain have been placed, C_i can only be placed on a circle if bond length and bond angles are fixed. Fixing the distance between C_i and the fixed C_α atom $C_{\alpha,t}$ of the residue at the end of the loop (i.e. the residue to reach to close the loop) further restricts the position of C_i to 0, 1 or 2 positions. In the method, the distance $d_{C_iC_{\alpha,t}}$ between C_i and $C_{\alpha,t}$ is sampled from the conditional distribution $P(d_{C_iC_{\alpha,t}}|d_{C_{\alpha,i}C_{\alpha,t}})$ where $d_{C_{\alpha,i}C_{\alpha,t}}$ is the distance between $C_{\alpha,i}$ (already placed) and $C_{\alpha,t}$. This distribution is extracted from available structural data, making this a DiSGro a hybrid method. Sampling from such a distribution guarantees that there is at least one possible position for C_i and that the loop closes. An extension of this method was designed to enable the sampling of multiple interacting loops [Tang 2015].

Many hybrid loop sampling methods actually make use of fragment databases that they combine or adjust to form suitable loop conformations. Note that the structural databases that are employed in these methods are decisive. In H3 modeling for example, the 'kinked' conformation (typical of this loop) is relatively rare in other protein systems. Therefore, knowledge-based methods may have to employ databases of antibody fragments to increase the probability to sample the right fragments. This idea is also supported by results in [Choi 2011], in which the use of a database comprising non-antibody fragments greatly decreased the quality of H3 predictions using FREAD. For example, Sphinx [Marks 2017] is a hybrid loop modeling methods which combines ideas from FREAD and *ab initio* methods, including a fragment database and loop closing using a method inspired from CCD. Sphinx includes a H3-specific version, that uses a specific fragment database.

1.3.3 The scoring phase

Energy evaluation for molecular systems is presented later, in Section 1.4.2. The present section introduces alternatives that can be applied to loop fragments.

An accurate evaluation of the quality of a protein loop structural model is important for many applications. It is primarily useful in the context of loop structure prediction, to determine the stable conformation(s) that the loop is most likely to adopt. It may also be used as a filter to eliminate high-energy conformations before costly downstream steps, for instance in ensemble docking [Amaro 2018]. Loop structure evaluation is furthermore employed in loop design, either to verify that the proposed loop will preferentially adopt the desired conformation [Kundert 2019], or, on the contrary, that it will not adopt an undesirable one (negative design) [Jin 2003, Hu 2007, Koga 2012].

The growing need to quickly and accurately assess loop structures has led to the development of various scoring functions [Ponder 2003, Yang 2008, Rata 2010, Dong 2013, Maier 2015, Alford 2017, Karasikov 2018, López-Blanco 2019], that greatly differ both from a conceptual and from a computational point of view. The different approaches can be divided into three groups: *physics-based*, *knowledge-based* and *hybrid* methods.

Atomistic physics-based methods aim at estimating an energy based on the calculation of forces involved in the structure [Ponder 2003, Vanommeslaeghe 2010, Maier 2015]. Such calculations include bonded energy terms (such as the energy between bonded atoms, energy following bond twisting, ...) and nonbonded energy terms, such as those related to van der Waals interactions (usually through the Lennard-Jones potential) or electrostatic interactions. The weights of the different terms are optimized to fit experimental data. Although physics-based methods are relatively general, they must be adapted to the molecule whose energy is to be estimated. Depending on the type of molecular system, the terms included in the calculation or their associated weights may vary. Physics-based methods are, in general, computationally demanding and highly sensitive to slight conformational changes.

Conversely, knowledge-based (mostly statistics-based) methods [Yang 2008, Rata 2010, Dong 2013, Karasikov 2018, López-Blanco 2019] are a tempting options due to their usually lower computational cost. These methods exploit available experimental data to assess the quality of a given structure. Naturally, the performance of these methods is highly related to the data from which they derive. Just like for most knowledge-based methods in any field, a common worry is that these methods may not be able to recognize unusual folds. In addition, since most of the published data come from X-ray crystallography, which captures only a snapshot of a protein's structure, the behavior of knowledge-based methods is unpredictable when dealing with a very flexible protein or region. Performance of these methods may also suffer from potential artifacts in the data resulting from crystal packing interactions.

The last category of scoring methods contain hybrid approaches. These techniques, such as ROSETTA scoring functions [Alford 2017], combine statistical terms with physics-based terms to assess the quality of a structural decoy.

A description of different loop scoring methods is provided in Section 4.1.3.

1.3.4 Loop modeling in this thesis

Loop modeling is an important part of this thesis work. Chapter 3 focuses on a novel loop sampling methods aimed at exhaustively exploring the conformational space accessible to a loop. Chapter 4 then investigates the ability of several loop scoring methods to correctly assess the quality of sampled conformations.

1.4 Exploring energy landscapes

Energy landscape exploration is an important problem in molecular modeling, for which numerous methods exist. This section only provides a brief description. For more details about molecular modeling and landscape exploration, the interested reader is referred to several textbooks [Leach 2001, Frenkel 2002, Wales 2003].

1.4.1 Definition

The potential energy landscape (PEL) of a protein consists of the potential energy value of the protein at each point of its accessible conformational space [Wales 2003, Zuckerman 2010]. The conformational entropy is also an important component to explain the structure and dynamics of a given molecule that is implicitly contained in the landscape. Given the high dimensionality of the conformational space, visualizing the energy landscape is a delicate task. For this reason, visualization is often performed through a projection in 2 dimensions of the protein conformation. The energy landscape is thus represented as a surface in this space. In such a representation, the depth of the surface would give the potential energy, and the conformational entropy would correspond to the width of basins.

It is assumed that molecular systems are in their stable conformation when their Gibb's free energy is at a minimum [Anfinsen 1973, McQuarrie 1999]. Therefore, energy considerations are very important in problems such as protein folding, structure prediction and complex formation. The fact that proteins could fold quickly and consistently reach the conformation corresponding to their free energy minimum, whatever the initial conditions were, was seen as a paradox called Levinthal's paradox. Later theories suggested that the folding energy landscape actually has a funnel-like shape [Dill 1997, Onuchic 1997, Onuchic 2004] (Figure 1.7(a)). The existence of such a landscape would explain the observed folding properties of proteins.

More precisely, proteins actually adopt an ensemble of conformations, with probabilities related to the energy of these conformations. The equilibrium thus reached is called Boltzmann equilibrium, with low-energy conformations more likely than higher-energy ones. In a perfect funnel-like energy landscape, the predominant conformation is obviously the conformation corresponding to the energy minimum. However, energy landscapes are rougher in reality, with several coexisting conformations for a single protein [Dill 1997] (Figure 1.7(b)).

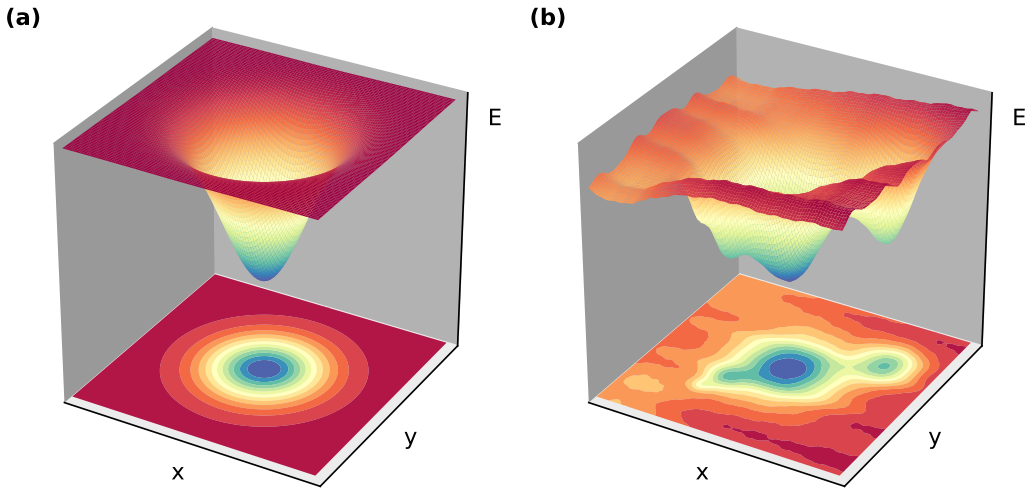


Figure 1.7. Energy landscape projected in 2D. (a) Perfect funnel-like landscape, with one unique stable conformation. (b) Realistic landscape, with several meta-stable conformations.

The exhaustive exploration of the energy landscape or of the structural space is infeasible in practice, especially for large molecular systems. Advanced methods thus have to be employed for the efficient exploration of this search space when dealing with problems such as conformational changes (e.g. in protein loops) and docking. A review of possible exploration methods of molecular energy landscapes or dynamics can be found in reference [Maximova 2016].

Energy landscape exploration requires three essential elements: (1) accurate energy evaluation, (2) efficient exploration, (3) visualization. In the following, Section 1.4.2 describes the different methods for energy evaluation, Section 1.4.3 presents methods for landscape exploration, and Section 1.4.4 gives options for landscape visualization.

1.4.2 Energy estimation of a given conformation

Exact energy computation is infeasible for large and complex systems, and approximations thus have to be made. Methods to assess the energy of whole molecular systems are similar to the methods used to score loop fragments that were presented in Section 1.3.3.

Most approaches make use of molecular force fields [Pillardy 2001, Herges 2004] (originally developed for molecular dynamics), that focus on potential energy and usually leave out entropic considerations. However, some do include solvation effects [Verma 2009]. Different molecular force fields exist, depending on the nature and representation of the molecular system. All-atom force fields are based on a full representation of the system while coarse-grained force fields use a simplified representation that can focus on the most important properties of the system at hand [Clementi 2008].

For some problems, reducing the free energy to its enthalpic contribution may be an overly simple approximation. For instance, the change in conformational entropy upon binding has been shown to be an important part of the binding free energy [Chang 2007].

Just like for loop systems (Section 1.3.3), other estimation methods include statistics-based potentials, whose idea is to provide a score to a given state using data from known structures.

1.4.3 Exploration methods

1.4.3.1 Exploration using continuous physics-based methods

To explore the energy landscape accessible to a molecule, and more specifically to a protein, a possibility would be to solve Newton's equations of motions. This is the idea behind Molecular Dynamics (MD) [McCammon 1977, Brooks 1983, Rapaport 2007]. Although very accurate, this method is limited by computational power to simulations of a few nanoseconds to a few microseconds. By modeling the actual movements of a molecule, MD simulations allow the calculation of thermodynamic properties using results from the field of statistical mechanics, provided they are performed on sufficient time scales.

Metadynamics [Laio 2002, Iannuzzi 2003] is a method derived from molecular dynamics which employs a time-dependent potential. This method allows for a faster exploration, while still enabling the calculation of the free energy of the system.

Replica Exchange Molecular Dynamics (REMD) [Sugita 1999] is another widely used variant of MD. It consists in running several simulations of the system in parallel, at different temperatures, and regularly attempting to exchange conformations between the different simulations. Exchanges between the replicas are accepted based on a stochastic test related to the Boltzmann probabilities of the configurations at the given temperatures.

1.4.3.2 Discrete methods, Monte Carlo and variants

Discrete methods emerged as an alternative to molecular dynamics, aiming at a faster exploration of energy landscapes.

Monte Carlo methods circumvent the problem of the long simulation times required in molecular dynamics by using stochastic processes to sample within the search space [Landau 2005]. In particular, Markov Chain Monte Carlo (MCMC) methods consist in a random walk in the search space. For example the Metropolis algorithm [Metropolis 1953, Hastings 1970] makes a trial move at each step, which is accepted according to the so-called Metropolis criterion related to the probability distribution, the idea being that if the move leads to a more likely (i.e. more energetically favorable) state it is always accepted while if it leads to a less likely state, the move is accepted according to a carefully chosen probability distribution. The

condition of 'detailed balanced' ensures that the sampled states follow the theoretical distribution after a certain simulation time. The problem of these methods is that they may get trapped in local minima when there are large energetic barriers to cross.

The POSH method [Nilmeier 2009] is an example of a Metropolis MCMC method aimed at exploring very rugged landscapes where local minima are sparse.

1.4.3.3 Robotics-inspired methods

The field of robotics uses motion planning algorithms to plan the trajectory of a robot system in a complex space with obstacles. Molecules, and in particular proteins, can be regarded as articulated mechanisms and many motion-planning algorithms can therefore be adapted to molecular-related problems [Al-Bluwi 2012]. Such an adaptation is necessary because contrary to robotic mechanisms, for which a configuration can be either allowed or forbidden, proteins are in more or less favorable conformations and there is no clear-cut definition of an accessible state.

In order to be applied to molecular simulation problems, these algorithms must therefore integrate the notion of energy in some way. For instance, the PDST algorithm was adapted to predict the relative movements of secondary structure elements that are necessary to go from an initial protein conformation to a final conformation [Haspel 2010]. Jaillet and co-workers introduced a variant of the RRT algorithm (transition-RRT) [Jaillet 2011] to explore the energy landscapes of molecules and find stable states and transitions between these states. Other variants of RRT can also be used for molecular simulations [Devaurs 2014]. Another tree-based method using fragment replacement has been used to predict transitions paths between conformations [Molloy 2013]. Motion-planning algorithms can also be used to predict protein loop motions [Cortés 2005, Paës 2012].

1.4.4 Landscape visualization

The high dimensionality of the energy landscape makes it hard to represent, or to describe. A possible way to visualize it is to use a dimensionality reduction approach such as PCA to plot the energy as a function of 2 coordinates [Jolliffe 2002, Mu 2005, Altis 2007]. Although this enables an easy visualization of the landscape, it does not provide an exhaustive description and the dimensionality reduction is likely to hide interesting features such as critical points. Therefore, other approaches have emerged that build graphs from critical points in the landscapes. For example, transition or disconnectivity graphs [Becker 1997, Cazals 2015] provide informative descriptions of the landscape.

Chapter 2

Flexible loops in antibodies

Contents

2.1 Introduction

2.1.1 Context and objective

This chapter analyzes from a structural point of view the flexible loops that ensure the recognition of the antigen by an antibody. These CDR loops, described in the previous chapter (Section 1.2.3) are determinant for antibody specificity and affinity. More precisely, these two crucial features of antibody-antigen recognition are allowed by the complex flexibility of CDR fragments (Section 1.2.4). This conformational variability poses a major problem to antibody-antigen docking pose prediction, that *a priori* ignores the conformations adopted by the loops in the bound conformation of the antibody (Section 1.2.7).

Conformational changes are often analyzed on a case by case basis, when the structures of both the complexed and unbound forms of an antibody become available. Although Sela-Culang and co-workers [Sela-Culang 2012] analyzed conformational changes upon antibody-antigen binding in a large dataset of 49 different antibodies, their analysis focused on significant conformational changes over all antibodies. The work presented in this chapter, however, focuses on the nature of the different movements that can be observed, even when those concern only a few antibodies in our dataset. The prime goal of our analysis is to gather quantitative information about conformational changes for a relatively large number of cases, providing a basis for estimating the expected amplitude of the different possible rearrangements and for defining the overall range of conformational changes.

The structures of bound and unbound antibodies from the protein-protein docking benchmark version 5.0 [Vreven 2015] were compared using root-mean-square deviations (RMSDs) of atom positions and of dihedral angles, calculated for different subparts of the Fab fragments. Analyzing these deviation measures, we aim to better characterize conformational changes undergone by antibody structures upon binding. Vreven and co-workers [Vreven 2015] also tested four docking algorithms on the newly added cases of their benchmark. Their results are further analyzed here, in combination with our measures, in order to understand the effects of the different types of conformational changes on the performance of current docking prediction algorithms.

The work presented in this chapter was published in [Barozet 2018].

2.1.2 Dataset employed in this study

The recently updated protein-protein docking benchmark version 5.0 [Vreven 2015] provides structures for both complexed and free proteins. This dataset contains 40 antibody-antigen systems, 12 of which do not contain a structure for the unbound antibody. One of the remaining cases, PDB entry 2I25 (New antigen Receptor PBLA8), a shark single-domain (IgNAR) antibody, was excluded from this study due to its major structural differences with 'conventional' antibodies.

The structural files for the remaining 27 antibodies (Table 2.1) were renumbered with the Martin (enhanced Chothia) scheme, using the ANARCI [Dunbar 2016] software and custom scripts. These files do not contain hydrogen atoms. RMSD calculations as well as alignments were therefore performed on heavy atoms only. Table 2.2 gives the lengths and sequences of the CDRs in those 27 antibodies.

2.2 The limits of canonical structures

2.2.1 Bound or unbound antibody structure?

Canonical structures (see Section 1.2.3) are known to provide very accurate predictions for the structures of H1, H2, L1, L2 and L3 CDR loops, given only their length and sequence. However, in articles defining canonical classes [Chothia 1987,

Table 2.1. Antibodies in this study. The docking score is defined in Section 2.4.1. BKB=backbone, SC=side-chain, conf.=conformational. [1] Elbow angle variation (°).

Complex/Free PDBID (resolution (Å))	ΔElb.[1]	Docking		Conformational changes – Observations
		Score	Rank	
1AHW/1FGN (3/2.5)	2.9		NA	Small BKB changes and shifts, internal changes in some SCs.
1BGX/1AY1 (2.3/2.2)	9.7		NA	Large shifts (all CDRs). Some BKB internal conf. changes (esp. H3, L1, L2). SCs reorganized in all CDRs.
1BVK/1BVL (2.7/2.87)	NA		NA	Large shift of H2's BKB. Internal BKB changes in L3, which remains globally in place. SC rearrangements in L1, L3 and H3, located at the center of the binding site.
1DQJ/1DQQ (2/1.8)	0.9		NA	Small BKB shifts with SC rearrangements of H2 and H3. Slight internal changes in H1.
1E6J/1E6O (3/1.8)	0.9		NA	Loops shift slightly, no major internal variation. FR does not align well. Some SC rearrangements in H3.
1JPS/1JPT (1.85/1.85)	2.8		NA	Only a few SCs conf. changes.
1MLC/1MLB (2.5/2.1)	5		NA	H2 changes conformation and shifts. The rest is stable overall. Some rearrangements in position and orientation for SCs in L1, and a few rotamer changes in L2.
1VFB/1VFA (1.8/1.8)	NA		NA	Slight changes at the tip of the H3 loop, where some SCs change conformations.
1WEJ/1QBL (1.8/2.26)	6.7		NA	Only SCs conf. changes in L1 and H2.
2FD6/2FAT (1.9/1.77)	5.1		NA	Internal conf. changes with SC rearrangements for H2. No global move.
2VIS/1GIG (3.25/2.3)	48.6		NA	BKB conf. changes at the tip of H3 and in L3, and in other loops, but minor. SCs rearrangements, especially where BKB changes conformations.
2VXT/2VXU (1.49/2.36)	13.1	0.17	4	Small internal conf. change for H3 and for some SCs in L1.
2W9E/2W9D (2.9/1.57)	2	0.0049	9	Some BKB conf. changes for H1 and H2, along with associated SCs global and internal rearrangements. No global move. A few SCs rearrangements.
3EO1/3EO0 (3.1/1.75)	17.2	0.0024	11	BKB internal conf. changes in all loops, slight for some. Only H1 moves globally, with rearranged SCs. H2 and L1 SCs also move following the BKB movement. SCs in L2 change their internal conformation.
3EOA/3EO9 (2.8/1.8)	49.1	0.0002	14	Small internal BKB changes without global movement at the tip of the H3 loop. The SCs in L1 and in the region of the tip of H3 are rearranged.
3G6D/3G6A (3.2/2.1)	67.9	0.0015	12	Slight shifts everywhere. Some minor conf. changes in H3 with some SCs moves and conf. changes.
3HI6/3HI5 (2.3/2.5)	22.8	0.0012	13	Large conf. changes in H3 BKB, with a large move. SCs in and surrounding H3 completely rearranged.
3HMX/3HMW (3/3)	29.4	0.25	3	Some limited conf. changes for H2. Small shifts of BKBs, especially for loops in the heavy chain. Some SCs in H2 and L2 change conformations.
3L5W/3L7E (2/2.5)	9.3	0.011	8	Small BKB shifts (all loops), especially in the heavy chain. Some SCs movements at the tip of H3.
3MXW/3MXV (1.83/1.9)	8.4	0.4	1	All 3 heavy chain hypervariable loops are slightly shifted but their conformations remain stable. Some SC rearrangements at the tip of H3.
3RVW/3RVT (1.95/2.05)	2.5	0.0024	10	Heavy chain hypervariable loops are slightly shifted, but all loops keep their overall conformations. A few SCs rearrangements, especially in H3 and L2.
3V6Z/3V6F (3.34/2.52)	2.1	0	16	Large conf. changes in the BKB and in the SCs of H3 and L3, that both move. L1 shifts. Some rearrangements in the tip of L1 and in L2.
4DN4/4DN3 (2.8/2.6)	28.8	0.049	6	Some limited internal BKB changes for H1 and L1, with a shift for H1. H2 shifts as well while maintaining its internal structure. SCs are rearranged everywhere except in H3.
4FQI/4FQH (1.71/2.05)	26.8	0.024	7	Move of H2 loop, along with small internal changes. H3 is shifted as well, but only a subpart. conf. changes for the SCs in H3 and H2, and also for SCs in L2. Some other SC rearrangements.
4G6J/4G5Z(2.03/1.83)	25.5	0.13	5	Stable BKB conformation of all loops. Some minor changes but the overall conformations are maintained. H2 shows some SC rearrangements.
4G6M/4G6K (1.81/1.9)	2.2	0.34	2	The BKB of H1 shifts slightly.
4GXU/4GXV (3.29/1.45)	26	0	15	Some limited BKB changes for L3 and the tip of H3. Loops very slightly shifted. In H3 and L3, SCs move with the BKB but keep their internal conformation. Some changes in conformations for L2 SCs.

Table 2.2. Lengths and sequences of CDR loops as defined using enhanced Chothia numbering scheme. Len.=Length. [1] PDBIDs are given as COMPLEX/FREE.

PDBID[1]	Name	H1		H2		H3	
		Len.	Sequence	Len.	Sequence	Len.	Sequence
1AHW/1FGN	5G9-Fab	7	GFNIKDY	6	DPENGN	8	DNSYYFDY
1BGX/1AY1	TP7-Fab	8	GYSITSDY	5	TYSGT	10	YYYGYWYFDV
1BVK/1BVL	HuLys-Fv	7	GFSLTGY	5	WGDGN	8	ERDYRLDY
1DQJ/1DQQ	HyHEL-63-Fab	7	GDSVTSD	5	SYSGS	5	WGGDV
1E6J/1E6O	13B5-Fab	7	GYTFTSY	6	NPSSGY	11	PVVRLGYNFDY
1JPS/1JPT	D3h44-Fab	7	GFNIKEY	6	DPEQGN	8	DTAAYFDY
1MLC/1MLB	D44.1-Fab	7	GYTFSTY	6	LPGSGS	7	GDGNYGY
1VFB/1VFA	D1.3-Fv	7	GFSLTGY	5	WGDGN	8	ERDYRLDY
1WEJ/1QBL	E8-Fab	7	GFNIKDT	6	DPASGN	8	YDYGNFDY
2FD6/2FAT	ATN615-Fab	7	GYSFTNF	6	FHGSDN	9	WGPHWYFDV
2VIS/1GIG	HC19-Fab	7	GFLLISN	5	WAGGN	14	DFYDYDVFYYAMDY
2VXT/2VXU	125-2H-Fab	7	GYSFTDY	6	DPYNGD	4	GLRF
2W9E/2W9D	ICSM18-Fab	7	RNTFTDY	6	YPNNGV	7	YYYDVSY
3EO1/3EO0	GC-1008-Fab	7	GYTFSSN	6	IPIVDI	11	TLGLVLDAMDY
3EOA/3EO9	Efalizumab-Fab	7	GYSFTGH	6	HPSDSE	12	GIYFYGTTYFDY
3G6D/3G6A	CNTO607-Fab	7	GFTFNSY	6	AYDSSN	13	GLGAFHWDMQPDY
3HI6/3HI5	AL-57-Fab	7	GFTFSRY	6	WPSGGN	11	SYDFWSNAFDI
3HMX/3HMW	Ustekinumab-Fab	7	GYSFTTY	6	SPVDSD	10	RRPGQGYFDF
3L5W/3L7E	ch836-Fab	9	GFSLSTYGM	5	WWDDV	11	MGSDYDVWFDY
3MXW/3MXV	ch5E1-Fab	7	GYTFIDE	6	RPYSGE	10	DWERGDFFDY
3RVW/3RVT	4C1-Fab	8	GYSITSDY	5	SYSGT	12	TGVYRYPERAPY
3V6Z/3V6F	e6-Fab	7	GFTFSSY	6	SSGGNY	14	EGAYSGSSSYPMDY
4DN4/4DN3	CNTO888-Fab	7	GGTFSSY	6	IPIFGT	10	YDGIYGELDF
4FQI/4FQH	CR9114-Fab	7	GGTSNNY	6	SPIFGS	12	HGNYYYYSGMDV
4G6J/4G5Z	Canakinumab-Fab	7	GFTFSVY	6	WYDGDN	9	DLRTGPFDY
4G6M/4G6K	Gevokizumab-Fab	9	GFSLSTSGM	5	WWDGD	10	NRYDPPWFVD
4GXU/4GXV	1F1-Fab	7	GFTFSSY	6	SYDGRN	17	ELLMDYYDHIGYSPGPT

PDBID[1]	L1		L2		L3	
	Len.	Sequence	Len.	Sequence	Len.	Sequence
1AHW/1FGN	11	KASQDIRKYLN	7	YATSLAD	9	LQHGESPYT
1BGX/1AY1	10	SASSSVSYMY	7	DSTNLAS	9	QQWSTYPLT
1BVK/1BVL	11	RASGNIHNYLA	7	YTTTLAD	9	QHFWSTPRT
1DQJ/1DQQ	11	RASQSISNNLH	7	YASQSIS	9	QQSNSWPYT
1E6J/1E6O	10	SASSSVSYMH	7	EISKLAS	8	QQWNYPFT
1JPS/1JPT	11	RASRDIKSYLN	7	YATSLAE	9	LQHGESPWT
1MLC/1MLB	11	RASQSISNNLH	7	YVSQSSS	9	QQSNSWPRT
1VFB/1VFA	11	RASGNIHNYLA	7	YTTTLAD	9	QHFWSTPRT
1WEJ/1QBL	11	RASGNIHNYLA	7	NAKTLAD	9	QHFWSTPWT
2FD6/2FAT	10	SASSSVSYMH	7	EISKLAS	8	QQWNYPFT
2VIS/1GIG	14	RSSTGAVTTSNYAN	7	GTNNRAP	9	ALWYSNHWV
2VXT/2VXU	11	RASQDIGSKLY	7	ATSSLDS	9	LQYASSPYT
2W9E/2W9D	10	SASSSVSYMH	7	DTSKLAS	9	HQWRSNPYT
3EO1/3EO0	12	RASQSLGSSYLA	7	GASSRAP	9	QQYADSPIT
3EOA/3EO9	11	RASKTISKYLA	7	SGSTLQS	9	QQHNEYPLT
3G6D/3G6A	11	SGDNIGGTFVS	7	DDNDRPS	10	GTWDMVTNNV
3HI6/3HI5	11	RASQSIGSYLN	7	AASSLQS	8	QQSYSTPS
3HMX/3HMW	11	RASQGISSWLA	7	AASSLQS	9	QQYNIYPYT
3L5W/3L7E	11	RASKSISKYLA	7	SGSTLQS	9	QQHNEYPYT
3MXW/3MXV	11	KASQSVSNDLT	7	YASNRYT	9	QQDYGSPPT
3RVW/3RVT	11	KASQDIYSYLS	7	RANRLIT	9	LQYDEFPYT
3V6Z/3V6F	17	KSSQSVLYSSNQKNYLA	7	WASTRES	10	HQYLSSYMYT
4DN4/4DN3	12	RASQSVSDAYLA	7	DASSRAT	10	HQYIQLHSFT
4FQI/4FQH	13	SGSDSNIGRRSVN	7	SNDQRPS	11	AAWDDSLKGAV
4G6J/4G5Z	11	RASQSIGSSLH	7	YASQSFS	9	HQSSSLPFT
4G6M/4G6K	11	RASQDISNYLS	7	YTSKLHS	9	LQGKMLPWT
4GXU/4GXV	13	SGSSSNIGSYTVN	7	SLNQRPS	12	AAWDDSLSAHVV

Chothia 1989, Martin 1996], a mix of bound and unbound antibody structures are used to make the classifications. Therefore, it is unclear how canonical structures compare with the crystal structures in the case of large conformational changes.

2.2.2 When canonical structures cannot be assigned

We used Martin lab's tool [Dr. Andrew C.R. Martin's Group at UCL 1995] to assign canonical structure to the 5 loops (L1, L2, L3, H1 and H2) in all of the 27 antibodies in our dataset (135 loops were thus classified), and then measured Cartesian and dihedral backbone RMSD to the class representative, after alignment of the framework residues of V_H for H1 and H2, and alignment of the framework residues of V_L for L1, L2 and L3.

Martin's lab tool uses 3 different method to assign canonical class. Out of the 135 loops to classify, 15 could not be classified by any of the three methods, and 1 loop was ambiguously classified. Overall, 10 over 27 antibodies have at least one unclassified loop, and 1 has an ambiguously classified loop, which was considered unclassified for the rest of the analysis. This means that in more than one third of the cases in our dataset, we cannot have a model for all five loops using canonical structures. This constitutes an issue for antibody modeling since alternative methods have to be used for loop prediction in these cases.

The results show that unbound structures are overall better predicted than bound structures. Indeed, out of the 120 classified loops, 71 have a higher bound than unbound Cartesian RMSD to the class representative (63 when using ϕ/ψ angular RMSD, 61 when using $\omega/\phi/\psi$ angular RMSD). The mean of RMSD of unbound loop to representative is lower than the mean of RMSD of bound loop to representative for all three metrics (Cartesian RMSD, ϕ/ψ angular RMSD, and $\omega/\phi/\psi$ angular RMSD).

Moreover, defining a rather lenient threshold of 1.5 Å to define a good CDR loop model, there are 33 out of the 240 modeled loops that are not correctly predicted, coming from 14 different antibodies. More than half of the antibodies in our dataset have at least one not correctly predicted loop, either in their bound or their unbound conformation (or both). Looking only at bound antibodies, 11 have at least one not correctly predicted loop.

Looking at the 11 non-H3 hypervariable loops whose backbone undergo a displacement of more that 1.5 Å Cartesian RMSD upon binding, 3 could not be assigned a canonical class (2W9E-H1, 3EO1-H1, 3V6Z-L3), 3 have not correctly predicted bound structures using canonical class representatives (1BGX-H1, 1BGX-L2, 1MLC-H2), 1 has a not correctly predicted unbound structure (1BVK-H2), 3 have not correctly predicted bound and unbound conformations (2W9E-H2, 4DN4-H1, 4FQI-H2), and only 1 (4DN4-H2) has relatively correct predictions (1.3 Å RMSD for the bound conformation, and 1.1 Å for the unbound one).

These results confirm that canonical classes give remarkably good results for hypervariable loop structure prediction. However, they also point to a few limitations. First, for a third of the antibodies in our dataset, at least one of the five loops

cannot be assigned a canonical class. This indicates that alternative methods are still required in a non-negligible number of cases. Second, although the prediction is accurate in most cases, they are still some incorrect predictions. In particular, loops displaying large movements are mostly incorrectly predicted and since canonical structures do not predict bound conformations better, we cannot rely on them to predict conformational changes upon binding.

2.2.3 Flexibility in CDRs

The next section (Section 2.3) provides a detailed analysis of the flexibility of antibodies in our dataset, and in particular of their CDR loops. Results show that these loops can be highly flexible, further denying the existence of a single canonical structure.

2.3 Detailed analysis of flexibility in antibodies

2.3.1 Methods

For each antibody, the Fv fragments of the bound and unbound domains were structurally aligned on all heavy atoms using the 'align' method from PyMOL [Schrödinger, LLC 2015], with default parameters. It may be argued that FR alignment would provide a better alignment of structures, since CDRs are known to be more flexible. However, aligning on Fv provides an equal treatment of FR and CDRs, allowing the comparison of FR and CDRs RMSDs. In addition, PyMOL 'align' method with default parameters contains 5 cycles of outlier rejection. Therefore, CDR loops displaying large movements should not influence the alignment of Fv domains. Comparison of FR Cartesian RMSD after FR or Fv alignment confirmed that aligning on Fv rather than FR provides the same quality of alignment on FR in practice (Table 2.3).

Custom python scripts and PyMOL were then used to calculate RMSDs of atom positions between the bound and unbound structures. Initially, the RMSDs were calculated on backbone atoms (C, C_α, N and O) for different subparts of the antibody: FRs, H1, H2, H3, L1, L2 and L3, while keeping the Fv fragments aligned between the bound and unbound structures. This RMSD was meant as a measure of whole loop movements.

Then, for each CDR, a new alignment was performed between the bound and unbound loops' backbones and Cartesian RMSD of side-chain atoms only was calculated. The same operation was done for FRs (the alignment was performed on backbone atoms of residues within FRs). These RMSD measures were meant as a measure of whole side-chain movements.

Angular RMSDs were also calculated for each subpart of each antibody, on backbone dihedral angles (ϕ, ψ and ω) [Brändén 1999] on the one hand, and on side-chain dihedrals on the other. These values were meant as a measure of internal conformational change, for the backbone and for the side-chains.

Table 2.3. Comparison of alignments on Fv or FR on FR all-atom Cartesian RMSD. The alignments were performed using PyMOL "align" method with default parameters (5 cycles of outlier rejection). The alignments do not show substantial differences: the differences in RMSD are below the precision level that can be expected from the measure of atomic positions. The difference in alignments is therefore negligible.

PDB	All-atom FR RMSD (Å)		Variation (Å)	Absolute variation (Å)
	After FR alignment	After Fv alignement		
1AHW	0.79715	0.79660	-5.50E-04	5.50E-04
1BGX	1.34271	1.34351	7.98E-04	7.98E-04
1BVK	1.31529	1.31789	2.59E-03	2.59E-03
1DQJ	0.99034	0.99185	1.52E-03	1.52E-03
1E6J	1.19938	1.20381	4.43E-03	4.43E-03
1JPS	0.82862	0.82899	3.75E-04	3.75E-04
1MLC	1.02216	1.02450	2.33E-03	2.33E-03
1VFB	0.95838	0.95918	8.06E-04	8.06E-04
1WEJ	0.80961	0.81012	5.02E-04	5.02E-04
2FD6	1.10940	1.10769	-1.71E-03	1.71E-03
2VIS	1.09164	1.09842	6.78E-03	6.78E-03
2VXT	1.34265	1.34376	1.11E-03	1.11E-03
2W9E	0.95412	0.95695	2.83E-03	2.83E-03
3EO1	1.05729	1.05803	7.44E-04	7.44E-04
3EOA	1.06013	1.06054	4.11E-04	4.11E-04
3G6D	0.63604	0.63751	1.47E-03	1.47E-03
3HI6	0.98537	0.98769	2.31E-03	2.31E-03
3HMX	1.01561	1.02857	1.30E-02	1.30E-02
3L5W	1.05444	1.05427	-1.71E-04	1.71E-04
3MXW	0.92467	0.92543	7.52E-04	7.52E-04
3RVW	0.82847	0.82700	-1.47E-03	1.47E-03
3V6Z	0.50735	0.50718	-1.77E-04	1.77E-04
4DN4	0.91967	0.92114	1.47E-03	1.47E-03
4FQI	0.86853	0.87946	1.09E-02	1.09E-02
4G6J	1.01086	1.01155	6.88E-04	6.88E-04
4G6M	0.71789	0.71548	-2.41E-03	2.41E-03
4GXU	0.69689	0.69807	1.18E-03	1.18E-03
Min	0.50735	0.50718	-2.41E-03	1.71E-04
Max	1.34271	1.34376	1.30E-02	1.30E-02
Mean	0.96462	0.96649	1.87E-03	2.35E-03
Median	0.98537	0.98769	8.06E-04	1.47E-03

Table 2.4. Dihedrals used per residue type for side-chain angular RMSD calculations.

	Dihedral 1	Dihedral 2	Dihedral 3	Dihedral 4	Dihedral 5
Ala	-	-	-	-	-
Arg	N-C_α-C_β-C_γ	C_α-C_β-C_γ-C_δ	C_β-C_γ-C_δ-N_ε	C_γ-C_δ-N_ε-C_ζ	C_δ-N_ε-C_ζ-$N_{\eta 1}$
Asn	N-C_α-C_β-C_γ	C_α-C_β-C_γ-$O_{\delta 1}$	-	-	-
Asp	N-C_α-C_β-C_γ	C_α-C_β-C_γ-$O_{\delta 1}$	-	-	-
Cys	N-C_α-C_β-S_γ	-	-	-	-
Glu	N-C_α-C_β-C_γ	C_α-C_β-C_γ-C_δ	C_β-C_γ-C_δ-$O_{\varepsilon 1}$	-	-
Gln	N-C_α-C_β-C_γ	C_α-C_β-C_γ-C_δ	C_β-C_γ-C_δ-$O_{\varepsilon 1}$	-	-
Gly	-	-	-	-	-
His	N-C_α-C_β-C_γ	C_α-C_β-C_γ-$C_{\delta 2}$	-	-	-
Ile	N-C_α-C_β-$C_{\gamma 1}$	C_α-C_β-$C_{\gamma 1}$-$C_{\delta 1}$	-	-	-
Leu	N-C_α-C_β-C_γ	C_α-C_β-C_γ-$C_{\delta 1}$	-	-	-
Lys	N-C_α-C_β-C_γ	C_α-C_β-C_γ-C_δ	C_β-C_γ-C_δ-C_ε	C_γ-C_δ-C_ε-N_ζ	-
Met	N-C_α-C_β-C_γ	C_α-C_β-C_γ-S_δ	C_β-C_γ-S_δ-C_ε	-	-
Phe	N-C_α-C_β-C_γ	C_α-C_β-C_γ-$C_{\delta 1}$	-	-	-
Pro	-	-	-	-	-
Ser	N-C_α-C_β-O_γ	-	-	-	-
Thr	N-C_α-C_β-$O_{\gamma 1}$	-	-	-	-
Trp	N-C_α-C_β-C_γ	C_α-C_β-C_γ-$C_{\delta 1}$	-	-	-
Tyr	N-C_α-C_β-C_γ	C_α-C_β-C_γ-$C_{\delta 1}$	-	-	-
Val	N-C_α-C_β-$C_{\gamma 1}$	-	-	-	-

Table 2.4 gives the list of dihedral angles involved in the calculation of side-chain angular RMSD per residue type. The distance between two angles was taken as the shortest distance on the trigonometric circle. While the definition of RMSD on backbone dihedrals is straightforward, the RMSD on side-chain dihedrals demands some clarification. In this study, the side-chain angular RMSD of a subpart is defined as the average of its side-chains' individual dihedral RMSDs. i.e., with $d(\alpha, \beta)$ the angular distance between angles α and β, R_X the side-chain angular RMSD of subpart X, N_X the number of side-chains in X, $n_{X,i}$ the number of dihedrals of the i-th side-chain of subpart X, $\chi_{X,i,j,c}$ the j-th dihedral of the i-th side-chain of subpart X in conformation c ($c = u$ for the unbound conformation, $c = b$ for the bound conformation), we have:

$$R_X = \frac{1}{N_X} \sum_{i=1}^{N_X} \left(\frac{\sum_{j=1}^{n_{X,i}} d\left(\chi_{X,i,j,u}, \chi_{X,i,j,b}\right)^2}{n_{X,i}} \right)^{1/2}$$

Note that glycines, alanines and prolines were excluded from the calculation, and that the deviation of the last dihedral of arginine, phenylalanine, tyrosine, aspartic acid and glutamic acid residues were taken to be between 0 and 90 degrees due to the invariability of these side-chains following a 180° rotation of their last dihedral. To perform those calculations, an in-house C++ program was used.

Except for 1BVK and 1VFB, which only contain the Fv fragment of the antibody, the variations of elbow angle between the bound and unbound conformations were calculated using the web tool developed by Stanfield and coworkers [Stanfield 2006].

2.3.1.1 Antigens

Cartesian RMSDs were also calculated on the antigen side. Bound and unbound antigens were aligned on their interface. Cartesian RMSDs (calculated both on backbone atoms only and on all atoms) were calculated and are reported in (Table 2.5) both for the antigen interface and for the whole antigen.

We note that for a few antigens, large domain rearrangements happen away from the interface, leading to a large RMSD for the full antigen, while the RMSD for the interface stays low. It is the case for the antigens of 1BGX, 2FD6 and 4FQI for example. The antigen in 3G6D is the one that displays the largest conformational changes at the interface level, by far (the RMSD on backbone atoms of the interface is as high as 2.79 Å, while the second highest is 2.08 Å for the interface of the antigen in 1BGX).

Since the present work focuses on conformational changes of antibodies, these values are only provided as complementary information, and are not further analyzed.

2.3.1.2 Antibody-antigen contacts

Contacts between antibody and antigen were taken as pairs of residues (r_1, r_2) so that r_1 belongs to the antibody, r_2 belongs to the antigen and there exists one atom a_1 from r_1 and one atom a_2 from r_2 whose distance is less than 5 Å. The interface is then defined as the set of residues that are involved in at least one contacting pair. Attractive electrostatic contacts were taken as contacts between arginine or lysine on the one hand and aspartic or glutamic acid on the other. Repulsive contacts were taken as contacts pairs involving either arginines and lysines only or aspartic and glutamic acids only.

2.3.2 Results summary

Conformational changes upon antigen binding were analyzed in 27 antibodies. Table 1 lists all antibodies and gives a short summary of the main changes for each of them. Note that given the difference in the experimental conditions that were employed to obtain the structures, conformational changes are not necessarily all due to antigen binding. Some of them may be a result of the change in pH conditions, or an effect of crystal packing for example. Despite these other possible sources of conformational changes, the analysis performed in this section is accurate and meaningful. It provides an overview of the potential conformational changes in antibodies and their order of magnitude, which can be very helpful for antibody modeling from partial or inaccurate experimental data.

Table 2.5. Antigen conformational changes upon binding. Bound and unbound antigens were aligned on their interface (all residues within 5 Å of the antibody in the complex structure). Cartesian RMSD on backbone atoms (N, C, C_α and O) and on all atoms are reported both for the interface residues and for the whole antigen. Note that some antigens display large domains rearrangements that result in a large full antigen RMSD (1BGX and 4FQI).

PDB ID	Residue count		Backbone RMSD (Å)		All-atom RMSD (Å)	
Complex	Full Ag	Interface	Full Ag	Interface	Full Ag	Interface
1AHW	199	27	2.17	0.69	2.36	1.09
1BGX	799	70	49.73	2.08	50.09	2.98
1BVK	129	20	1.27	1.63	1.79	1.62
1DQJ	129	24	0.97	1.29	1.50	1.79
1E6J	71	13	1.87	1.04	2.42	1.53
1JPS	182	24	1.90	0.65	2.19	1.07
1MLC	129	19	1.13	0.71	1.44	1.11
1VFB	129	21	1.24	1.73	1.66	2.03
1WEJ	104	15	0.53	0.82	1.22	1.83
2FD6	247	17	8.22	0.67	8.48	1.22
2VIS	267	22	0.94	0.46	1.25	0.64
2VXT	152	24	2.65	1.94	3.58	2.93
2W9E	99	19	2.60	1.73	3.03	2.63
3EO1	224	17	4.68	1.27	4.99	2.17
3EOA	178	16	1.34	0.38	1.69	0.87
3G6D	106	19	3.02	2.79	3.61	3.46
3HI6	178	25	2.11	1.35	2.72	2.20
3HMX	406	25	3.63	1.07	3.97	1.55
3L5W	101	11	2.77	0.42	3.22	1.43
3MXW	150	24	0.57	0.75	1.05	1.06
3RVW	219	19	0.69	0.70	1.18	1.36
3V6Z	137	22	4.28	0.92	4.69	1.56
4DN4	61	14	1.46	0.51	1.92	1.24
4FQI	1419	24	29.33	1.31	29.38	1.54
4G6J	149	23	0.88	1.02	1.44	1.88
4G6M	149	24	0.75	0.55	1.52	1.09
4GXU	1446	28	3.67	0.76	3.74	0.90

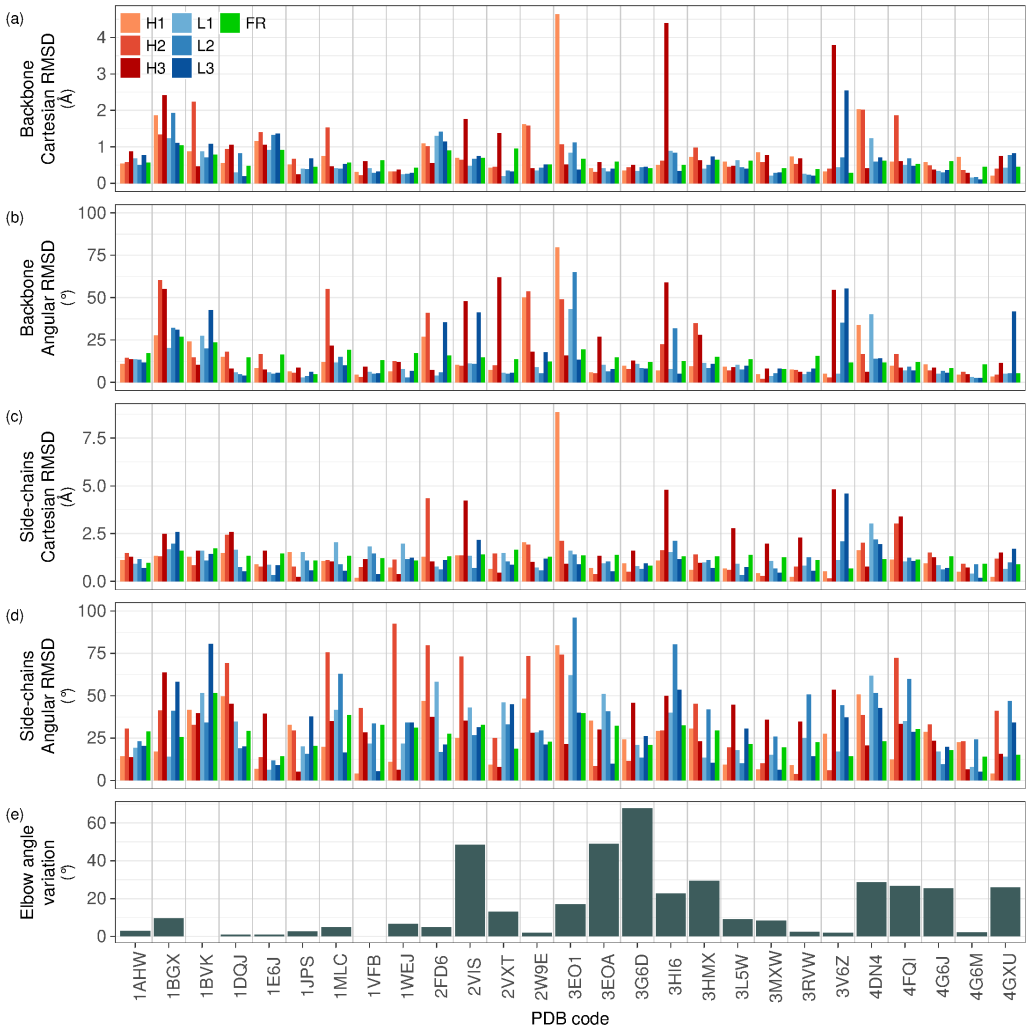


Figure 2.1. Conformational changes. (a) Backbone Cartesian RMSDs after alignment of bound and free Fv. (b) Backbone dihedral RMSDs. (c) Side-chains Cartesian RMSDs after alignment of bound and free backbones of each antibody subpart. (d) Side-chains dihedral RMSDs. (e) Elbow angle variation between free and bound conformations.

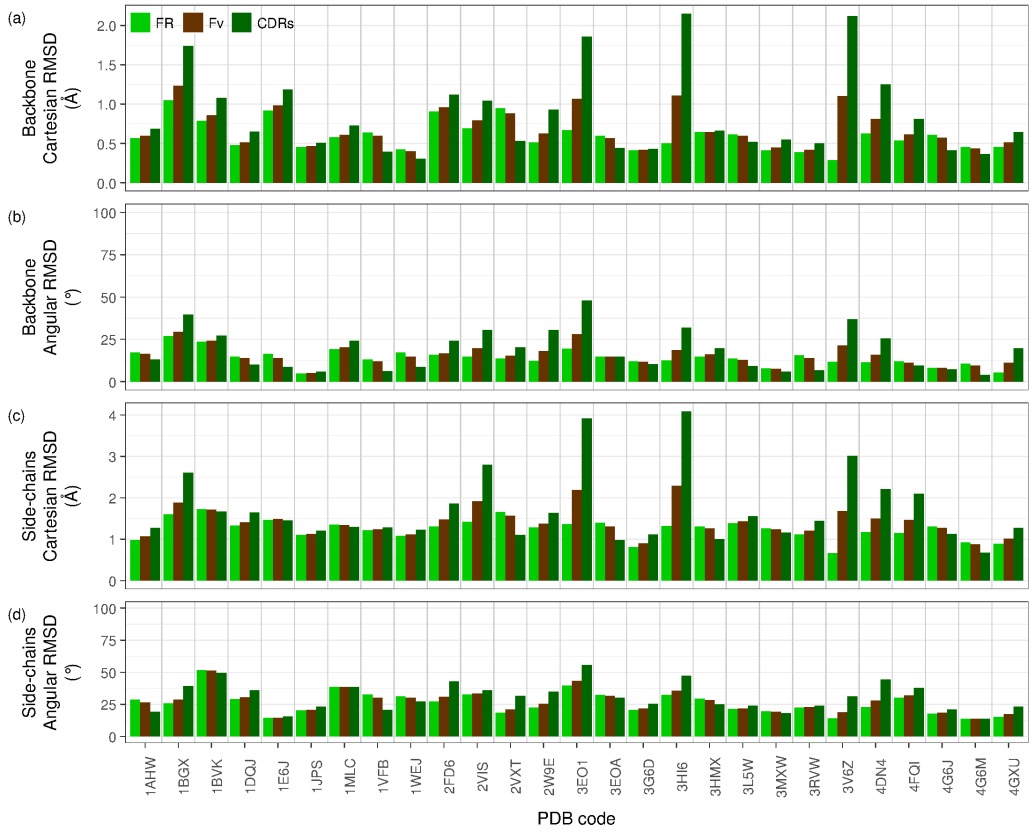


Figure 2.2. Conformational changes for Fv as a whole, for FR and for the CDRs in concert. (a) Backbone Cartesian RMSDs after alignment of bound and free Fv. (b) Backbone dihedral RMSDs. (c) Side-chains Cartesian RMSDs after alignment of bound and free backbones of each antibody subpart. (d) Side-chains dihedral RMSDs.

Three main types of conformational changes can be analyzed using our results: backbone movements within the Fv domain, particularly of hypervariable loops, side-chain movements in CDRs, and articulation between the variable and constant domains. Results for each individual antibody can be found in Figures 2.1 and 2.2. Note that Cartesian RMSDs are subject to biases depending on the model used for generating the structure from X-ray crystallography's electronic density. The results reported here are meant to be taken as general trends rather than analyses of conformational changes in individual antibody-antigen pairs.

2.3.3 Backbone movements

2.3.3.1 Framework regions

FRs are the regions of Fv outside CDR loops. In agreement with previous work, backbone movements are found to be very limited for residues in FRs as can be observed in Figure 2.3(a). The highest backbone Cartesian RMSD calculated was 1.05 Å and corresponds to antibody from complex 1BGX (this RMSD was calcu-

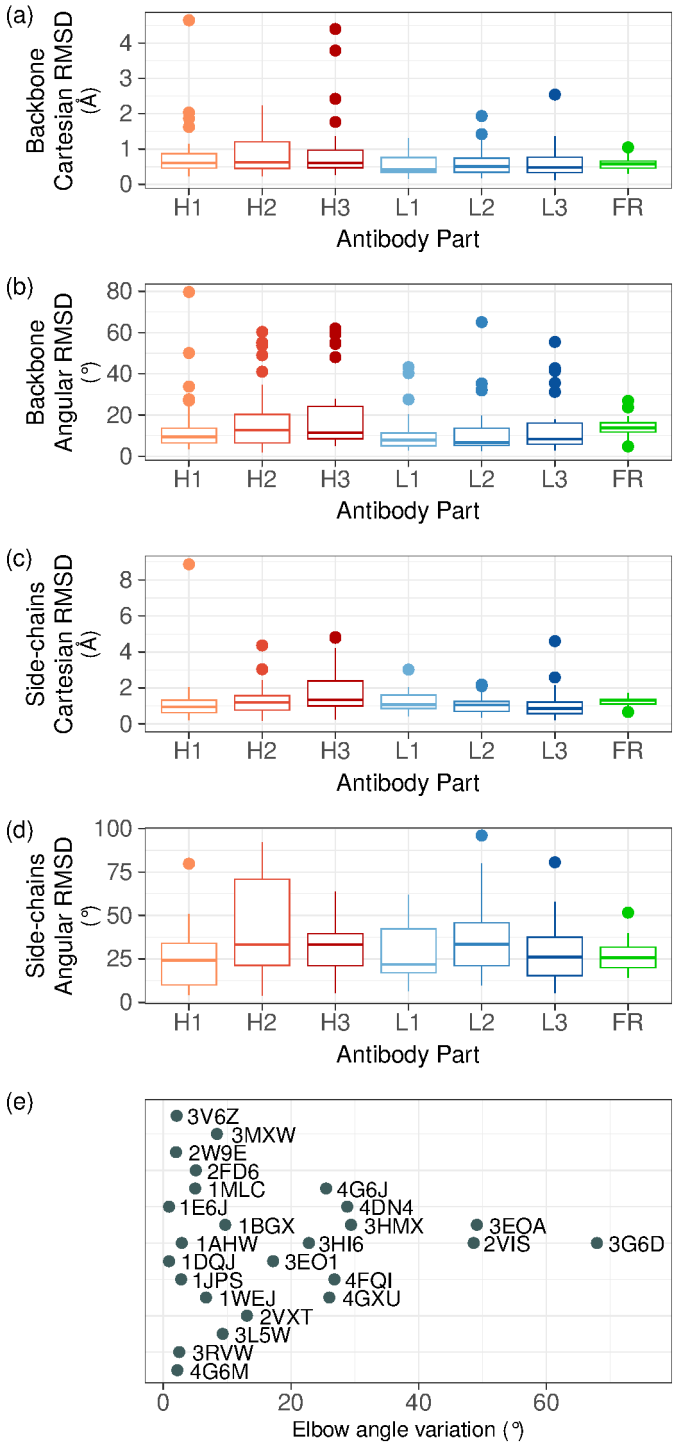


Figure 2.3. RMSDs and elbow angle distributions. (a) Boxplot of RMSDs on backbone atom positions. (b) Boxplot of backbone dihedral RMSDs. (c) Boxplot of RMSDs on side-chain atom positions. (d) Boxplot of side-chain dihedral RMSDs. (e) Elbow angle distribution.

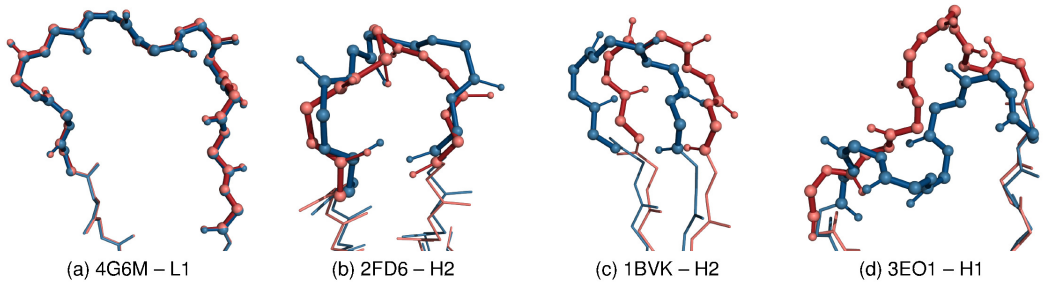


Figure 2.4. Different types of backbone movements of hypervariable loops upon binding. The bound antibody is represented in blue, the unbound antibody in red. The loop under focus is in a ball and stick representation. (a) Stable backbone for the L1 loop of 4G6M antibody. (b) Local movements of the loop backbone for the H2 loop of 2FD6 antibody. (c) Global backbone movement with stable internal conformation for the H2 loop of 1BVK antibody. (d) Local and global backbone movements for the H1 loop of 3EO1 antibody.

lated on 688 atom pairs from 172 residue pairs). For 20 complexes out of 27, the backbone positional RMSD measured for the FRs was below 0.7 Å, indicating that Framework Regions have a very stable structure. Notwithstanding, the median of backbone dihedral RMSDs for FRs is high compared with the medians of backbone dihedral RMSDs in each CDR loop (Figure 2.3(b)). The backbone dihedrals in FRs might be less restricted than those in short loops since they do not have to satisfy loop closure constraints, or this might be an artifact of the method used for obtaining the atom coordinates from electron density. Nevertheless, FRs present almost no outlier for backbone dihedral RMSDs: all values are below 20° except for the antibodies in 1BVK and 1BGX, with values of 23.8° and 26.9°, respectively. The dihedral changes in the backbone of FRs thus remain limited. Taken together, these observations suggest that the backbone dihedrals in Framework Regions are flexible but compensate each other so that FRs keep their overall structure with very limited changes in atom positions.

2.3.3.2 CDRs

Many loops show no backbone movement between their bound and unbound conformations, with very low atomic and dihedral RMSDs. It is the case for instance of the L1 loop from 4G6M (Figure 2.4(a)). Other loops show large backbone conformational changes upon binding. Disparities between atomic and dihedral RMSDs distributions for CDRs indicate that loop movements can be decomposed into internal conformational changes and global shifts. The former are responsible for the shape of the loops, while the latter correspond to a displacement of the loop in 3D space. Some loops such as the H1 loop of antibody in 3EO1 (Figure 2.4(d)) combine both movements. Others, such as the H2 loop of antibody in 2FD6 (Figure 2.4(b)), change their conformation while retaining their overall position. Finally, a few loops such as the H2 loop of the antibody in 1BVK (Figure 2.4(c)) shift but keep their internal shape.

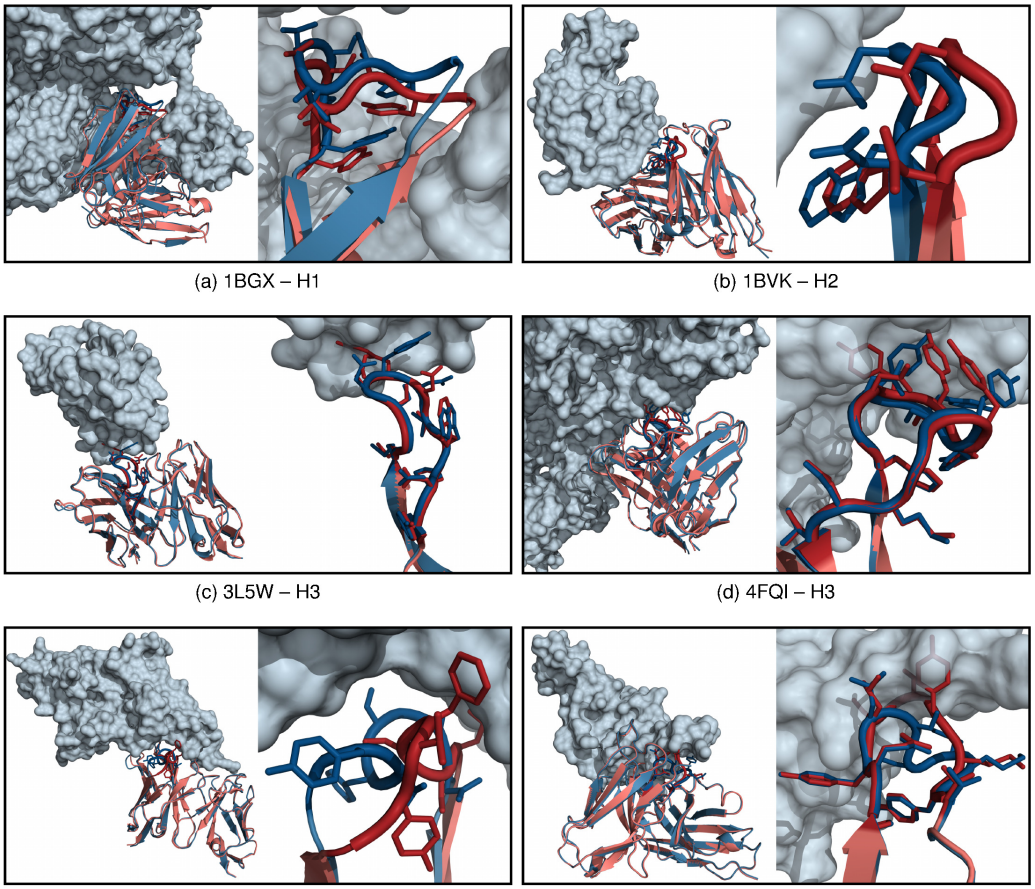


Figure 2.5. Different types of backbone and side-chains movements of hypervariable loops upon binding. The antigen is represented as a blue transparent surface. The bound antibody is represented in blue, the unbound antibody in red. The loop under focus is represented with thicker tubes and has its side-chains represented as sticks. (a)(b) Large movement of the loop backbone with limited side-chain conformational changes for the H1 loop of 1BGX antibody and the H2 loop of 1BVK antibody. (c)(d) Stable backbone with side-chain conformational changes for the H3 loop of 3L5W antibody and the H3 loop of 4FQI antibody. (e)(f) Large movements of both the backbone and the side-chains for the H1 loop of 3EO1 antibody and the L3 loop of 3V6Z antibody.

Hypervariable loops were found to be more flexible than FRs upon binding. More than half the complexes (14) have at least one CDR with a backbone Cartesian RMSD larger than 1 Å. Overall, hypervariable loops in the heavy chain were found to be more flexible than those in the light chain. No L1 or L2 and only one L3 loop exhibit a backbone positional RMSD larger than 2 Å, while 2 H1, 2 H2 and 3 H3 loops do. The largest loop motion is observed in the H1 loop of the antibody in the complex 3EO1 (Figures 2.4(d), 2.5(e)). The loop's backbone Cartesian RMSD is 4.65 Å and corresponds to a large loop rearrangement.

Distributions of backbone dihedral RMSDs for each CDR show two or more peaks: one major peak centered below 10° and one or more smaller peaks for higher dihedral RMSD values (Figure 2.6). This multimodal distribution may come from loop closure constraint: variation of one dihedral angle causes the other dihedrals to vary in order to keep satisfying loop closure. Therefore, either the loop presents very little dihedral changes, or it has a high dihedral RMSD. Large dihedral RMSD is not always correlated with high Cartesian RMSD, as we can see for the L3 loop of 2VIS. A pair of dihedral angles may compensate each other and leave a small local change if the residues they are found in are close to one another in the loop.

Most CDRs loops maintain their unbound conformation, yet contrary to FRs, CDRs backbone dihedral RMSDs show a few high outliers, suggesting major internal conformational changes in some loops. L1, L2, L3, H1 and H2 loops have been shown to adopt canonical conformations, determined by their length and sequence [Chothia 1987, Chothia 1989, Al-Lazikani 1997, North 2011]. Our results show that a few of these loops display large conformational changes upon binding, and therefore do not seem to adopt canonical conformations. Even though this is only a minority in our dataset, and canonical structures show remarkable accuracy for most loops, these cases expose the limits of the reliability of predictions based on canonical conformations.

H3 is known to be the most variable loop in antibodies, in terms of length, sequence and conformation. This makes it difficult to predict its structure. The distributions of positional and angular RMSDs for H3 do show some variability, but are similar to those of the H2 loop in our dataset. However, these results may not be comparable due to the length difference between the two loops, H3 being on average much longer than H2 (on our antibody set, the average lengths in residues are 10.0 and 5.7 for H3 and H2, respectively).

We analyzed further one of the most striking cases of non-H3 loop movement: the conformational change of H1 loop in 3EO1. Analysis of the antigen interface in the vicinity of H1 shows a hydrophobic patch constituted of two leucine residues (LEU-28 and LEU-64), one glycine (GLY-29) and two tryptophan residues (TRP-30 and TRP-32) (Figure 2.7(a)). Upon binding, H1 moves and changes its conformation to cover this hydrophobic patch (Figure 2.7(b)(c)). Aligning the unbound antibody to its bound position against the antigen leaves a cavity against this patch (Figure 2.7(d)-(f)). This cavity is filled by H1 in the complex. Global and local conformational changes allow H1 to be closer to the patch and to fill the cavity (Figure 2.7(g)-(i)). The local conformational changes also bury the hy-

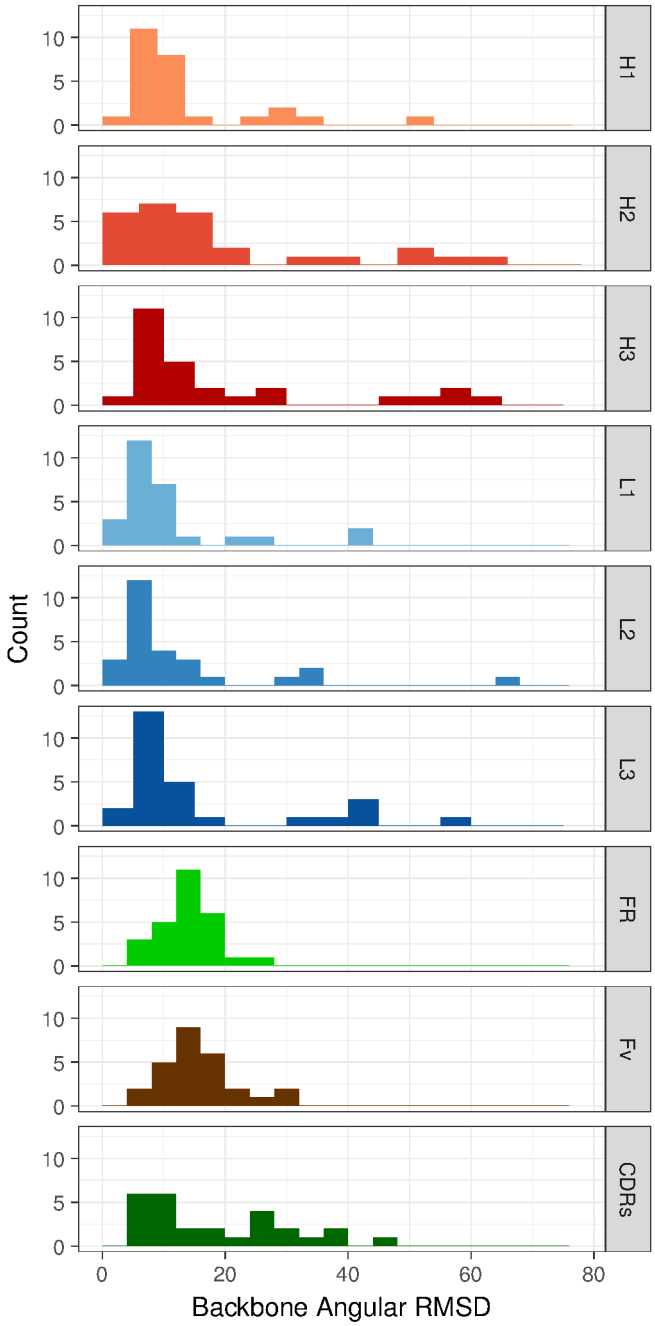


Figure 2.6. Histogram of angular RMSDs for individual CDRs and for FRs, Fv and CDRs in concert. Backbone angular RMSD is given in degrees.

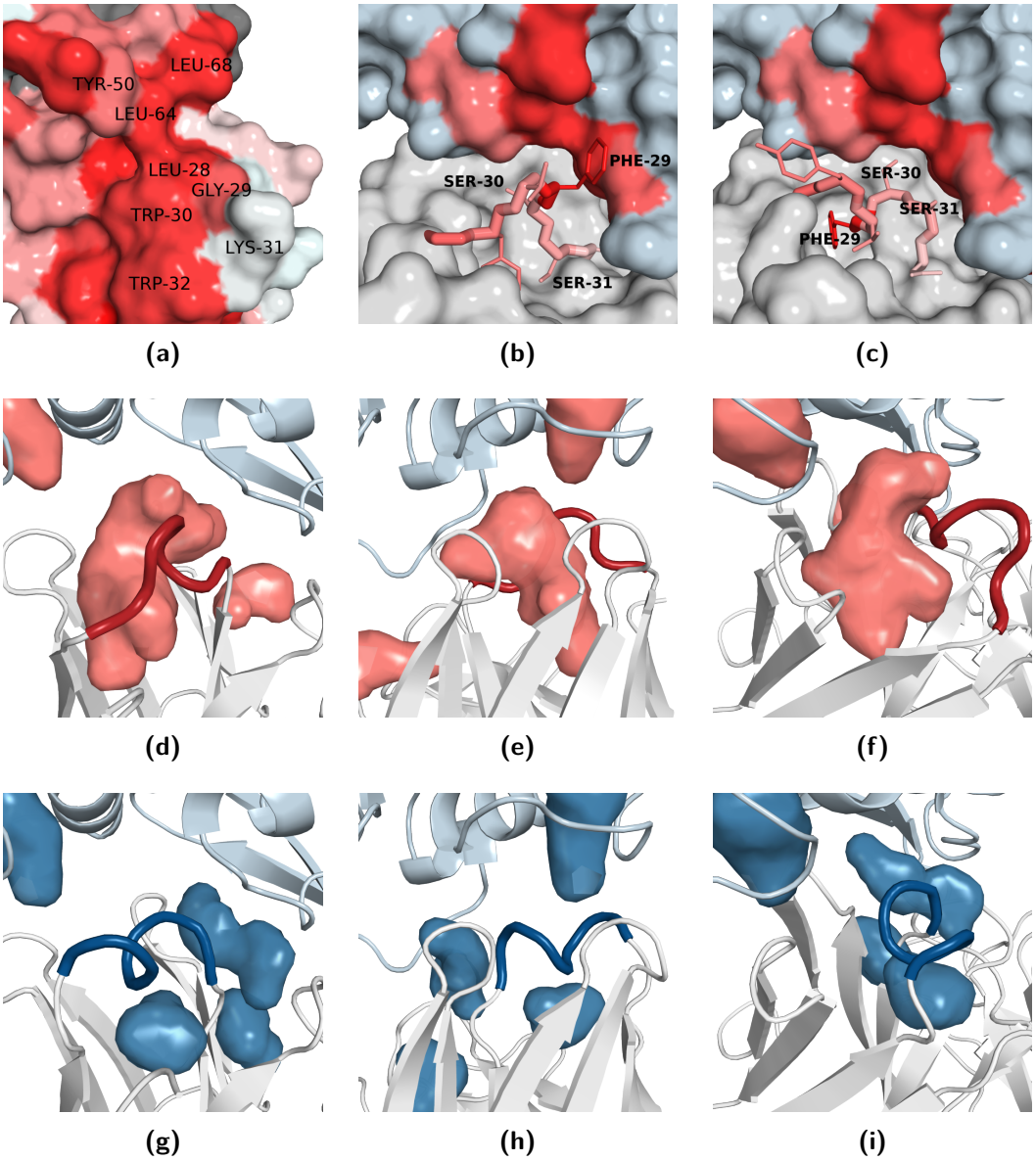


Figure 2.7. Conformational changes in 3EO1. (a) Hydrophobic patch in contact with H1 in the antigen (deeper red indicates greater hydrophobicity). (b) Simulated interface between unbound antibody and bound antigen. The antibody is in grey, the antigen in light blue. The hydrophobic patch on the antigen is in red (deeper red indicates greater hydrophobicity). SER-30 and SER-31 are turned away from the antigen, and PHE-29 is in contact with the solvent. (c) Bound antibody in complex with antigen. The antibody is in grey, the antigen in light blue. The hydrophobic patch on the antigen is in red (deeper red indicates greater hydrophobicity). SER-30 and SER-31 are directed towards the hydrophobic patch, and PHE-29 is buried. (d)(e)(f) Three different views of the cavities at the simulated interface between unbound antibody and bound antigen. The unbound antibody is in grey, the antigen in light blue. The cavities and the H1 loop are shown in red. This position of the H1 loop would leave a large cavity. (g)(h)(i) Three different views of the cavities at the antibody-antigen complex interface. The antibody is in grey, the antigen in light blue. The cavities and the H1 loop are shown in blue. The cavity is partly filled by the H1 loop with bound conformation.

drophobic PHE-29 of H1. All these conformational changes contribute to limiting the hydrophobic surface exposed to the solvent.

No correlation could be observed between the amplitude of a loop movement upon binding and the amount of contacts if forms with the antigen. There are loops that form very few contacts and yet display large movements (H1 loop from 4DN4 forms 1 contact - out of 42 for the whole interface- , but its backbone shows a Cartesian RMSD over 2 Å upon binding). Conversely, some loops form a large number of contacts but do not show any movement (H3 loop from 4GXU contributes to 38 contacts out of 61 for the whole interface but barely moves upon binding). The relationship between contacts with antigen and loop movements is further analyzed in Section 2.3.5.

2.3.4 Side-chain movements

To measure the amplitude of side-chain movements, we calculated the Cartesian side-chain RMSDs of each antibody subpart after backbone alignment of this subpart in the bound and unbound conformations (Figure 2.3(c)), as well as side-chain dihedral RMSDs (Figure 2.3(d)). Not surprisingly, these values are larger than backbone RMSDs previously calculated, indicating that side-chain movements are larger than backbone movements.

Remarkably, FRs side-chain atomic RMSDs are very similar in all cases (they are all comprised between 0.67 Å and 1.73 Å), and relatively high compared to CDRs, whereas the inverse trend would have been intuitively expected. Visual observation of a few structures unsurprisingly reveals that side-chains located within the core volume move very little, whereas surface side-chains are quite flexible, although the cause of their movement may be other than antigen binding (crystal packing effects, data quality, coexistence of alternative side-chain conformations unresolved at the resolution of the data...). Side-chain angular RMSDs for FRs also show a tight distribution compared to the side-chain angular RMSDs for CDR loops, with a median that is neither below nor above the distributions of CDR loops'. FRs side-chain conformational variations are comparable from one antibody to another, whereas CDR side-chain variations can have very different levels.

CDR side-chain Cartesian RMSDs are far more dispersed than FRs side-chain Cartesian RMSDs, but not always higher. Although positional RMSDs show higher variations for H3 side-chains, angular RMSDs suggest that H2 side-chains are more flexible. This indicates that H3 side-chains movements are mainly due to dihedral changes within the backbone, while H2 side-chains movements are also due to dihedral changes within the side-chains. Contrary to backbone Cartesian RMSDs, side-chains Cartesian RMSDs after backbone alignment do not reveal striking differences between the heavy and light chains.

Some correlation between backbone and side-chain atomic RMSDs after backbone alignment can be observed (Figures 2.1(a) and 2.1(c)). More precisely, we find that when backbone Cartesian RMSD is high, side-chain Cartesian RMSD is generally also high. An obvious explanation is that large backbone rearrangements

cause imperfect alignment of loop backbones, thus increasing side-chain positional RMSD after alignment. Notwithstanding, another reason could be that a displacement of the loop causes a change of environment for the side-chains, that are consequently repacked. This is confirmed by a similar yet looser correlation between backbone and side-chain angular RMSDs (Spearman correlation test gives $\rho = 0.46$, p-value$= 1.36 \times 10^{-9}$), showing that large backbone rearrangements are accompanied by side-chain repacking. This is the case e.g. for the H1 loop of the antibody in 3EO1 (Figure 2.5(e)), and the L3 loop of 3V6Z (Figure 2.5(f)). In the former case, the bound loop adopts a different conformation in order to fit in a cavity between the rest of the antibody and the antigen. The side-chains are then repacked to fill the cavity. In the latter case, the bound loop has to change its conformation to avoid major steric clashes with the antigen. The side-chains also have to move to avoid collisions. Both these loops adopt entirely different backbone conformations in the bound case, and the side-chains are packed to create favorable contacts with the antigen according to this new conformation, regardless of their previous positions.

However, a few antibodies loops show relatively high backbone movement with little side-chain repacking: this is the case for the H1 loop of the antibody in 1BGX (Figure 2.5(a)), whose backbone moves slightly upon binding to avoid collision with the antigen, and the H2 loop of the antibody in 1BVK (Figure 2.5(b)), which gets closer to the antigen upon binding, supposedly maximizing positive contacts. These loops are only slightly shifted or displaced, and their conformation is overall maintained, which explains why most of their side-chains remain approximately in the same place.

It is also common to observe large side-chain movements with limited or no backbone rearrangement. This happens to the H3 loop of the antibody in 3L5W where the Tyr-99 moves to allow the binding of the antigen (Figure 2.5(c)). Likewise, in the H3 loop of 4FQI, the Tyr-98 moves to allow the binding of the antigen, thus triggering a cascade of conformational changes for the other side-chains (Figure 2.5(d)).

After the observation of the different types of loop movements illustrated by Figures 2.4 and 2.5, we decided to classify all the loops in our dataset and assign them a “backbone” class and a “whole loop” class, corresponding to the classes observed in the figures. The classification is reported in Table 2.6. Although these results give interesting insight into the number of loops representing each class, the number of complexes in our dataset is too limited to draw any final conclusion about the actual frequency of each class. However, this analysis seems to confirm the high structural diversity at the interface level. Indeed, among antibodies that display conformational changes at the interface, no pattern was discernible.

2.3.5 Loop movements and contacts with the antigen

We investigated the correlation between the conformational changes in a CDR loop upon binding and its contribution to the antibody-antigen interface.

Table 2.6. Classification of CDR loops into classes. A classification was made after the definition of thresholds. A global movement was considered above 1.5 Å of Cartesian RMSD and a local movement was considered above $40°$ angular RMSD. These thresholds are arbitrary but work well when visually checking class assignments. For each loop, one class was assigned for the backbone and another for the whole loop. For the "backbone" classes:

- class 0: absence of conformational changes (backbone Cartesian RMSD lower than 1.5 Å and backbone angular RMSD lower than $40°$)
- class 1: local conformational changes without displacements (backbone Cartesian RMSD lower than 1.5 Å and backbone angular RMSD greater than $40°$)
- class 2: hinge motion without local conformational changes (backbone Cartesian RMSD greater than 1.5 Å and backbone angular RMSD lower than $40°$)
- class 3: global displacement with internal conformational changes (backbone Cartesian RMSD greater than 1.5 Å and backbone angular RMSD greater than $40°$)

For the "whole loop" classes:

- class 0: loops without any conformational changes, ("backbone" class is 0, side-chain Cartesian RMSD is lower than 1.5 Å after backbone alignment and side-chain angular RMSD is lower than $40°$)
- class 1: backbone conformational changes without side-chain movements ("backbone" class is not 0, side-chain Cartesian RMSD is lower than 1.5 Å after backbone alignment and side-chain angular RMSD is lower than $40°$)
- class 2: stable backbone with side-chain movements ("backbone" class is 0 and either side-chain Cartesian RMSD is higher than 1.5 Å after backbone alignment or side-chain angular RMSD is above $40°$)
- class 3: both backbone and side-chains conformational changes ("backbone" class is not 0 and either side-chain Cartesian RMSD is higher than 1.5 Å after backbone alignment or side-chain angular RMSD is above $40°$)

PDB Code of complex	Backbone classes (H1-H2-H3-L1-L2-L3)	Whole loop classes (H1-H2-H3-L1-L2-L3)
1AHW	0-0-0-0-0-0	0-0-0-0-0-0
1BGX	2-1-3-0-2-0	1-3-3-2-3-2
1BVK	0-2-0-0-0-1	2-1-2-2-0-3
1DQJ	0-0-0-0-0-0	2-2-2-2-0-0
1E6J	0-0-0-0-0-0	0-0-2-0-0-0
1JPS	0-0-0-0-0-0	2-0-0-2-0-0
1MLC	0-3-0-0-0-0	0-3-0-2-2-0
1VFB	0-0-0-0-0-0	0-2-0-2-0-0
1WEJ	0-0-0-0-0-0	0-2-0-2-0-0
2FD6	0-1-0-0-0-0	2-3-0-2-0-0
2VIS	0-0-3-0-0-1	0-2-3-2-0-3
2VXT	0-0-1-0-0-0	0-0-1-2-0-2
2W9E	3-3-0-0-0-0	3-3-0-0-0-0
3EO1	3-1-0-1-1-0	3-3-0-3-3-2
3EOA	0-0-0-0-0-0	0-0-0-2-2-0
3G6D	0-0-0-0-0-0	0-0-2-0-0-0
3HI6	0-0-3-0-0-0	0-2-3-2-2-2
3HMX	0-0-0-0-0-0	0-2-0-0-2-0
3L5W	0-0-0-0-0-0	0-0-2-0-0-0
3MXW	0-0-0-0-0-0	0-0-2-0-0-0
3RVW	0-0-0-0-0-0	0-0-2-0-2-0
3V6Z	0-0-3-0-0-3	0-0-3-0-2-3
4DN4	2-2-0-1-0-0	3-3-0-3-2-2
4FQI	0-2-0-0-0-0	0-3-2-0-2-0
4G6J	0-0-0-0-0-0	0-2-0-0-0-0
4G6M	0-0-0-0-0-0	0-0-0-0-0-0
4GXU	0-0-0-0-0-1	0-2-2-0-2-3

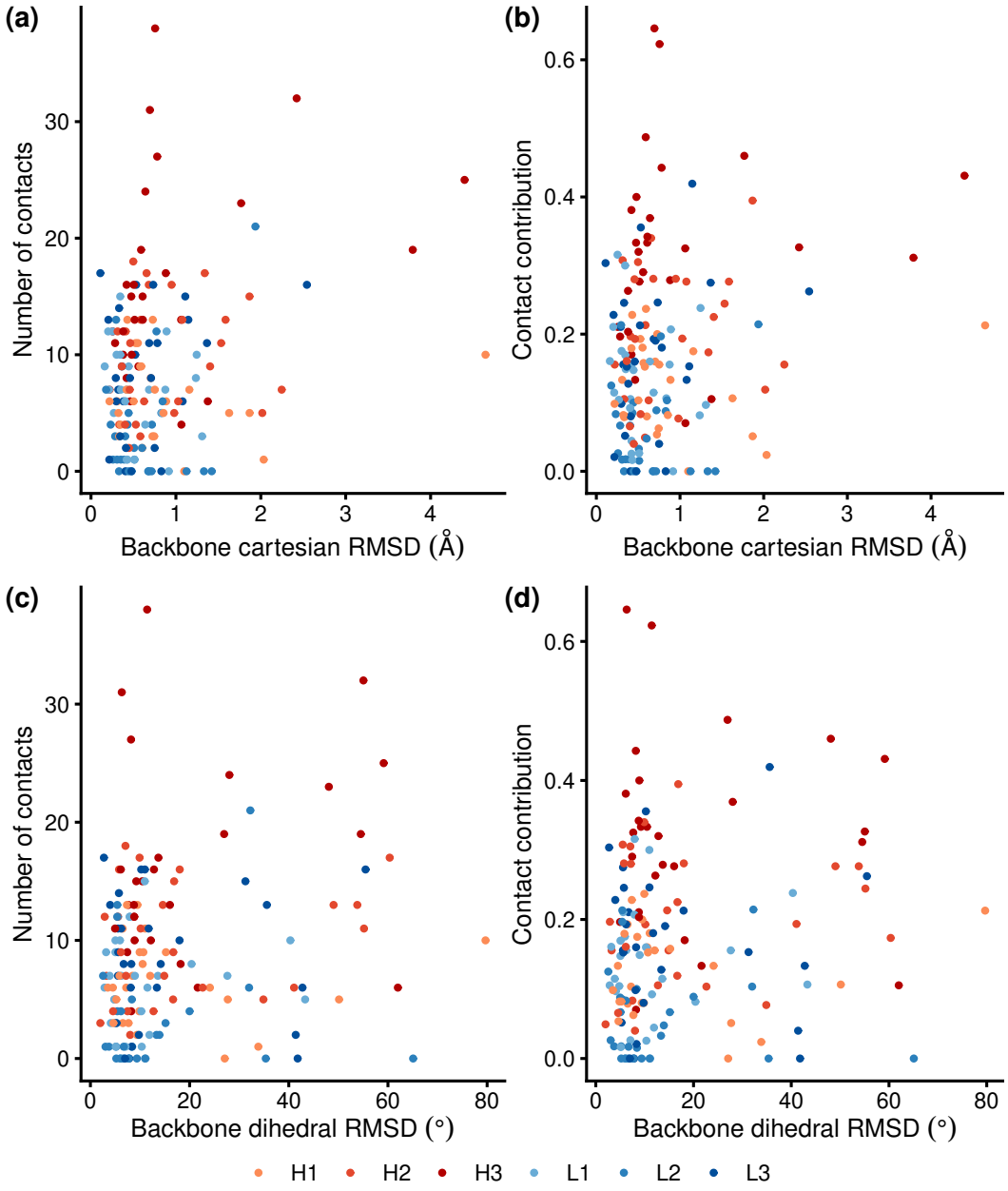


Figure 2.8. Contacts with antigen and backbone movements for CDR loops in the dataset. The number of contacts corresponds to the number of contacting residue pairs that the loop is involved in. The contact contribution corresponds to the number of contacts made by the loop divided by the total number of contacts formed by all 6 CDR loops. (a) Number of contacts and backbone Cartesian RMSD. (b) Contact contribution and backbone Cartesian RMSD. (c) Number of contacts and backbone Cartesian RMSD. (d) Contact contribution and backbone Cartesian RMSD.

The connection between contribution to the interface and loop movements is delicate to study. Indeed, the contribution to the interface can be measured in a variety of ways, and so can loop movement (this work actually provides 4 different metrics to describe loop conformational changes). Moreover, comparing different CDR loops is complex since these have different positions relative to the interface, and different lengths and sequences giving them a variable flexibility. Here, we chose to study the relationship between the number of contacts that each loop forms with the antigen and the conformational changes of the backbone of this loop upon binding.

The definition of contacts was defined earlier (Section 2.3.1) and we define the contact contribution as the number of contacts formed by the loop divided by the number of contacts formed by all six CDR loops.

Fig 2.8 gives the number of contacts or the contact contribution of the loops as a function of their backbone Cartesian or dihedral RMSD upon binding. No obvious trend is visible from these plots: a greater contribution to the antibody-antigen interface does not correlate with a larger movement, or on the contrary with a higher conformational stability. Unsurprisingly, H3 loops form more contacts than other loops, probably due to their privileged position at the center of the binding site. Most loops are found in the leftmost part of the graphs: they show little or no backbone movement, regardless of the contacts they form.

From this analysis we conclude that the link between the contribution of a CDR loop to the antibody-antigen interface and its movement upon binding is complex. In addition, the movement of a loop that does not directly contribute to the interface may still be relevant for binding, because of a change in energy, or simply because the loop makes way for other loops to move upon binding.

2.3.6 Elbow angle variation

The distribution of elbow angle variation presents two major clusters: one between 1° and 11° (14 antibodies), and another between 24° and 29° (6 antibodies) (Figure 2.3(e)). Therefore, although for most antibodies the elbow angle barely varies, there is still a sizable number of antibodies for which relatively large elbow angle variations occur. Only 2 antibodies show a variation of the elbow angle between 11° and 24°, and 3 above 29°. These three outliers indicate that very large variations of the elbow angle can also be observed, which should be borne in mind while building a model for a Fab structure. The largest elbow angle variation is measured for the antibody in 3G6D (67.9°). Structural alignment of the bound and free Fv showed a clear displacement of the constant domain between the two conformations.

2.4 Docking success and conformational flexibility

This section analyzes the relationship between conformational flexibility and success of docking methods. We first define a docking success score and then detail the results.

2.4.1 Docking success score

Vreven and co-workers report the results of four docking algorithms on all new cases of the protein-protein docking benchmark version 5 [Vreven 2015]. This concerns 16 antibodies of our dataset. They report the presence of high, medium or acceptable quality solutions (according to CAPRI's criteria) among the top 1, 5, 10, 50 and 100 predictions as ranked by the scoring function for each of the four docking algorithms (SwarmDock [Moal 2010, Li 2010], PyDock [Cheng 2007], ZDOCK [Chen 2003a, Chen 2003b] and HADDOCK [Dominguez 2003]). For one target, we define $S_{x,A}$ as the score associated with the top x predictions of algorithm A, $T_{x,A}$. $S_{x,A}$ is 3 if $T_{x,A}$ contains at least one high quality prediction, 2 if $T_{x,A}$ contains no high quality prediction but at least one medium quality prediction, 1 if $T_{x,A}$ contains no high or medium quality prediction but at least one acceptable prediction, 0 otherwise. We then define the score P_A related to the performance of algorithm A on this docking target:

$$P_A = \frac{S_{100,A} + 4 \times (S_{50,A} + 4 \times (S_{10,A} + 4 \times (S_{5,A} + 4 \times S_{1,A})))}{4 + 4^2 + 4^3 + 4^4 + 4^5}$$

The docking success score is then taken as the average of the score of the 4 algorithms. It is a value between 0 and 1, with 0 describing a failure of all 4 docking algorithms, without any acceptable solution in the top 100 predictions, and 1 describing a situation where the top prediction for all 4 algorithms is of high quality.

2.4.2 Flexibility, particularly in loops, perturbs antibody docking pose prediction

Docking success score defined in Section 2.4.1, as well as values for conformational changes for each new antibody case in the docking benchmark version 5, are displayed on Figure 2.9.

None of the four algorithms seems to perform distinctly better than the others. Each of them performs better than all others on at least one case (4DN4 for SwarmDock, 2VXT for pyDock, 3G6D for ZDOCK and 4G6J for HADDOCK), and fail in cases where others succeed. ZDOCK never performs undeniably worse than all three other algorithms: for each case, at least one of SwarmDock, pyDock and HADDOCK performs worse than ZDOCK. However, since ZDOCK rarely yields better results than all other algorithms (at least on the antibody-antigen cases we consider in this study), and since the first solution it predicts is never even acceptable, it cannot be considered to perform better overall.

Results reveal that large backbone movements in CDRs are correlated with low docking performance. Indeed, 3V6Z, 3HI6 and 3EO1 show substantial backbone movements of CDRs and docking algorithms perform badly on those cases. 2W9E, 4FQI and 4DN4 also show backbone movements in CDRs, although not as large as those three previous cases. Docking performance is increased for those cases, but

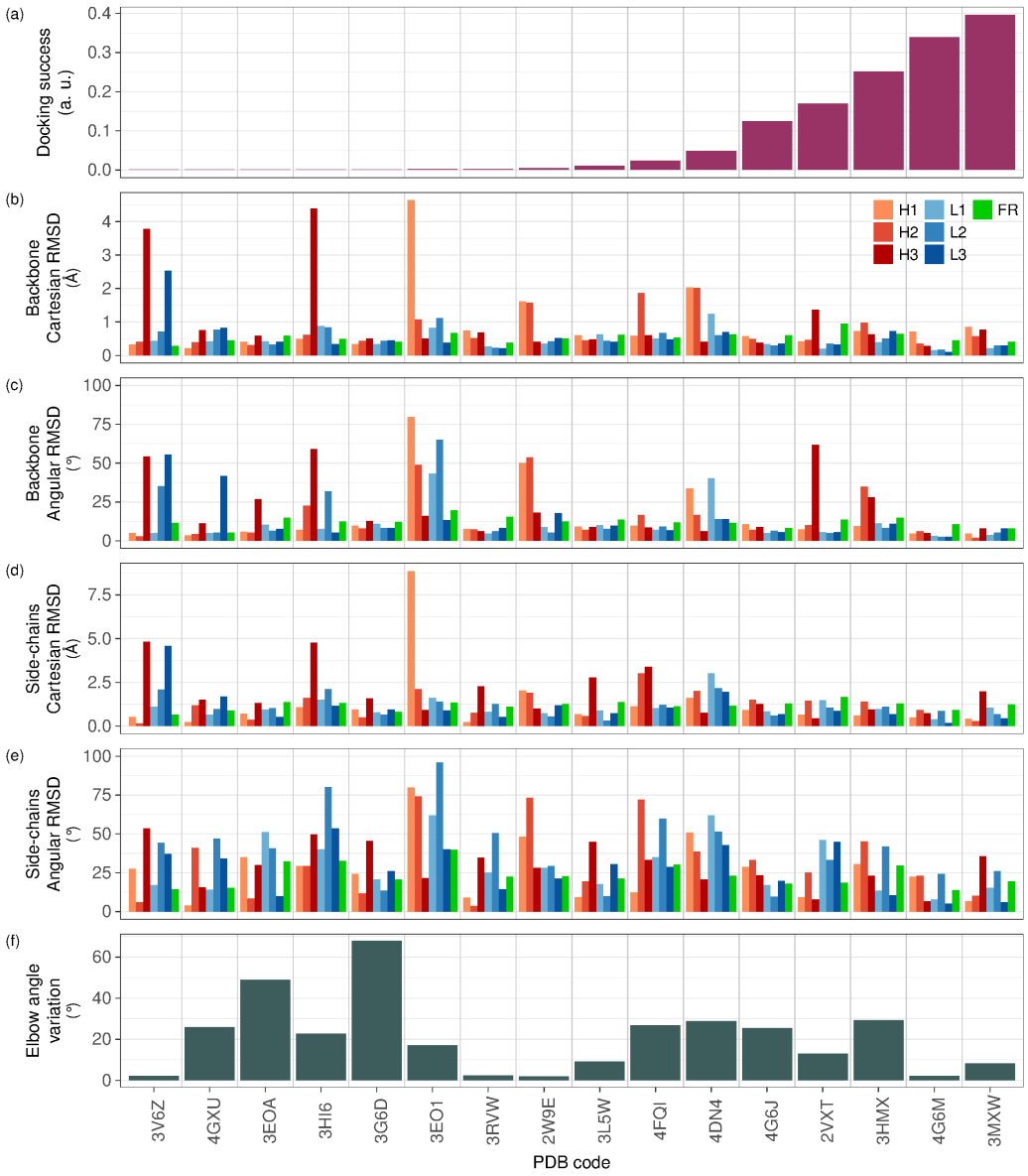


Figure 2.9. Docking success relative to conformational changes. (a) Docking success score. (b) Backbone Cartesian RMSDs after alignment of bound and free Fv. (c) Backbone dihedral RMSDs. (d) Side-chains Cartesian RMSDs after alignment of bound and free backbones of each antibody subpart. (e) Side-chains dihedral RMSDs. (f) Elbow angle variation between free and bound conformations.

remains limited. Side-chain repacking also seems to be correlated with docking difficulty. Indeed, 3L5W shows limited backbone movements but substantial side-chain repacking in H3 and the docking performance for this case is also limited. Conversely, in the five easiest cases (3MXW, 4G6M, 3HMX, 2VXT and 4G6J), only one CDR loop presents a backbone atomic RMSD of more than 1 Å (2VXT–H3, 1.38 Å), and no CDR exhibits a atomic side-chain RMSD of more than 2 Å. Angular RMSD does not reveal such a clear correlation with docking success, suggesting that the actual displacement of atoms has higher impact on docking algorithms efficiency than their local reorganization. However, this has to be tempered by the fact that neither backbone angular RMSDs nor side-chain angular RMSDs show any notable outlier for the best-working cases. These results suggest that conformational changes in CDRs, whether whole backbone movements or side-chains rearrangements, are poorly handled by docking algorithms, which tend to fail when those become too large.

Nevertheless, the negative correlation between conformational changes in CDR and docking performance is far from being perfect. Indeed, 4GXU, 3EOA and 3G6D show very limited conformational changes at the binding site yet docking algorithms perform very poorly for these cases. This may be due to the large variation in the elbow angle that occurs for these three cases. Such a large conformational change, although remote from the binding site, may be important for binding. Indeed, constant regions have been shown to stabilize antibody-antigen binding, in particular during Molecular Dynamics simulations [Knapp 2017]. Similarly, correct modeling of the constant regions may be necessary in order to improve the accuracy of the scoring functions in docking algorithms.

Still, a large variation of the elbow angle upon binding may not necessarily cause computational docking algorithms to fail. For example, the elbow angles of 4G6J and 3HMX vary of 26° and 29° upon binding, respectively. Despite these relatively large variations, these two cases are among the most successful for computational docking.

The poor docking performance for the case of 3RVW is harder to explain with arguments related to conformational changes. This case presents some side-chain rearrangements at the interface, in particular in the H3 loop, yet these remain limited. Conformational changes in the antigen exist but are also limited (Figure 2.10). This example shows that docking remains a difficult problem, even for some almost rigid cases. In such cases, for which conformational changes are limited, sampling is generally not an issue and the difficulty more probably lies in the ability of the scoring function to determine the right pose. Although the analysis of the nature and size of the interface did not provide any interesting information that would let us hypothesize on the reason for docking prediction failure, the strength of the interaction do. Vreven and co-workers provide values of free energy and dissociation constants for all of the antibody-antigen complexes also tested for docking, except for 3EO1 and 3HMX. These values reveal that 3RVW is the complex with the second highest measured dissociation constant (after 3HI6, which displays large conformational changes at the interface). The weakness of the antibody-antigen

interaction may be one reason why docking algorithms are unable to find the right docking pose.

Vreven and co-workers also assessed the expected difficulty of docking on each case based on the interface atoms RMSD and the number of non native contacts in aligned unbound structures. Of course, the result is highly correlated to the level of conformational changes at the interface, even though it is smoothed over the whole binding site. Nevertheless, separating the different components of the antibody gives a more precise insight into what actually constitutes a challenge for docking algorithms. Indeed, among the new antibody-antigen cases, none is expected to be difficult according to the classification by Vreven and co-workers yet many of them yield very poor or no results. Looking at the conformational changes of each component of the Fab fragment, we can see that a very large movement of a single CDR loop will more likely make the case difficult than a higher positional RMSD over the whole Fv fragment, even though both are obviously correlated.

Other than conformational changes at the interface (which may prevent the sampling of the right pose), the failure of some cases can be explained by the difficulty to design an accurate scoring function able to discriminate between the actual docking pose and other poses. By focusing on the RMSD and contacts at the interface, the classification provided by Vreven and co-workers ignores changes away from the binding site (which may be important for the accuracy of scoring functions of docking algorithms).

The antibody-antigen complexes on which the four docking algorithms were tested can be divided into three classes. The "medium difficulty" cases as defined by Vreven *et al.*, showing major conformational changes either in the antibody or the antigen (3V6Z, 3HI6, 3EO1, 3L5W), the "easy - low success" cases (easy according to the classification by Vreven and co-workers, but with poor success when tested by the docking algorithms: 4GXU, 3EOA, 3RVW, 2W9E) and the "easy - high success" cases (easy according to Vreven et al. and with high success with docking algorithms: 4FQI, 4DN4, 4G6J, 2VXT, 3HMX, 4G6M, 3MXW). When tested on "medium difficulty" cases, the docking algorithms unsurprisingly fail or yield poor solutions. However, perhaps less evidently, the RMSD profiles of the "easy - low success" and the "easy - high success" are similar, and it seems that conformational changes alone cannot account for the difference in the performance of docking algorithms. Using Capri's definition of contacts (two residues on both sides of the interface are in contact if we can find one atom belonging to the first residue and one atom belonging to the other within a distance below 5 Å), we tried to relate the docking success difference between these two categories with the contacts formed by the CDR loops. In particular, we looked at electrostatic contacts. Attractive electrostatic contacts (between arginine or lysine on the one hand and aspartic or glutamic acid on the other) greatly outnumber repulsive contacts (at least 5 more contacts) in the interface of 4 "easy - high success" cases (4DN4, 2VXT, 4G6M, 3MXW) and 2 "medium difficulty" cases (3G6D, 3L5W) but not in any "easy - low success" case. On the contrary, repulsive electrostatic contacts outnumber attractive electrostatic contacts in one "easy - low success" case, 2W9E.

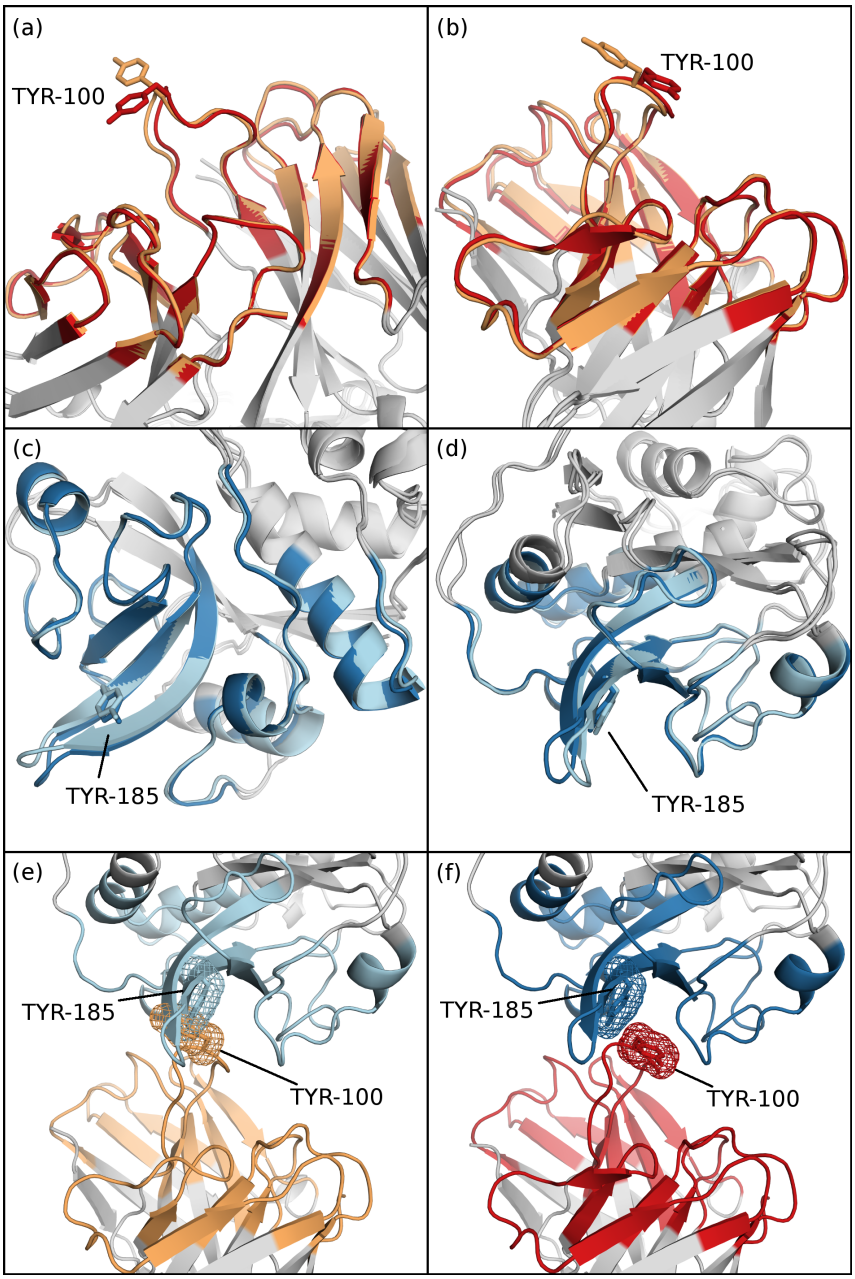


Figure 2.10. Conformational changes of the antibody in 3RVW and its antigen upon binding. The interface is displayed in color. The bound antibody is in red, the unbound antibody is in yellow. The bound antigen is in dark blue, the unbound antigen is in light blue. The unbound antibody was aligned on the bound antibody Fv. The unbound antigen was aligned on the bound antigen. (a) and (b) Limited conformational changes of the antibody interface upon binding, shown in two different orientations. (c) and (d) Limited conformational changes of the antigen interface upon binding, shown in two different orientations. (e) The antibody-antigen interface in its unbound conformation shows a major steric clash of side-chains TYR-185 of antigen and TYR-100 of the antibody's heavy chain (enhanced Chothia numbering). (f) Displacement of the TYR-100 side-chain of the antibody's heavy chain prevents this major clash and allows binding.

These results suggest that scoring functions may rely too strongly on electrostatic interactions to determine the right pose. Indeed, when sampling does not seem to be an issue (easy cases as classified by Vreven *et al.* show limited conformational changes at the interface), complexes showing many more attractive than repulsive electrostatic contacts appear to be more easily predicted, suggesting that scoring functions better discriminate this type of profile. Looking at the number and nature of the residues at the interface, this constitutes the only difference we could observe between "easy - low success" and "easy - high success" interfaces.

We also note that 3EOA, one of the "easy - low success" cases, exhibits a large elbow movement upon binding. When this is not predicted by the docking algorithm, this may contribute to the inaccuracy of the scoring function, even though this conformational change happens away from the interface.

2.5 Conclusion

In this work, we have analyzed the conformational changes in 27 antibodies upon binding. Results show that Framework Regions are structurally stable, despite some side-chains movements on the antibody surface. More importantly, their variability is similar in the different antibodies, with very few outliers. Hypervariable loops are much more flexible overall, and much more heterogeneous. Some are extremely stable and rigid, while others display large conformational changes upon binding. Conformational changes may consist in large backbone motions, while others may be large side-chains rearrangements. Some movements are local only, while some create a large displacement of the loop. A classification of the loop movements into classes showed that there is an even higher diversity of conformational changes at the antibody interface level: each antibody shows a different profile from any other. The limited size of the dataset did not allow us to draw any further conclusion, but extending this classification to a larger number of cases constitutes an interesting lead for future work. The orientation between the Fv and the constant domain in the Fab fragment is also very variable. A considerable number of antibodies present variations of the elbow angle larger than 25 °.

Conformational changes were found to partly account for docking difficulty in most cases. Antibodies presenting large CDR loop motions or substantial side-chains rearrangements upon binding appear harder to dock, implying that changes in the topology of the binding site are a major obstruction to successful computational docking. However, conformational changes at the interface only could not explain the lack of success of some cases, suggesting that scoring remains an important issue in antibody/antigen docking. Some antibodies with rigid binding sites but large elbow angle variations yielded poor docking results, suggesting that large movements of the constant domains may hinder docking in some cases, possibly through the inaccuracy of scoring functions. The nature of the antibody-antigen interface was also found to play a role in the success of docking algorithms. Interfaces with a much larger number of positive rather than negative electrostatic contacts

appeared to be better predicted in the absence of large conformational changes, which suggests that scoring functions used in docking greatly focus on electrostatic contacts to assess the correctness of a pose. The improvement of antibody-antigen docking predictions will thus require the design of more accurate scoring functions that can work better when the interfaces are not of an electrostatic nature.

By underlining the conformational changes happening in CDR loops and the poor results caused by the incapacity of docking prediction methods to correctly account for those changes, these conclusions emphasize the importance of accurate methods to model flexible loops. Loop modeling involves two major steps (see Section 1.3): sampling and scoring. The former is the focus of Chapter 3, through the presentation of a sampling method exhaustively exploring the conformational space of protein loops. The latter is at the heart of Chapter 4, with an assessment of state-of-the-art loop scoring methods and of their ability to model the energy landscape of flexible protein loops.

Chapter 3

Loop sampling

Contents

3.1 Introduction

The previous chapter highlighted the importance of an accurate modeling of antibody loops, especially when these loops exhibit flexibility. As mentioned in Section 1.3.1, this problem is not antibody-specific and can be generalized to loops from other protein systems. Existing modeling methods mostly focus on determining one conformation for the missing loop. Therefore, assessing the performance of loop sampling methods usually consists in coupling them with a scoring method and searching top-scoring samples for near-native conformations. However, the major flaw of this process is that it makes the underlying assumption that the loop can only adopt a single conformation, thereby ignoring its potential flexibility. Yet, lack of structural data in some portions of proteins resolved via X-ray crystallography usually indicates that the concerned regions are too flexible to be observed using this commonly used technique. Considering this fact, representing missing loops from crystal structures by using a single conformation appears inherently contradictory.

This chapter presents a new loop sampling method, called MoMA-LoopSampler, that is not focused on the prediction of a single conformation, but is aimed at a more thorough exploration of the loop's conformational space. The proposed method is

not antibody-specific and can be applied to any protein loop. It employs a hybrid approach to loop modeling that constructs candidate loops utilizing an extensive structural database of small protein fragments and a closed-form inverse kinematics (IK) solver. The loop is divided into a set of consecutive three-residue fragments. All but one of the fragments are iteratively sampled from the database, where each sample is slightly perturbed in order to increase the size of the explored loop configuration space. The last fragment is completed by closing the kinematic chain utilizing an IK solver. MoMA-LoopSampler varies the assignments of which fragment is solved by inverse kinematics to further increase the sampled space. As each fragment is placed in the candidate loop, collision detection and forward reference checking are performed to prune the search space. This chapter also investigates a more advanced version of the sampler, which incorporates a reinforcement learning (RL) strategy. For each fragment within the loop, similar structural pieces are clustered in a low-dimensional projection. Clusters that lead to successfully closed loops are sampled from more frequently.

Validating the method and testing its performance constitutes a consequential part of this work. Several features of MoMA-LoopSampler are thoroughly studied, including the database and the projection employed for RL. As for the performance, it is assessed on multiple aspects. First, we tested the ability of the method to sample near-native conformations. Although it is not its only goal, accurately sampling known conformations is a requirement of the method. Moreover, this work involves a comparison to other state-of-the-art methods whereas other performance aspects are not so easily comparable. Then, using the example of a flexible loop from the streptavidin protein, the capacity of MoMA-LoopSampler to sample multiple stable states, along with intermediate states, is investigated. The effects of activating RL are also extensively tested, with a focus on the preservation of conformational diversity in generated ensembles.

The remainder of this chapter is organized as follows. Section 3.2 describes the details of the MoMA-LoopSampler method and lingers over the validation of several of its components. Section 3.3 showcases the ability of both the learning and non-learning approaches to generate a set of diverse loops, and also applies MoMA-LoopSampler to several loop prediction benchmark datasets. A summary of these results along with potential future work are discussed in Section 3.4.

This work has been accepted by and will be published in *Bioinformatics* [Barozet 2019b].

3.2 Methods

The MoMA-LoopSampler method operates in two modes. The first mode, referred to as the basic method, utilizes a structural database combined with a technique from robotics to generate loop samples. The second mode, referred to as the learning method, builds upon the first by employing a low-dimensional projection to organize

the information extracted from the structural database. Reinforcement learning is then utilized to speed up future sampling.

3.2.1 Protein representation

Proteins are represented using an all-atom model where the degrees of freedom are the backbone dihedral angles ϕ, ψ, ω (Figure 1.4). Bond lengths and bond angles are held constant as per the idealized model [Engh 1991]. The loop portion of each protein is decomposed into a set of n tripeptides (continuous segments of a protein comprised of 3 amino acid residues). Each tripeptide is represented by 3 sets of ϕ, ψ, ω angles, and one set of values for these 9 angles will be referred to as a tripeptide "state" throughout the chapter. Loops where the number of residues is not divisible by 3 can also be handled by including additional residues at either side of the loop and by restraining the dihedral angles of these residues. However, for the sake of simplicity, we only explain the case of loops with a number of residues divisible by 3. Sampling of ϕ, ψ, ω angles is performed through the selection of tripeptide states from an appropriate database. Initially, the side-chains of the loop are omitted in the model. They can be added once a closed conformation of the backbone has been found.

3.2.2 MoMA-LoopSampler without reinforcement learning

3.2.2.1 Loop sampling

Algorithm 1 showcases the basic MoMA-LoopSampler. To facilitate using samples from the structural database for each tripeptide in the loop, function `ConstructLoopPlans` (line 2), constructs a set of building plans. A plan corresponds to one possible order in which the tripeptides of the loop can be assembled. At each iteration, the function `SelectPlan` (line 4) randomly picks a plan that the recursive function `BuildLoopPos` will follow to build a conformation, starting from the tripeptide with index 1 in the plan (line 5). Among the numerous possibilities, our implementation only considers a subset of n possible plans, where n is the number of tripeptides in the loop. Plan number p assembles the loop starting from tripeptide 1 to $p-1$, then from the end of the loop working backwards from tripeptide n to $p+1$. The method attempts to close the loop by utilizing inverse kinematics for the last tripeptide in the plan, p. This technique allows all of the positions within the loop to be sampled from the database.

C contains the working conformation of the protein. Initially, $\mathsf{C}_{\mathsf{init}}$ includes all the atoms for the non-loop portion of the protein. Once a plan is selected in line 4, tripeptides in the loop are recursively assembled employing a backtracking method. For each tripeptide in the loop, the function `SampleTripeptide` utilizes the amino acid sequence of the tripeptide to access the structural database and randomly sample a tripeptide state (line 11). The nine corresponding angles are then slightly perturbed to enable MoMA-LoopSampler to sample the loop conformational space more finely (function `PerturbState`, line 12). Function

`InstallTripeptide` (line 13) then installs the sampled and perturbed tripeptide into the working conformation `C` and checks that it respects all the required constraints (see Section 3.2.2.3). If it does, the function `BuildLoopPos` is called to continue building a conformation from the next tripeptide in the plan (line 18). When the last tripeptide in the plan is reached (line 16), loop closure is attempted (function `CloseLoop`). If this is successful, the conformation is added to the ensemble Ω (line 30). The method finishes when it reaches an iteration limit or has sampled a pre-defined number of successful loop conformations.

Due to the backtracking search, this method has linear space complexity and exponential time complexity $\mathcal{O}({max_{\text{atts}}}^{n-1})$, where max_{atts} is the maximum number of attempts to be made in each of the $n-1$ tripeptide positions. However, in practice, failures are detected early in the search process, making this a practical approach. Additionally, a timer (not shown in Algorithm 1) limits the total duration of the `BuildLoopPos` procedure to prevent the algorithm from getting trapped.

3.2.2.2 Database construction

A database of tripeptide states (indexed by their corresponding amino-acid sequence) was constructed using the structures of protein domains obtained from SCOP 2.06 [Fox 2014]. This collection contains 244,326 domains, extracted from 77,439 PDB entries. The 95% ID filtered subset of the domains, consisting of PDB-style files for 28,011 domains, was utilized to build the structural database. DSSP [Kabsch 1983] was employed to assign secondary structure labels to each residue in these files.

Each structure file was processed by passing a sliding window of size 3 along the amino acid sequence. Each resulting tripeptide was added to the database if all of its 3 residues had a DSSP code of T, S, B, G or no code (which corresponds to an unclassified structural type). In other words, no portion of the tripeptide participates in an alpha-helix or a beta-strand. The tripeptide state, which corresponds to its 9 backbone dihedral angles (3 sets of ϕ, ψ, and ω), was recorded in the database and indexed by its corresponding amino acid sequence.

A slightly different treatment was applied when the provided domain structure file originated from NMR data. For each structural file that contained more than one model, a distance filter was applied to corresponding tripeptides in each model to avoid redundancy in the database. A tripeptide state was considered sufficiently distant from another tripeptide state, and was thus added to the database, if it met at least one of the two following criteria: the RMSD on ω, ϕ and ψ angles is above 0.3 radians, or one of the nine dihedral angles differs by more than 1 radian.

MoMA-LoopSampler constructs loops by concatenating tripeptide states. The concatenation of two tripeptides t_1 and t_2 contains four tripeptides in total: t_1, t_2 and two implicit tripeptides t_3 and t_4 straddling t_1 and t_2. Sampling the states of t_1 and t_2 implicitly sets the states of the two intermediary tripeptides t_3 and t_4. These intermediary states are referred to as synthetic states, as they were not directly extracted from the set of experimentally solved structures within SCOP.

Algorithm 1: Build Loop

```
1  void BuildLoop(C_init, L_start, L_end)
2      Plans ← ConstructLoopPlans(L_start, L_end)
3      for i ← 1 to Iterations do
4          plan ← SelectPlan(Plans)
5          BuildLoopPos(plan, C_init, 1)
6  bool BuildLoopPos(plan, C, pos_tri)
7      attempts ← 0
8      success ← false
9      while attempts < max_atts and success = false do
10         attempts ← attempts +1
11         tripeptide ← SampleTripeptide(plan, pos_tri)
12         tripeptide ← PerturbState(tripeptide)
13         C', success ← InstallTripeptide(C, tripeptide)
14         if success then
15             if pos_tri = plan.lastIndex then
16                 success ← CloseLoop(plan, C')
17             else
18                 success ← BuildLoopPos(plan, C', pos_tri +1)
19     return success
20 bool CloseLoop(plan, C)
21     attempts ← 0
22     success ← false
23     while attempts < max_IK and success = false do
24         attempts ← attempts +1
25         tripeptide_IK ← PerturbOmegas(plan, plan.lastIndex)
26         Solutions_IK ← SolveIK(C, tripeptide_IK)
27         foreach sol_IK ∈ Solutions_IK do
28             C', success_sol ← InstallTripeptide(C, sol_IK)
29             if success_sol then
30                 Ω ← Ω ∪ C'
31             success ← (success ∨ success_sol)
32         if success then
33             return success
34     return success
```

To validate the use of these states, an analysis was performed to find structural neighbors (Section 3.2.4.1).

3.2.2.3 Tripeptide placement constraints

When a tripeptide is appended to the loop being constructed within C, its acceptance is subject to two constraints. First, a common AI approach known as forward checking is employed to help improve performance [Russell 2009]. This approach consists first in recording the maximum length of a tripeptide from end to end when the database is loaded at the beginning of the process, for each amino-acid sequence key. Then, upon appending a tripeptide, the distance between the two working loop ends is measured. If this distance cannot be closed by concatenating the longest tripeptide states from the database for the remaining positions, the current loop configuration is considered invalid. This enables backtracking as early as possible in the construction process.

To avoid steric clashes in the final structure, the second constraint validates that the installed tripeptide's backbone atoms do not penetrate any of the van der Waals spheres in C. In MoMA-LoopSampler, the C_β atoms are placed simultaneously with the backbone, since their positions are fully determined by the dihedral angles of the backbone and the fixed bond lengths and angles. In order to eliminate conformations that do not leave room for side-chains, the van der Waals radii for the C_β atoms are artificially increased during the collision detection process. The radii are set depending on the type (and thus size) of the associated side-chains, as originally proposed by Levitt [Levitt 1976].

Another feature of MoMA-LoopSampler is the ability to add a constraint on the position of an atom. Although this feature is not showcased in this chapter, it may be interesting in situations where the position of an atom needs to be fixed (e.g. in order to contact another residue), or its position has been experimentally determined.

3.2.2.4 Loop closure

The function `CloseLoop` (Algorithm 1) attempts to close the loop by computing the dihedral angles of the last tripeptide (all the other tripeptides are held fixed during this process). The method starts by randomly sampling the three ω angles (function `PerturbOmegas`, line 25). First, a *cis* or *trans* configuration is selected for each ω angle. The probability to generate a *cis* ω angle is taken as the frequency of this event in the database for the corresponding residue type. Depending on the selected configuration, the angle is then sampled around the value 0 or π with a Gaussian distribution whose standard deviation follows that measured in the database for this type of residue.

The six remaining ϕ and ψ angles of the backbone are solved via an in-house IK solver [Cortés 2004] (function `SolveIK`, line 26), although the design of MoMA-LoopSampler allows for other inverse kinematics solvers to be employed

(e.g. [Manocha 1994, Dinner 2000, Coutsias 2004]). With only 6 degrees of freedom, this problem can be solved very efficiently in closed form and may yield up to 16 potential solutions. All the collision-free solutions are recorded, and the resulting loop conformations are added to the set of sampled loops. If no solution exists or if all of them are in collision, the process (ω sampling followed by IK) is repeated until at least one collision-free solution is found, or the maximum number of attempts, max_{IK} is reached. In the latter case, backtracking is employed and the construction process continues.

The IK solution yields tripeptide states that may not exist within the structural database. Section 3.2.4.2 describes a study showing that for the loops with a very low energy after relaxation, the IK solved tripeptide generally has a close structural neighbor within the database (Figure 3.4). This analysis also shows that upon loop relaxation, the distance of the final tripeptide to the database tends to lower, suggesting that the final tripeptide acts as a buffer that "absorbs" the rigidity of the other tripeptides by being more "lenient" on the distance to the database, and thus more flexible (Figure 3.5). Therefore, setting a threshold on the distance of the final tripeptide to the database may lead to the generation of higher quality loops, but care has to be taken in order not to be too restrictive on accepted loop conformations.

3.2.3 MoMA-LoopSampler with reinforcement learning

3.2.3.1 Objectives and principle

The loop construction method detailed so far effectively discretizes the conformational space of the loop by sampling from the structural database. For small loops (≤ 9 residues), an exhaustive search of all combinations can be completed within a few hours. However, larger loops (≥ 15 residues) present a formidable challenge, as computational requirements increase exponentially. In this section, a new method is proposed which utilizes reinforcement learning to improve the naive sampling strategy presented in Algorithm 1. The goal of utilizing RL for short loops is to provide a more efficient and exhaustive characterization of the loop. For longer loops, especially in highly constrained environments, RL can quickly prune infeasible areas of the search space, resulting in a more computationally efficient search.

3.2.3.2 Learning approach

A new approach that incorporates RL is shown in Algorithm 2. For each loop plan, a learning tree is built (function `ConstructRLTrees`, line 3). This data structure (presented in Section 3.2.3.3) records statistics about prior tripeptide state selection and their associated participation in successfully closed loops. On line 13, a new function is used to sample a tripeptide state (`SampleTripeptideRL`). This function uses statistics from the appropriate learning tree to guide tripeptide sampling towards zones that have a higher chance of generating successful conformations, or

Algorithm 2: Build Loop Reinforcement Learning

```
1  void BuildLoopRL(C_init, L_start, L_end)
2      Plans ← ConstructLoopPlans(L_start, L_end)
3      Trees ← ConstructRLTrees(Plans)
4      for i ← 1 to Iterations do
5          plan ← SelectPlan(Plans)
6          tree ← SelectTree(Trees, plan)
7          BuildLoopPosRL(plan, C_init, 1, tree)
8  bool BuildLoopPosRL(plan, C, pos_tri, tree)
9      attempts ← 0
10     success ← false
11     while attempts < max_atts and success = false do
12         attempts ← attempts +1
13         tripeptide ← SampleTripeptideRL(plan, pos_tri, tree)
14         tripeptide ← PerturbState(tripeptide)
15         C', success ← InstallTripeptide(C, tripeptide)
16         tree ← RecordSuccessPlacement(tripeptide, tree, success)
17         if success then
18             if pos_tri = plan.lastIndex then
19                 success ← CloseLoop(plan, C')
20             else
21                 success ← BuildLoopPosRL(plan, C', pos_tri+1, tree)
22     tree ← RecordSuccessClosure(tripeptide, tree, success)
23     return success
```

which have not been explored yet. Details about tripeptide state selection can be found in Section 3.2.3.4.

On line 16, the success or failure of the tripeptide placement is recorded in the learning tree to update the statistics about the tripeptide state's ability to form a successful loop (function `RecordSuccessPlacement`). Similarly, the closure of the loop (or absence thereof) is recorded by function `RecordSuccessClosure` (line 22), updating the statistics of all the tripeptide states used in the successfully generated loop sample.

3.2.3.3 Learning data structure

Tripeptides are projected into a space of low dimension m and organized in a m-dimensional tree data structure (see section 3.2.3.5). This structure is called an *octree* when $m = 3$. For the sake of simplicity, this m-dimensional tree data structure will be merely referred to as a *tree* throughout the chapter. For each tripeptide amino-acid sequence, all corresponding states from the database are projected to compute a bounding box $\mathcal{B}$ within the lower-dimensional space. Box $\mathcal{B}$ is divided

into 2^m sub-boxes, or *cells*, by passing orthogonal hyperplanes through its center. Each cell can then be subdivided using the same method, and so on, thus creating a tree structure. Final cells (those that are not subdivided) are called *leaves*. Initially, all trees are of depth 1, meaning that the highest level box is divided once, and all of its 2^m sub-boxes are leaves.

This data structure groups together tripeptide states that are close to one another in the chosen lower-dimensional space. Each leaf holds statistics about the group of tripeptide states it contains, namely their ability to form a closed loop without collisions. A leaf is subdivided (and thus is no longer a leaf) when the statistics about the states within it are too heterogeneous, indicating that these states have very different behaviors with regard to loop construction and closing (see Section 3.2.3.6). This new subdivision is aimed at separating the groups of tripeptide states into homogeneous groups with respect to their participation in successfully constructed loops.

The complete data structure involves chained trees. A loop plan gives the order in which the tripeptides will be built. In our implementation, plan i first builds tripeptides 1 to $i-1$, followed by tripeptides n to $i+1$ before ending with the final tripeptide i. Each loop plan is associated with a chain of m-dimensional trees organized as follows: the first tree (the *root tree*) contains the possible states for the first tripeptide to be built in the plan (the first tripeptide of the loop if $i > 1$, or the last one if $i = 1$). Each leaf within the first tree points to another tree containing the possible states for the second tripeptide to be built in the plan, whose leaves each point to another tree containing the states for the third tripeptide to be built, and so on until obtaining the trees for the penultimate tripeptide to be built.

For example, assume that tripeptides 1 to $k-1$ in the plan have been sampled and assembled so far, and that we are sampling a state for tripeptide k. For j from 1 to $k-1$, let us call c_j the state chosen for tripeptide j, and $\mathcal{L}_j$ the leaf that holds the statistics about state c_j. $\mathcal{L}_1$ is the leaf of the root tree containing c_1. $\mathcal{L}_2$ is the leaf of the tree pointed to by $\mathcal{L}_1$ containing c_2, and so on: $\mathcal{L}_{k-1}$ is the leaf of the tree pointed to by $\mathcal{L}_{k-2}$ containing c_{k-1}. The statistics about the sampling and placement of tripeptide k will be recorded in the tree $\mathcal{T}_k$ pointed to by $\mathcal{L}_{k-1}$.

$\mathcal{T}_k$ contains the states available in the database for tripeptide k. This structure stores statistics about the ability of tripeptide k's states to participate in forming a successful loop once the states for tripeptides 1 to $k-1$ have been chosen in leaves $\mathcal{L}_1$ to $\mathcal{L}_{k-1}$, respectively. Each leaf in $\mathcal{T}_k$ holds statistics about the attempts that were performed with the states it contains, i.e. the times when a state for tripeptide k was chosen in this leaf while the states for tripeptides 1 to $k-1$ had been selected in leaves $\mathcal{L}_1$ to $\mathcal{L}_{k-1}$, respectively. The statistics collected are: the number of times a state it contains could be placed while respecting all constraints (forward checking and collisions), the number of times it could not be placed, and the number of successfully closed loops containing tripeptide states within this leaf.

To reduce the size of the data structure, the tree is not expanded in dead-ends. Only the leaves containing at least one state that could be successfully placed (even if the loop it formed could not be closed) point to another tree for the next position.

3.2.3.4 Tripeptide state selection

When a state is sampled from the database for a given tripeptide, the statistics about the previously sampled states are used to guide the choice. The learning tree corresponding to the current tripeptide is selected given the states of the already placed tripeptides. A score is associated to each leaf of the tree. For the tests presented here, the score S for leaf $\mathcal{L}$ is set as:

$$S = \begin{cases} N \cdot S_{\max} & \text{if at least one state in } \mathcal{L} \text{ has been used to build a successfully closed loop} \\ N \cdot \min\left(S_{\max}, T\right) & \text{otherwise, with } T = \max\left(S_{\min}, \frac{t^{n-k}}{a}\right) \end{cases}$$

In this formula, $S_{\min}$ and $S_{\max}$ are parameters setting lower and upper limits on the score, respectively. N is the number of tripeptide states in $\mathcal{L}$, and a is the number of times a state from $\mathcal{L}$ has been sampled earlier in the sampling process. k is the position of the tripeptide in the plan and n is the number of tripeptides in the loop. Finally, t is a positive real number parameter setting the learning rate (lower values of t correspond to higher learning rates, i.e. to a greedier learning process). It is an important parameter as it is determinant for the diversity of the loop ensemble and the speed at which loops will be generated. The influence of t is investigated in Section 3.3.3.

When sampling a state for a given tripeptide with `SampleTripeptideRL` (Algorithm 2, line 13), a leaf is randomly picked among all the tree's leaves using the probabilities corresponding to the normalized scores. Then, a state is selected from the selected leaf by uniform random sampling.

T acts as a threshold score. Ideally, the score of leaves containing no working states (successful loop closures) should be zero. In practice, it is impossible to guarantee that no state in a leaf is able to lead to a successful loop, even when those have all been tested (because of the small perturbations and the fact that the tripeptide states sampled upstream may be different from one attempt to another). Therefore, this method maintains the score so that the leaf has a non-zero chance of being explored even when it has failed to lead to a successful loop closure. After each failed attempt, the threshold score decreases until reaching a lower limit. However, as soon as a state is found that leads to a successful loop, the score is set back to its maximum value, which it maintains for the remainder of this search.

Note that this scoring approach does not involve a cumulative reward, characteristic of standard RL methods. Therefore, it would be more accurate to refer to it as a RL-based heuristic method rather than a RL method. Nevertheless, we prefer this abuse of language for the sake of simplicity. Other scoring approaches, possibly involving cumulative reward, can be applied within this method. However, among the options we tested, this score is the one that best preserved the diversity among sampled conformations.

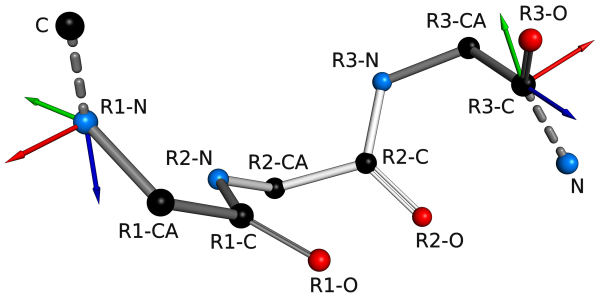


Figure 3.1. Frames attached to the beginning and the end of a tripeptide. R1, R2 and R3 designate the first, second and third residues of the tripeptide, respectively. The x-, y- and z-axes are represented in red, green and blue, respectively. Only backbone atoms are represented for clarity. The C atom of the preceding residue and the N atom of the following residue are also visible.

3.2.3.5 Tripeptide projection

The definitions of the different tripeptide state projections used to organize the tripeptides in the learning data structure are based on the tripeptide geometry obtained after applying the nine dihedral angles constituting the state. We associate a reference frame to the beginning and the end of each tripeptide. The frame associated to the beginning of a tripeptide is centered on the N atom of the first residue, while the frame associated to the end is centered on the C atom of the last residue (see Figure 3.1). The length of a tripeptide is defined as the distance between the first and the last atom of the tripeptide backbone.

Several options were considered for tripeptide state projection. A comparative analysis of the different options, detailed in Section 3.2.4.3, motivated the choice of the projection named *Position*. This projection is the vector of the translational part of the transformation between the beginning and the end of the tripeptide (as defined by the associated frames). It corresponds to the coordinates of the last C atom of the tripeptide backbone in the frame attached to the first N atom. It is a projection in dimension $m = 3$.

3.2.3.6 Leaf subdivision

A leaf in the learning structure can be split if the results obtained for the tripeptide states it contains become too heterogeneous with regard to construction success. The states contained in a leaf should be similar enough to have comparable levels of placement and subsequent loop closing success. When the results show otherwise, the leaf is split around its center. Since each placement event (successful placement of a tripeptide, steric clash, loop closure) is recorded along with its corresponding tripeptide state, it is easy to recompute the statistics in each of the 2^m newly created child leaves in case the leaf splits.

Each child leaf obtained after a split is assigned a new tree corresponding to the next position in the plan. Consequently, the statistics from the formerly pointed to tree are also distributed among the newly created trees and their leaves. After this process, the scores of all newly created leaves obtained after the subdivision are the same as what would have been obtained if the tree had utilized this structure from the beginning of the sampling process.

Typically, a leaf split can happen when a state leads to a steric clash between 25% and 75% of the time. Splitting can help isolate the states that are responsible for steric clashes, while regrouping the states that lead to a successful placement. Another criterion for splitting the leaf is based on the frequency of forming a closed loop.

3.2.4 Analysis and consistency of the sampling methods

3.2.4.1 Tripeptide database analysis

Employing a database of protein configurations to discretize the sampled conformational space capitalizes on the prior knowledge that the backbone dihedral angles only occupy a limited range of values, which are dependent on their neighboring amino acids. The technique of sampling from databases has been utilized in many structural biology problems, including loop sampling and de novo structure prediction.

While the success of these applications may show this approach has merit, we extend this idea by validating the resulting tripeptide states that are formed by joining two tripeptides into a structure. Note that these tripeptides may be concatenated in one direction or in another, so that both directions have to be tested. Figure 3.2 shows two tripeptide sequences (MVK and PGT) that have been extracted from our database and joined together to construct a larger protein structure. The red lines highlight new tripeptides that are formed from the overlaps. We refer to the states of these tripeptides as synthetic states, since they were not sampled from the database, and thus, their structural validity is unknown.

We propose validating these synthetic states to strengthen the theoretical basis for our proposed approach. This validation occurs as follows. For each pair (i, j) of tripeptides in the database, we extract 4 synthetic states (as shown in Figure 3.2), formed by the concatenations ij and ji. Next, for each synthetic state, we search the database using the resulting amino acid keys. For the example in Figure 3.2, these keys would be VKP, KPG, GTM, and TMV respectively. If we are able to locate a similar state in the database, we label the synthetic state as valid. Given that our database is built from a small subset of the protein universe, we can not say anything about synthetic states for which we do not find a similar neighbor.

In this work, similarity is measured as the RMSD of the 3 sets of ϕ, ψ and ω backbone dihedral angles that define each tripeptide state. Similarity is established when the RMSD is less than some threshold ε.

Figure 3.2. Synthetic tripeptide states creation by sampling two states from the database. The four red lines represent the 4 synthetic tripeptide states extracted.

The database validation analysis was performed with a threshold of 0.5 radians for the dihedral RMSD. To speed up the process, only 1 out of 1,000 synthetic states were randomly selected and tested for a close neighbor. This represents around 9.5 billion synthetic states tested. On average, 85 % of synthetic states had a close neighbor in the database. Figure 3.3 gives more precise results by sequence. Unsurprisingly for a database of tripeptides involved in coils, sequences containing glycines are the most populated. However, sequences containing rarer amino acids like cysteines, histidines, methionines and tryptophans contain much fewer states. The distribution of synthetic states with a close neighbor in the database shows that sequences containing these amino acids are also the ones for which the proportion of synthetic tripeptides with a close neighbor in the database are the lowest. This may point to a lack of data for these relatively rare sequences. This fact is also supported by Figure 3.3(d), in which sequences with few representatives in the database are also the ones for which the average distance of synthetic states to the database is the highest.

This analysis contributes to the validation of the method consisting in concatenating the tripeptides to build the loop. Indeed, most synthetic states without a close neighbor in the database actually coincide with a rare tripeptide sequence for which data is insufficient. However, the results reveal a limitation of the method, which is the availability of experimental data for rare sequences. Future improvements of MoMA-LoopSampler will therefore include enriching the database with additional states, potentially sharing states across similar sequences.

Cases in which the absence of close neighbors for a synthetic state is not due to lack of data, but to their very low probability to exist in physiological conditions, do not constitute an important issue if one wants to exhaustively sample the conformational space. Generating loops with such states does not prevent any acceptable conformations from being sampled. Conversely, it might be an issue if one's goal is to preserve the distribution of structural preferences encoded in the database when generating a loop ensemble. Nevertheless, generating statistically-meaningful conformational ensembles goes beyond the scope of this chapter.

3.2.4.2 IK-solved tripeptide: distance to database and loop quality

The state of the last tripeptide used to close the loop has not been sampled from the database. In this section, we investigate how its distance to the database relates to the quality of the sampled conformation. We define the distance between this tripeptide state and another state in the database as the RMSD of their nine

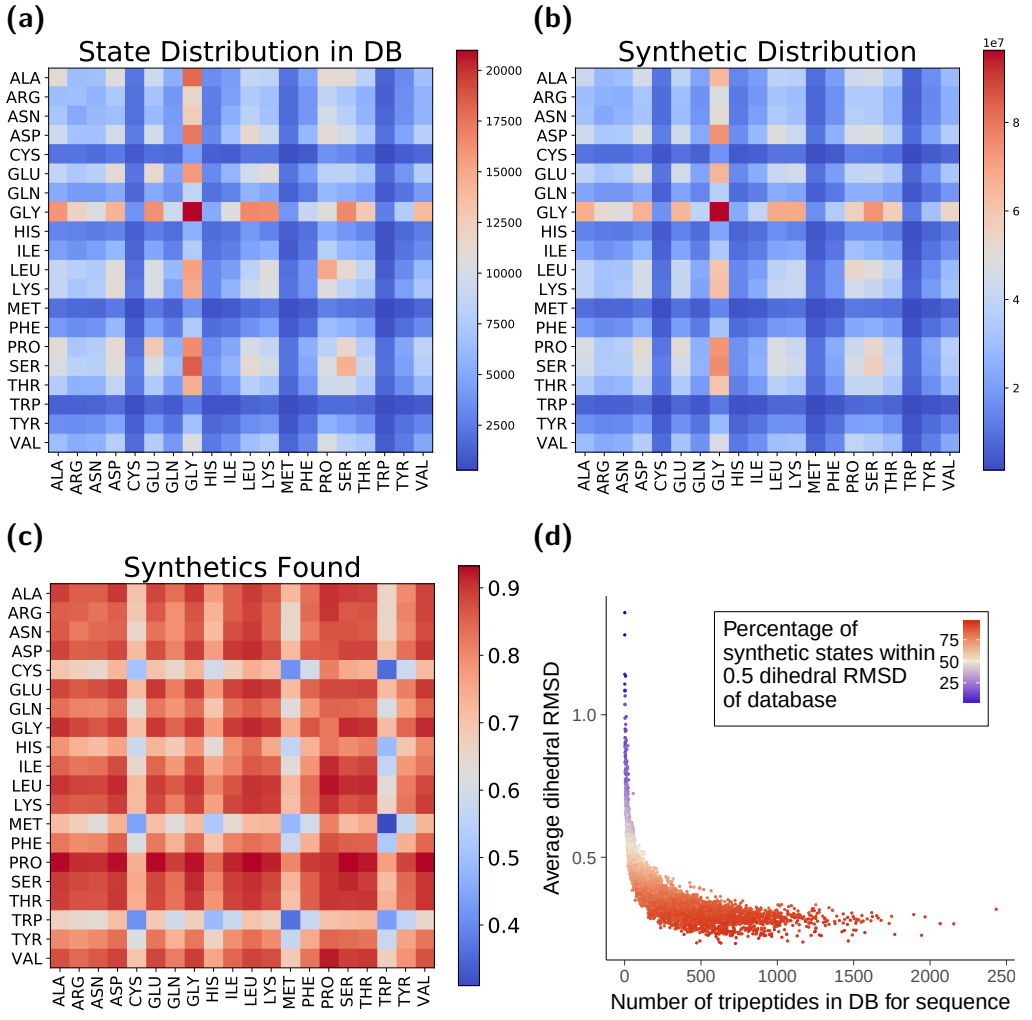


Figure 3.3. Each tripeptide state in the database in projected into a 20 by 20 map based on its first two amino acid residues. (a) The top left heatmap shows the distribution of tripeptides using this projection for the SCOP 95% similarity database filtered to exclude states that participate within a secondary structure element. This database contains 2.2 million states. (b) The top right plot is the distribution of synthetic states created by concatenating tripeptide pairs in the database (approximately 9.5 billion states). (c) The bottom left plot shows for each of the synthetic states created, what percentage of these had close structural neighbors in the original database (neighbor distance < 0.5 dihedral RMSD). (d) The bottom right plot shows for each tripeptide sequence the average dihedral RMSD to database obtained for the tested synthetic states depending on the number of actual states in the database.

backbone dihedral angles. The distance of the last tripeptide state to the database is defined as the lowest distance between this tripeptide state and a state in the database for the corresponding amino-acid sequence.

For two loop systems, we performed brute force searches of loop conformations with side-chain placement. We then measured the distance to the database for the closing tripeptide state of each sampled conformation and recorded the total energy of the system for each loop conformation using AMBER's ff14SBonlysc force field [Maier 2015] and a simple Generalized Born implicit solvent model ($igb = 1$ and the *mbondi* radii sets, as recommended in the AMBER manual). The generated samples were then relaxed using AMBER 16 biosimulation package [Case 2005, Case 2016], and their energy after relaxation was measured again. The new distance to database of the relaxed closing tripeptide state was also calculated.

Figure 3.4 shows the relationship between the energy of generated loops and the distance of their closing tripeptide to the database. We see clear positive correlation between energy and the distance of the closing tripeptide state to the database before relaxation. After relaxation, the profile changes considerably. Although the correlation is still present, it is mainly apparent for low energy loop samples. Setting a threshold on the distance to database for the closing tripeptide state would thus have to be done very carefully if one does not want to exclude statistically-likely conformations.

Figure 3.5 shows the distribution of the distance to database differences after minus before relaxation for the closing tripeptide. Results show that the distribution is skewed to negative values, indicating a tendency to lower the distance of the closing tripeptide state to the database upon relaxation. This suggests that, by being unconstrained with regard to the database, the closing tripeptide acts as a buffer that absorbs the rigidity of the other tripeptides. Relaxation causes the other tripeptides to relax and the closing tripeptide sate to move closer to states in the database.

Setting a threshold on the distance of the closing tripeptide state to the database is delicate since even loops with closing tripeptides far from the database may have a low energy. Therefore, such a threshold should not be set too high. Moreover, looking at Figure 3.4(d), we can see that even after relaxation, a loop conformation with a closing tripeptide state at distance 1 radian of any state in the database may be relatively likely compared to the generated loop ensemble. In our method, no threshold was set on the distance of the closing tripeptide state to the database. This choice follows from the objective to exhaustively sample the loop's conformational space.

3.2.4.3 Comparison of tripeptide projections

When using reinforcement learning, tripeptide states are organized into an m-dimensional tree according to a pre-defined projection of the tripeptides. Choosing an appropriate tripeptide state projection is of crucial importance to obtain an effective RL. If the projection is ill-chosen, it may group tripeptide states that yield

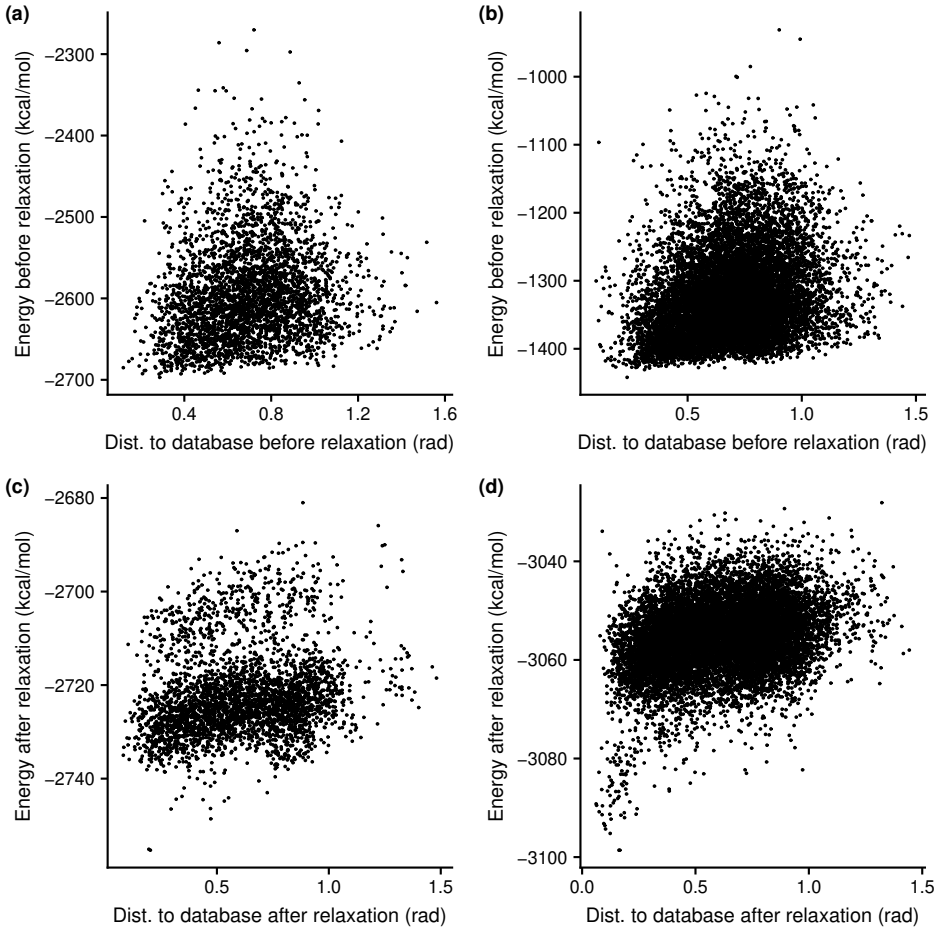


Figure 3.4. AMBER energy vs. distance to database for the closing tripeptide state. (a) and (c) are the graphs for loop 45 (9 residues) while graphs (b) and (d) are the graphs for loop 31 (9 residues). (a) and (b) show the energy and distance before relaxation while (c) and (d) show the energy and distance to database after relaxation.

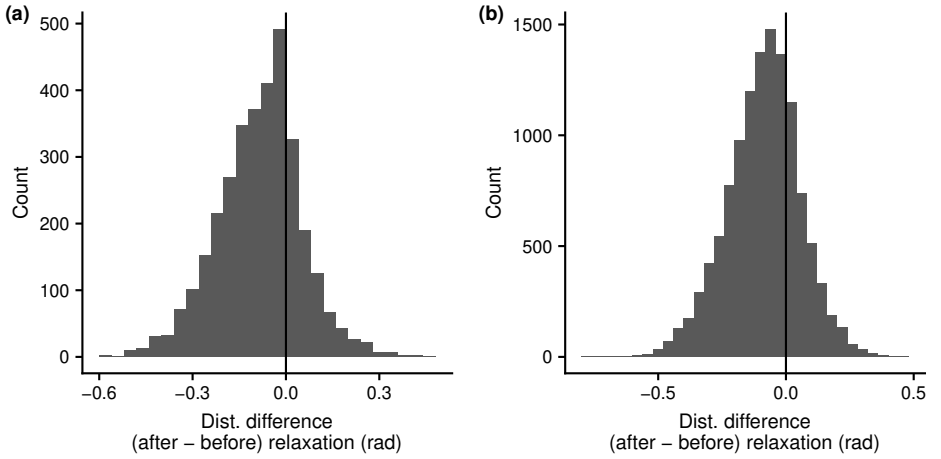


Figure 3.5. Histogram of distance to database difference after and before relaxation. Negative values indicate that the distance to database lowered upon relaxation. (a) Loop 45. (b) Loop 31.

heterogeneous outcomes with respect to loop closure, in which case no correct assumption can be made of one state even if the success of other close states in the projection space is known. The idea of the tripeptide projection is thus to group similar states together so that a state's success can be inferred from the results obtained by its neighbors.

Several criteria were considered in the choice of tripeptide state projection. A good projection should provide a good distribution of the tripeptide states in space, and more importantly group together tripeptide states that are almost interchangeable in the loop building process. For example, if two states are very close to one another in the projection space and one is in a collision when placed, then the other should have a strong likelihood to be in collision as well. Besides, a high value of the dimension m would result in a data structure that has very high memory requirements, with undesirable sparsity in the leaves. Indeed, when a tripeptide state is sampled, its results are exploited to make predictions about the success of all its neighbors in the leaf. If tripeptide states are isolated in the data structure, reinforcement learning will not be very effective.

Several options were tested for tripeptide state projection, and compared on the basis of the previous criteria. These projections mostly involve the relative position or orientation of the two ends of the tripeptide. For the orientation, several representations were tested. More precisely, the tested tripeptide state projections are (with the associated dimension m in parentheses):

Position ($m = 3$): The vector of the translational part of the transformation between the beginning and the end of the tripeptide (as defined by the associated frames).

Euler angles ($m = 3$): The vector of the Euler angles of the rotational part of the transformation between the beginning and the end of the tripeptide.

Euler angles and length ($m = 4$): The vector containing the three Euler angles of the rotational part of the transformation between the beginning and the end of the tripeptide and the length of the tripeptide.

Quaternion ($m = 4$): The quaternion representing the rotational part of the transformation between the beginning and the end of the tripeptide.

Quaternion and length ($m = 5$): The vector containing the quaternion representing the rotational part of the transformation between the beginning and the end of the tripeptide, and the length of the tripeptide.

Axis-angle ($m = 4$): The axis-angle representation of the rotational part of the transformation between the beginning and the end of the tripeptide.

Axis-angle and length ($m = 5$): The axis-angle representation of the rotational part of the transformation between the beginning and the end of the tripeptide, and the length of the tripeptide.

We ran brute force searches for six loop systems using each of the different projections. We compare the first level of the octrees in all the runs performed, in terms of distribution of tripeptides and success probability of each cell. The first level of the octrees corresponds to the cells obtained after dividing the bounding box containing all tripeptide states once in each dimension. The number of cells at this level is thus 2^m where m is the dimension of the projection. Our method employs several loop construction plans for each loop system (one per tripeptide in the loop, corresponding to a plan ending with this tripeptide). There are therefore 3 plans for the 9-residue loops, and 4 plans for 12-residue loops. The success probability of a cell is defined as the number of successful combinations of tripeptides using a state from this cell, divided by the theoretical number of tripeptide combinations that use a state from this cell.

Figure 3.6 shows the results for the first levels of all loop systems, for each of their loop construction plan employed. Looking at the different heatmaps, it is clear that no projection performs consistently better than all others. However, we will try to analyze the differences in the results. First, looking at the 3-dimensional projections *Position* and *Euler angles*: tripeptide state distributions seems to be satisfying in both cases, with very few empty cells. However, the distribution of solutions is more heterogeneous with *Position*. Indeed, with this projection, more of the first level cells are void of working solutions. This is particularly striking for 1dim-12 (A213-A224) (loop 55), for the three last plans. *Euler angles* is not able to gather the solutions into only a few cells, whereas *Position* concentrates the solutions into two or three cells. The same observation can be made for the other systems. *Position* therefore seems to be a better predictor of tripeptide success in building a loop than *Euler angles*. Comparing the 4-dimensional projections is more delicate. Indeed, *Axis-angle*, *Euler angles and length* and *Quaternion* behave differently between the different systems. It seems that *Euler angles and length* better gathers the solutions for 1dim-12 (A213-A224) (loop 55) than other projections, but on 153l-12 (A98-A109) (loop 61), *Quaternion* is the one that better gathers solutions, while *Axis-angle* gathers solutions relatively well in all the cases. For the distribution of tripeptides, the same observation can be made. The quality of the distribution differs depending on the tripeptide sequence. However, 4-dimensional projections do not seem to perform better than *Position*. For most systems and most loop plans, there are more than 8 cells left with solutions. Of course those cells contain fewer tripeptide states on average, which is why comparing 3-dimensional and 4-dimensional projections is delicate. In 5-dimensional projections, the distribution of tripeptides results in an undesirable sparsity, where the solutions are unsurprisingly found where tripeptides are located. Therefore these do not stand out from all the projections either.

This comparison is only possible on first level cells with these plots. Many other levels are left to explore and each projection is likely to better separate working and non working states in lower levels. Based on our analysis, we decided to retain a 3-dimensional projection in order to limit the size of the learning tree in memory.

In light of the results, we thus selected *Position* as the tripeptide projection for our tests.

Figure 3.6. Comparison of projections. Several different projections were tested for reinforcement learning. The following heatmaps show the ability of each projection to distribute tripeptides and to regroup states that succeed in creating a closed loop together. The results are presented for 6 different systems. For each system, the heatmaps of the learning trees corresponding to the different loop plans are shown. A loop plan is designated by the position of the last tripeptide used to close the loop (IK position). The first of the two heatmaps for a given learning tree gives the distribution of tripeptides in the top level cells of the root tree. The second shows the success probability of the leaf, meaning the number of successful loop conformations that start with a state from this leaf divided by the theoretical number of tripeptide combinations starting with a state from this leaf.

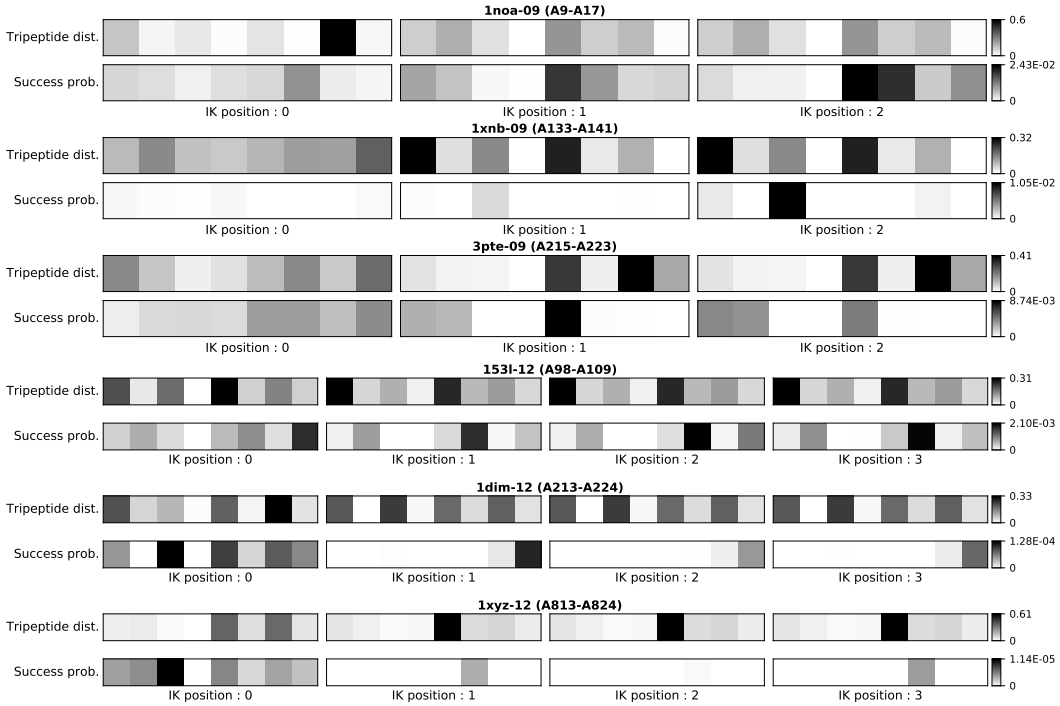


Figure 3.6(a): Position

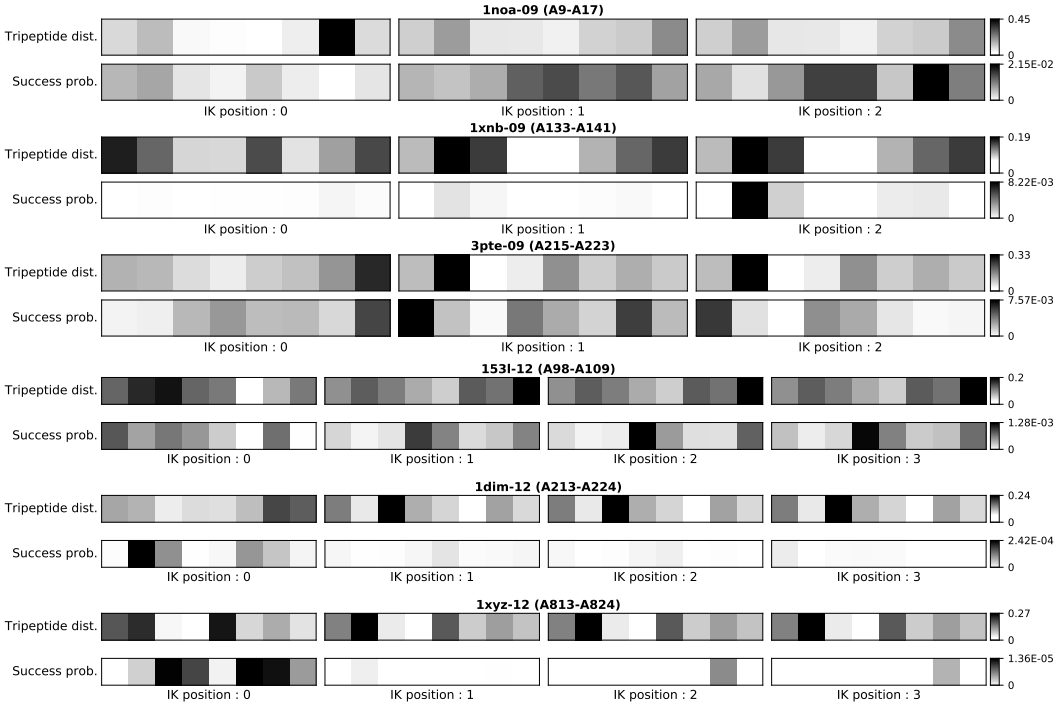


Figure 3.6(b): Euler angles

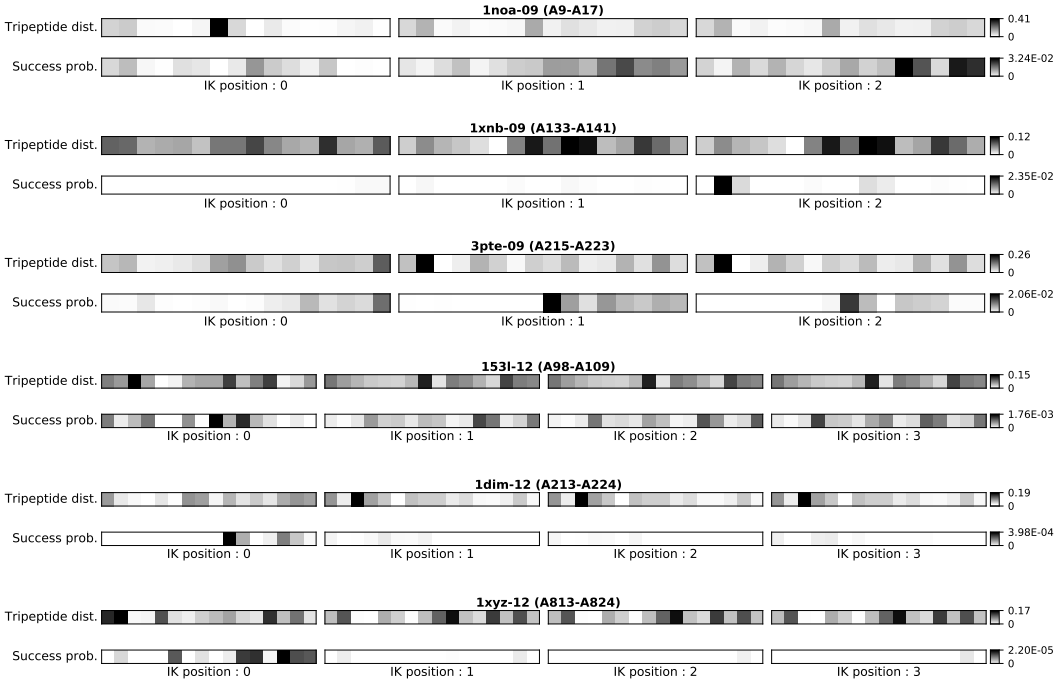


Figure 3.6(c): Euler angles and length

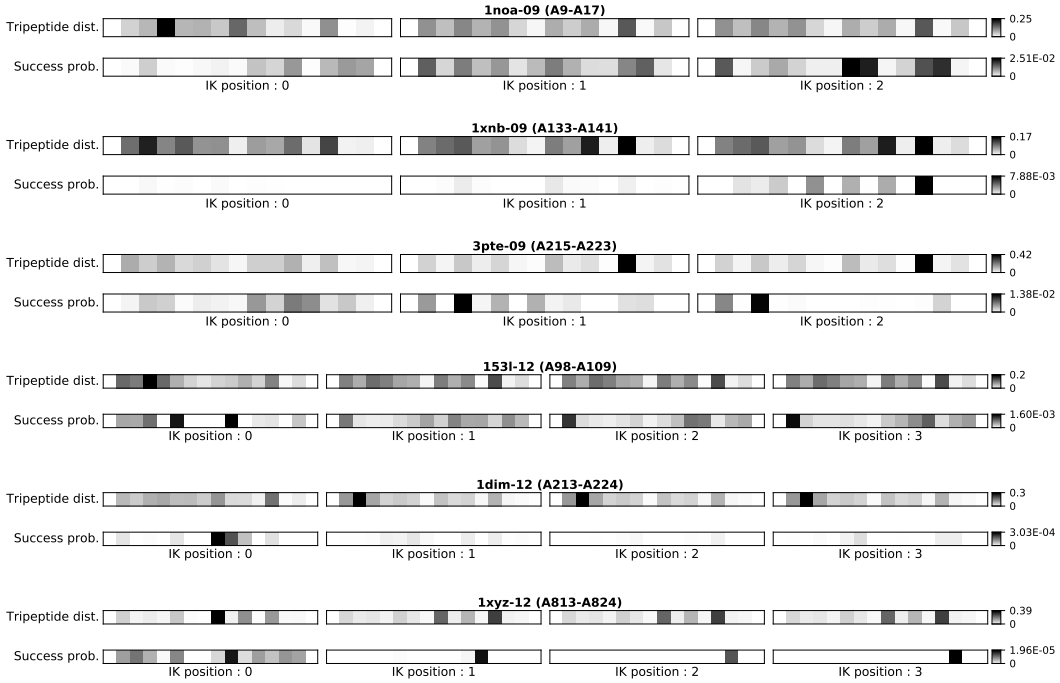


Figure 3.6(d): Quaternion

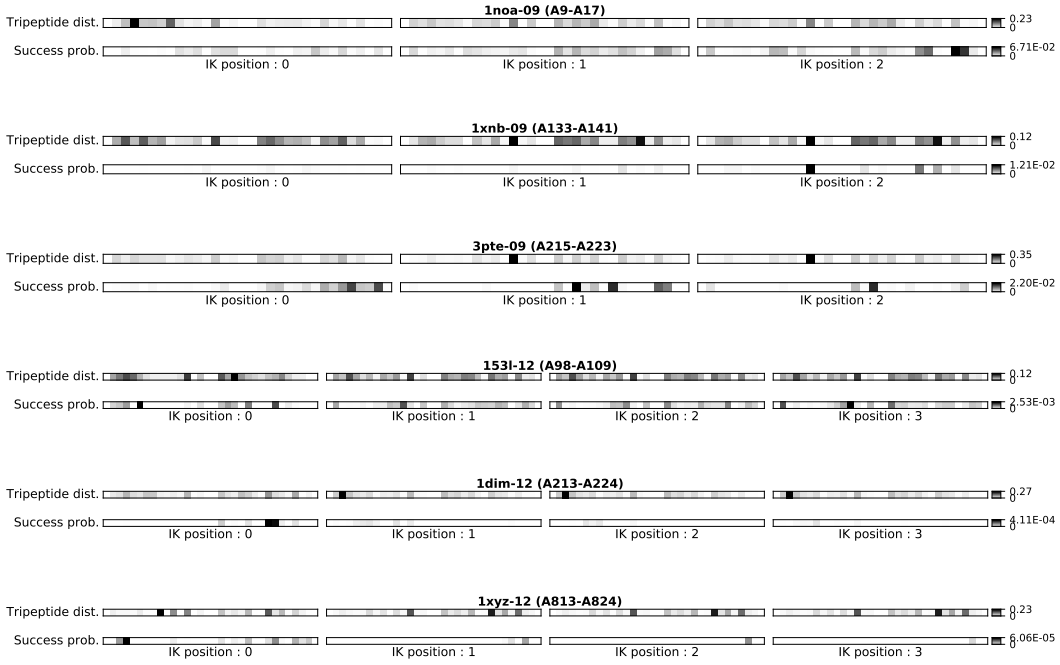


Figure 3.6(e): Quaternion and length

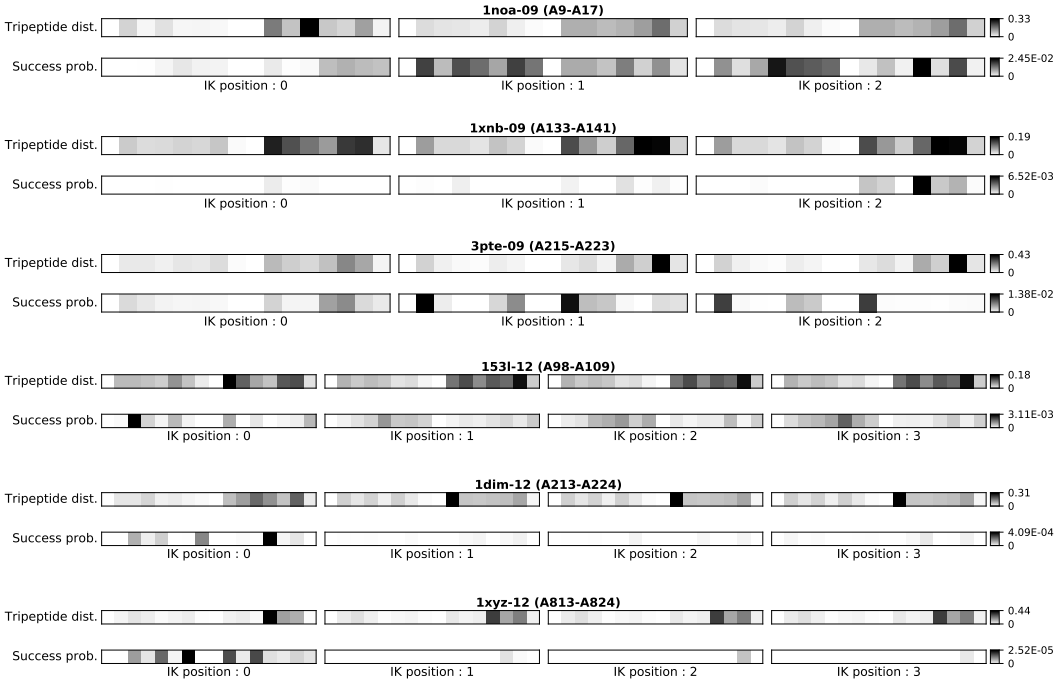


Figure 3.6(f): Axis-angle

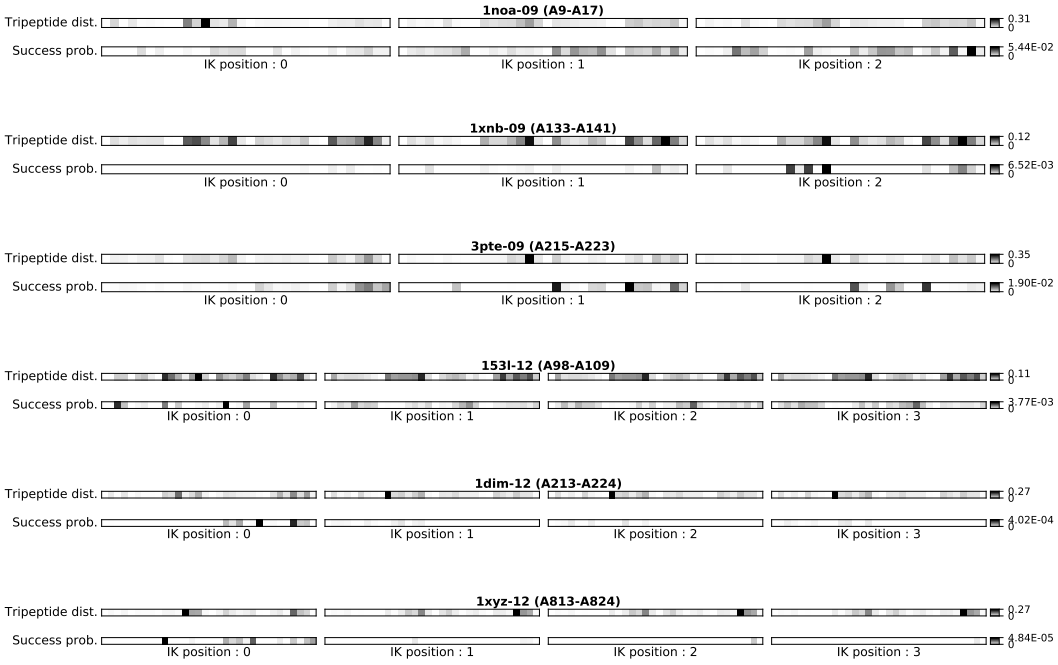


Figure 3.6(g): Axis-angle and length

3.3 Results and discussion

We applied our loop modeling framework to a few benchmark sets of proteins. We first present the results obtained without activating reinforcement learning, showcasing the ability of basic MoMA-LoopSampler to sample near-native loops, its computational efficiency, and its ability to sample intermediate states along the transition path between two conformations. The performance of reinforcement learning is then detailed.

In all that follows, the "native" loop is the name given to the loop conformation that is provided from the Protein Data Bank. Nevertheless, it is important to stress that due to their inherent flexibility, protein loops rarely have one single stable conformation. MoMA-LoopSampler is precisely aimed at discovering every conformation that the loop may adopt.

The distance to native is used as a means to compare MoMA-LoopSampler to other sampling methods whose final purpose it to "predict" the most probable loop conformation. This distance is also used in our analysis of the diversity of the ensemble of generated loops. An absence of the native conformation from the generated ensemble would indicate that the conformational ensemble is not fully sampled by the method. Conversely, the ability to sample the native loop conformation is by no means evidence that the method is capable of sampling all relevant conformations. However, it is an encouraging indication in that direction. We define the RMSD_{min} of a loop ensemble as the lowest backbone RMSD between a loop conformation from the ensemble and the native loop conformation, after alignment of the fixed portion of the protein.

3.3.1 Tests performed and visualization of results

3.3.1.1 Brute force exploration

A brute force variant of MoMA-LoopSampler was developed for evaluation purposes. By combining all the possible tripeptide states from the database, this variant allows to capture the full conformational ensemble that is accessible under the geometric constraints enforced by MoMA-LoopSampler. This variant has two main applications: (1) to evaluate the exhaustiveness of the database, and (2) to verify the distributions of conformations sampled using the basic and RL versions of MoMA-LoopSampler, and compare them to the reachable space.

Algorithm 3 shows this Brute Force version of MoMA-LoopSampler. Since the construction method perturbs the sampled tripeptides, every brute force search in our tests is carried out three times to obtain a set of loop conformations that is representative of the conformational space reachable by MoMA-LoopSampler. Reinforcement learning (RL) can be activated to obtain the final learning tree and analyze the distribution of the solutions.

Algorithm 3: Build Loop Brute Force

```
1  BuildAllLoops(C_init, L_start, L_end)
2      Plans ← ConstructLoopPlans(L_start, L_end)
3      Trees ← ConstructRLTrees(Plans)
4      foreach plan ∈ Plans do
5          tree ← SelectTree(Trees, plan)
6          BuildLoopsPosBF(plan, tree, C_init, 1)
7  int BuildLoopsPosBF(plan, tree, C, pos_tri)
8      nb_sols ← 0
9      Tripeptides ← GetAllTripeptidesStates(plan, pos_tri)
10     foreach tripeptide ∈ Tripeptides do
11         tripeptide ← PerturbState(tripeptide)
12         C', success ← InstallTripeptide(C, tripeptide)
13         tree ← RecordSuccessPlacement(tripeptide, tree, success) if
               success then
14             if pos_tri = plan.lastIndex then
15                 success ← CloseLoop(plan, C')
16                 tree ← RecordSuccessClosure(tripeptide, tree, success)
17             else
18                 nb_closed ← BuildLoopsPosBF(plan, C', pos_tri +1)
19                 nb_sols ← nb_sols + nb_closed
20                 tree ← RecordSuccessClosureNb(tripeptide, tree, nb_closed)
21     return nb_sols
```

3.3.1.2 Test sets

The performance of our method was tested on three benchmark sets of 9-residue loops, 12-residue loops and 15-residue loops. The 9-residue test set is a subset of the loops gathered by Jacobson and colleagues [Jacobson 2004]. 2alp-139 and 8ruc-79 were removed because they are not included in the set by Soto and co-workers [Soto 2008]. 1ivd-244 and 1pda-108, which were excluded from the modified Fiser set by DePristo and colleagues for either poor quality or missing side-chain atoms and gaps, were also removed [DePristo 2003]. Finally, 4gcr-94 was removed because it only contains 3 turn residues surrounded by β-sheet and α-helix residues, and is not strictly speaking a loop. Thus, the 9-residue test set involves 53 loops. The 12 residue test set contains the ten 12-residue loops gathered by Jacobson and colleagues [Jacobson 2004], and the 15 residue test set contains the 30 15-residue loops used in the analysis by Zhao *et al.* [Zhao 2011]. The list of loops utilized in the tests as well as the corresponding identifiers employed throughout the chapter can be found in Tables 3.1 to 3.3.

Table 3.1. 9-residue Test Set

# Loop	Corresponding PDB numbering
1	3pte-09 (A107-A115)
2	1xyz-09 (A795-A803)
3	1lkk-09 (A193-A201)
4	1mla-09 (A194-A202)
5	2ayh-09 (A169-A177)
6	1arb-09 (A90-A98)
7	1mrp-09 (A284-A292)
8	1cse-09 (E95-E103)
9	1tca-09 (A217-A225)
10	2eng-09 (A172-A180)
11	1xyz-09 (A568-A576)
12	1aba-09 (A69-A77)
13	1nif-09 (A266-A274)
14	1arp-09 (A127-A135)
15	1noa-09 (A99-A107)
16	1lkk-09 (A142-A150)
17	1xif-09 (A59-A67)
18	1rhs-09 (A216-A224)
19	1fus-09 (A91-A99)
20	1php-09 (A91-A99)
21	2cpl-09 (A24-A32)
22	1nls-09 (A131-A139)
23	1xnb-09 (A133-A141)
24	1btl-09 (A102-A110)
25	1mrj-09 (A92-A100)
26	1gpr-09 (A63-A71)
27	1csh-09 (A252-A260)
28	1sgp-09 (E109-E117)
29	3pte-09 (A78-A86)
30	1isu-09 (A30-A38)
31	1noa-09 (A9-A17)
32	2hbg-09 (A18-A26)
33	1nfp-09 (A12-A20)
34	1pgs-09 (A117-A125)
35	1tca-09 (A170-A178)
36	1ptf-09 (A10-A18)
37	1npk-09 (A102-A110)
38	3pte-09 (A215-A223)
39	1ra9-09 (A142-A150)
40	1mrk-09 (A53-A61)
41	1wer-09 (A942-A950)
42	3tgl-09 (A56-A64)
43	2sil-09 (A183-A191)
44	1amp-09 (A57-A65)
45	1aac-09 (A58-A66)
46	1arb-09 (A168-A176)
47	1fus-09 (A31-A39)
48	1byb-09 (A246-A254)
49	1xnb-09 (A116-A124)
50	1ede-09 (A257-A265)
51	1aru-09 (A36-A44)
52	1onc-09 (A70-A78)
53	1noa-09 (A76-A84)

Table 3.2. 12-residue Test Set

# Loop	Corresponding PDB numbering
54	1arb-12 (A74-A85)
55	1dim-12 (A213-A224)
56	1xyz-12 (A813-A824)
57	1bkf-12 (A9-A20)
58	2ayh-12 (A21-A32)
59	1akz-12 (A181-A192)
60	1luc-12 (A158-A169)
61	153l-12 (A98-A109)
62	1cex-12 (A40-A51)
63	1ixh-12 (A160-A171)

Table 3.3. 15-residue Test Set

# Loop	Corresponding PDB numbering
64	2v3v-15 (A382-A396)
65	1qqf-15 (A1112-A1126)
66	1h4a-15 (X19-X33)
67	2aeb-15 (B156-B170)
68	3a3p-15 (A286-A300)
69	1wui-15 (L454-L468)
70	1qaz-15 (A298-A312)
71	3css-15 (A95-A109)
72	3a64-15 (A350-A364)
73	2o2k-15 (A1220-A1234)
74	1ju3-15 (A486-A500)
75	2h3l-15 (A1339-A1353)
76	2cjp-15 (A58-A72)
77	1ryo-15 (A172-A186)
78	1ah7-15 (A157-A171)
79	1s95-15 (A477-A491)
80	1ra0-15 (A361-A375)
81	1wb4-15 (A1033-A1047)
82	1y12-15 (A10-A24)
83	3f1l-15 (A99-A113)
84	2pkf-15 (A26-A40)
85	2oit-15 (A290-A304)
86	2dsj-15 (A354-A368)
87	1bhe-15 (A121-A135)
88	3ea1-15 (A136-A150)
89	2b0t-15 (A701-A715)
90	1ra0-15 (A283-A297)
91	1zhx-15 (A392-A406)
92	3bb7-15 (A231-A245)
93	3bf7-15 (A49-A63)

3.3.1.3 Test parameters

In all tests, the max_{atts} parameter was set to 10 for 9-residue loops, 7 for 12-residue loops and 5 for 15-residue loops. The max_{IK} parameter was set to 100, and the maximum time for the `BuildLoopPos` and `BuildLoopPosRL` functions was set to 20 seconds. Two atoms separated by more than 3 bonds were considered in collision if the distance between them was below 0.7 times the sum of their van der Waals radii [Bondi 1964]. This threshold is called the van der Waals scaling factor.

For runs performed with RL, *very low learning rate* results correspond to the results obtained with parameter t (see Section 3.2.3.4) set to 15, *low learning rate* results with $t = 10$, *high learning rate* results with $t = 2$, and *very high learning rate* results with $t = 1$.

3.3.1.4 2D loops projections

A convenient way to visualize the sampled regions in the conformational space is to plot two-dimensional projections based on two meaningful descriptors for each loop. The generated plots can give insight into the density of the sampled conformations in different regions. It is especially convenient to compare the conformational ensembles obtained under two different conditions.

The first chosen descriptor (x-axis) is the distance between an atom located in the middle of the loop and a fixed atom in the protein. The second descriptor (y-axis) is the angle formed by three atoms: an atom at approximately one quarter, one half, and three quarters of the way down of the loop. The first descriptor (d_1) is given in Ångström (Å), while the second descriptor (d_2) is given in degrees (°).

3.3.2 Results obtained without reinforcement learning

3.3.2.1 Distance to native loop

To test the ability of MoMA-LoopSampler to sample loop configurations close to the native state, we ran a series of tests on each test dataset. The run time for each test was determined by the loop length being sampled: 2 hours for 9-residue loops, 4 hours for 12-residue loops, and 6 hours for 15-residue loops. Each experiment was repeated 4 times. [1]

Figure 3.7 gives the distribution of the lowest RMSD to the native loop for both the best and worst of the four executions. In all four tests, at least one conformation within 2 Å of native (using backbone RMSD as a distance metric) was sampled for each of the 9- and 12-residue loops, and for a minimum of 23 (up to 26 depending on the test) of the 15-residue loops (out of 30 loops). Decreasing the threshold to test for a sample within 1 Å of the native state, the results are still very good for 9- and 12-residue loops: with 51 9-residue loops (out of 53) and 9 12-residue loops (out of 10). For the 15-residue loops however, this number drops to 9 loops

[1] Computing times reported throughout this chapter correspond to runs on a single core of a 2.30 GHz Intel® Xeon® E5-2695 v3 processor.

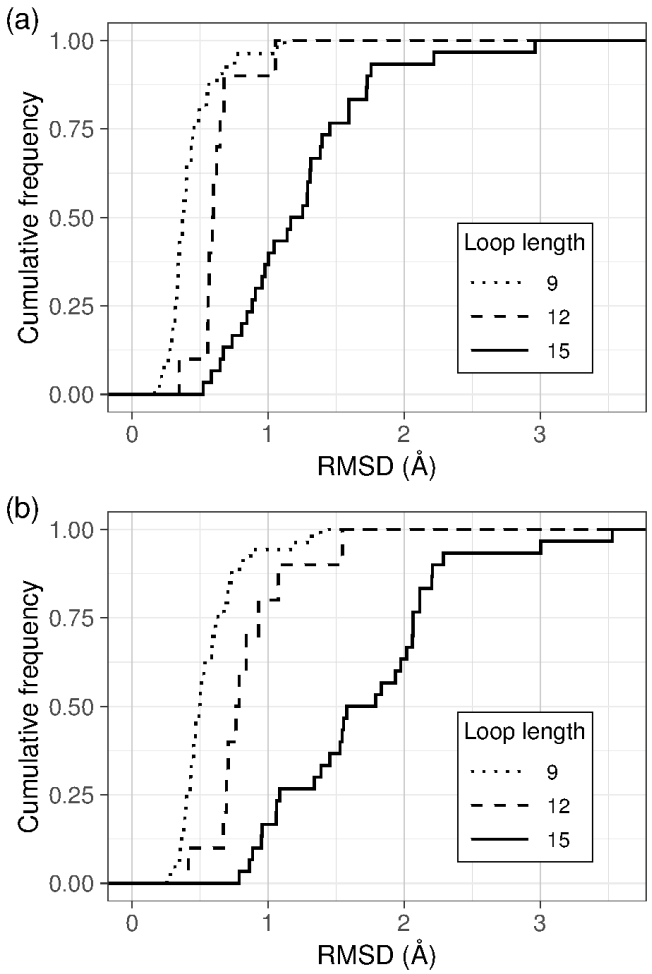


Figure 3.7. Cumulative distribution of the lowest backbone RMSD to native among sampled conformations for the 93 loops in our test sets. Since sampling was performed four times, (a) shows the distribution of the best of the four results for each loop, while (b) shows the distribution of the worst of the four results.

(over all four tests, this was achieved for 11 loops). The fact that these results vary from one test to another suggests that the sampling time is not sufficient for some 15-residue loops. With such a loop length, the number of possible conformations may be very large when the loop environment is not strongly constrained, and thus, the sampling time must be adapted to the size of the conformational space.

The observed times to sample a loop within 1 or 2 Å of the native are reported in Table 3.4. Computational requirements increase with loop length. Given that the native loop is one of many valid conformations, the relatively high variance in the time to generate a nearby configuration for long loops is not surprising. This is due to the stochasticity of the method. However, the variation of RMSD_{min} for 15-residue loops suggests that the provided sampling time is not adequate to perform a sufficient sampling of the loop's conformational space.

Table 3.4. Time needed to generate a conformation within one or two Ångström of the native loop. Statistics are calculated on all the runs for which such a distance was reached. MoMA-LoopSampler sampled a loop within 2 Å of native in at least one of the four tests for all 9-residue loops and 12-residue loops and 28 15-residue loops. It sampled a loop within 1 Å of native in at least one of the four tests for 51 9-residue loops, 9 12-residue loops and 11 15-residue loops.

Test Set	Time to reach...							
	2 Å to native				1 Å to native			
	Min	Median	Mean	Max	Min	Median	Mean	Max
9 residues	0.3 sec	4.5 sec	21.4 sec	11.8 min	1.0 sec	28.6 sec	4.1 min	1.7 h
12 residues	0.89 sec	23.4 sec	2.5 min	28.9 min	3.26 sec	5.6 min	16.5 min	2.2 h
15 residues	2.84 sec	19.6 min	1 h	5.9 h	4 min	1 h	1.6 h	5.1 h

These results show that MoMA-LoopSampler can construct loop ensembles for the 9 and 12-residue cases that include the native state with high precision, while 15-residue loops still present a formidable challenge, potentially because of the dimensionality of the search space coupled with a less constrained environment.

3.3.2.2 Comparison of MoMA-LoopSampler with state-of-the-art loop prediction methods

We compare the sampling performance of MoMA-LoopSampler to that of DiSGro [Tang 2014] and of the updated version of RCD [Chys 2013, López-Blanco 2016]. Note that, based on the results from references [Soto 2008] and [Tang 2014], DiSGro performs better than earlier loop closure or loop prediction methods such as CCD [Canutescu 2003], Wriggling [Cahill 2003], PLOP-build [Jacobson 2004], LOOPY$_{bb}$ [Xiang 2002], Random Tweak [Shenkin 1987], or Direct Tweak [Xiang 2002, Xiang 2006]. Therefore, we do not compare directly to these older methods.

Source code for DiSGro was obtained from `http://tanto.bioe.uic.edu/DiSGro/download.html`. The code had to be slightly modified to output exactly the required number of clash-free conformations (instead of the subset of clash-free conformations among a required number of closed ones), and conformations were generated without side-chains. Binaries for RCD version 1.40 were downloaded from `http://chaconlab.org/modeling/rcd/rcd-download`. MoMA-LoopSampler uses stricter constraints than DiSGro and RCD, in particular for steric clash detection. Therefore, in order to more adequately compare running times, we also tested a variant of MoMA-LoopSampler (Soft MoMA-LoopSampler) that uses collision constraints comparable to that of DiSGro and RCD. This variant uses a van der Waals scaling factor of 0.6 (instead of 0.7 for the other tests), does not use enlarged C_β atoms, and uses a lower max_{IK}. These changes are expected to lower the quality of the ensemble and its exhaustiveness, but also to considerably decrease sampling

time, allowing a more straightforward time comparison with DiSGro and RCD. The computational time and ability to generate near-native loops are compared, using the same computational resources for all four methods.

Difference between the sampling methods

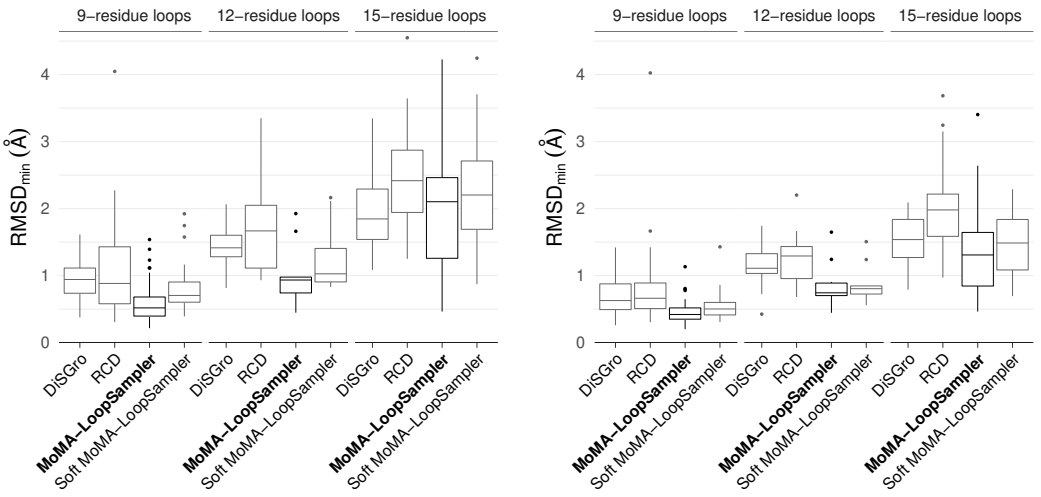
Different sets of constraints are enforced by the three sampling methods. Concerning collisions, DiSGro employ an energy function that makes steric clashes unlikely. The maximum allowed ratio between non-bonded atom pair distances and the sum of their van der Waals radii (called van der Waals scaling factor) was set to 0.6 in DiSGro (value found in the source code) and to 0.5 for intra-loop backbone collisions in RCD. Collisions with the rest of the protein are handled differently in RCD: this method uses a grid and considers that there is a collision if an atom of the loop is placed in a non-empty cell. This collision detection method is much faster but also less accurate than considering the actual distance between atoms. In MoMA-LoopSampler, steric clash avoidance is a crucial component: backbone atoms and the C_β atoms (with enlarged volumes to account for side-chain placement) of the sampled loops are placed without major steric clash among themselves or with the rest of the protein. While the van der Waals scaling factor was set to 0.6 in the more collision-tolerant version of MoMA-LoopSampler (Soft MoMA-LoopSampler), we set this cutoff at 0.7 to test the basic and RL MoMA-LoopSampler.

Structural knowledge is included in all three methods. Although RCD can be considered an *ab initio* method, it samples dihedral angles following neighbor-dependent Ramachandran probability distributions. DiSGro includes a stronger knowledge-based component, with a more complex dihedral angle sampling that follows distributions extracted from a structural database. Finally, MoMA-LoopSampler is strongly dependent on structural knowledge, since it directly uses fragments from experimentally-solved protein structures.

Results

Among 5000 sampled conformations, RMSD_{min} obtained by MoMA-LoopSampler are much lower than that obtained by the other methods for 9- and 12-residue loops (Figure 3.8a). Soft MoMA-LoopSampler also obtains RMSD_{min} lower than DiSGro and RCD, but higher than the basic version of MoMA-LoopSampler. For the 15-residue loops, RMSD_{min} obtained on 5,000 sampled conformations are comparable for the four methods. However, looking at 100,000 sampled conformations for these longer loops, MoMA-LoopSampler obtains a much lower RMSD_{min} (Figure 3.8b). Note that the RMSD_{min}s obtained for DiSGro on the 15-residue loop test set are lower than the ones the authors report for 100,000 sampled conformations on the same test set [Tang 2014].

Generating 100,000 conformations instead of 5,000 lowered the RMSD_{min} for all the methods (Figure 3.10). However, this decrease varies from one method to another. It is very limited for MoMA-LoopSampler on the 9- and 12-residue loops, while being considerable for other methods. Note that despite this, the RMSD_{min}



(a) 5,000 sampled conformations. **(b)** 100,000 sampled conformations.

Figure 3.8. Minimum distance to native (RMSD$_{min}$) obtained among sampled conformations, without side-chains.

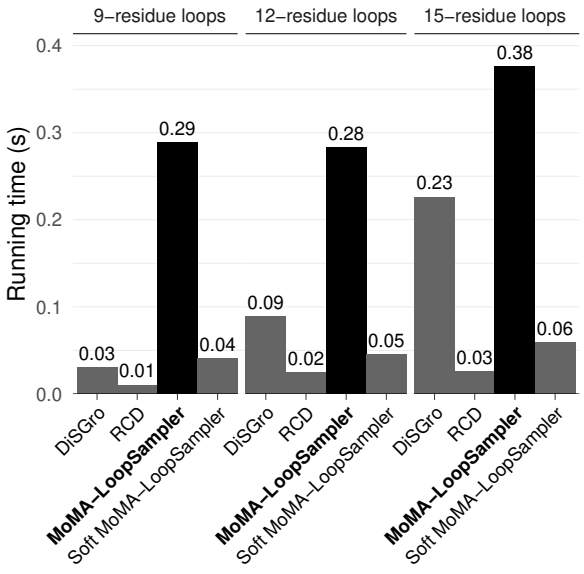


Figure 3.9. Median time per sampled conformation (estimated on 5000 sampled conformations, without side-chain placement). Computations were performed using a single core of a 2.30 GHz Intel® Xeon® E5-2695 v3 processor.

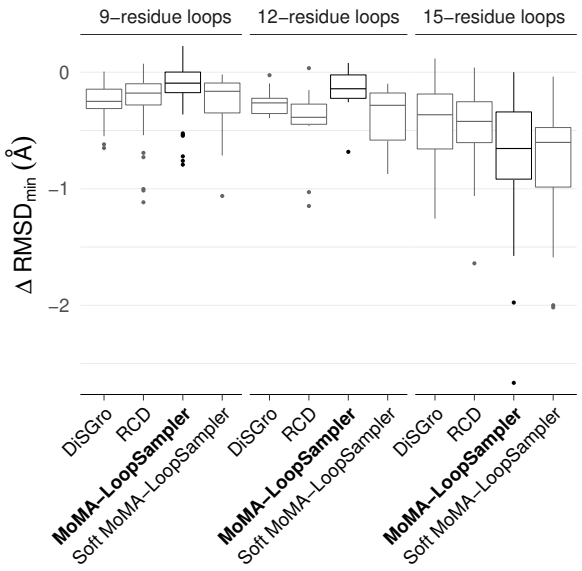


Figure 3.10. Difference in RMSD$_{min}$ obtained when generating a larger number of conformations (100,000 conformations instead of 5,000).

obtained by MoMA-LoopSampler on these loops is still much lower compared to the other tested methods. The 15-residue loops show the opposite trend: the decrease in RMSD$_{min}$ is much larger for MoMA-LoopSampler and its "soft" version than for the other methods on 15-residue loops, resulting in a lower RMSD$_{min}$ for MoMA-LoopSampler compared to RCD and DiSGro.

Concerning running times, MoMA-LoopSampler and its soft version are much less sensitive to the length of the loop than the other two methods. The soft version of MoMA-LoopSampler is slightly slower than other methods on 9-residue loops but faster than DiSGro for 12- and 15-residue loops. We note that the running times obtained for DiSGro are higher that those reported in [Tang 2014]. In some tests, DiSGro got blocked during the sampling process, in which case we started the run again. These failed sampling attempts are not counted in the running times we report. The basic version of MoMA-LoopSampler is unsurprisingly slower than the other methods, due to the stronger constraints it enforces. Overall, results show that the "soft" version of MoMA-LoopSampler has running times comparable to that of other methods and provides slightly lower RMSD$_{min}$. The basic version of MoMA-LoopSampler on the other hand, trades computational efficiency off for better filtering of sampled ensembles.

Discussion

The difference in the collision constraints enforced by the different methods can help to explain two major observations in the results: (1) MoMA-LoopSampler (especially the basic version) is less sensitive to loop length than RCD and DiSGro, from a running time point of view; (2) while MoMA-LoopSampler obtains a much

lower RMSD_{min} than other methods for 9- and 12-residue loops on ensembles of 5,000 conformations, a higher number of sampled conformations is necessary for 15-residue loops to observe a difference.

To explain the first point, we hypothesize that MoMA-LoopSampler is more sensitive to the environment of the loop, and whether it is constrained or not, than to the length of the loop. First, an essential component of the method consists in checking that the distance between the two working loop ends can be covered by the tripeptides left to place. Making this verification after a tripeptide is added facilitates the closing of long loops. Second, shorter loops generally have a more constrained environment than longer loops. As previously mentioned, MoMA-LoopSampler is more intolerant to collisions compared to other methods. Therefore, generating a conformation for a loop in a more constrained environment is a problem with a difficulty comparable to that of sampling a longer unconstrained loop for MoMA-LoopSampler, which explains why median running times per conformation vary little with loop length.

To explain the second point, a similar reasoning can be conducted. MoMA-LoopSampler only samples the accessible conformational space by carefully avoiding collisions in generated conformations. Therefore, it finds the native conformations using fewer samples than other methods. With more constrained environments, 9- and 12- residue loops have a resulting conformational space that is particularly reduced, which is why MoMA-LoopSampler finds the native conformation very early into the search. Longer loops are usually more flexible, and a larger portion of the conformational space is allowed. Therefore, methods that are overall more tolerant to collisions sample fewer bad-quality conformations in proportion. For these loops, more conformations need to be sampled to achieve a better coverage of the conformational space. The benefit in RMSD_{min} for MoMA-LoopSampler is thus logically observable when sampling a higher number of conformations (100,000) for 15-residue loops (Figure 3.8b).

In simpler terms, this means that MoMA-LoopSampler better explores the conformational space, performing an exhaustive exploration using fewer samples than other methods. The difference in RMSD_{min} observed after generating larger ensembles further supports this idea. Indeed, 5,000 samples from MoMA-LoopSampler are enough to explore the conformational space of 9- and 12-residue loops, explaining why the RMSD_{min} barely decreases when generating a much larger number of conformations. For other methods, the RMSD_{min} considerably decreases upon generating more conformations, showing that these methods keep discovering relevant conformations among the extra conformations. Conversely, for 15-residue loops, the difference in RMSD_{min} obtained upon generating 100,000 conformations instead of 5,000 is considerable for all four methods, with MoMA-LoopSampler and Soft MoMA-LoopSampler showing a much larger decrease than RCD and DiSGro. This confirms (1) that 5,000 conformations are not enough to cover the much larger conformational space of these longer loops, and (2) that MoMA-LoopSampler performs a more efficient exploration, discovering relevant conformations using fewer samples than RCD and DiSGro.

Obtaining an ensemble of conformations of good quality is essential considering the costly downstream processing steps of applications involving loop sampling, in particular in the context of stable states prediction. These steps include side-chain addition, relaxation, scoring, clustering or filtering, and can be extremely time-consuming. In that regard, generating fewer conformations, but which are more representative of the ensemble overall, is perfectly satisfactory. This suggests that MoMA-LoopSampler is a good candidate for the sampling stage of many structural bioinformatics applications, including stable states prediction, since it obtains the same RMSD_{min} as other methods (or a lower one) without needing to sample as many conformations as these methods do.

3.3.2.3 Application to a multi-state loop

We demonstrate the exhaustive sampling ability of MoMA-LoopSampler by generating relevant loop conformations using the streptavidin protein. Streptavidin is a homotetramer protein that strongly binds biotin. Each monomer exhibits a biotin binding site and a flexible loop L (between residues 44 and 52) that stabilizes the complex by "closing" upon binding. Two conformations are known for L: "open" and "closed". Sampling loop conformations from several structures of the protein with MoMA-LoopSampler, we intend to explain the presence of a conformation or another in the crystal structure, and to determine which of the known conformations are accessible to the loop.

Three high-resolution structures of streptavidin were extracted from the PDB: 2F01 [Le Trong 2006], 3RY1 [Le Trong 2011] and 3RY2 [Le Trong 2011]. 3RY2 and 2F01 both contain two subunits in the asymmetric unit, whereas 3RY1 contains four subunits. The subunits from 2F01 and 3RY2 are all bound to a ligand (either biotin or epi-biotin). Their L loops are thus in the "closed" conformation. The subunits from 3RY1 are all unbound, however, one of them shows L in the "closed" conformation, while the others have L in the "open" conformation. The fact that L is found in the "closed" conformation while unbound is likely due to crystal packing interactions, showing the intrinsic flexibility of this loop and the dependence of its conformation on its environment.

We separated the different subunits (by separating the chains of the PDB files) and used MoMA-LoopSampler to perform a brute force exploration of L's conformation space from each subunit of the three starting crystallographic structures, after removing the ligand (if present). For all the conformations obtained after three rounds of brute force search, side-chain placement was attempted with an in-house method using continuous rotamers from BASILISK [Harder 2010]. When the side-chains could not be placed without collisions, the backbone conformation was discarded. The brute force sampling with side-chain placement finally yielded 5338 conformations from scaffold 2F01(A), 3825 from 2F01(B), 4042 from 3RY1(A), 1304 from 3RY1(B), 702 from 3RY1(C), 2820 from 3RY1(D), 5320 from 3RY2(A) and 3611 from 3RY2(B). Those conformations were relaxed using AMBER 16. Total energies were calculated using the ff14SBonlysc force field and a simple Generalized

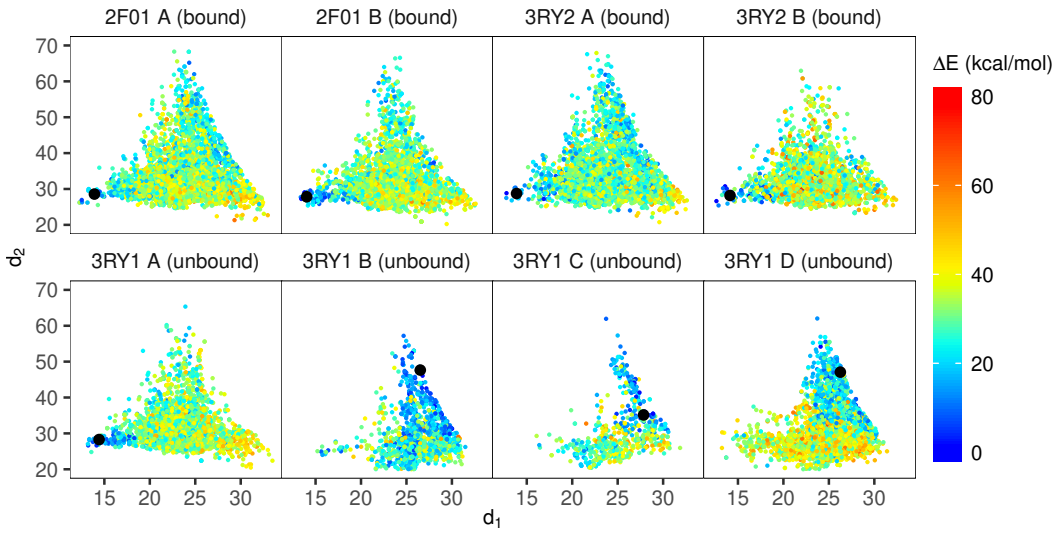


Figure 3.11. 2D projections of conformations sampled using MoMA-LoopSampler in brute force mode for a loop in streptavidin, from eight starting X-ray protein structures. The first dimension, d_1 (x-axis), is the distance (Å) between an atom located in the middle of the loop and a fixed atom in the protein. The second dimension, d_2 (y-axis), is the angle (°) formed by three atoms: an atom at approximately one quarter, one half, and three quarters of the way down of the loop. The conformations from the crystallographic structures are shown in black. For each system, the loop with the lowest energy was identified and each conformation was then colored according to the difference between its energy and this lowest energy.

Born implicit solvent model ($igb = 1$ and the *mbondi* radii sets, as recommended in the AMBER manual) [Case 2005]. No constraints were applied for the relaxation, so that both backbone and side-chain movements were allowed. Bond length, bond angles and dihedral angles were all free to vary. The relaxation took 133 s per loop on average. The first 250 cycles used steepest descent minimization, while the remaining cycles applied conjugate gradient. The maximum number of cycles was set to 500 and the minimization was considered to have converged when the root-mean-square of the cartesian elements of the gradient was lower than 0.1 kcal/(mol·Å). Finally, the loops were projected in 2D space (see Methods). The results thus obtained are shown in Figure 3.11.

The projected conformations adopt an overall triangular shape. The "closed" conformations are projected in the lower left vertex, while the "open" conformations are more diverse and are projected on the opposite side of the triangle. The crystallographic conformations are always found in a low-energy basin. The projection plots show major differences depending on the starting structure. Sometimes, both the "open" and "closed" conformations appear to be in low-energy basins (2F01(A), 2F01(B), 3RY2(A), 3RY1(B) and 3RY1(C)). Other projection plots show only one low-energy basin, around the crystallographic conformations (3RY2(B), 3RY1(A) and 3RY1(D)). The energies were calculated without a ligand, showing the strong influence of the conformation around the loop. In the case of 2F01(A), 2F01(B), and 3RY2(A), the environment surrounding the loop allows it to adopt both confor-

mations, but the presence of the ligand probably stabilizes the loop in one of the two basins. Energy barriers of different heights separate the "closed" and "open" basins. In 3RY2(B), it seems that only the "closed" conformation is stable, suggesting that the environment of the loop also changes (due to crystal packing or ligand binding), and stabilizes this conformation. A profile similar to that of 3RY2(B) is found for 3RY1(A), although this subunit is unbound in the crystallographic structure. This is an indication of the conformational changes that occur around the loop in the crystal. For 3RY1(D), the "open" conformation is in a large low energy basin. A few low energy conformations are found in the "closed" loop region but a large, high energy barrier separates the two regions. In the case of 3RY1(B) and 3RY1(C), the conformational space appears tighter. The loop environment is probably more constrained sterically. Nevertheless, both energy basins are found, with a lower energy barrier separating them, and a much lower energy minimum for the "open" basin, explaining the "open" conformation adopted by these subunits in the crystal.

In addition to analyzing the energy landscape of a loop, MoMA-LoopSampler can be used to sample intermediate states along the path between two stable conformations. For example, in the cases in which the "open" and "closed" basins are both present and of low energy. This could then enable the analysis of the docking mechanism of streptavidin and biotin in great detail. Nevertheless, the analysis of conformational transitions goes beyond the scope of this chapter.

Landscapes obtained with other sampling methods

Using the same relaxation, projection and scoring protocol, the landscapes can be obtained for other sampling methods. This was done for DiSGro (Figure 3.12) and using the server version of RCD , RCD+ (Figure 3.13), after generating the same number of conformations as was done by MoMA-LoopSampler in brute force mode. Side-chain placement was activated for DiSGro, and sampling with RCD+ was performed using the dedicated web server [López-Blanco 2016], which performs side-chain placement and refinement.

The landscapes obtained using DiSGro and RCD+ are very different from those obtained using MoMA-LoopSampler. Although some common features can be observed (such as the basin around the "open" conformations from scaffolds 3RY1(B), 3RY1(C) and 3RY1(D)), the landscapes are much rougher and harder to interpret. They are also more spread out than landscapes obtained when sampling with MoMA-LoopSampler. The fact that these methods are more collision-tolerant may explain these observations. Indeed, many statistically unlikely conformations are generated, perturbing the analysis of the landscape. By creating a better filtered ensemble, MoMA-LoopSampler clarifies the analysis of energy landscapes for this loop.

3.3.3 Performance of reinforcement learning

Results of MoMA-LoopSampler with RL are described and compared to results obtained utilizing the basic method. We analyze the benefits of using RL, as well

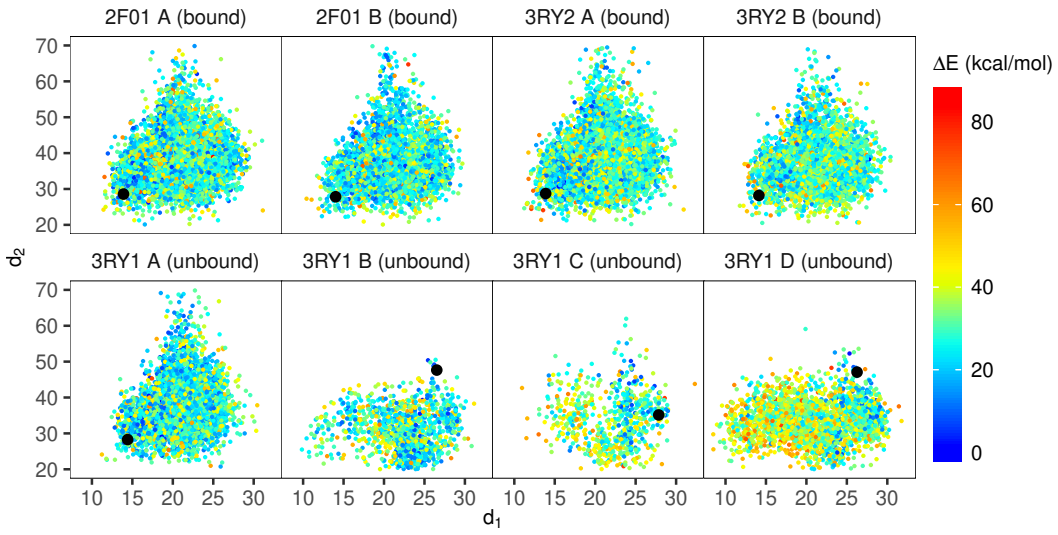


Figure 3.12. 2D projections of conformations sampled using DiSGro for a loop in the streptavidin protein, from eight starting X-ray protein structures. The first dimension, d_1 (x-axis), is the distance (Å) between an atom located in the middle of the loop and a fixed atom in the protein. The second dimension, d_2 (y-axis), is the angle (degrees) formed by three atoms: an atom at approximately one quarter, one half, and three quarters of the way down of the loop. The conformations from the crystallographic structures are shown in black. For each system, the loop with the lowest energy was identified and each conformation was then colored according to the difference between its energy and this lowest energy.

as the potential downfalls, mainly in terms of loop diversity. Four different learning rates were tested (see Section 3.3.1.3). These tests with RL were performed in the same conditions as the four tests performed without learning: 9-residue loops were sampled for 2 hours, 12-residue loops for 4 hours and 15-residue loops for 6 hours.

3.3.3.1 Number of conformations sampled

Basic versus RL mode

The main interest of RL is that it enables faster generation of loop conformations, as shown by Figures 3.14 and 3.15 and by the higher densities in Figure 3.18. The *very high* learning rate generates 127% more conformations on average compared to the basic mode, but this percentage is highly variable across loop systems. Loop 73, which is located in a very constrained environment, constitutes an extreme case in which activating RL can multiply by over 41 the number of conformations sampled over 6 hours. Loops 68, 21 and 85 are other very successful examples, for which RL multiplies by 12, 5.4 and 5.2 the number of sampled conformations, respectively. However, for a few loops, RL may decrease (by up to 12%) the number of sampled conformations. This is actually due to the overhead of the learning process itself: each time a loop is sampled, statistics are updated and the learning trees are maintained. In a few cases, the time saved during conformation sampling itself

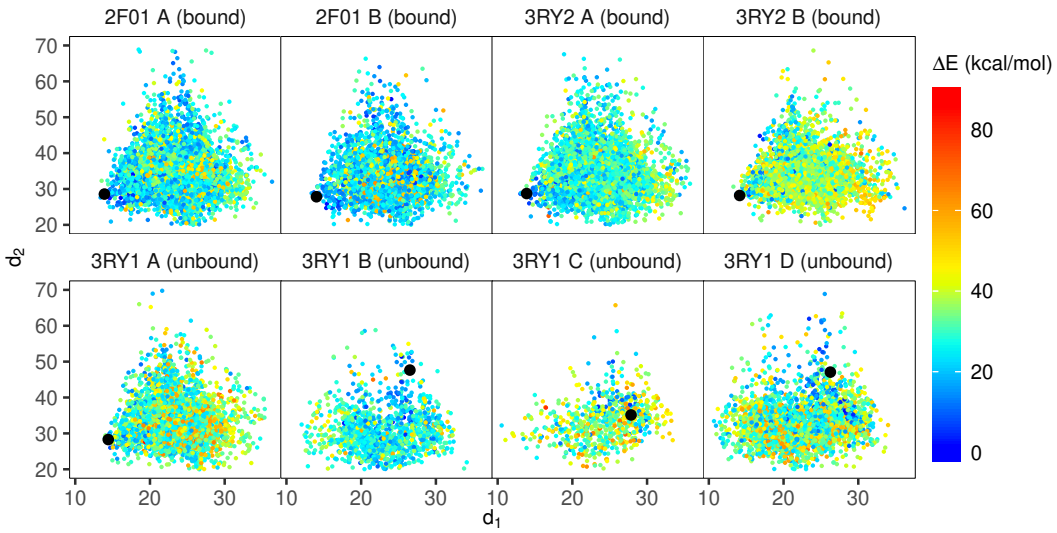


Figure 3.13. 2D projections of conformations sampled using RCD+ for a loop in the streptavidin protein, from eight starting X-ray protein structures. The first dimension, d_1 (x-axis), is the distance (Å) between an atom located in the middle of the loop and a fixed atom in the protein. The second dimension, d_2 (y-axis), is the angle (degrees) formed by three atoms: an atom at approximately one quarter, one half, and three quarters of the way down of the loop. The conformations from the crystallographic structures are shown in black. For each system, the loop with the lowest energy was identified and each conformation was then colored according to the difference between its energy and this lowest energy.

(especially with lower learning rates) is not high-enough to compensate for the time lost in maintaining the learning data structures.

Influence of loop lengths and learning rates

Overall, higher learning rates tend to produce a higher number of conformations. This is true for many systems such as loops 14 and 68 (Figure 3.15), which show very different sampling speed depending on the learning rate. However, the effect of RL depends on both the length of the loop and the loop/protein system itself. While most 9- and 12-residue loops (and a few 15-residue loops) exhibit this expected behavior, for other loops (and for many 15-residue loops), only runs performed with *high* and *very high* learning rates are capable of generating a larger number of conformations (e.g. loop 61). As previously mentioned, activating RL with a *very low*or *low* learning rate can even reduce the number of sampled conformations. This is mainly observed for 15-residue loops (see Figure 3.14), such as loop 76 (Figure 3.15(c)). Loop 42 illustrates yet another case: all learning rates generate loops at comparable speed, but still much faster than MoMA-LoopSampler in basic mode does.

Loops 68 and 73 also illustrate an interesting phenomenon. The curves for *high* and *very high* learning rates show some plateaus spanning 10 minutes or more, which are due to the overhead of maintaining a very large learning tree. When

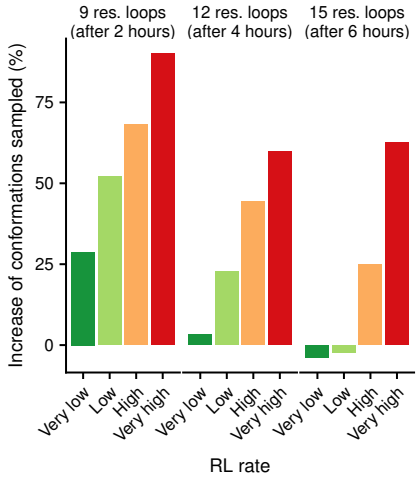


Figure 3.14. Median percentage of increase in number of conformations sampled in RL modes compared to the basic mode across loop systems.

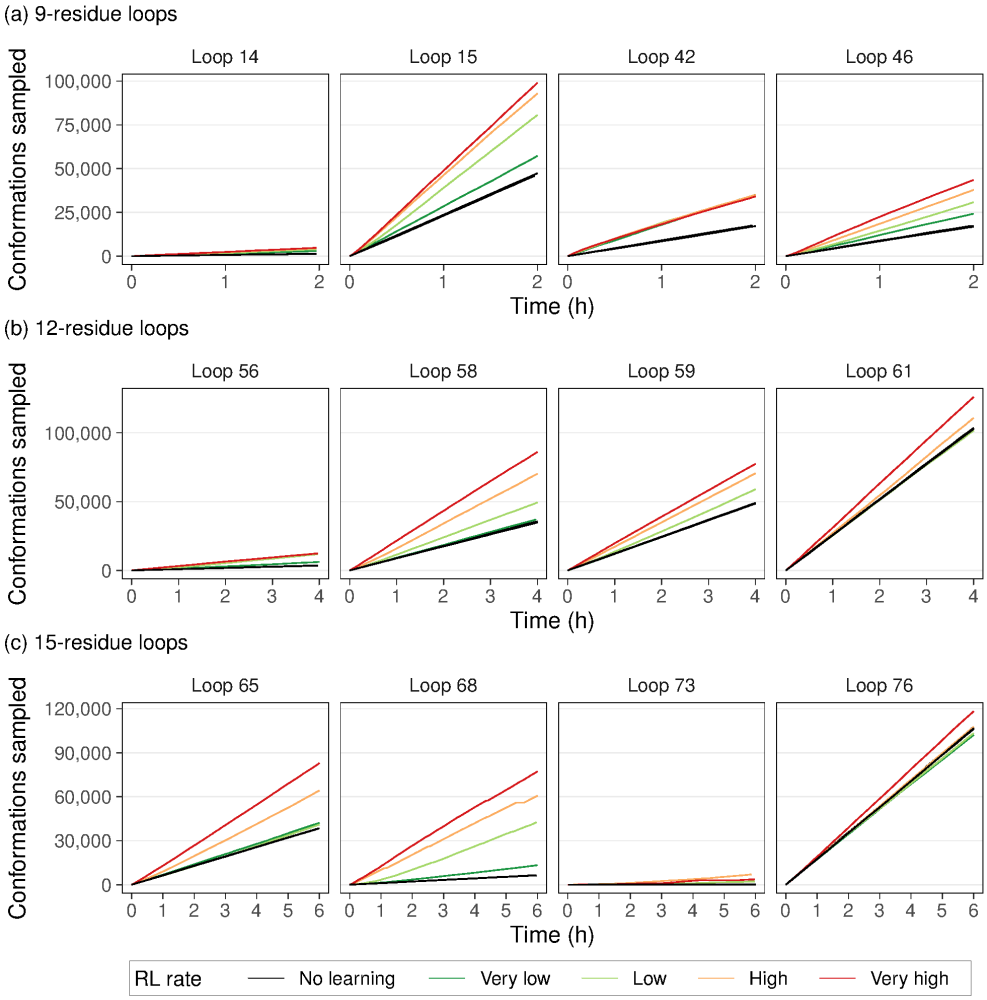


Figure 3.15. Number of conformations sampled as a function of time for different levels of reinforcement learning and for a few representative loops.

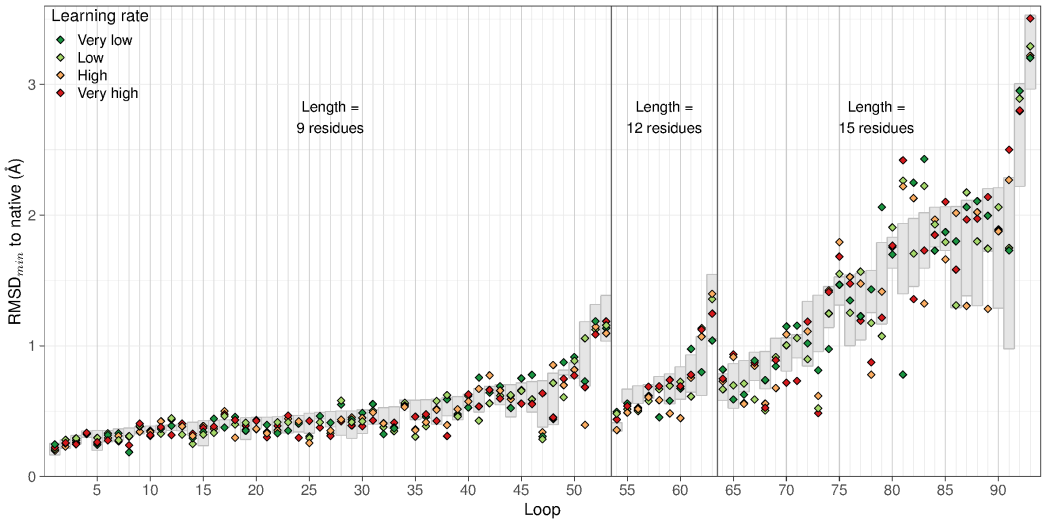


Figure 3.16. RMSD$_{min}$ of MoMA-LoopSampler on each loop ensemble. Colored data points correspond to the results obtained for each level of reinforcement learning. The gray bars show the RMSD$_{min}$ range obtained by the four tests performed without learning. A data point above the error bar shows that the corresponding run with learning misses some closer-to-native conformations that can be sampled when turning off reinforcement learning. A data point below the error bar indicates that the corresponding run found conformations that are even closer to native than the runs without learning did.

the loop is long, a split in the initial levels of the tree triggers downwards leaves splitting and data reassignment. This issue is mainly observed in longer loops, since the dimensionality of the tree is exponential in the length of the loop.

3.3.3.2 Sampling near-native conformations

The RMSD$_{min}$ per loop is shown in Figure 3.16 (and summarized in Table 3.5) for the different learning rates and for the sampling performed without RL. For 9-residue and 12-residue loops, runs with RL are able to generate conformations as close to native as runs utilizing the basic mode. A tendency to generate loops slightly closer to native can also be observed overall. For 15-residue loops, the effect of RL is much less clear. For some loops, learning enables the generation of loops much closer to native (e.g. for loops 73 and 78). But the contrary effect is also observed (e.g. loops 75 and 81). As previously mentioned, 6 hours appear insufficient for some 15-residue loops. Indeed, the conformational space of 15-residue loops that are not strongly constrained is too large to be exhaustively sampled in only a few hours (using a single CPU core). The difference in RMSD$_{min}$ between ensembles generated with and without RL is not significant.

The times to generate the first conformations with a RMSD to native lower than 1 Å are shown in Figure 3.17. Although these values are generally of the same order of magnitude for runs with or without RL, some observations can still be made. For 9-residue loops, using RL may considerably delay the sampling of

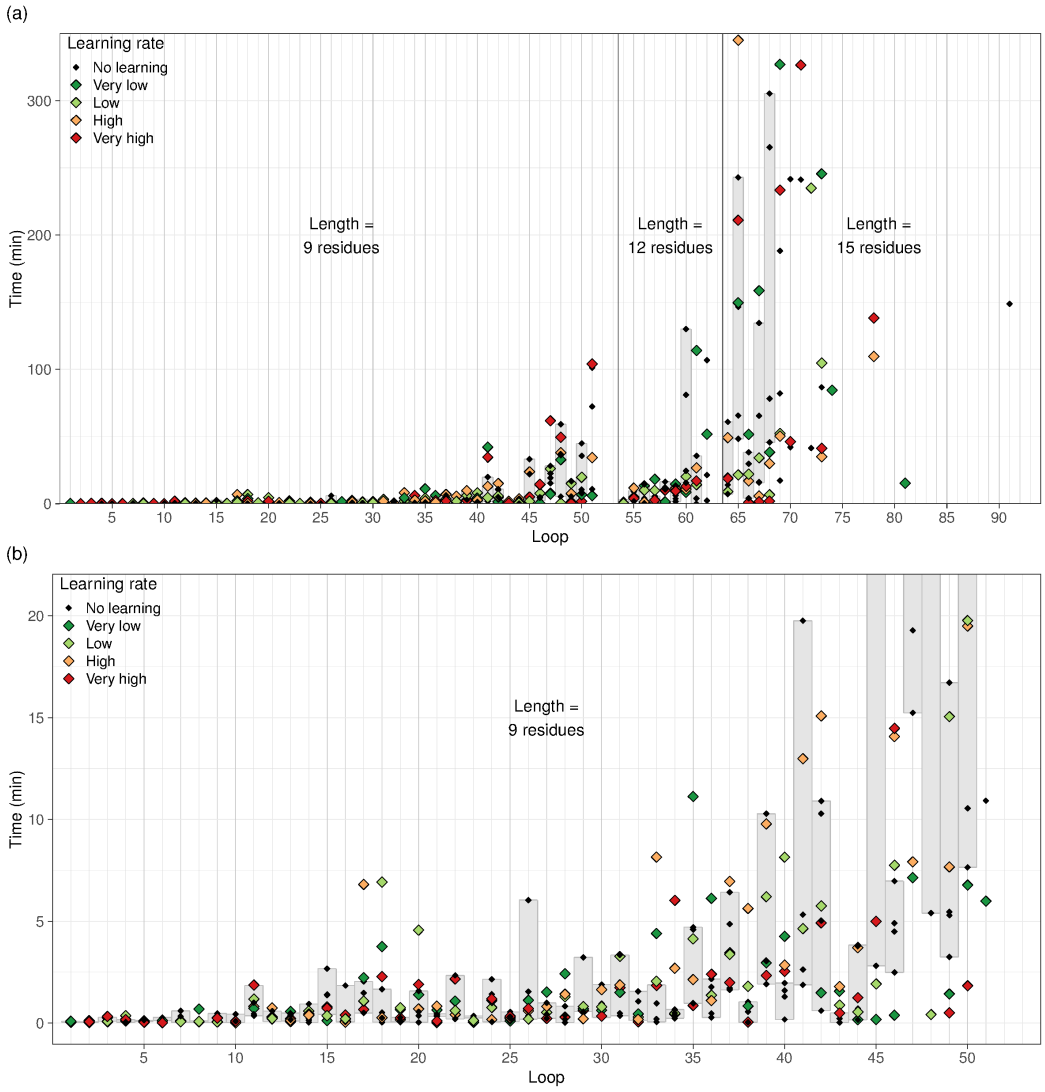


Figure 3.17. Time needed to generated the first conformation with a RMSD to native below 1 Å. Smaller black data points correspond to the 4 tests performed without learning. Colored data points correspond to runs performed with learning. A failure in generating a loop within 1 Å of native results in a missing point. The gray bars represent the range of values obtained by the tests performed without learning, when all four of them succeeded in finding a conformation that close to native. (a) All loops. (b) Zoom on the 9-residue loops.

Table 3.5. $RMSD_{min}$ obtained for different learning rates.

Length		Basic mode	RL mode learning rate			
			Very low	*Low*	*High*	*Very high*
9 res.	Mean (Å)	0.49	0.48	0.47	0.46	0.46
	Median (Å)	0.43	0.41	0.41	0.39	0.41
	SD (Å)	*0.21*	*0.21*	*0.21*	*0.20*	*0.19*
12 res.	Mean (Å)	0.73	0.67	0.73	0.67	0.75
	Median (Å)	0.67	0.60	0.64	0.55	0.69
	SD (Å)	*0.26*	*0.20*	*0.30*	*0.32*	*0.26*
15 res.	Mean (Å)	1.50	1.52	1.48	1.48	1.50
	Median (Å)	1.45	1.45	1.43	1.42	1.44
	SD (Å)	*0.62*	*0.68*	*0.70*	*0.67*	*0.73*

a conformation that is close to native. A possible reason is that introducing RL modifies the probability for selecting tripeptides. While in the basic mode, MoMA-LoopSampler picks tripeptide states at each step with a uniform distribution, the RL mode offers MoMA-LoopSampler the possibility to adjust the sampling of states so that a suitable distribution is obtained. Considering the tree used by RL to organize the tripeptides, MoMA-LoopSampler in the basic mode chooses each cell with a distribution that is directly proportional to the number of tripeptides it contains. Conversely, in the learning mode, MoMA-LoopSampler samples each cell according to their score. The score is currently set so as to sample effectively as many diverse conformations as possible, but other strategies may be contemplated, for example in order to obtain a loop ensemble that follows the density of the tripeptide database for each tripeptide position. Such an ensemble could provide a more statistically accurate representation of the loop conformational space, which would be interesting to analyze entropic effects.

3.3.3.3 Diversity of sampled loops

As mentioned in the previous section, using RL changes the distribution used to sample tripeptides. As sampling progresses, the method learns which cells in the tree have not led to successful loop conformations so far, and starts sampling these cells less frequently. If the learning process is too greedy, this may happen even though the states explored in the cell are not adequately representative. Therefore, a careful parameterization of RL is crucial to get an exhaustive sampling.

We explored the diversity of the ensembles sampled with different learning rates in order to determine if all areas of conformational space are adequately covered. Figure 3.18 shows the 2D projections of the loop samples obtained by employing various learning rates on four different systems.

The most obvious observation is that all projections corresponding to the same system look similar, in the sense that they have the same overall shape. Even with a *very high* learning rate, there does not seem to be major areas of conformational

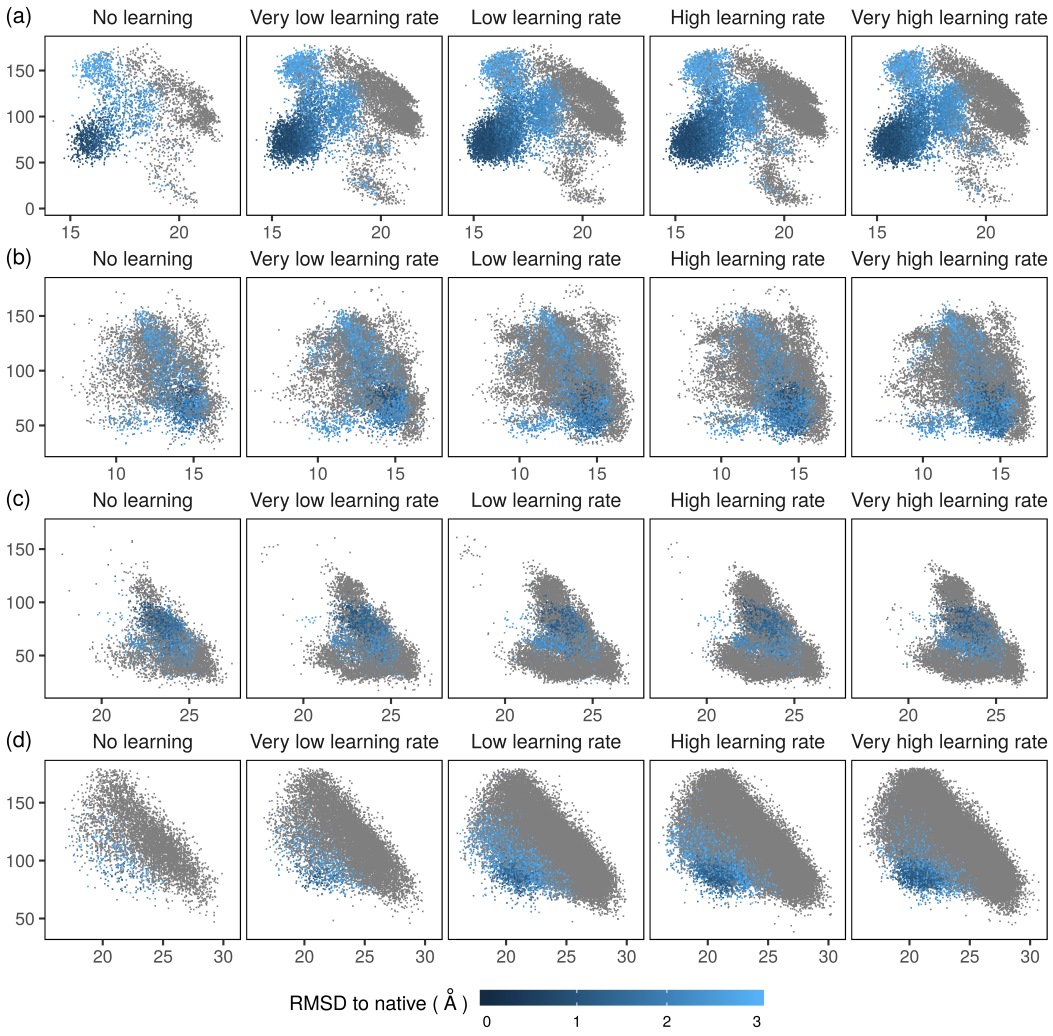


Figure 3.18. Two-dimensional projections of the sampled loops at different levels of reinforcement learning. Each point represents a loop. Points are colored according to the distance to native of the corresponding loop, or in grey if their RMSD to native is above 3 Å. The x-axis gives the first projection descriptor d_1 (in Å), while the y-axis gives the second projection descriptor, d_2 (°). (a) Loop 21, (b) Loop 26, (c) Loop 40, (d) Loop 68.

space that are ignored. However, with *high* or *very high* learning rates, some sparse areas of the 2D projection space may no longer be sampled. For example, in Figure 3.18(c), the area at the top left-hand corner is void of conformations with the *very high* learning rate. The same observation can be made for the top-most region of Figure 3.18(b).

A very striking observation is that the 2D projection plots are denser at higher RL rates. This is a natural result of sampling more conformations in the same amount of time. However, the distribution of points within these projections indicates that in the RL mode, MoMA-LoopSampler samples the conformational space with a higher resolution than in the basic mode and that it does not create tiny clusters of conformations. RL provides greater diversity in the areas that get sampled. When the area around the native loop gets explored more densely, the native loop can be found with better accuracy. This is clearly the case for loops 21 and 68 (Figures 3.16 and 3.18(a)(d)).

We also show the evolution of the distribution of sampled loops with RL activated in Figure 3.19. These heatmaps, shown in the same two-dimensional projection, showcase the density of the sampled loops in the first and the last ten minutes of the exploration. Two effects are observed when comparing the beginning and the end of the sampling process:

(1) The first effect of learning is that the number of sampled conformations increases, and the coverage of the conformational space improves. In other words, the projection of sampled loop conformations in 2D appears to be more homogeneous and continuous. The effect is similar to that observed for the different learning rates. Indeed, the algorithm progressively stops exploring the regions of space where it does not find any solution. The probability to sample a tripeptide in a cell from which all attempts have failed so far decreases with running time. Consequently, the success rate becomes higher since MoMA-LoopSampler focuses on the vicinity of regions that are successful. This is very clear in Figure 3.19(d), for *very low* and *low* learning rates. It can also be observed for other systems, although to a lesser degree. The ability of the learning process to quickly identify areas where no solution exists depends on the positions of solutions in the conformational space and how they cluster, on the projection chosen to organize the tripeptide states, and on the speed of learning.

(2) The second effect is the sudden discovery of whole regions of the conformational space. As previously mentioned, the probability to explore a region invariably found unsuccessful so far decreases based on the number of attempts and the learning rate. If the number of attempts is too low (the learning process is too greedy), MoMA-LoopSampler can fail to explore some regions in which a few successful conformations could have been found. However, these regions may suddenly get "unlocked" after a closed loop is finally sampled. One successful conformation is necessary and sufficient to set the score of the cell leading to sampling in that region back to its maximum. This is the case for the region of the conformational space that is projected in the bottom right-hand corner of the heatmaps shown in Figure 3.19(a) at a *very high* learning rate. This phenomenon is also observed

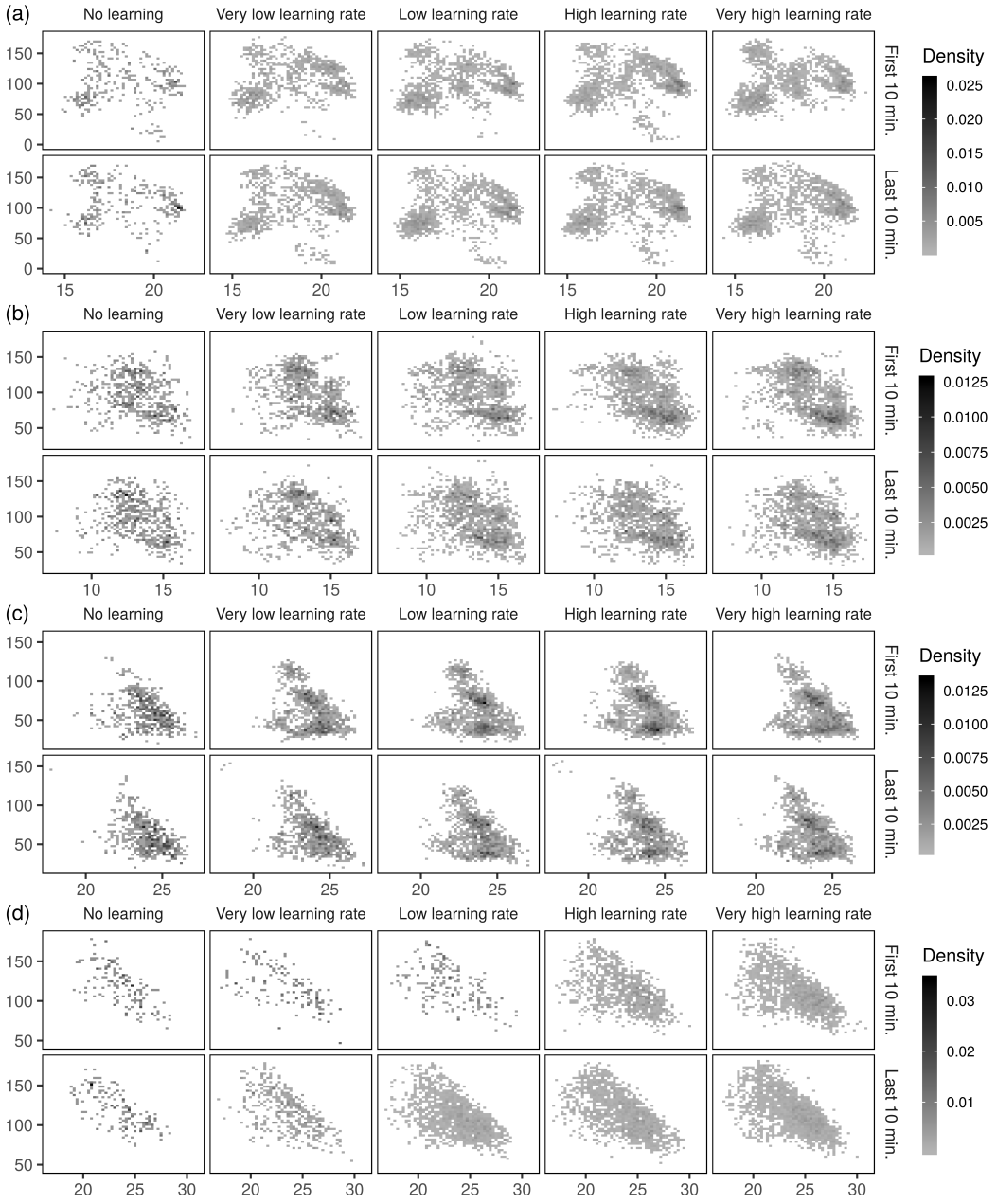


Figure 3.19. Evolution of the distribution of sampled loops due to reinforcement learning. The plots show heatmaps of two-dimensional projections of the sampled loops for different levels of reinforcement learning, during the first and the last ten minutes of sampling. The x-axis gives the first projection descriptor d_1 (in Å), while the y-axis gives the second projection descriptor, d_2 (°). (a) Loop 21, (b) Loop 26, (c) Loop 40, (d) Loop 68.

for other systems in Figure 3.19, albeit on smaller regions of the projection plot. This is a warning that learning is already too greedy. In theory, learning should only stop exploring regions when enough exploration has been carried out and the probability to find a successful loop conformation in that region is negligible. In practice, this can never be determined with certainty, as long as the full region has not been explored. For RL to have the desired behavior, a trade-off has to be found in order to enable faster exploration while not risking to lose large and/or relevant areas of the conformational space.

These results also show that the distribution of sampled conformations may be different depending on the learning rate and the running time: beyond the observation that the sampled conformational space is more continuous, the density observed in the heatmaps evolves between the different conditions. The samples appear to be more uniformly distributed when RL is involved. Another interesting observation is that, for all these four systems, some consequences of learning become visible very early in the sampling process, within the first 10 minutes.

The results provided here suggest that using RL allows for a much faster conformational sampling, although the generated ensembles may lose diversity if the learning process is too greedy. It is therefore of crucial importance to limit RL rates if one wants to preserve the diversity enabled by the tripeptide database.

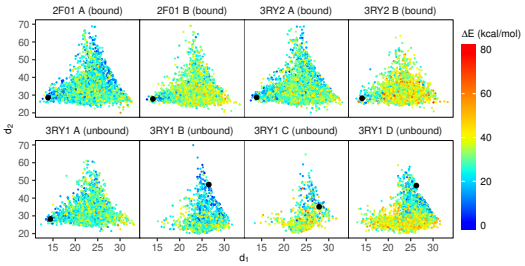
3.3.3.4 Energy landscape of a multi-state loop using RL

In order to verify that RL preserves the quality of the ensembles generated by the basic mode, it was applied on the earlier example of streptavidin. The landscapes resulting from sampling with MoMA-LoopSampler in RL mode were generated for the flexible loop in streptavidin, using the same protocol as in Section 3.3.2.3 (Figure 3.20). The number of sampled loop conformations was the same as in brute force mode.

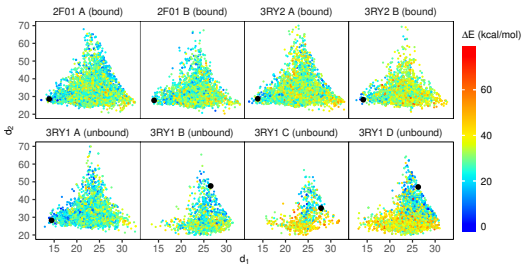
The landscapes look very similar to those obtained in brute fore mode, both with MoMA-LoopSampler in basic mode and in RL mode (at any learning rate), showing that RL parameterization does not majorly impact the sampling diversity for this loop system. Only the landscape resulting from MoMA-LoopSampler in RL mode with a *very high* learning rate shows slight differences from the other landscapes. In the landscape built from scaffold 2F01(A), the basin around the "open" conformation is not as clearly apparent as it is from other RL levels.

3.3.3.5 Memory requirements

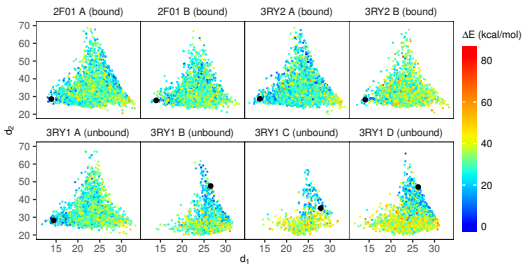
A practical downfall of RL is its potentially high memory requirements. In the case of an unconstrained loop environment, the size of the learning tree is expected to increase exponentially with the size of the loop. However, in practice, the tree will not grow in some directions because of unsuccessful states for positions early in the plan. Therefore, the current implementation is not usable for very long loops, unless the systems are highly constrained. Designing and implementing another RL



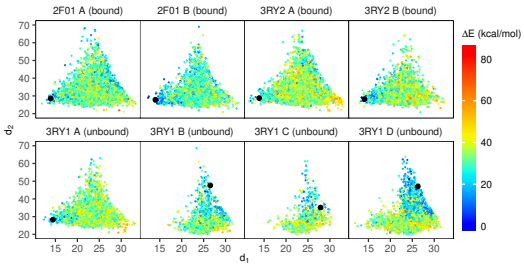
(a) Basic mode



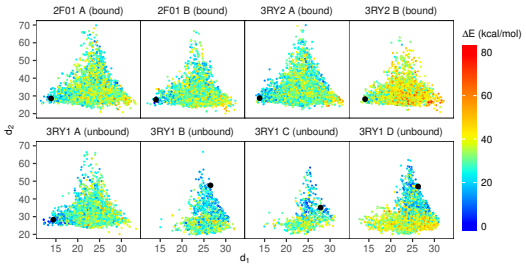
(b) *Very low* learning rate



(c) *Low* learning rate



(d) *High* learning rate



(e) *Very high* learning rate

Figure 3.20. 2D projections of conformations sampled using MoMA-LoopSampler at different learning rates for a loop in the streptavidin protein, from eight starting X-ray protein structures. The first dimension, d_1 (x-axis), is the distance (Å) between an atom located in the middle of the loop and a fixed atom in the protein. The second dimension, d_2 (y-axis), is the angle (°) formed by three atoms: an atom at approximately one quarter, one half, and three quarters of the way down of the loop. The conformations from the crystallographic structures are shown in black. For each system, the loop with the lowest energy was identified and each conformation was then colored according to the difference between its energy and this lowest energy.

approach with lower memory requirements would thus be a sensible development for the future.

3.4 Conclusions

This chapter has introduced MoMA-LoopSampler, a new method that employs local sequence-dependent structural knowledge and geometric techniques combined with reinforcement learning to exhaustively and efficiently sample protein loop conformations. The results show that this new method performs similarly to (or better than) existing computational methods in terms of computational efficiency and that the ensemble of sampled loop conformations includes those found in experimental structures (the "native" state of the loop). The implemented reinforcement learning approach allows MoMA-LoopSampler to accelerate sampling while maintaining conformational diversity (avoiding "over-learning"), and is scalable to large loop regions (15 residues). This work has also shown that MoMA-LoopSampler enables modeling loops present in several low-energy basins, thus being a useful tool when investigating energy landscapes and studying conformational transitions.

Further enhancements to the method include improving the learning component to limit its memory requirements. Another area to investigate is adjusting the scores of leaves within the learning structure so that the distribution of sampled conformations corresponds to the distribution of tripeptide states present in the database. Database adjustments may also improve the quality of the results. For example, filtering the database to only keep one representative among very similar tripeptides would speed up the sampling, and adding the states of similar sequences to the states of rare tripeptide sequences may allow the sampling of conformations currently inaccessible due to a potential lack of data.

Finally, building relevant loop ensembles requires both a sampling and a scoring components. While MoMA-LoopSampler is aimed at providing a diverse ensemble of possible conformations, it does not evaluate the sampled loops or estimate their likelihood. Designing an appropriate scoring function, or integrating existing ones into MoMA-LoopSampler, constitutes an interesting direction for future work.

The last item motivated a comparative assessment of multiple state-of-the-art loop scoring methods. This study was aimed at determining a scoring method that could smoothly complement MoMA-LoopSampler while providing an accurate evaluation of the quality of sampled conformations. The detailed analysis is presented in the next chapter.

Chapter 4

Loop scoring and landscape reconstruction

Contents

4.1 Introduction

4.1.1 Motivation

Loop modeling usually operates in two stages: a first stage of conformational sampling, to generate an exhaustive ensemble of loop conformations, followed by a scoring step to filter out infeasible conformations and select the most probable ones. The previous chapter introduced a novel method that performs exhaustive loop sampling in a computationally efficient manner. In this chapter, we focus on the second step, *scoring*. Namely, we compare existing state-of-the art methods

on their ability to accurately score diverse sampled loop states and to model a consistent conformational landscape.

Similarly to loop sampling methods, loop scoring methods are mostly evaluated on their ability to identify the "native" conformation among a pool of decoys. However, many applications (e.g. loop design [Kundert 2019]) require to accomplish more complex tasks such as: determining the effective conformational space of a loop, identifying not one but all of its meta-stable states, or even understanding its motions. More than a single stable conformation, these tasks require an accurate description of the energy landscape of the loop. With that in mind, a scoring method must be capable of adequately scoring any loop conformation in accordance with its associated energy.

Physics-based scoring methods (Section 1.3.3) aim at directly approximating potential energy. However, these methods are computationally expensive and prone to modeling energy landscapes that are too rough to be correctly interpreted. Conversely, statistics-based methods (Section 1.3.3) are more computationally efficient and known to model smoother landscapes. However, they rely on data from X-ray crystallography, which only provides a single conformation for a protein loop. In this context, the ability of statistical potentials to correctly assess flexible loops and to identify alternative conformations is uncertain.

The work presented in this chapter compares the performance of several scoring methods on multiple systems comprising a loop known to be flexible. Various scoring methods are used to score loop samples from exhaustive ensembles generated with MoMA-LoopSampler. Results are then used to assess the ability of these methods to identify one or several of the known conformations. Building upon the idea implemented in the previous chapter (Section 3.3.1.4), appropriate 2D projections are employed to visualize and analyze the implicitly modeled conformational landscapes. Consistency of theses landscapes with known loop conformations is verified, and the influence of the surrounding protein conformation is detailed.

By analyzing the produced landscapes in addition to the agreement between known structures and top-scoring loop states, this work aims at identifying the qualitative differences between the results obtained by the various scoring methods. In turn, by examining these differences, we aim at providing guidelines as to which methods to employ depending on the problem at hand, and what conditions should be gathered to expect accurate results. In particular, such a comparison is intended to verify whether the trade-off between computational cost and accuracy offered by faster statistics-based potentials remains appealing.

The remainder of this section lists the loop systems used in this work, and describes the scoring methods that will be compared. Section 4.2 presents the *in silico* protocol employed. Section 4.3 details the results obtained by the different scoring methods on the different systems, while Section 4.4 attempts to summarize the various results and to draw more general trends about the behavior of individual scoring functions.

This work is currently under revision by *Proteins: Structure, Function and Bioinformatics* [Barozet 2019a].

Throughout this chapter, we make the distinction between a *loop*, set of residues and atoms forming a flexible protein fragment; a loop *state*, fully determined by the values of its internal degrees of freedom; and a loop *conformation*, defined as a consensus state, or a limited set of similar states.

4.1.2 Loop systems

Eight flexible protein loops that have been crystallized in at least two different conformations were gathered for this work. The list, together with relevant information, is provided in Table 4.1. Visualization of the different known conformations, along with distances between them are provided in Figure 4.1 and Table 4.2, respectively. All systems, except #1 and #7, were also used in a related publication that motivated our work [Marks 2018].

These systems were chosen so as to provide a variety of loop lengths and protein sizes. The set includes loops with two or three known conformations. Systems #1 (streptavidin) and #7 (triosephosphate isomerase, TPI) were included in this study because they are well-known proteins in which the flexibility of the loop has been shown to play a functional role.

4.1.3 Scoring methods

Different scoring methods were tested in this work (Table 4.3).

The first method employs the AMBER force field ff14SBonlysc [Maier 2015] to score loop states. It is a physics-based method that works on an all-atom protein model, including side-chains. The second method uses the ref2015 ROSETTA scoring function [Alford 2017]. It is a hybrid method that combines physics-based terms such as those employed by AMBER force fields with other statistical terms. This scoring function also uses an all-atom structure with side-chains. Although ROSETTA includes an option to perform relaxations after replacing side-chains with their centroid, this was not tested in this work.

Both AMBER force fields and ROSETTA scoring functions are very sensitive to slight divergences from the 'ideal' geometry or to minor steric clashes. Consequently, the associated methods model very rough molecular energy landscapes where low-energy states border on high-energy ones. In order to better assess the stability of a structural model, a relaxation has to be performed so as to allow the modeled state to fall into the closest basin. Those relaxations can turn out to be prohibitively computationally expensive, especially when the number of loop states to score increases.

All the other tested methods are statistical potentials, which do not require relaxation. Two of them, DFIRE2 [Yang 2008] and SOAP-Loop [Dong 2013] are all-atom potentials that require the side-chains to be placed in the structure to score. DFIRE2 is a simple distance-dependent potential. It uses a single descriptor for each heavy atoms pair: the triplet (a_1, a_2, r) where a_1 and a_2 are the residue-specific atom types of the first and second atoms in the pair, respectively, and r

Table 4.1. Protein loops studied in this work. The table gives the list of PDB-IDs of the structures used as scaffolds for loop sampling. These PDB-IDs are classified according to the conformation of the loop in the corresponding X-ray structure.

#	Name[1]	Loop[2]	Length[3]	PDB scaffolds crystallized in conformation			Scaffold size	Sampled states[4]
				1	2	3		
1	**Streptavidin**	44-52	**9**	2F01(A) 2F01(B) 3RY1(A) 3RY2(A) 3RY2(B)	3RY1(B) 3RY1(D)	3RY1(C)	1,745 to 1,812 atoms 120 to 126 residues	702 to 5,338
2	**MR-MLE**	115-125(126)	**11** (12)	3N4F(A)	3QPE(D) 3VCC(A)	-	5,962 to 6,041 atoms 385 to 390 residues	5,000
3	**NTPase**	41-50(51,52)	**10** (12)	4KFR(B)	4KFU(A)	-	3,422 to 3,473 atoms 205 to 208 residues	5,000
4	**Pot1pC**	109-118(119,120)	**10** (12)	4HID(A) 4HIK(A) 4HIM(A) 4HIO(A) 4HJ9(A)	4HJ7(A)	-	2,298 to 2,310 atoms 138 to 139 residues	5,000
5	**PTPN9**	466-477	**12**	2PA5(A) 4GE2(A) 4GE6(B)	4ICZ(A)	-	4,710 to 4,860 atoms 292 to 304 residues	5,000
6	**RNU2**	29-40	**12**	3AGN(A)	3AGO(A)	3AHW(A)	1,621 atoms 114 residues	5,000
7	**TPI**	165-179	**15**	1YPI(A) 1YPI(B)	2YPI(A) 2YPI(B)	-	3,778 atoms 247 residues	5,000
8	**UTB**	(65)66-76	**11** (12)	3IRS(C)	3K4W(L)	-	13,182 atoms 843 residues	5,000

[1] MR-MLE: Mandelate racemase - muconate lactonizing enzyme; Pot1pC: Protection of telomeres protein 1; PTPN9: Protein tyrosine phosphatase non-receptor type 9; RNU2: Ribonuclease U2; TPI: Triosephosphate isomerase; UTB: Uncharacterized Tim-Barrel protein BB4693.

[2] The residues in parentheses are added to allow MoMA-LoopSampler to work with a multiple of three residues, but their sampled angles are restricted to those found in the X-ray structure.

[3] The length is given in residues. The length in parentheses correspond to the length with added residues.

[4] Number of states sampled *per scaffold*. For streptavidin, the samples are those from the previous chapter (Section 3.3.2.3), using MoMA-LoopSampler in brute force mode.

Table 4.2. Backbone heavy atom RMSD between known loop structures of 8 different proteins. PDB-IDs are colored according to the conformation they contain. The loop is defined as the extended (multiple of 3 residues) loop used for sampling.

Strept.	2F01(A)	2F01(B)	3RY1(A)	3RY2(A)	3RY2(B)	3RY1(B)	3RY1(D)	3RY1(C)
2F01(A)	0	0.16	0.43	0.06	0.14	8.9	8.44	8.23
2F01(B)	0.16	0	0.36	0.17	0.04	8.8	8.35	8.13
3RY1(A)	0.43	0.36	0	0.44	0.37	8.59	8.13	7.93
3RY2(A)	0.06	0.17	0.44	0	0.15	8.9	8.44	8.23
3RY2(B)	0.14	0.04	0.37	0.15	0	8.81	8.35	8.14
3RY1(B)	8.9	8.8	8.59	8.9	8.81	0	0.81	2.48
3RY1(D)	8.44	8.35	8.13	8.44	8.35	0.81	0	2.16
3RY1(C)	8.23	8.13	7.93	8.23	8.14	2.48	2.16	0

MR-MLE	3N4F(A)	3QPE(D)	3VCC(A)
3N4F(A)	0	4.14	4.13
3QPE(D)	4.14	0	0.22
3VCC(A)	4.13	0.22	0

NTPase	4KFR(B)	4KFU(A)
4KFR(B)	0	4.09
4KFU(A)	4.09	0

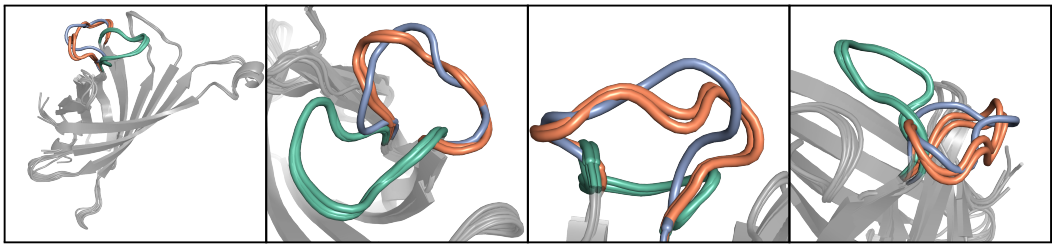
RNU2	3AGN(A)	3AGO(A)	3AHW(A)
3AGN(A)	0	6.93	5.81
3AGO(A)	6.93	0	5.29
3AHW(A)	5.81	5.29	0

Pot1pC	4HID(A)	4HIK(A)	4HIM(A)	4HIO(A)	4HJ9(A)	4HJ7(A)
4HID(A)	0	0.78	0.7	0.21	0.71	2.22
4HIK(A)	0.78	0	0.18	0.73	0.21	2.41
4HIM(A)	0.7	0.18	0	0.63	0.17	2.42
4HIO(A)	0.21	0.73	0.63	0	0.66	2.31
4HJ9(A)	0.71	0.21	0.17	0.66	0	2.35
4HJ7(A)	2.22	2.41	2.42	2.31	2.35	0

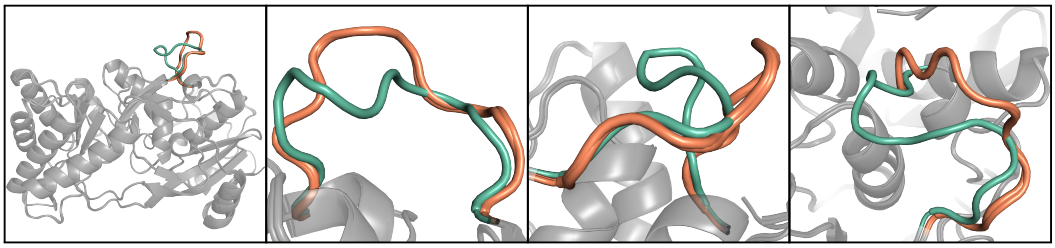
PTPN9	2PA5(A)	4GE2(A)	4GE6(B)	4ICZ(A)
2PA5(A)	0	0.13	0.76	3.45
4GE2(A)	0.13	0	0.8	3.49
4GE6(B)	0.76	0.8	0	3.18
4ICZ(A)	3.45	3.49	3.18	0

UTB	3IRS(C)	3K4W(L)
3IRS(C)	0	4.26
3K4W(L)	4.26	0

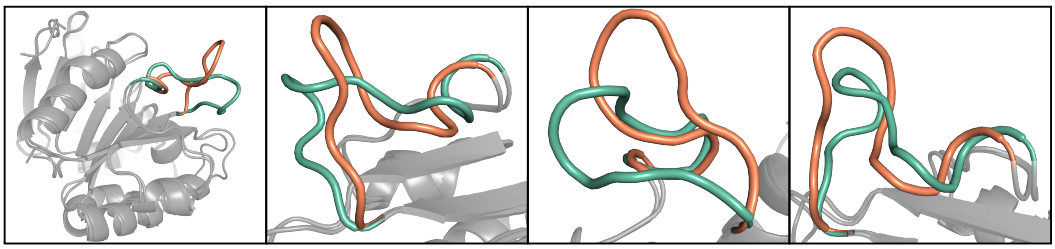
TPI	1YPI(A)	1YPI(B)	2YPI(A)	2YPI(B)
1YPI(A)	0	0.37	4.24	4.15
1YPI(B)	0.37	0	4.28	4.19
2YPI(A)	4.24	4.28	0	0.53
2YPI(B)	4.15	4.19	0.53	0



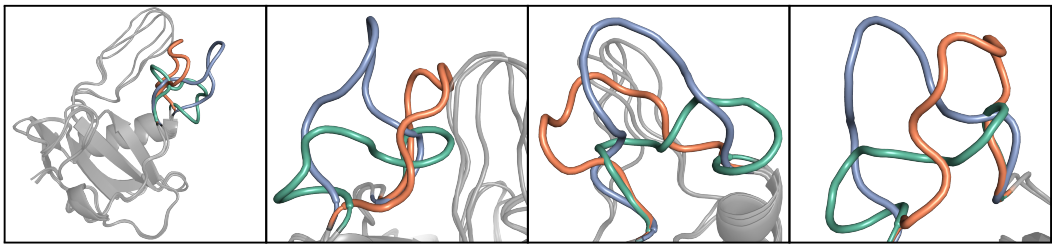
(a) Streptavidin



(b) MR-MLE

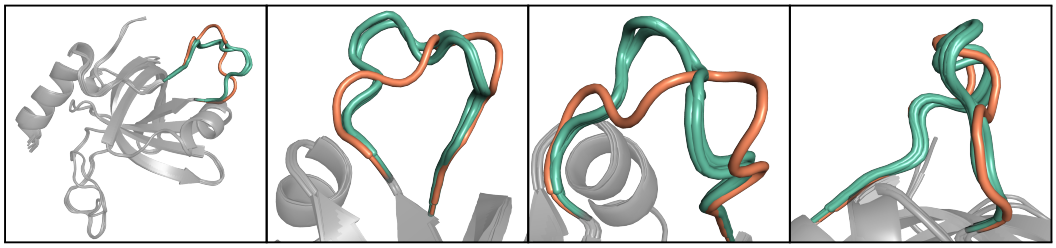


(c) NTPase

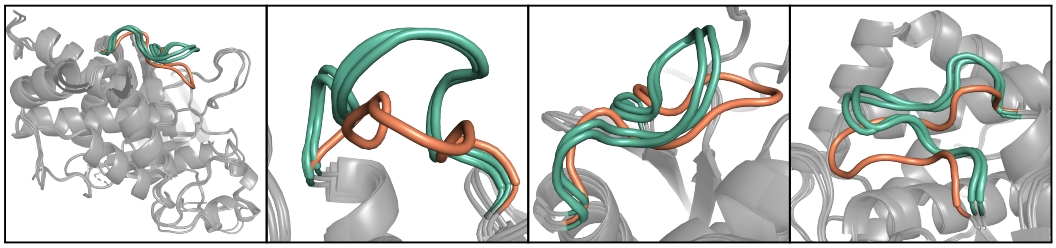


(d) RNU2

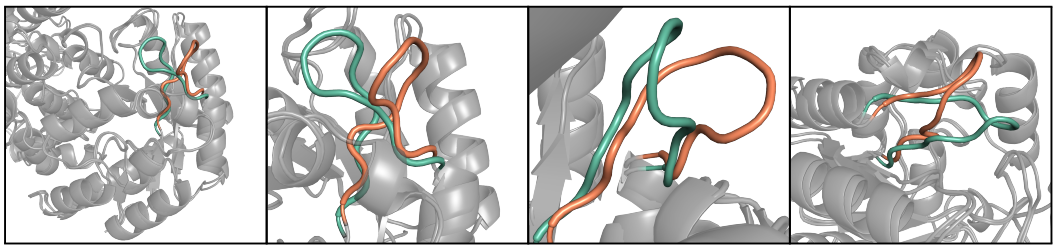
Figure 4.1 (first part). Cartoon views of the crystallographic structures of the different flexible loops studied in this work.



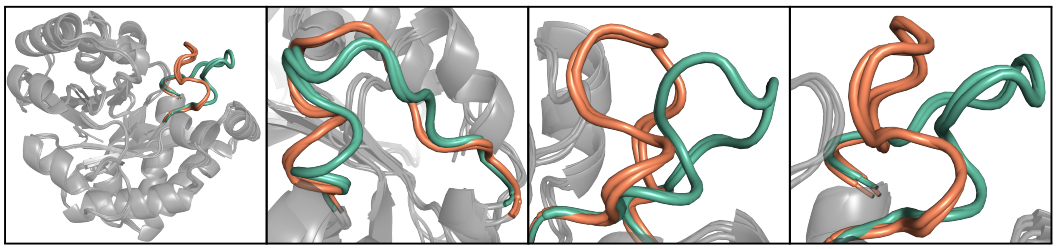
(e) Pot1pC



(f) PTPN9



(g) UTB



(h) TPI

Figure 4.1 (cont.). Cartoon views of the crystallographic conformations of the different flexible loops studied in this work.

Table 4.3. Scoring methods compared in this work.

Method	Type	Relaxation	Side-chains
AMBER	Physics-based	Needed	Needed
ROSETTA	Hybrid	Needed	Needed
DFIRE2	Statistical	-	Needed
SOAP-Loop	Statistical	-	Needed
KORP	Statistical	-	-
SBROD	Statistical	-	-
Torsions-only	Statistical	-	-

is the distance between them. SOAP-Loop also employs inter-atomic distances as descriptors, but includes terms related to the orientation between pairs of covalent bonds and the relative atomic surface accessibility.

The last three methods tested in this work are KORP [López-Blanco 2019], SBROD [Karasikov 2018] and a very simple statistical potential solely based on the loop's ϕ and ψ dihedrals [Rata 2010], which will subsequently be called *Torsions-only*. They are coarse-grained potentials and none of them requires the side-chains to be modeled. KORP only needs the positions of the N, C_α and C backbone atoms. It uses one distance and five angular features to describe the relative position and orientation of each amino-acid pair. Torsions-only uses the amino-acid type, along with the ϕ and ψ angles of the residues in the loop as the only descriptors. SBROD uses many features gathered into four groups: residue-residue pairwise features, backbone atom-atom pairwise features, hydrogen bonding features and solvent-solvate features. While DFIRE2, SOAP-Loop, KORP and Torsions-only are all Bayesian-based, SBROD uses the *Ridge Regression* machine learning technique to optimize the weights in the linear model. Another major difference between SBROD and the other statistical potentials compared in this work is the training dataset. DFIRE2, SOAP-Loop, Torsions-only and KORP use non redundant sets of protein structural data to derive observed frequencies of the different features. Conversely, SBROD is trained on sets of decoy models and is designed to discriminate between well-folded and misfolded structures, which makes it fundamentally different from the other knowledge-based potentials.

4.2 Methods

4.2.1 Preprocessing of structure files

Structural data corresponding to the IDs listed in Table 4.1 were extracted from the PDB. Only one monomer was kept from the assymetric unit. Ligands, ions, water molecules and other non-protein elements were removed. Hydrogen atoms that were originally present in the model were also stripped, and all hydrogen atoms were re-placed using AMBER 16 [Case 2005, Case 2016]. In order to remove major steric clashes and idealize the scaffold geometry, the structures were relaxed

using AMBER 16's energy minimization protocol [Case 2005, Case 2016], with a resulting median RMSD of 0.2 Å on all heavy backbone atoms upon relaxation.. The first 1000 cycles used steepest descent, while the remaining cycles were run using conjugate gradient. The maximum number of cycles was set to 2000 and the minimization was considered to have converged when the root-mean-square of the Cartesian elements of the gradient was lower than 0.1 kcal/(mol·Å). We used the ff14SBonlysc force field and a simple Generalized Born implicit solvent model ($igb = 1$ and the *mbondi* radii sets, as recommended in the AMBER manual).

4.2.2 Sampling loop states

Loop sampling was performed using MoMA-LoopSampler (Chapter 3). Once a state was sampled for the backbone, the side-chains were placed employing the following protocol.

Side-chains are sampled one after the other in a random order. Dihedral χ angles are randomly sampled following the probability distributions associated with the continuous rotamers implemented in BASILISK [Harder 2010]. Initially, soft spheres are employed to model the atoms, allowing some limited interpenetration. However, when a strong collision with the protein backbone or another placed side-chain is detected, the χ angles are re-sampled. After a certain number of unsuccessful tries, backtracking is employed to rearrange previously placed side-chains. After all the side-chains in the loop have been placed, strong collisions with surrounding side-chains in the protein are solved. χ angles of the involved residues are randomly perturbed until the collision is solved, if possible. A second pass is performed with harder atom spheres: one side-chain after the other is checked for collision and, if necessary, its dihedral angles are randomly perturbed to solve the steric clash. The whole process is attempted twice, and a timer limits the total duration of side-chain placement process. In case of failure, the sampled backbone state is rejected.

Each scaffold was employed to sample 5,000 states with side-chains using MoMA-LoopSampler, except for the 8 streptavidin scaffolds, for which the loop length and the constrained environment allowed an exhaustive brute force sampling, concatenating all possible fragments from the tripeptide library (see Section 3.3.2.3).

4.2.3 Scoring loop states

For every scoring method except AMBER and ROSETTA, the binaries to score loop states were either downloaded from the dedicated websites, or provided by the authors. After adapting the structural data to the input formats requested by the different methods, the binaries were used to score each individual loop state. For AMBER and ROSETTA, sampled states were first relaxed before being scored, with the following relaxation protocols.

AMBER: The full structures were relaxed and then scored using AMBER's ff14SBonlysc force field [Maier 2015] and a simple Generalized Born implicit sol-

vent model ($igb = 1$ and the *mbondi* radii sets, as recommended in the AMBER manual). The relaxation was performed using the energy minimization protocol provided by the AMBER 16 biosimulation package [Case 2005, Case 2016]. The first 250 cycles of energy minimization used steepest descent, while the remaining cycles were run using conjugate gradient. The maximum number of cycles was set to 500 and the minimization was considered to have converged when the root-mean-square of the Cartesian elements of the gradient was lower than 0.1 kcal/(mol·Å).

ROSETTA: The full structures were relaxed and then scored using ROSETTA's ref2015 scoring function. The relaxation was performed on the full atom model, but the backbone dihedrals outside the loops were fixed. The default number of cycles (5) was employed. Two repeats of the relaxations were performed for each sampled state, and only the best final energy was kept.

4.2.4 Landscape reconstruction

The combination of the sampled states and their associated scores can be used to represent the energy landscape modeled by a given scoring method. Each sampled state corresponds to a point in an n-dimensional space where n corresponds to the number of degrees of freedom of the sampled loop. In order to visualize these points, they were projected in 2D space using two rational descriptors, one related to the global position of the loop, and another related to its internal configuration, following the idea proposed in Section 3.3.1.4. The choice of a specific projection for a given loop system was based on the readability of the obtained landscape.

The 2D space was then discretized using a grid (using 40 bins on each axis), where each cell was colored according to the best score of all the states projected within it. The presence of a few cells with a score far above the mean tends to make the landscape appear flat. To circumvent this, cells whose associated score was more than three standard deviations above the mean were considered empty (as if no sampled state was projected there). Empty cells were considered to have the maximum score observed among the populated cells. A bicubic interpolation method was used to smooth the landscapes.

4.3 Results

This section presents results from several perspectives. We start by analyzing the quality of the sampled conformational ensembles (Section 4.3.1), because sampling is the first bottleneck of landscape modeling. Indeed, if statistically likely conformations are missing in the generated ensemble, the energy landscape will be inaccurate, whatever scoring method is subsequently used. Next, since they are a determining criterion in the choice of a scoring method, the running times observed for the different functions are detailed (Section 4.3.2). The relationship between known conformations and scores is then examined. Such an analysis is complex for two reasons. The first reason is that known conformations are not necessarily

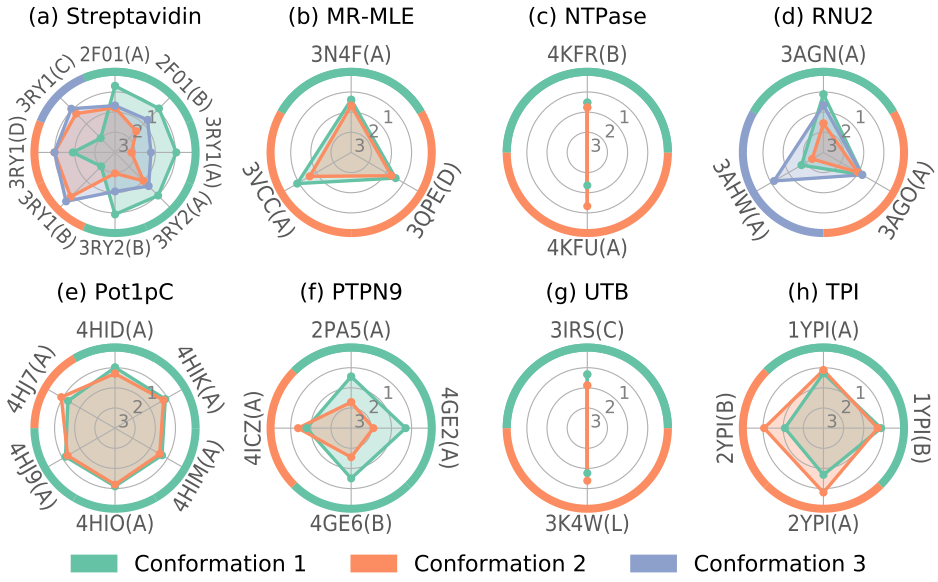


Figure 4.2. Lowest RMSD to each known stable conformations among sampled states, for each scaffold. RMSDs are calculated on the heavy atoms of the backbone. Scaffolds are distributed around the disk and their names indicated outside the disk. The outer circle is colored according to the conformation present in the X-Ray structure the scaffold originated from. The outer circle corresponds to a RMSD of 0 Å, while the center corresponds to RMSDs of 4 Å and above: a larger colored area indicates that the corresponding stable conformation is found with greater accuracy in the sampled ensemble.

the most stable conformations: instead, they may be stabilized by ligands, crystal contacts or metallic ions. However, the fact that they are observed does indicate that they are statistically likely conformations, which should not be eliminated by a filtering method. The second reason why this analysis is challenging is that the methods we compare provide scores and not binary good/bad classification, so that eliminating unlikely conformations requires setting an arbitrary threshold, either on the score itself or on the rank. Given these considerations, we analyze how known conformations are scored from two different points of view. We first report the ranks measured for the the sampled states that are closest to known conformations (Section 4.3.3), and then provide the distances between top-scoring states and known conformations (Section 4.3.4). In order to determine whether the different functions can agree despite their fundamental differences, correlations between the scoring methods were also analyzed. The results are provided in Section 4.3.5. Finally, the landscapes implicitly modeled by the different scoring functions are presented, aiming to provide a more exhaustive comparison of the results they provide (Section 4.3.6).

4.3.1 Sampling known conformations

Figure 4.2 shows the distance of the closest sampled state to each experimentally determined conformation, from each scaffold. Note that throughout the results section, distances between two loop states are given as the RMSD of the heavy atoms of the backbone. To better understand what the figure shows, let us illustrate with the example of scaffold 3AHW(A) from RNU2. The scaffold comes from a crystal structure with the loop in conformation 3 (as indicated by the color of the outer circle). From this scaffold, a state within 1 Å of conformation 3 was sampled. However, no state was sampled closer than 2.7 Å from conformation 1 or 3.2 Å from conformation 2.

Overall, the results suggest that the known conformations are usually retrieved from the sampled ensemble. One can however notice a limitation: the crystallographic structure corresponding to the employed scaffold is almost always sampled more closely than the other experimentally-determined conformations. Although the presence of such a bias is not surprising, the RMSD difference can be substantial in some cases (e.g. conformation 1 for streptavidin, or conformation 2 for PTPN9). A closer observation of the concerned structures reveals large rearrangements in the loop environment that accompany the loop conformational change and thus hinder the sampling of some regions of the loop's conformational space, where other known conformations could be found. For example in streptavidin, the position of the backbone of GLU-51 in conformation 1 coincides with the position of the side-chain of ARG-84 in conformations 2 and 3. As a consequence, loop conformation 1 cannot be sampled very closely from scaffolds originating from conformations 2 or 3. Similarly in RNU2, there are major backbone collisions between the loop in conformation 2 (GLY-33 - ASP-34) and the scaffold in conformation 3 (SER-74 - ARG-75).

This problem illustrates a limitation common to all loop sampling methods that consider the rest of the protein as a rigid body, and underlines the necessity to better account for flexibility outside the loop. A possibility could be to remove all the side-chains in a large surrounding of the loop anchors before sampling, although that may unnecessarily broaden the space accessible to the loop and make an exhaustive sampling harder. Another alternative could be to employ a scaffold ensemble instead of a single one. This is further discussed in Section 4.4.

4.3.2 Running times

Running times differ by several order of magnitude from one scoring method to another, and obviously depend on the system's size. Figure 4.3 reports the average time required to score one sampled state for three systems of different sizes. ROSETTA and AMBER are by far the most costly methods, with comparable running times using the relaxation protocol adopted in this work. SOAP-Loop is the slowest statistical method, possibly due to its evaluation of the atomic surface accessibility. SBROD is the next method in terms of running time, followed by

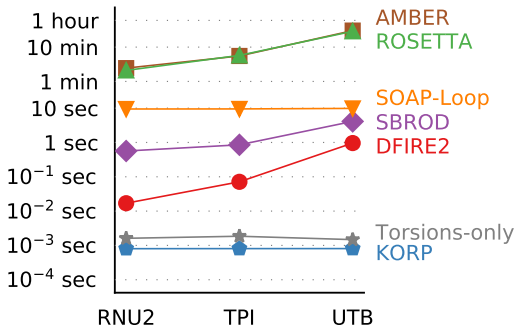


Figure 4.3. Running times per scored sampled state measured on three different systems. These were obtained on a single core of a 2GHz Intel® Xeon® processor. Note that the scale on the y-axis is logarithmic. RNU2, TPI and UTB scaffolds contain 114, 247, and 843 atoms each, respectively.

DFIRE2. Torsions-only and KORP are the least expensive methods, presumably because the former only employs the provided ϕ and ψ angles of the loop and the latter offers a convenient batch mode, where the structure of the whole protein is only provided once.

4.3.3 Ability to rank near-native conformations

Figure 4.4 gives the ranks of the five closest sampled states to each known conformation. As an example to read this figure, let us consider the sixth line corresponding to known conformation 1 of NTPase loop. The upper parts of the disks relate to the five closest states to conformation 1 sampled from the first scaffold (4KFR(B)). The ranks obtained using AMBER place the third closest state to conformation 1 among the ten states with the lowest energies, while the other four closest states have a rank above 100. None of the other scoring method are able to place any of the five states closest to conformation 1 among the ten best scoring states.

There are global trends regarding how well the states in the vicinity of the different known conformations are scored. These trends are unsurprisingly system-dependent: sampled states similar to known conformations of PTPN9 are well identified, while states similar to known conformations of NTPase are not. The accuracy with which known conformations are sampled can partially explain this trend, but other factors are needed to explain these inter-system differences, such as the intrinsic flexibility of the loop, or how favorable its surroundings are to the creation of attractive or repulsive contacts. Disparities are also observed between the different conformations of a single system: the states close to conformation 1 of UTB are better scored than those close to conformation 2. The nature of the conformation (clear stabilizing contacts, “canonical” shape, ...) may be a determining factor.

Overall, DFIRE2 and KORP are the ones that best identify states in the vicinity of known conformations. KORP is the method that gives the best scores to

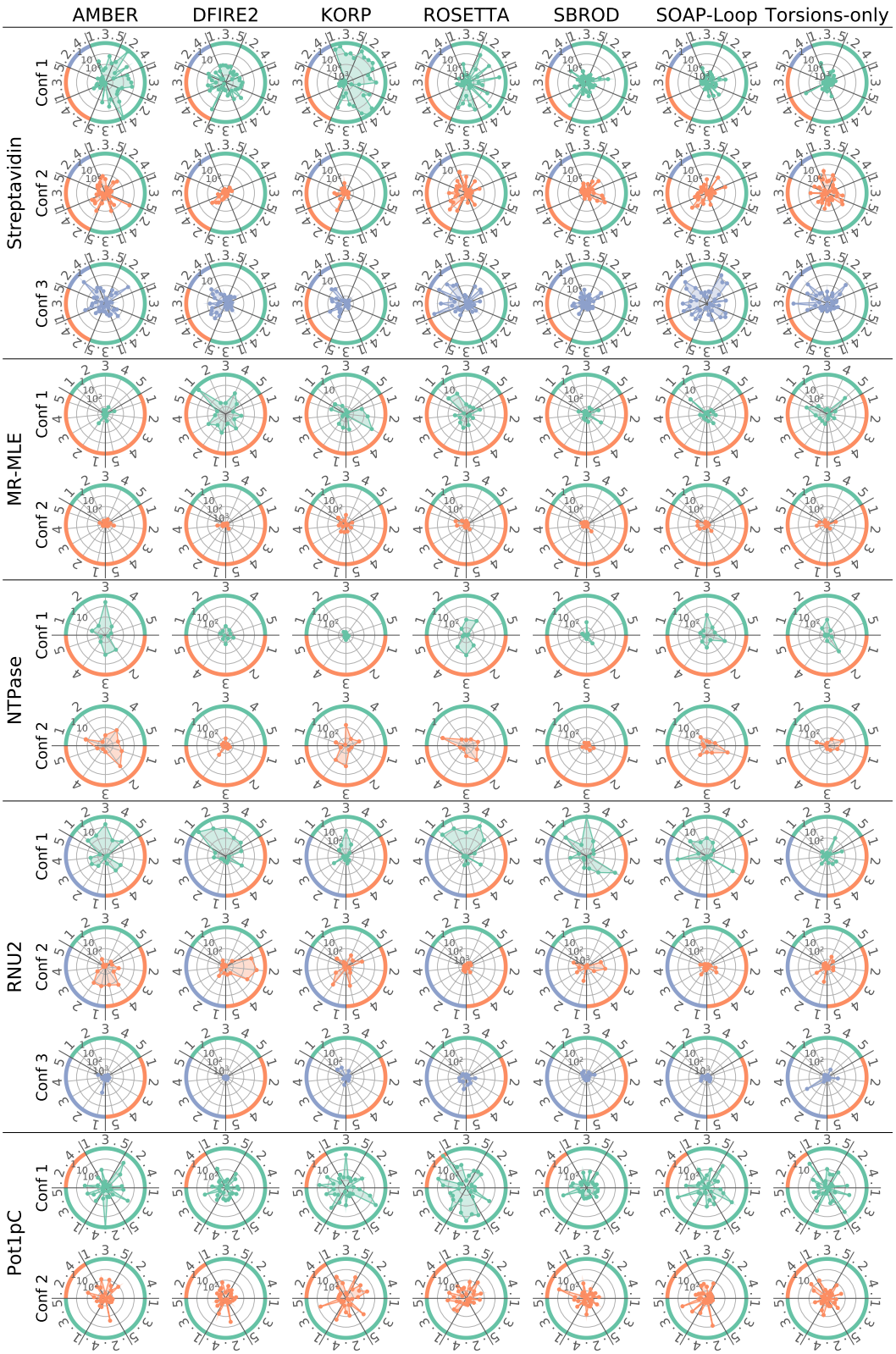


Figure 4.4 (first part)

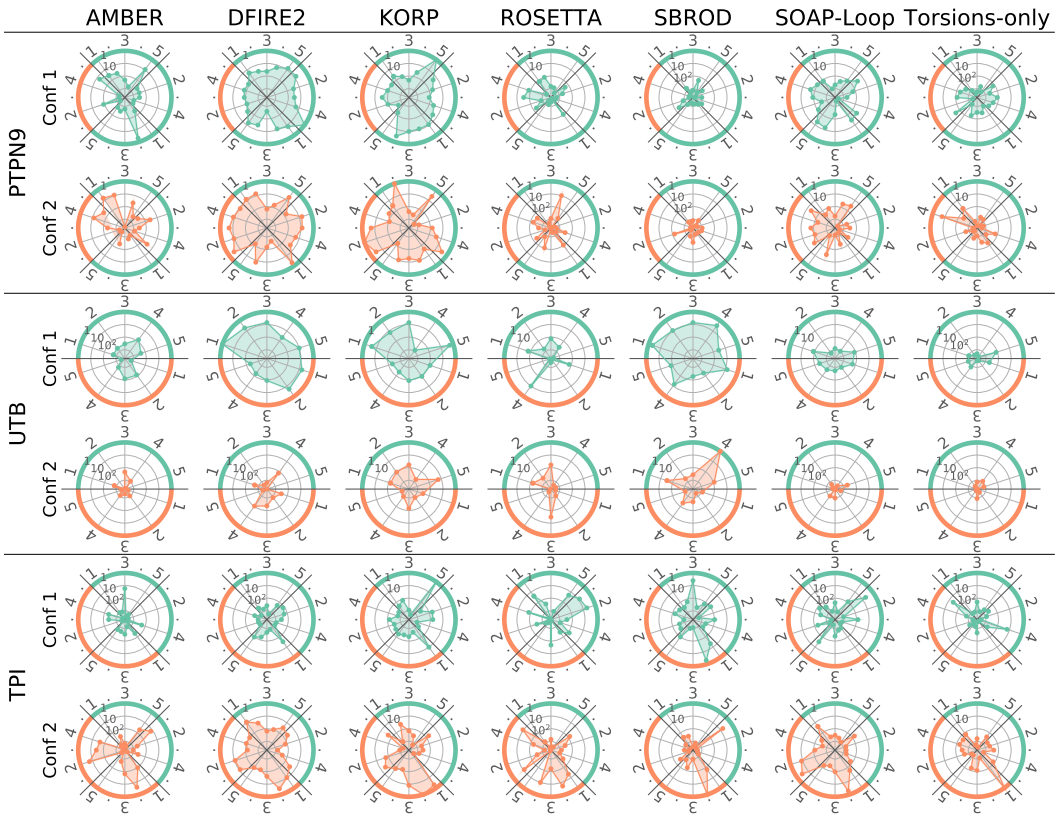


Figure 4.4 (cont.). Ranks of the five closest sampled states to each known conformations, from each scaffold. Scaffolds are distributed around the disk and separated by thicker dark grey lines. Their names have been omitted for clarity but they are arranged as in Figure 4.2. The outer circle is colored according to the conformation present in the X-Ray structure the scaffold originated from. Each line corresponds to a given known conformation. The ranks are represented by the proximity to the outer circle or to the center. The outer circle corresponds to the best scoring ranks, while the center corresponds to the worst rank. Note that the radial axis has a logarithmic scale. With such a representation, points far from the center of the circle correspond to well-scored sampled states.

sampled states around conformation 1 of streptavidin. It is rather consistent across the different conformations to identify, but fails to detect any state close to conformations 2 and 3 of streptavidin, conformation 1 of NTPase, conformation 2 of MR-MLE and conformation 3 of RNU2 (but these last two conformations are not well identified by any of the tested scoring methods). DFIRE2 performs slightly better than KORP on conformation 3 of streptavidin but otherwise misses the same conformations and conformation 2 of NTPase. Despite their similar overall success, KORP and DFIRE2 do not identify the same near-native states, perform differently depending on the sampling scaffold and can be of different precision when identifying a conformation.

AMBER rarely fails completely for a system, but it is less robust than KORP or DFIRE2, with ranks varying substantially from one of the 5 closest states to another. This may come from a lack of convergence of the relaxations, or from the

roughness of the conformational landscape modeled by this function. ROSETTA has a similar performance and the same shortcoming concerning robustness. This is not surprising since ROSETTA also models a rather rough landscape and heavily depends on the prior relaxation. SOAP-Loop, despite not needing a relaxation, like AMBER and ROSETTA do, obtains similar results to these two methods. SBROD rarely scores near-native states better than other methods and performs badly overall. This may be due to the major differences in the way this method was designed compared to other statistical methods. Finally, Torsions-only is the method with the least satisfying results if compared to other scoring functions. However, taking into account its extreme simplicity, results are still remarkable, and it turns out to be as good as other methods at identifying states similar to known conformations for Pot1pC or NTPase.

4.3.4 Top scoring loop states

For each of the five top-scoring sampled states determined by each method (for each system and from each scaffold), we identified the closest known conformation by computing the cartesian RMSD for the backbone heavy atoms (Figure 4.5).

Both KORP and DFIRE2 perform well in identifying known conformations among top-scoring states. The main difference appears for streptavidin: KORP ranks near-native states as top-scoring while states best ranked by DFIRE2 are far from any known conformation. Again AMBER, ROSETTA and SOAP-Loop show similar results, placing states similar to known conformations as top-scoring for streptavidin, Pot1pC, and TPI. SBROD only identifies known conformations among the five top scoring states for TPI, UTB and Pot1pC, with a milder success than other methods for this last system. Torsions-only does so for streptavidin, Pot1pC and for one state of the TPI loop.

It should be noted that there may exist statistically probable conformations different from those observed in the available crystallography structures. Indeed, the fact that no known conformation appear among the top-ranked states may be due to the identification of other locally stable conformations. However, this is impossible to confirm in the absence of additional structural data. The current accessible information designates KORP and DFIRE2 as the most reliable methods from the top-scoring-states point of view.

4.3.5 Correlation between scoring methods

Agreement between the different scoring methods was assessed by looking at correlations between rankings. Each measure consists of the ranking of one sampled state and was assigned a weight to balance the influence *i)* of the different loop systems, *ii)* of the different starting conformations of a same loop, *iii)* of the different scaffolds containing the same starting conformation, *iv)* of the different states sampled from the same scaffold.

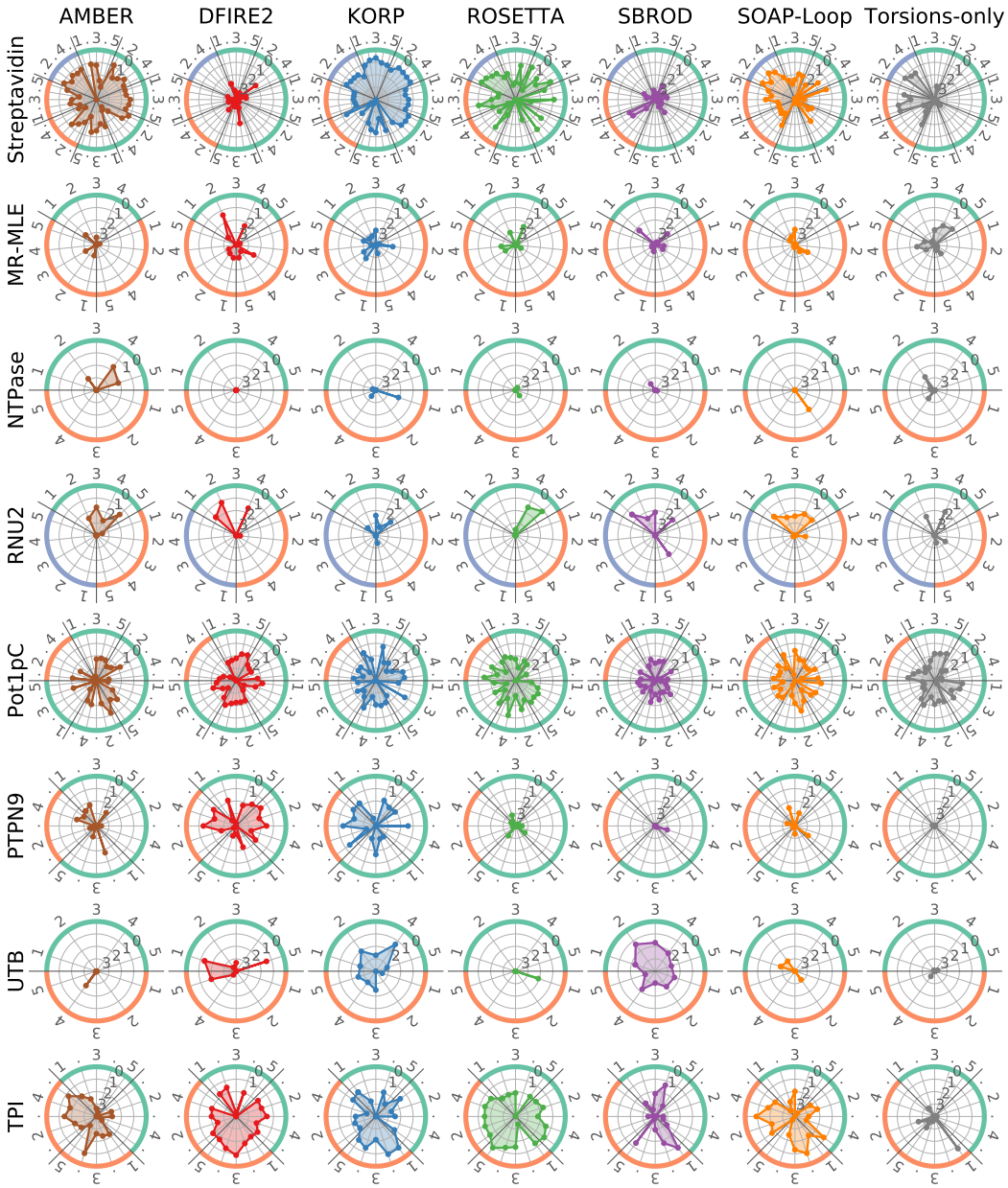


Figure 4.5. Backbone RMSD to the closest known conformation of the five top-scoring states among those sampled from a unique scaffold. Scaffolds are distributed around the disk and separated by thicker dark grey lines. Their names have been omitted for clarity but they are arranged as in Figure 4.2. The outer circle is colored according to the conformation present in the X-Ray structure the scaffold originated from.
The outer circle corresponds to a RMSD of 0 Å, while the center corresponds to RMSDs of 4 Å and above: a larger colored area indicates that the top scoring states are close to a known stable conformation.

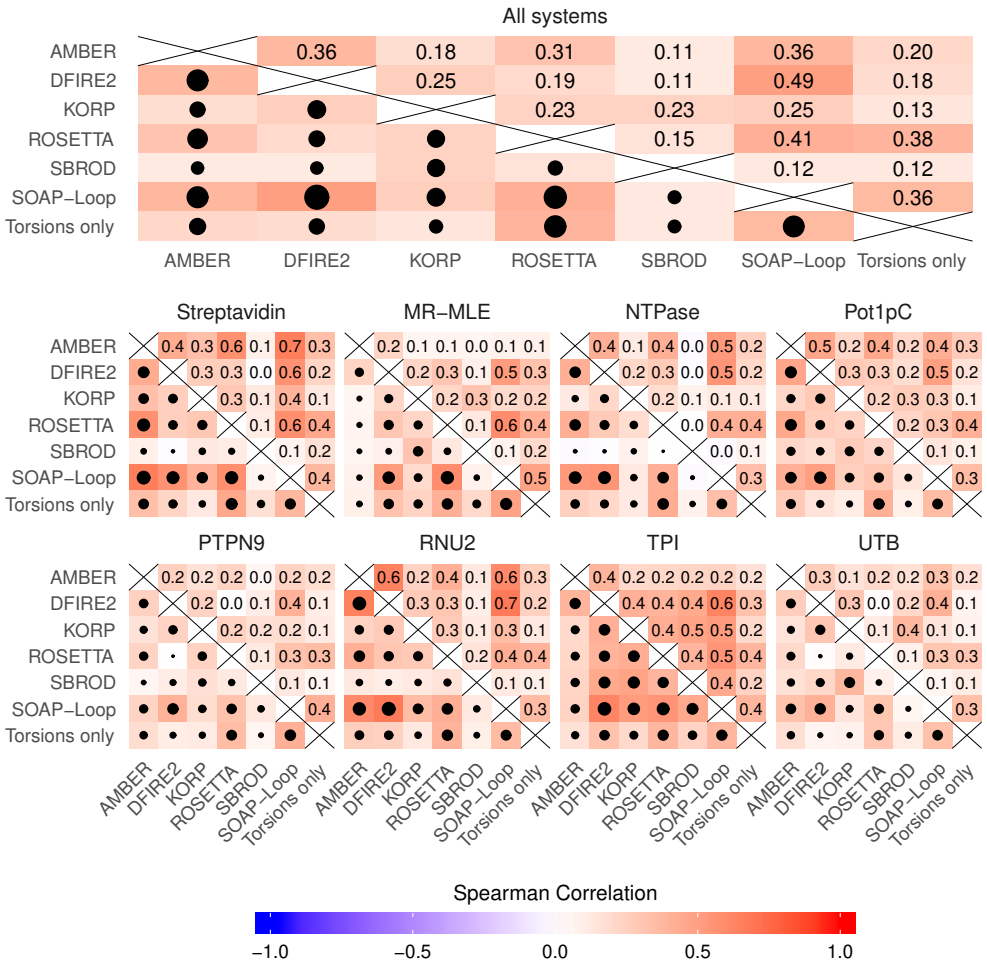


Figure 4.6. Correlations between the different scoring methods calculated using the Spearman correlation test and weighted measures.

Correlations were first calculated using the Spearman coefficient with rank differences weighted as described above (Figure 4.6). SOAP-Loop is the method that best agrees with all others. Conversely, SBROD does not correlate very well with other methods, which is not surprising given the difference in the underlying principles. The greatest correlation between two methods is obtained between DFIRE2 and SOAP-Loop. Interestingly, ROSETTA and Torsions-only have a relatively high correlation coefficient, showing the strong weight given to the dihedral angles term in the ROSETTA scoring function.

Figure 4.7 contains heatmaps showing more precisely for which ranks the methods are in agreement (using the same weights as previously). All pairs of methods show a higher agreement on the very first ranked and the very last ranked states. In other words, they agree on most very 'good' and most very 'bad' states. Yet, most pairs of methods disagree for the states that are not in the extremities of the

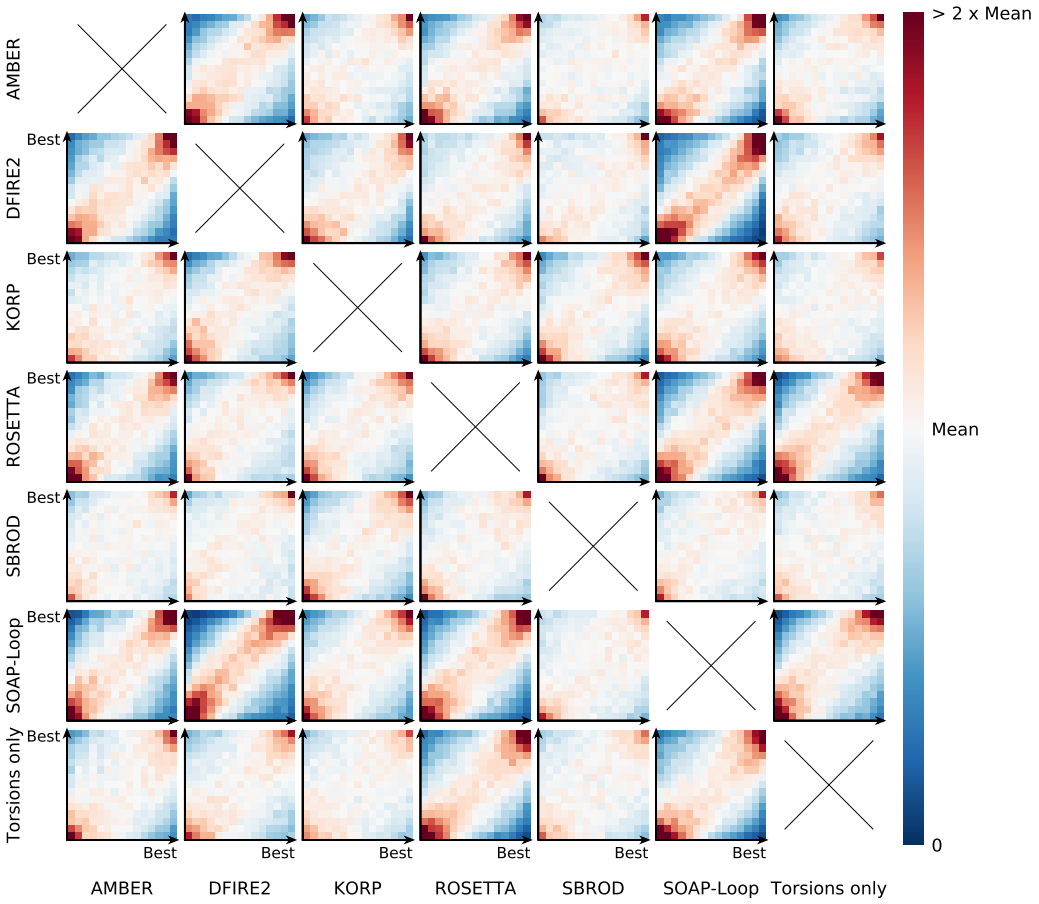


Figure 4.7. Distribution of ranks for the different pairs of methods. x-axes give the ranks as scored by the method indicated at the bottom. y-axes give the ranks as scored by the method indicated on the left. Ranks in both directions are binned so as to obtain a heatmap giving the distribution of ranks for each method pair.

rankings. However, the agreement between DFIRE2 and SOAP-Loop is particularly striking, at all levels of rankings.

Despite small differences in the precise objective that these methods are designed for, they are all meant to give good scores to structures that are statistically likely, while badly scoring those that are highly improbable. Thus, it is expected and even desirable for the methods to show correlation. Still, the major differences in the underlying principles of these methods make their correlation remarkable, confirming some common trends between physics- and statistics-based methods.

4.3.6 Modeled energy landscapes

A direct analysis of the energy landscapes modeled by the different scoring methods in combination with MoMA-LoopSampler allows to gain a more global insight into

the topography induced by these methods. The produced landscapes are depicted on Figures 4.8 through 4.15.

For streptavidin (Figure 4.8), only AMBER manages to clearly identify the different basins. The other methods sometimes identify the basin around conformation 1 or the one around known conformations 2 and 3, but not both of them.

In MR-MLE modeled landscapes (Figure 4.9), DFIRE2 places known conformation 1 in a basin, whatever the scaffold conformation is. KORP, however, places known conformation 2 in a basin for all scaffolds. Even though both conformations can be identified as stable conformations for this system, none of the methods clearly identifies both basins. Both ROSETTA and AMBER model a rather flat landscape.

Landscapes obtained for Pot1pC (Figure 4.11) are consistent with the known conformations for this system. Depending on the method, a basin is identified in the area around conformations 1 or 2 or between them.

The case of PTPN9 (Figure 4.12) is a very good illustration of the power of statistical methods. One of the two known conformations is not in an area sampled by the method but the other is located on the edge of the projected landscape. KORP and DFIRE2 very clearly identify that area as a deep basin whereas landscapes obtained by other methods, and in particular AMBER and ROSETTA, are much less consistent with the crystallographic structures of the loop. The landscapes modeled by KORP or DFIRE2 could guide a more thorough sampling around that basin, possibly indicating which side-chains to remove to allow sampling in that area.

KORP is the method that produces the most consistent and precise landscapes for TPI (Figure 4.14). While most methods place both stable conformations in a very vast basin, KORP models one relatively narrow landscape, deeper around the crystallographic conformation originally present in the scaffold. The example of TPI clearly shows the influence of the starting scaffold. While for some systems the landscape remains unchanged when other scaffolds are used, landscapes produced for TPI from scaffolds originally in conformation 1 (1YPI(A), 1YPI(B)) are clearly different from those produced from scaffolds in conformation 2 (2YPI(A), 2YPI(B)), with a clear displacement of the main basin towards one or the other known conformation.

Landscapes modeled for UTB by AMBER and ROSETTA are very rough (Figure 4.15), making it difficult to draw any conclusion from them. Landscapes generated by statistical methods are again more consistent with experimental data.

The different landscapes overall confirm that statistical methods produce smoother landscapes. The main pitfall of this is that they lack precision and are fuzzy. While they usually manage to identify a main basin containing the known conformations, they are rarely capable of differentiating them. Overall, DFIRE2 and KORP produce the most consistent landscapes. Considering their extreme simplicity and speed compared to AMBER and ROSETTA, they constitute a very good choice for a first analysis of the sampled structures, to filter out very improbable conformations or to select an area to sample more thoroughly.

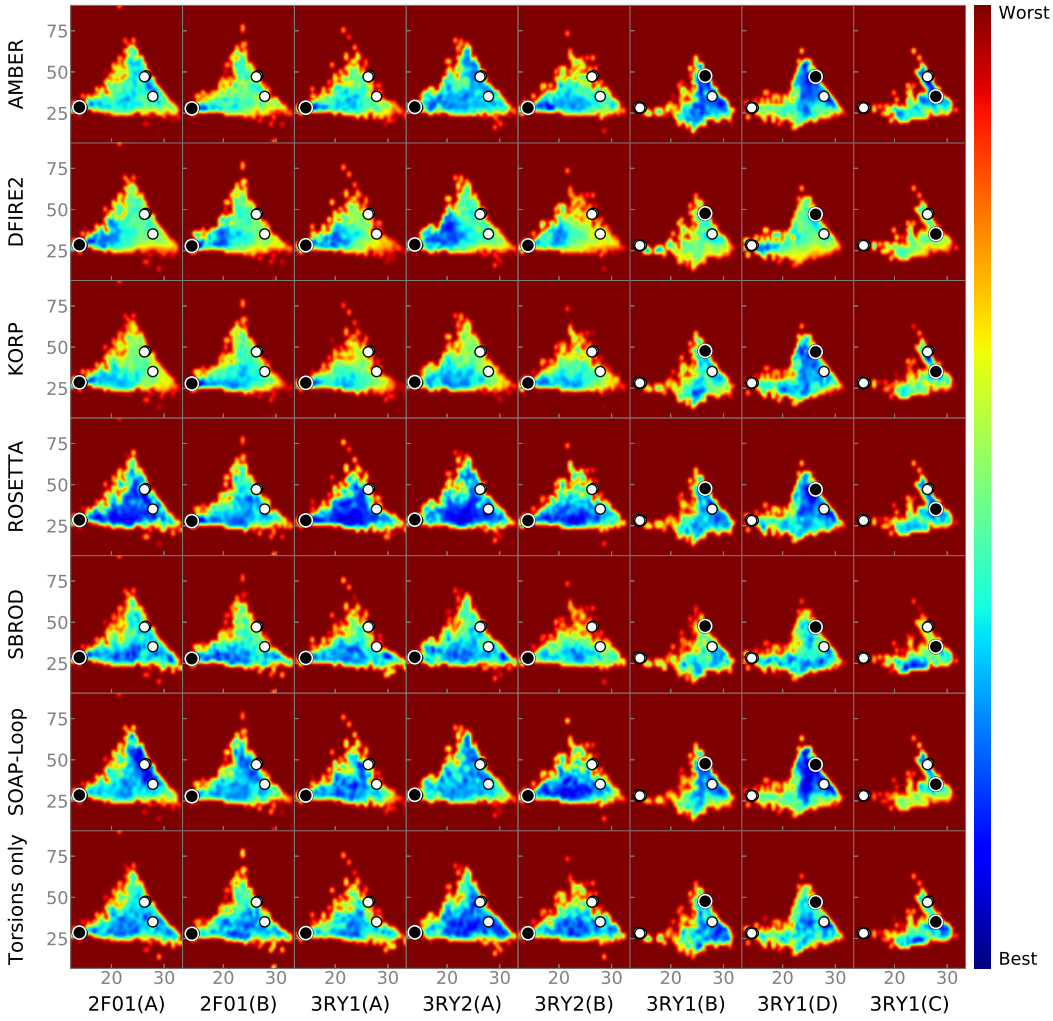


Figure 4.8. Energy landscape obtained by the different scoring methods on the Streptavidin loop.
Each row correspond to a single scoring method and each column to a starting scaffold used for the conformational sampling. The x- and y- axes correspond to the first and second dimensions of the projection, respectively (see Section 4.2.4). The first dimension is the distance between a given atom of the loop and another fixed atom of the protein in Å. The second dimension is the angle formed by three given backbone atoms of the loop in degrees.
In each landscape, the loop conformation initially in the crystal structure used as scaffold is projected as a black point while other known conformations are projected as white points.

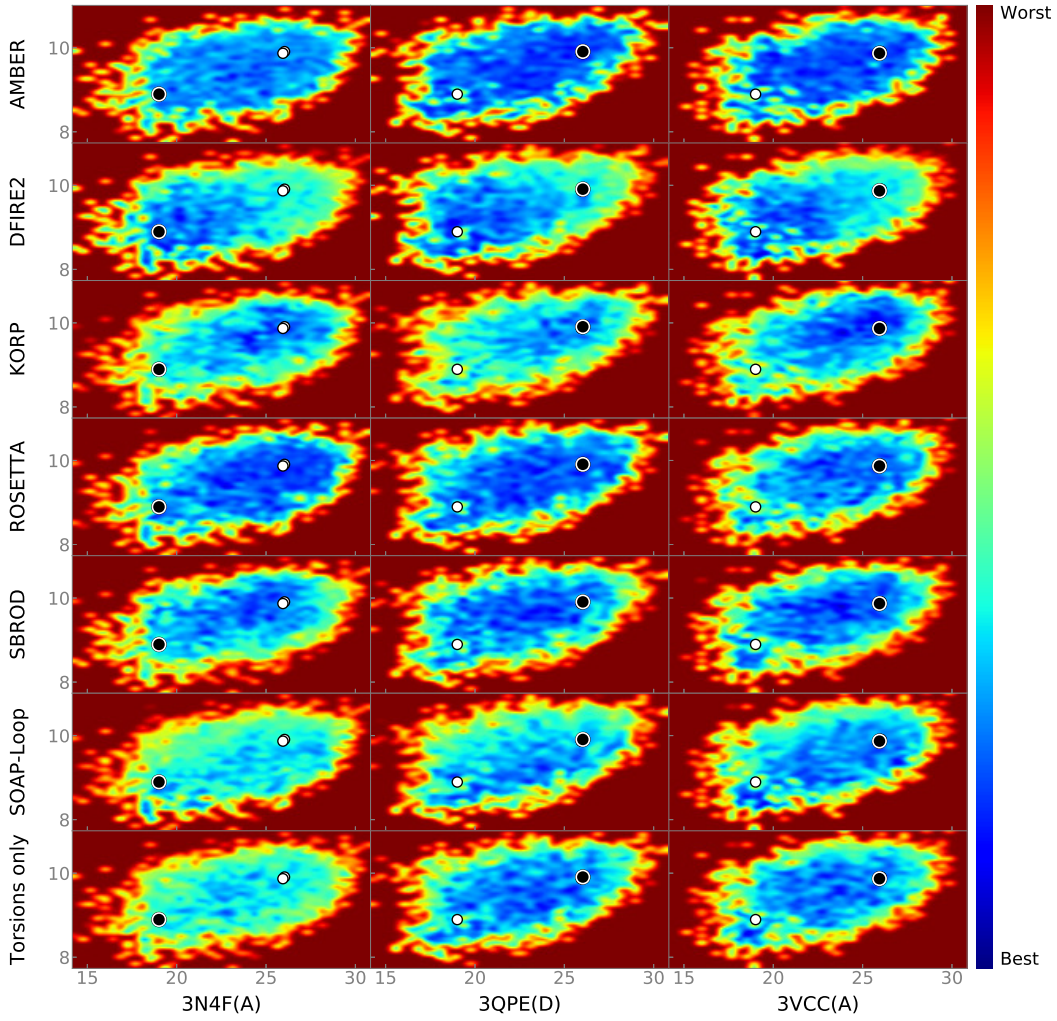


Figure 4.9. Energy landscapes obtained by the different scoring methods on MR-MLE loop. Each row correspond to a single scoring method and each column to a starting scaffold used for the conformational sampling. The x- and y- axes correspond to the first and second dimensions of the projection, respectively (see Section 4.2.4).
The first dimension is the distance between a given atom of the loop and another fixed atom of the protein in Å. The second dimension is the average distance between all pairs of C_α atoms of the loop, in Å.
In each landscape, the loop conformation initially in the crystal structure used as scaffold is projected as a black point while other known conformations are projected as white points.

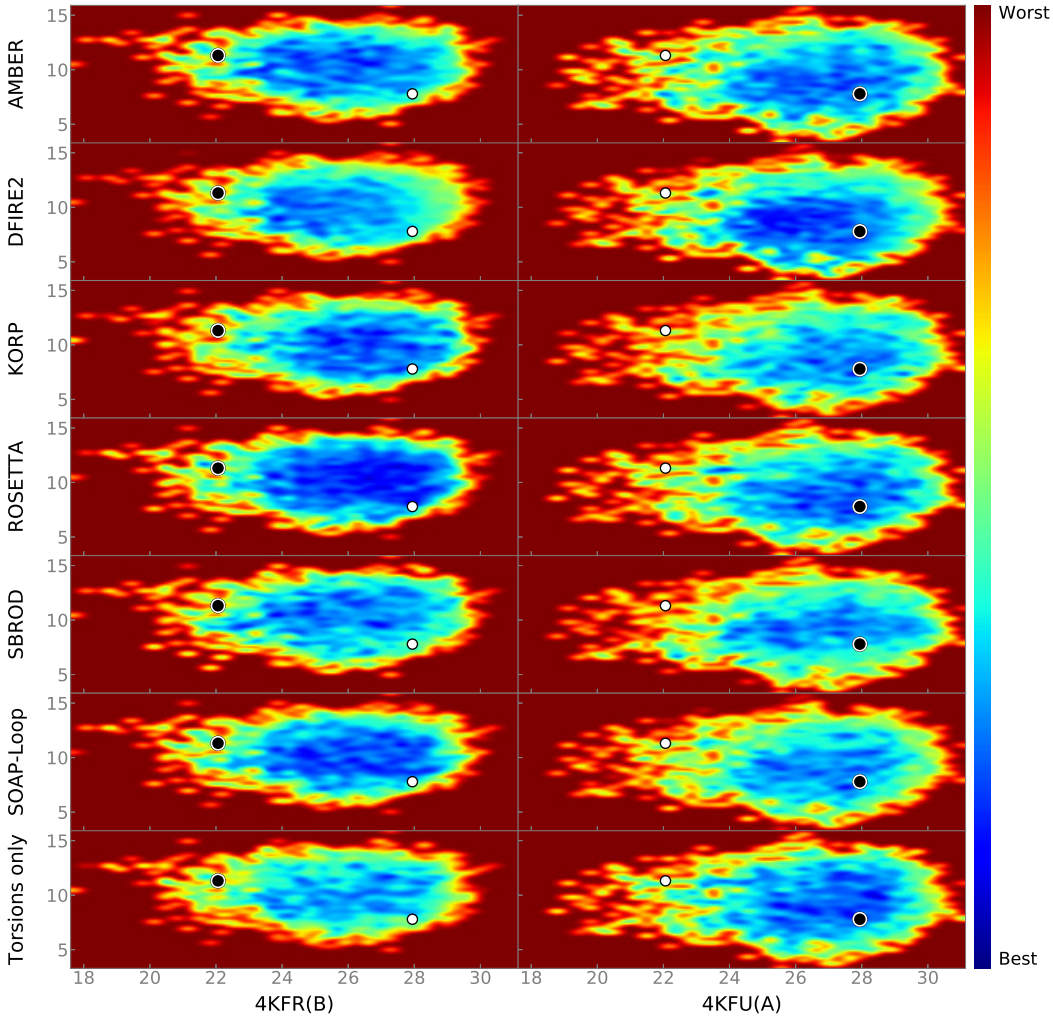


Figure 4.10. Energy landscape obtained by the different scoring methods on NTPase loop. Each row correspond to a single scoring method and each column to a starting scaffold used for the conformational sampling. The x- and y- axes correspond to the first and second dimensions of the projection, respectively (see Section 4.2.4). The first dimension is the distance between a given atom of the loop and another fixed atom of the protein in Å. The second dimension is the average distance between all pairs of C_α atoms of the loop, in Å.
In each landscape, the loop conformation initially in the crystal structure used as scaffold is projected as a black point while other known conformations are projected as white points.

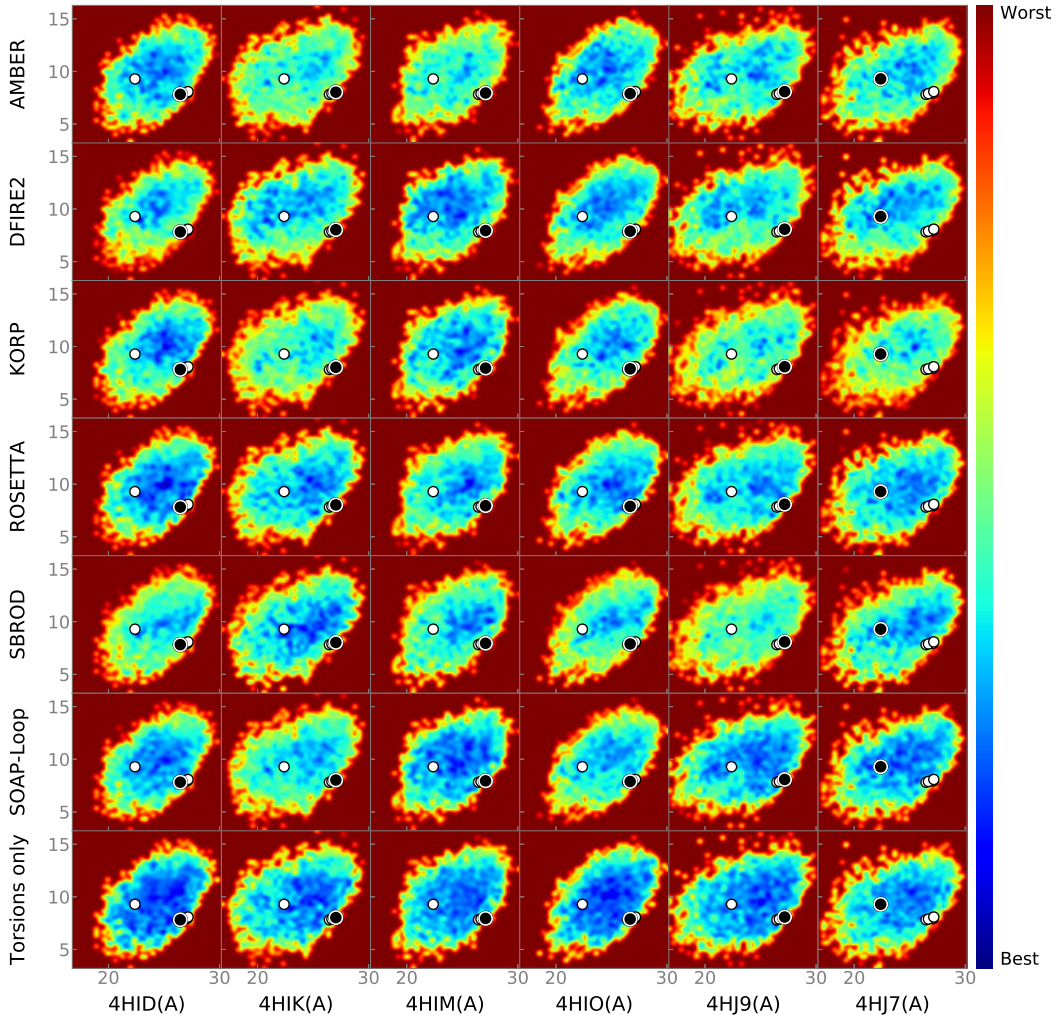


Figure 4.11. Energy landscapes obtained by the different scoring methods on Pot1pC loop. Each row correspond to a single scoring method and each column to a starting scaffold used for the conformational sampling. The x- and y- axes correspond to the first and second dimensions of the projection, respectively (see Section 4.2.4). The first dimension is the distance between a given atom of the loop and another fixed atom of the protein in Å. The second dimension is the distance between 2 backbone atoms respectively at one fourth and three fourths of the loop, in Å.
In each landscape, the loop conformation initially in the crystal structure used as scaffold is projected as a black point while other known conformations are projected as white points.

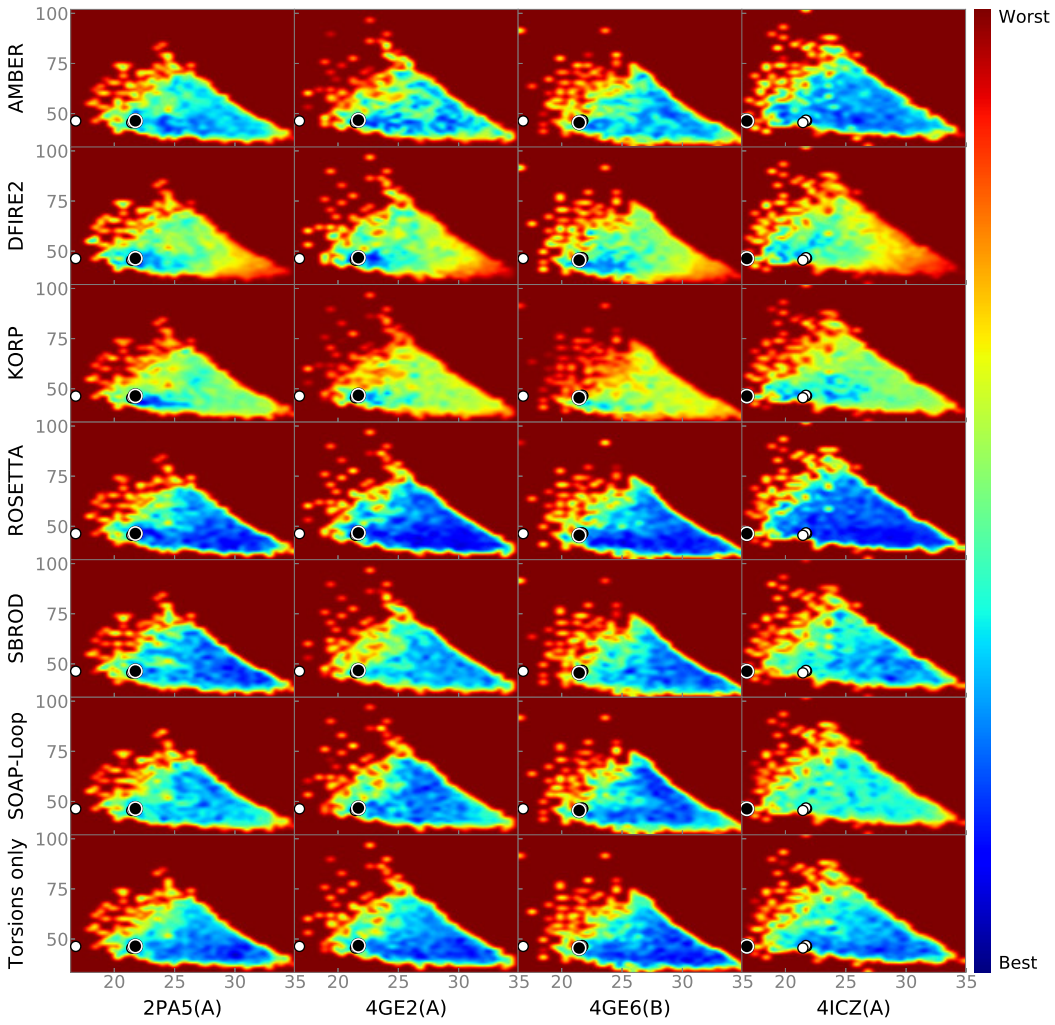


Figure 4.12. Energy landscapes obtained by the different scoring methods on PTPN9 loop. Each row correspond to a single scoring method and each column to a starting scaffold used for the conformational sampling. The x- and y- axes correspond to the first and second dimensions of the projection, respectively (see Section 4.2.4).
The first dimension is the distance between a given atom of the loop and another fixed atom of the protein in Å. The second dimension is the angle formed by three given backbone atoms of the loop in degrees.
In each landscape, the loop conformation initially in the crystal structure used as scaffold is projected as a black point while other known conformations are projected as white points.

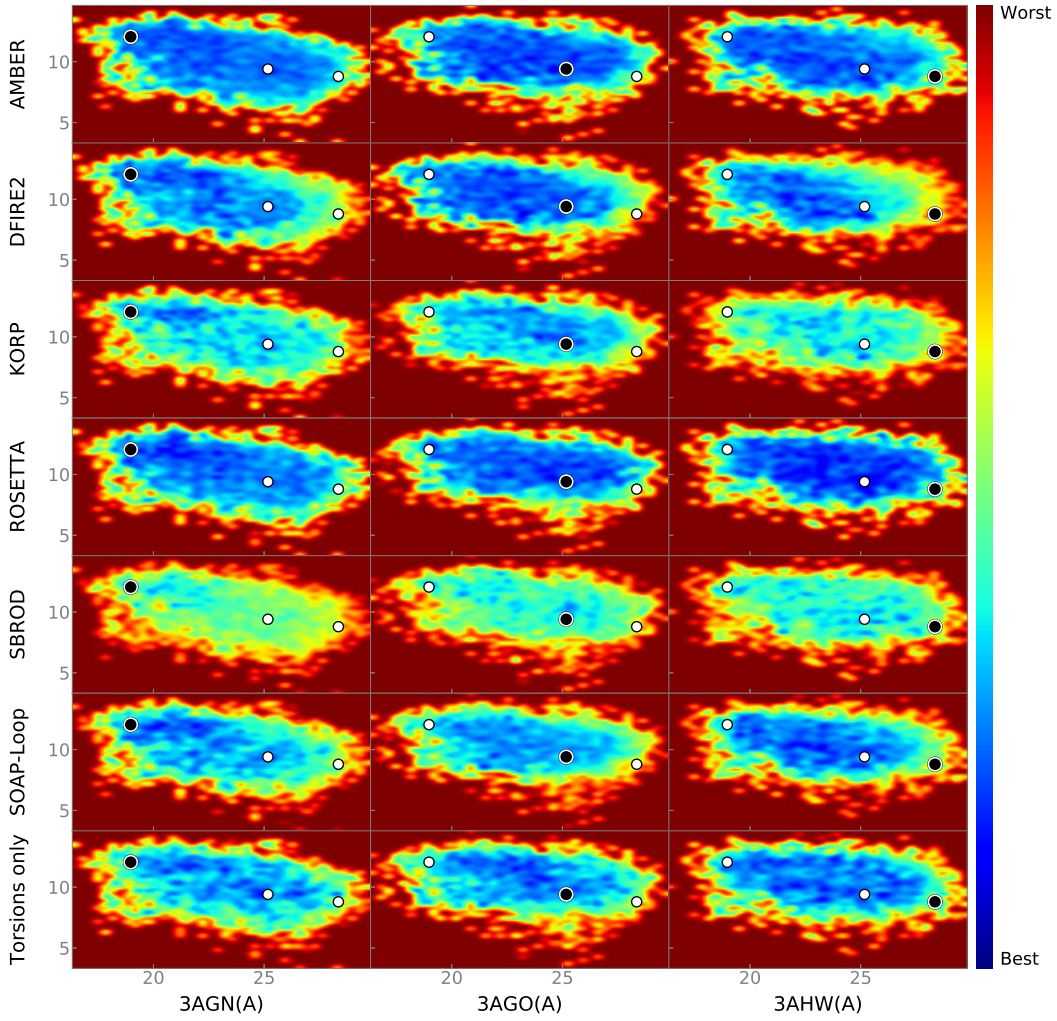


Figure 4.13. Energy landscape obtained by the different scoring methods on RNU2 loop.
Each row correspond to a single scoring method and each column to a starting scaffold used for the conformational sampling. The x- and y- axes correspond to the first and second dimensions of the projection, respectively (see Section 4.2.4). The first dimension is the distance between a given atom of the loop and another fixed atom of the protein in Å. The second dimension is the distance between 2 backbone atoms respectively at one third and two thirds of the loop, in Å.
In each landscape, the loop conformation initially in the crystal structure used as scaffold is projected as a black point while other known conformations are projected as white points.

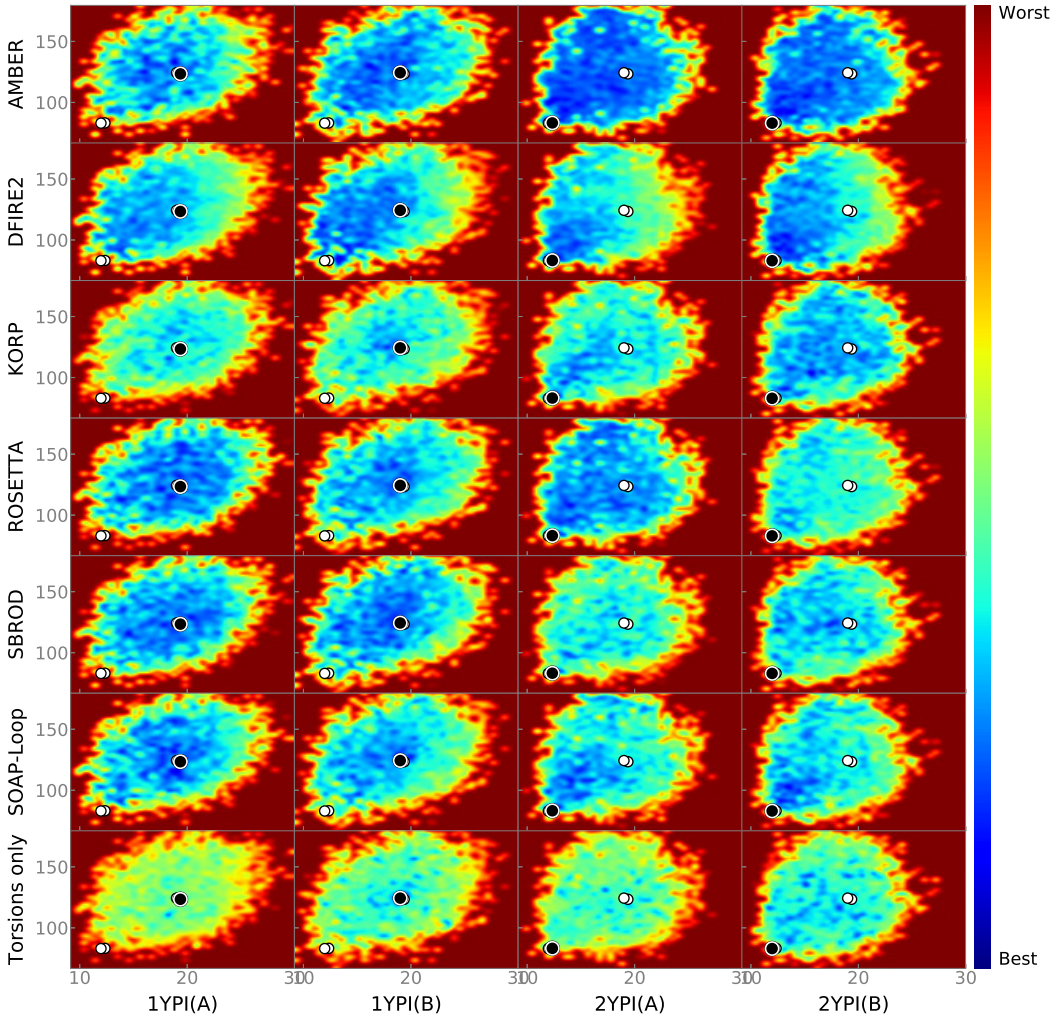


Figure 4.14. Energy landscapes obtained by the different scoring methods on TPI loop.
Each row correspond to a single scoring method and each column to a starting scaffold used for the conformational sampling. The x- and y- axes correspond to the first and second dimensions of the projection, respectively (see Section 4.2.4).
The first dimension is the distance between a given atom of the loop and another fixed atom of the protein in Å. The second dimension is the angle formed by three given backbone atoms of the loop in degrees.
In each landscape, the loop conformation initially in the crystal structure used as scaffold is projected as a black point while other known conformations are projected as white points.

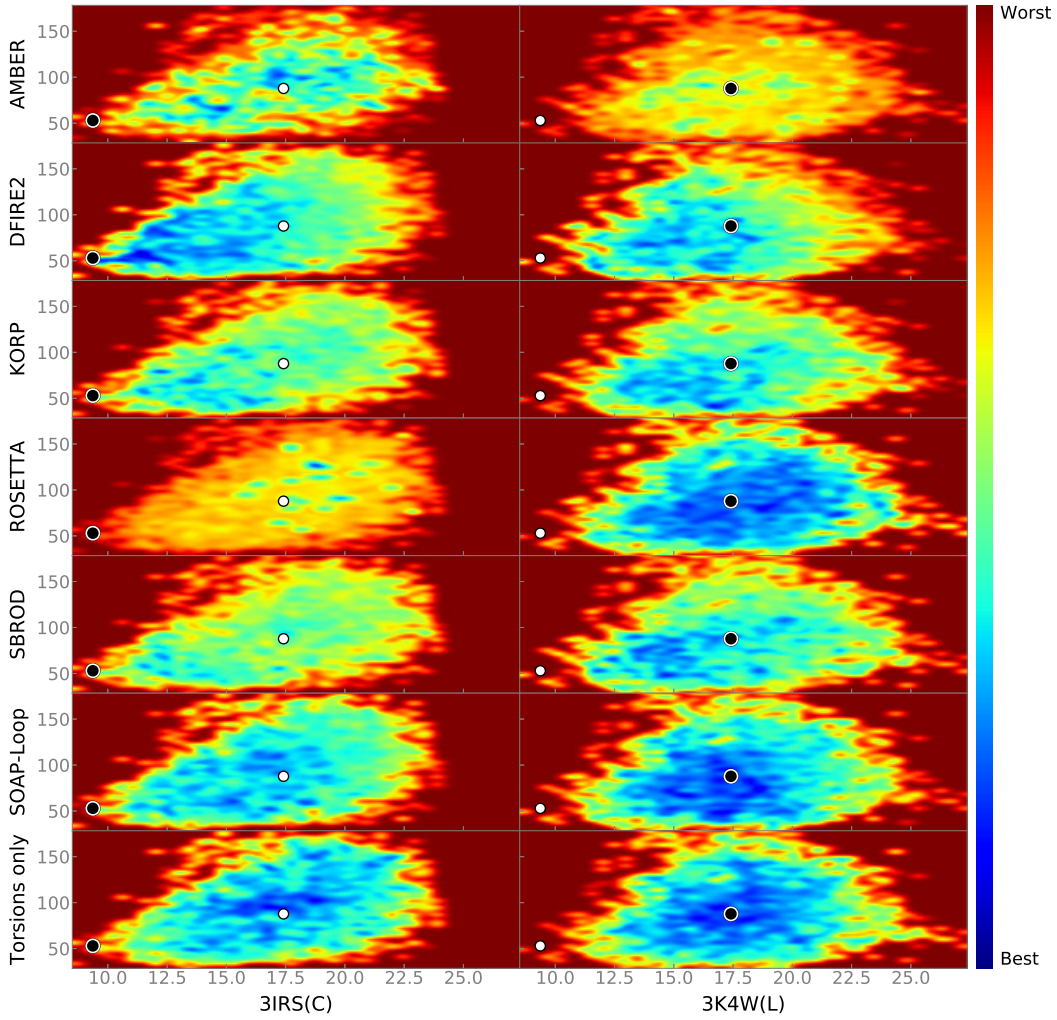


Figure 4.15. Energy landscapes obtained by the different scoring methods on UTB loop.
Each row correspond to a single scoring method and each column to a starting scaffold used for the conformational sampling. The x- and y- axes correspond to the first and second dimensions of the projection, respectively (see Section 4.2.4). The first dimension is the distance between a given atom of the loop and another fixed atom of the protein in Å. The second dimension is the angle formed by three given backbone atoms of the loop in degrees.
In each landscape, the loop conformation initially in the crystal structure used as scaffold is projected as a black point while other known conformations are projected as white points.

4.4 Discussion

The results presented in the previous section show that, both from the sampling and scoring points of view, the scaffold structures play a crucial role. If some surrounding side-chains are not modeled as flexible elements, they may prevent relevant alternative loop conformations from being sampled. Besides, even when the loop conformational space is properly covered during the sampling phase, the scaffold structure still greatly influences the predicted topography of the landscape, whichever scoring method is used. These observations underline the need of careful scaffold structure preparation for loop modeling. Specifically, due to the high sensitivity of the scoring functions to minor changes in the scaffold structures, the need to use several starting structures becomes clear (obtained e.g. by applying slight perturbations to a modeled or known conformation, or by gathering several known structures if those exist). From this perspective, employing fast sampling and scoring methods is appealing since it allows the exploration of several starting structure candidates using fewer computational resources. As shown in Chapter 3, MoMA-LoopSampler provides exhaustive ensembles while excluding highly statistically improbable conformations, thus generating better-filtered ensembles. This makes MoMA-LoopSampler a suitable sampling method to use on multiple starting scaffolds. Fast statistical scoring methods then constitute a natural complement to such a loop modeling process.

Estimating the exact contribution of an accurate side-chain placement in flexible loop modeling is not a straightforward task. However, one can expect that correctly modeling side-chains improves the quality assessment performed by scoring methods employing all-atom models. For this purpose, we could have used side-chain prediction methods such as SCWRL4 [Krivov 2009]. Although SCWRL4 is a good and popular technique, it does not enforce strict rules concerning steric clashes, thus producing infeasible side-chain placements when the environment is very constrained. Given the importance of steric clashes and side-chain rotamers in most all-atom methods, and for consistency in sampling, we decided to implement our own side-chain placement method in MoMA-LoopSampler. It follows similar ideas to those applied in the backbone sampling phase. It uses the same model with the dihedral angles as sole degrees of freedom, forbids major collisions and allows for some deviations from the rotamers while still sampling around them. Due to steric constraints, this side-chain placement process has a relatively high failure rate and rejects many backbone sampled states. In addition, it returns the first placement that respects these constraints, without evaluating its quality. To improve results, we are currently working on a method integrating energy minimization for side-chain positioning. Building upon the results obtained in this work, DFIRE2 may be a good option to guide this minimization since it represents a good trade-off between accuracy and rapidity while considering an all-atom model.

Unsurprisingly, statistical methods that do not require structural relaxation were found to be much faster than AMBER and ROSETTA. KORP and DFIRE2 are among the fastest methods while yielding remarkably satisfying results. They

model smoother and more consistent landscapes than other methods. A major downfall is that they are rarely able to model landscapes with several basins, but the use of several starting scaffold structures may circumvent this shortcoming. KORP does not even consider the side-chains in the structures. Given that their placement is delicate and time-consuming, this is a considerable advantage over DFIRE2.

Concerning the energy landscapes, known conformations are often found adjacent to a high energy barrier. We hypothesize that these conformations are stabilized by a certain number of atomic contacts, within the loop or with surrounding residues. Areas of high energy would then correspond to conformations having these contacting atoms in steric collision. Note that using a 2D representation that only displays the energy of the best-scored state at a projected position prevents some energy barriers from appearing in the projected landscape. Therefore, areas projected in the middle of basins may still be forming numerous contacts, even though the energy barrier corresponding to bringing these contacting atoms closer together until they overlap is not apparent.

While all other statistics-based methods are built using a Bayesian framework on structural data from proteins with non redundant sequences, SBROD is built with a regression method and is trained to distinguish the native fold among several decoys. This design enables SBROD to perform well for *ab initio* structure prediction, as demonstrated in the latest CASP13 experiment (`www.predictioncenter.org/casp13/`), but yields disappointing results on the flexible loops studied in this work.

SOAP-Loop is the method that best agrees with other methods, including ROSETTA. It still takes a considerable amount of time to score the different states and does not provide overall results as satisfying as those of DFIRE2 or KORP. Thus, it is less suited to landscape reconstruction and to the analysis of numerous sampled states as performed in this work.

The landscapes modeled by AMBER and ROSETTA are, as expected, too rough to enable a satisfying analysis of a loop system. It is likely that the relaxations performed in this work were insufficient: for AMBER, the number of cycles may have been too low, or the convergence criterion too high; for ROSETTA, unrestricting the backbone dihedral angles of the rest of the protein may have yielded better results. However, longer relaxations would have been extremely costly in terms of computational resources. Although those scoring methods are inappropriate for the global/ensemble modeling of flexible loops, results presented in this work suggest that they may perform well for the refinement of stable structures. Statistical methods have their precision limited by design. AMBER and ROSETTA, conversely, may be used on a more limited number of states, e.g. to discriminate among similar models.

4.5 Conclusion

In this work, we have investigated the capability of state-of-the-art sampling and scoring methods to model flexible protein loops, which may adopt different (meta-)stable conformations. Our analysis shows that, despite the promising results obtained during both sampling and scoring steps, substantial methodological work is still required to identify the most probable conformations in an accurate and reliable manner.

To begin with, the success of loop sampling methods is limited by the difficulty to efficiently account for flexibility outside the loop and to correctly place side-chains. Indeed, whatever the methods employed for scoring multiple states, the structural scaffolds over which the loops were modeled proved decisive for the topography of the implicit landscapes. The integration of a flexible component in loop sampling methods thus constitutes an important direction for future work. Regarding side-chains, DFIRE2, that considers the position of all atoms and is among the fastest methods, could be used to optimize side-chain placement before scoring. Although structural relaxation is not needed in theory for this method, local optimization of side-chains generated by a global search strategy (as applied in this work) could improve the results.

Concerning loop scoring methods, some of them can reliably identify unfeasible states and are capable of providing valuable insight into the global topography of a loop's energy landscape. However, the modeled landscapes are often too fuzzy to allow a precise modeling of the loop conformational space. In addition, most scoring methods provide erratic results from one loop to another, making their performance on a fully unknown system too unpredictable. In practice, such observations suggest that scoring methods can be reliably employed for applications requiring to coarsely filter loop states, but that their results are not accurate-enough for applications such as protein design. More precisely, the qualitative comparison of scoring methods for loop modeling presented in this chapter validates the use of fast statistical potentials such as KORP or DFIRE2 as primary filters or as overall quality assessment methods for large pools of loop structures. Indeed, these methods can identify states close to statistically-probably conformations, regardless of poor local geometry or inner collisions. However, their low sensitivity to small conformational changes prevents them from providing a more precise evaluation. For such cases, physics-based or hybrid methods would be more appropriate, provided that the necessary structural relaxations are carefully performed.

Taken together, Chapters 3 and 4 provide a general protocol to analyze the energy landscape of a flexible protein loop and indications related to the reliability of the different scoring methods. However, tests were only performed on benchmark systems and not on antibodies, which constitute our proteins of interest. For this reason, the next chapter will focus on applying the full protocol on an antibody from Sanofi with a flexible H3 loop, in order to verify that the previous conclusions also hold for such a system. The scoring methods tested in this chapter will all be applied on this system, in order to (1) verify their performance on unpublished

data and (2) analyze the different landscapes in the light of results presented in this chapter.

Chapter 5

H3 loop modeling in an antibody from Sanofi

Contents

5.1 Introduction

As revealed in Chapter 2, flexibility of CDR loops is crucial to understanding antibody-related mechanisms. However, as highlighted in this chapter, accurately modeling CDR loop plasticity remains an open problem, for which efficient methods are still required. In order to address this shortcoming, Chapter 3 then presented a method to sample diverse and high-quality loop conformational ensembles, while Chapter 4 focused on accurately scoring the generated conformations, in order to model meaningful energy landscapes. However, the methods presented in Chapters 3 and 4 are not antibody-specific and their performance on CDR loops has not been demonstrated yet. The present chapter focuses on testing these methods on an antibody which comprises a flexible H3 loop.

Following a project of humanization and artificial maturation of an antibody at Sanofi, several structures of the Fab fragment were obtained under different conditions (free, bound, humanized, maturated, ...). These structures revealed

flexibility in the H3 CDR loop, which was found to adopt two distinct conformations: an *apo* conformation and a *holo* conformation (identical for all bound structures of the antibody).

The different structures provide us with an ideal test case for the methods presented in this thesis. Although the humanization and maturation project did not focus on the H3 loop, the conditions that are necessary to follow a similar process to that of Chapter 4 are gathered:

- Knowledge of several conformations for the loop to study.
- Existence of several structures to use as scaffold for the Fab fragment.

In the following, Section 5.2 details the structures employed in this work and summarizes the methods employed to model the energy landscape; Section 5.3 describes the results of the different scoring methods, and the modeled energy landscapes, similarly to Section 4.3 in the previous chapter; Section 5.4 discusses these results, and finally, Section 5.5 concludes on the applicability of the general loop modeling methods developed in this thesis to CDR loops in antibodies.

5.2 Structures and methods

5.2.1 Note on data confidentiality

The structures utilized in this work have not been published yet. Therefore, this work does not disclose the nature of the antibody or of its antigen. Neither does it detail the mutations performed during the artificial maturation process.

However, the structures should be published soon, along with the analysis performed in this work.

5.2.2 Available structures for the antibody

This section lists existing structures for the variants of our antibody of interest (subsequently called Ab), through the description of the different steps in the humanization and maturation processes. Table 5.1 summarizes the different structures used in this work.

At the start of the process, antibody Ab1 was known to bind an undisclosed antigen. The structure for the free Fab was obtained at 1.7 Å resolution (Crystal ID 1), with one Fab in the asymmetric unit (Fab #1). The structure exhibits tight crystal packing which involves CDR loops.

The structure of the Fab fragment of Ab1 complexed with its antigen was also obtained, at lower resolution (3.25 Å) (Crystal ID 2). The asymmetric unit contained 4 complexes, so that 4 bound antibody structures could be obtained (Fab #2 to #5). Fab structures #2 to #5 are very similar, apart from the constant domain which is rotated in #5 compared to #2-#4. However, while the heavy and light chains of structures #2 to #5 superimpose very well independently with the heavy and light chains of Fab #1, their relative orientation changes, creating a rather different scaffold from Fab structure #1. The other major difference between Fab

Table 5.1. Antibody structures used as scaffold for H3 conformational sampling. "Bound" structures were obtained from the structure of the antibody complexed with the antigen. "Crystal structure ID" refers to the ID of a crystal structure, possibly containing several Fabs (with different associated "Fab structure #") in the asymmetric unit. "Scaffold code" refers to the name used for this scaffold structure in this chapter.

Fab structure #	Crystal structure ID	Resolution (Å)	Antibody molecule	Free/Bound	Humanized	Mutant	Scaffold code	H3 conformation
1	1	1.70	Ab1	Free	No	No	apo	*apo*
2	2	3.25	Ab1	Bound	No	No	holo,1	*holo*
3	2	3.25	Ab1	Bound	No	No	holo,2	*holo*
4	2	3.25	Ab1	Bound	No	No	holo,3	*holo*
5	2	3.25	Ab1	Bound	No	No	holo,4	*holo*
6	3	2.81	Ab2	Bound	Yes	No	H,holo,1	*holo*
7	3	2.81	Ab2	Bound	Yes	No	H,holo,2	*holo*
8	4	2.50	Ab3	Bound	Yes	Triple	M,holo,1	*holo*
9	4	2.50	Ab3	Bound	Yes	Triple	M,holo,2	*holo*

structures #1 and #2-#5 are the different conformations of the H3 loop. Although the loop keeps the same internal backbone conformation overall, it is shifted in Fabs #2-#5 compared to Fab #1.

The antibody was then humanized (Ab2). The structure of the Fab fragment of Ab2 with its antigen was obtained at 2.81 Å resolution (Crystal ID 3), with two complexes in the asymmetric unit (Fab structures #6, #7). Fabs in #6 and #7 differ by the orientation of their constant domain. Although H3 adopts a conformation very similar to the one in complexed Ab1 for both #6 and #7, the docking pose is slightly different, with a shift of up to 1.2-1.5 Å of the Fab fragment when aligning the epitope domain. Note that the mutations in the humanized antibody are located more than 6 Å away from the H3 loop.

The docked structures suggested 3 point mutations in the light chain to improve the affinity between the antibody and the antigen. An asparagine residue was replaced with a tyrosine in L1, a glutamine residue was replaced with a leucine in L2, and an asparagine residue was replaced with a aspartic acid in L3. The triple mutant humanized antibody (Ab3) was crystallized in complex with the antigen at 2.5 Å resolution (Crystal ID 4). The asymmetric unit contained 2 nearly identical complexes, yielding two other Fab structures #8 and #9. For these structures, the conformation of the H3 loop is very similar to that from Fab structures #2 to #7, and the point mutations are located more than 5.5 Å away from any atom of the H3 loop.

Table 5.2. Backbone heavy atom RMSD between the different H3 loop conformations. Scaffold names are colored according to their associated known conformation.

	apo	holo,1	holo,2	holo,3	holo,4	H,holo,1	H,holo,2	M,holo,1	M,holo,2
apo	0	2.26	2.22	2.2	2.03	2.27	2.33	2.26	2.17
holo,1	2.26	0	0.09	0.12	0.29	0.62	0.64	0.68	0.54
holo,2	2.22	0.09	0	0.05	0.26	0.61	0.64	0.66	0.5
holo,3	2.2	0.12	0.05	0	0.22	0.61	0.64	0.66	0.49
holo,4	2.03	0.29	0.26	0.22	0	0.64	0.69	0.7	0.5
H,holo,1	2.27	0.62	0.61	0.61	0.64	0	0.19	0.29	0.68
H,holo,2	2.33	0.64	0.64	0.64	0.69	0.19	0	0.35	0.71
M,holo,1	2.26	0.68	0.66	0.66	0.7	0.29	0.35	0	0.58
M,holo,2	2.17	0.54	0.5	0.49	0.5	0.68	0.71	0.58	0

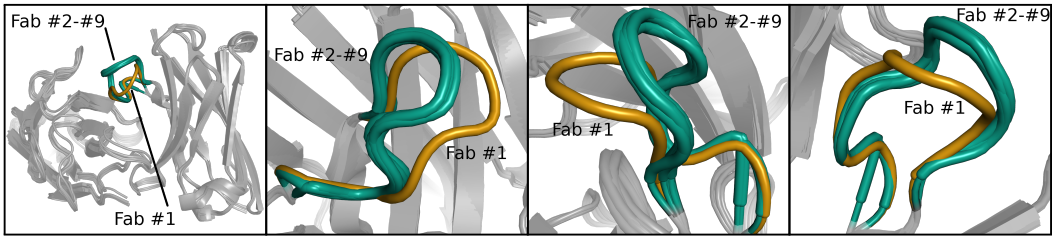


Figure 5.1. Cartoon views of the conformations of the H3 loop in the different crystal structures of Ab. The *apo* conformation is in orange, while the *holo* conformations are in blue-green.

The different structures of the H3 loop can be visualized on Figure 5.1 and the values of the RMSDs (calculated on backbone heavy atoms) between the different structures of the H3 loop are available in Table 5.2.

5.2.3 Methods to model H3 landscapes

To model the energy landscape of H3 in the different scaffold structures, the same protocol as in Section 4.2 was applied. Briefly, from each scaffold, H3 was sampled using MoMA-LoopSampler and scored using the different scoring methods listed in Table 4.3. The sampled states were projected in 2D using two adequate descriptors and the energy landscapes reconstructed from the 2D projections and the scores.

5.3 Results

5.3.1 Sampling

Figure 5.2 gives the best RMSD (calculated on heavy backbone atoms) among sampled H3 states to the known *apo* and *holo* conformations. Results show that from all the scaffolds originating from crystal structures of the antibody bound to

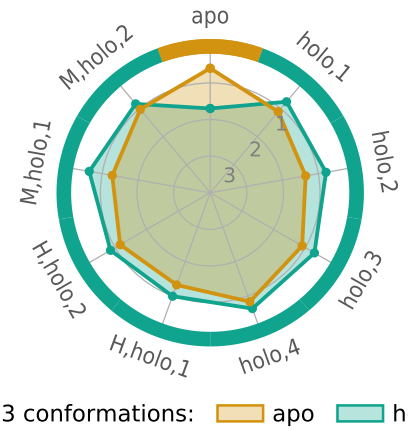


Figure 5.2. Lowest RMSD to the *apo* and *holo* conformations of the H3 loop, among sampled states from each scaffold. RMSDs are calculated on the heavy atoms of the backbone.
Scaffolds are distributed around the disk and their names indicated outside the disk. The outer circle is colored according to the conformation present in the X-Ray structure the scaffold originated from.
The outer circle corresponds to a RMSD of 0 Å, while the center corresponds to RMSDs of 4 Å and above: a larger colored area indicates that the corresponding stable conformation is found with greater accuracy in the sampled ensemble.

its antigen, both conformations are accurately sampled (within 1.2 Å). However, from the scaffold originating from the crystal structure of free Ab1, only the *apo* conformation is closely sampled. Indeed, the closest sampled state to the *holo* conformation is about 1.8 Å RMSD away.

Looking at the structures, it appears that L2 is slightly shifted between the *apo* and *holo* structures. In the *apo* conformation of the scaffold, a tyrosine from L2 is located in a position that coincides with the position of the H3 loop in the *holo* conformation. Following these results, a re-sampling of H3 was performed from the *apo* scaffold, allowing at most one of the surrounding side-chains to be in collision with the backbone of H3. Note that this could be any side-chain. The colliding side-chain was then added to the set of side-chains to place once a closed backbone was generated. Results following this re-sampling are presented in Section 5.3.5.

5.3.2 Top-scored and closest loop states to *apo* or *holo* conformations

The ranks of the 5 closest sampled states to each known conformation, from each scaffold are represented in Figure 5.3. Results show that scoring methods are capable of ranking the known conformations among the top states. Among the different scoring methods, DFIRE2 is the one that gives the most impressive results. Apart from a few exceptions, it ranks the 5 closest states to the *apo* conformation among the 100 top samples, for all the scaffolds, including those originating from the structure of a bound antibody. Although it does not rank the *holo* conformation as well

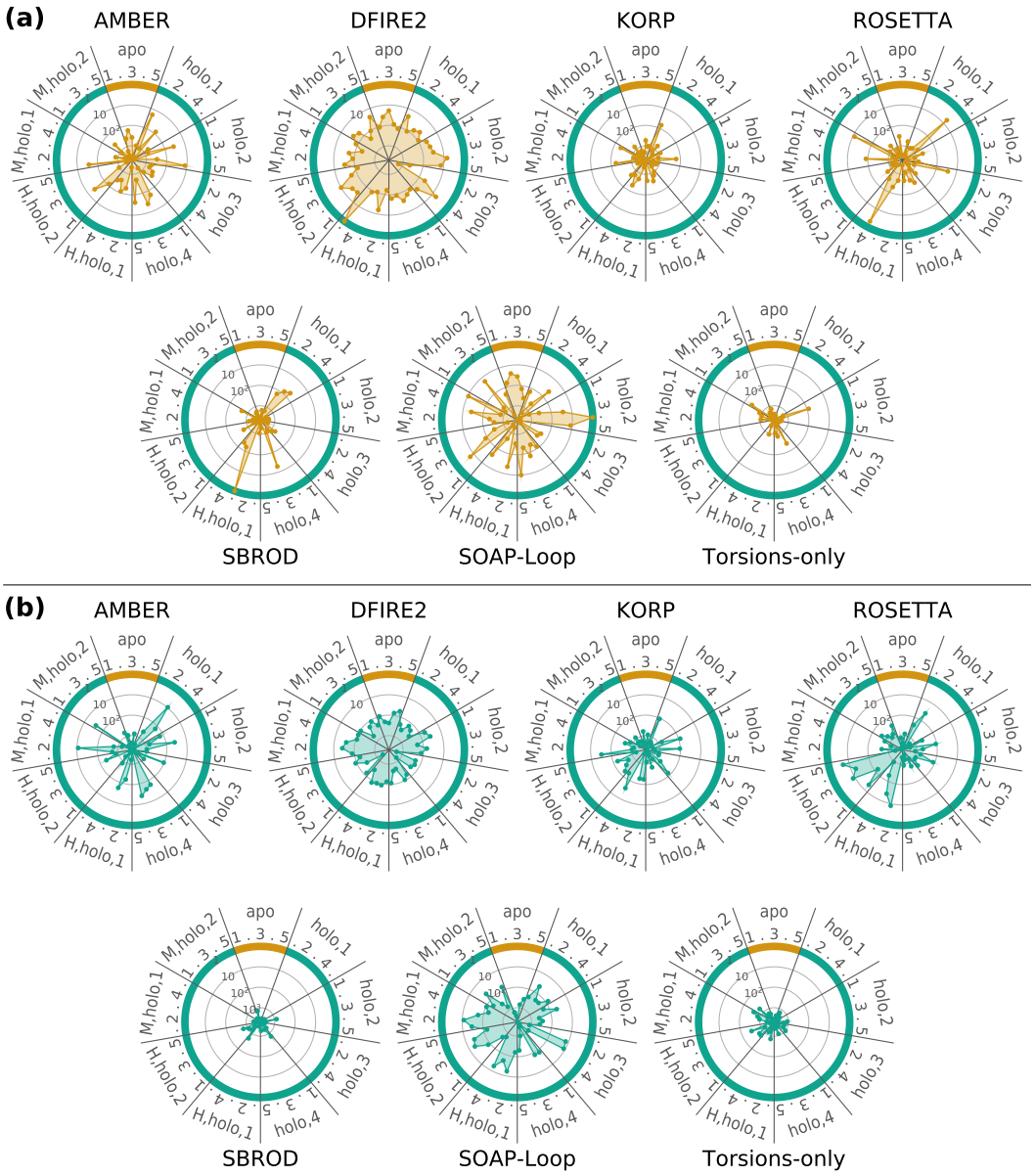


Figure 5.3. Ranks of the five closest sampled states to the *apo* (a) and *holo* (b) conformations of the H3 loop, from each scaffold.
Scaffolds are distributed around the disk and separated by thicker dark grey lines. The outer circle is colored according to the conformation present in the X-Ray structure the scaffold originated from. The ranks are represented by the proximity to the outer circle or to the center. The outer circle corresponds to the best scoring ranks, while the center corresponds to the worst rank. Note that the radial axis has a logarithmic scale.
With such a representation, points far from the center of the circle correspond to well-scored sampled states.

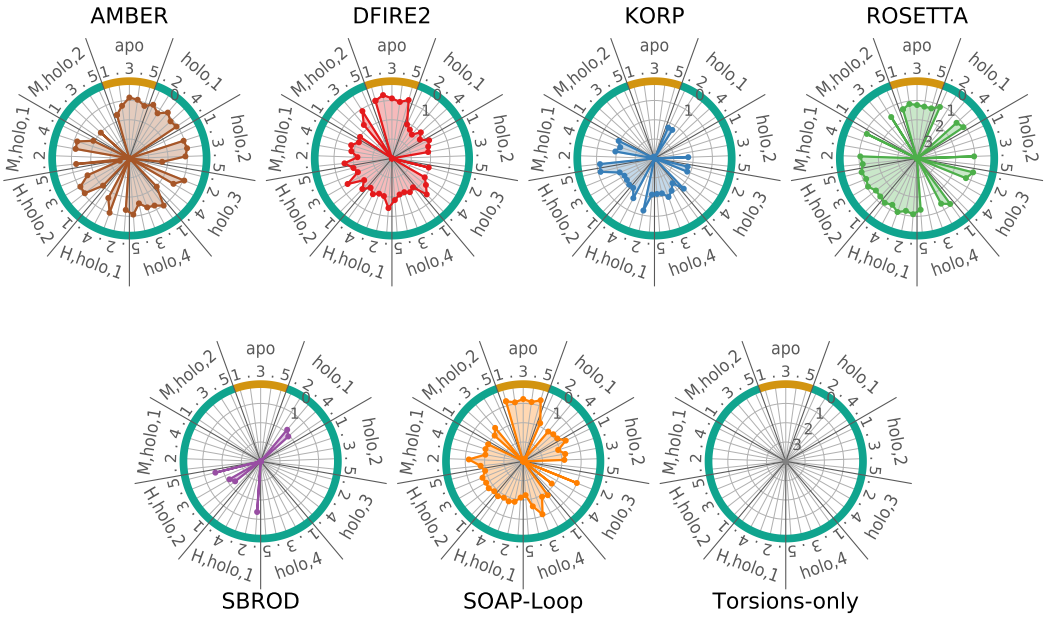


Figure 5.4. Backbone RMSD to the closest known conformation (*apo* or *holo*) of the five top-scoring states among those sampled from a unique scaffold. Scaffolds are distributed around the disk and separated by thicker dark grey lines. The outer circle is colored according to the conformation present in the X-Ray structure the scaffold originated from.
The outer circle corresponds to a RMSD of 0 Å, while the center corresponds to RMSDs of 4 Å and above: a larger colored area indicates that the top scoring states are close to a known stable conformation.

as the *apo* conformation, it still almost consistently places the 5 closest states to this conformation among the top 20% states. This result difference between *apo* and *holo* conformations can easily be explained by the fact that the antigen is not present for the scoring. This would suggest that the *apo* conformation is more statistically likely when the antibody is free, but that the *holo* conformation is still probable from a statistical point of view, and stabilized by the presence of the antigen.

Other methods, especially SOAP-Loop, but also AMBER and ROSETTA, manage to rank a few of the states closest to known conformations among the top-most states, but they are not as consistent as DFIRE2, with highly varying ranks from one of the 5 closest loops to another, or from a scaffold to another.

The distances to the closest of the *apo* and *holo* conformations of the 5 best-ranked states by the different scoring methods, from the nine scaffolds, are represented on Figure 5.4. The distances are given as RMSD on backbone heavy atoms.

Results show that most of these top states as ranked by AMBER are within 1 Å of either the *apo* of *holo* conformation, except from the *M,holo,2* scaffold, which also yields poor results for the other scoring methods. Most top states as ranked by DFIRE2 are within 2 Å RMSD of one of the two known conformations. For KORP, some top sampled states from the humanized scaffolds are close to a known

conformation, but results are not as good from the other scaffolds. For ROSETTA, the results are very different depending on the starting scaffold. From the scaffold of the free and of the humanized antibody structures, all top 5 states are within 1 Å RMSD of one of the two known conformations. However, the results from the other scaffolds vary strongly from one the of the 5 top samples to another. SBROD and Torsions-only yield disappointing results, with very few top samples close to any known conformation. SOAP-Loop, on the contrary, performs as well as DFIRE2, identifying most of the 5 top samples within 2 Å of either one of the two known conformations. It performs particularly well for the scaffold of the free antibody, with all 5 top scoring states within 1 Å of a known conformation.

5.3.3 Clustering

The top 10 sampled states, as scored by the five scoring methods showing the best results (AMBER, DFIRE2, KORP, ROSETTA, and SOAP-Loop), were gathered for each of the nine scaffolds. This represents $10 \times 5 \times 9 = 450$ elements, to which were added the nine crystal structures. These 459 elements were clustered using hierarchical cluster analysis (HCA), using RMSD on backbone heavy atoms as pairwise distance between the elements. Note that if two methods identify the same state among the 10 best models, the corresponding state was still treated as two distinct elements for the clustering. The HCA method employed single linkage, i.e. the distance between two clusters a and b was taken as $\min_{x \in a, y \in b} d(x, y)$ where d is the RMSD on the heavy atoms of the backbone of the H3 loop. The clustering stopped when the closest two clusters were more than 1 Å distant.

Figure 5.5 shows the distribution of the clusters in terms of sizes and content, for clusters with at least 2 elements. The clustering created 53 clusters, 19 of which contain at least two elements and only 5 of which contain states identified by only one scoring method. The largest cluster contains 152 elements identified by all five different scoring methods. The second largest cluster contains 105 elements, still identified by five scoring methods. The third largest cluster drops at 44 elements, identified by four scoring methods (all except KORP). The fourth largest cluster only contains 22 elements. The *holo* conformation is found in the second largest cluster, and the *apo* conformation in the third largest cluster. Figure 5.6 shows cartoon views of the states in the top three clusters.

The largest cluster contains a conformation, $\mathcal{C}$ (Figure 5.6(a)), identified by all 5 scoring methods and distinct from the *apo* and *holo* conformations. It is the conformation adopted by many of the top ten states identified by AMBER, DFIRE2, KORP, ROETTA, SOAP-LOOP, but also SBROD (data not shown). Since this conformation has not been experimentally determined, we cannot rule out the fact that conformation $\mathcal{C}$ may be a false positive, although such a concept is complex to define in loop modeling, especially without additional experimental data. Indeed, it cannot be excluded that this conformation exists for the free antibody, but that is was not crystallized.

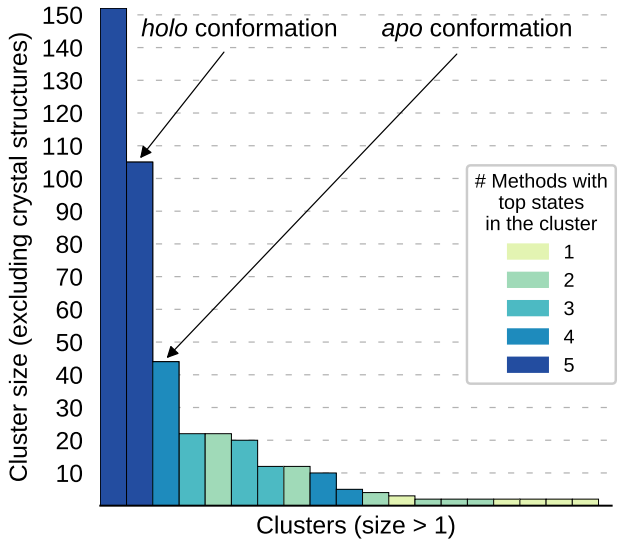


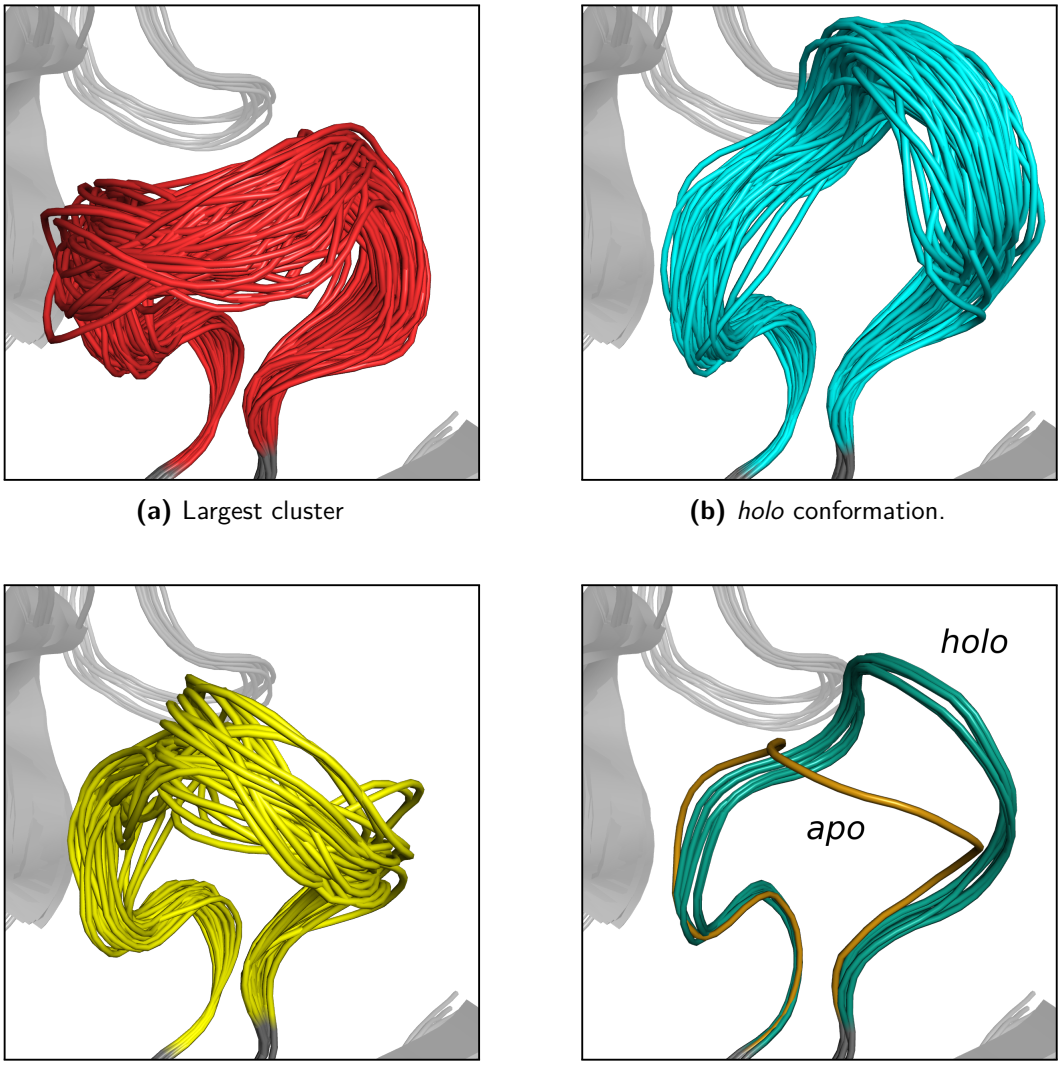
Figure 5.5. Distribution of cluster size obtained by single linkage HCA clustering on the crystal conformations and the 10 best-scored states by AMBER, DFIRE2, KORP, ROSETTA and SOAP-Loop. The maximum distance to merge two clusters was taken as 1 Å. Cluster sizes are given excluding crystal conformations. Arrows indicate the clusters containing the *apo* and *holo* conformations.

5.3.4 Modeled landscapes

Following the clustering results, the 2D projection chosen to visualize the energy landscapes for the antibody resulted from a principal component analysis (PCA) trained on the sampled states from the top three clusters. The PCA was applied to the 27 variables corresponding to the Cartesian coordinates of the 9 C_α of the H3 loop. All sampled states, as well as the crystal conformations, were projected on the two first principal components. The landscapes modeled by combining these 2D projections of the sampled states with the scores given by the different methods are given in Figures 5.7 and 5.8.

Using the *apo* scaffold, the *holo* conformation of the H3 loop is never in a basin, whatever the scoring methods. This is not surprising given that this conformation has not been closely sampled from this scaffold. However, the area corresponding to the cluster containing this *holo* conformation includes a basin, particularly visible with DFIRE2 and SOAP-Loop. This basin may correspond to a conformation that is similar to the *holo* conformation. According to AMBER, DFIRE2, ROSETTA and SOAP-Loop, the *apo* conformation is in an energy basin from the *apo* scaffold. Among these methods, DFIRE2 gives this basin with the best accuracy, with a landscape that is less "flat".

From most *holo* scaffolds, DFIRE2 accurately identifies the basin around the *holo* conformation. KORP does so for the *M,holo,1* scaffold. SOAP-Loop and ROSETTA also identify this basin from a few scaffolds, but not as accurately and unambiguously as DFIRE2, and not as consistently across scaffolds.



(a) Largest cluster

(b) *holo* conformation.

(c) *apo* conformation.

(d) Crystal structures.

Figure 5.6. Clustering: cartoon view of the three largest clusters, compared to the crystal structures.

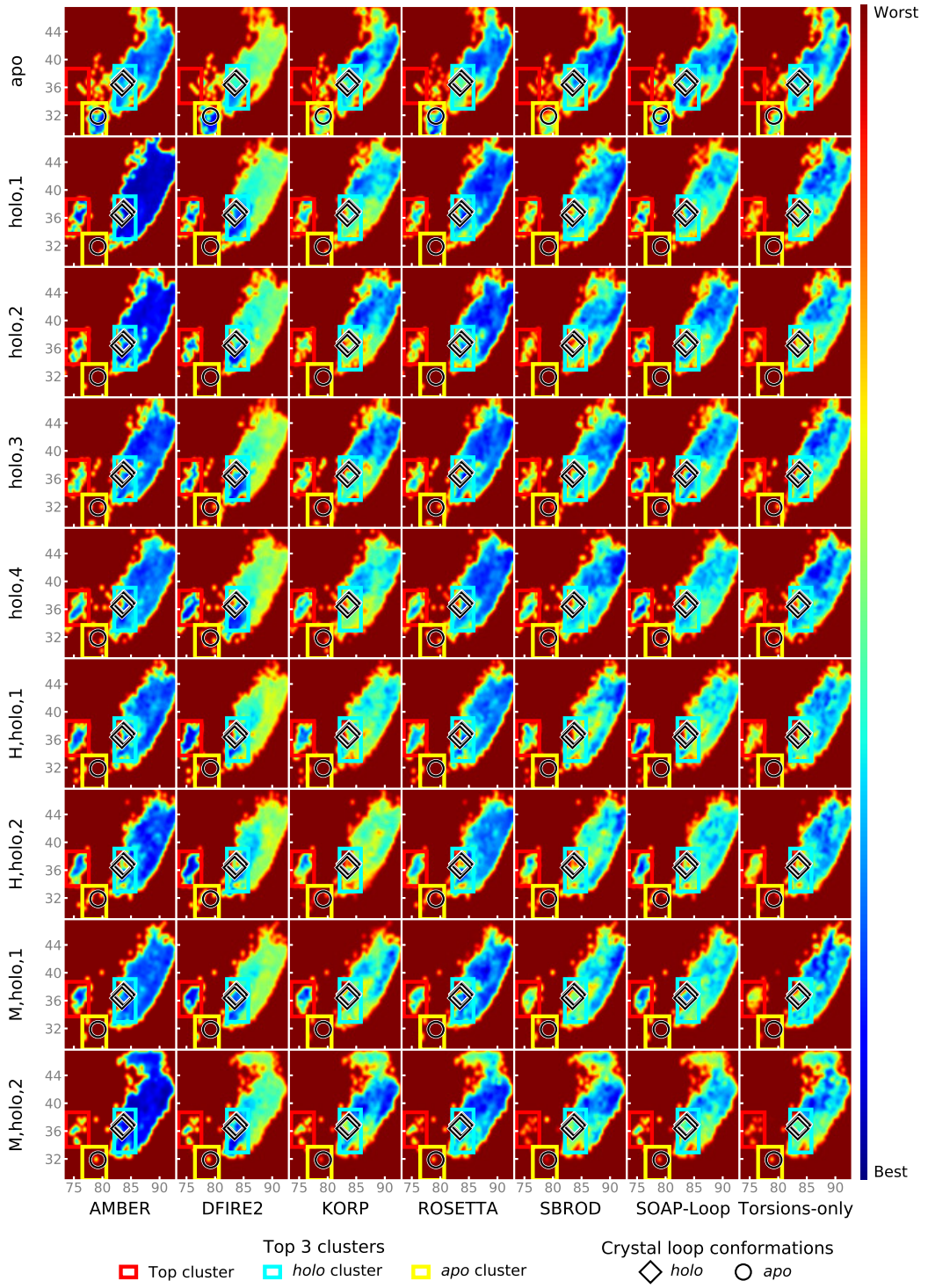


Figure 5.7. Modeled energy landscapes for H3. Rows and columns are reversed for readability, compared to Figures 4.8 to 4.15. Each row corresponds to a starting scaffold used for sampling, while each column corresponds to a scoring method. The x- and y- axes correspond to the first and second principal components of the PCA, respectively.

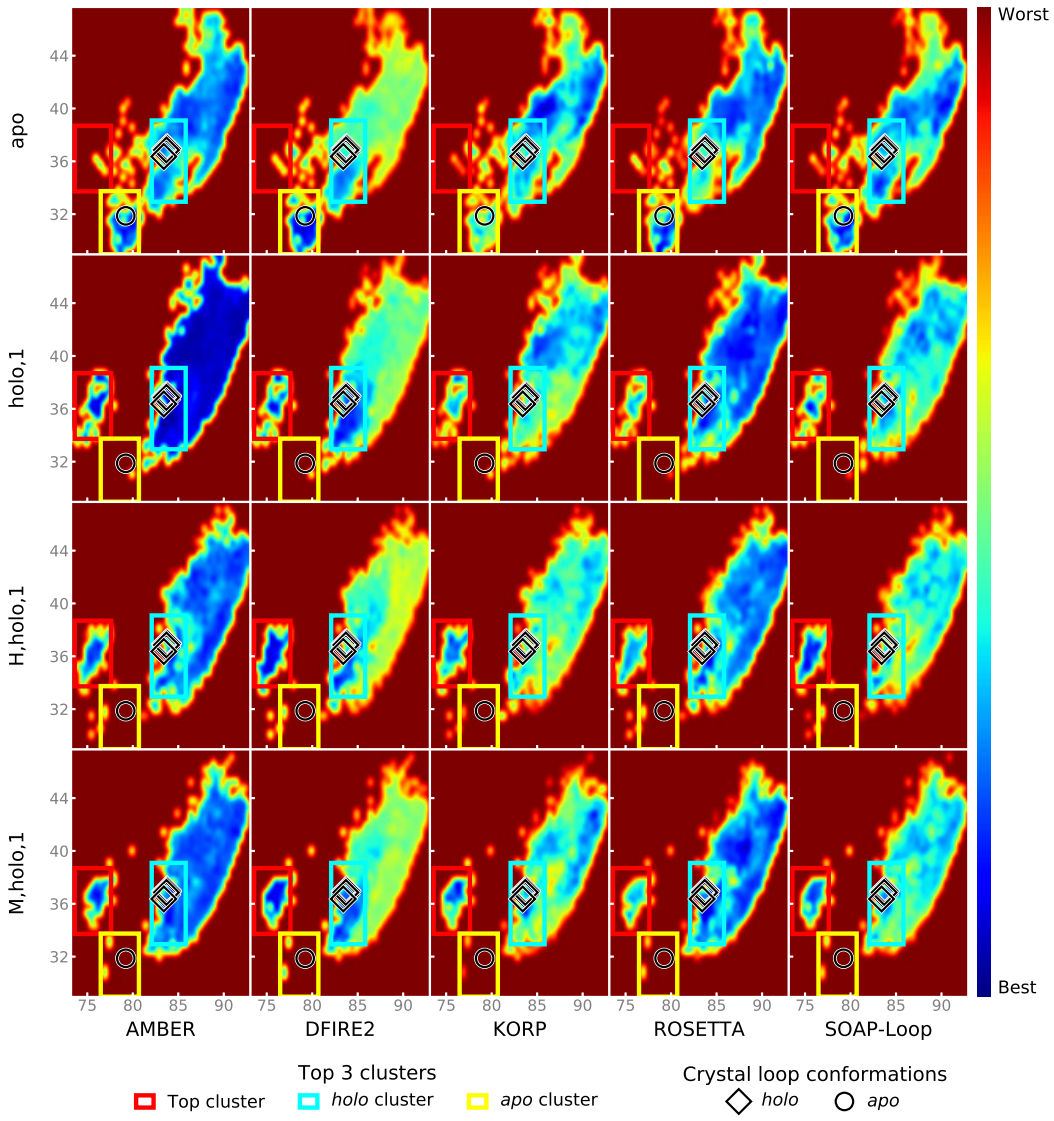


Figure 5.8. Subset of landscapes from Figure 5.7 with the five methods yielding the best results and a representative set of scaffolds. See Figure 5.7 for caption.

Looking at the different scaffold structures does not provide direct explanations for the differences between the landscapes obtained from the various *holo* scaffolds. The mutations for humanization and maturation are located in residues away from the H3 loop, and there is no major conformational difference in the immediate surrounding of the loop. However, the mutations may have an effect on the scoring despite not being directly in contact with the loop. In addition, a difference in the sampled ensembles cannot be excluded.

Concerning other methods, AMBER models flat landscapes, while Torsions-only and SBROD miss both basins.

Looking at the top cluster, it is identified as a basin only from the *holo* scaffolds, by all methods except Torsions-only. This basin is very deep for DFIRE2, KORP and SOAP-Loop, compared to the rest of the landscapes. However, this basin appears to correspond to a small 'island' of the conformational space, isolated from the rest of sampled states (which are located on a large connected area of the conformational space).

5.3.5 Results after re-sampling from the *apo* scaffold

Section 5.3.1 revealed that the *holo* conformation could not be closely sampled from the *apo* scaffold, due to the presence of an obstructive tyrosine side-chain from the L2 loop. Therefore, as mentioned in that section, a new sampling was performed from this *apo* scaffold, but allowing at most one side-chain in the loop environment to be in collision with the closed backbone. The colliding side-chain was then placed alongside the loop side-chains. If no clash-free state could be found for the side-chains, the sampled state was rejected. 5,000 clash-free states were thus generated for the H3 loop.

Figure 5.9 compares the RMSD_{min} to each known crystal structure, using the two sampling strategies. The *basic* strategy refers to the strategy considering all side-chains in the scaffold rigid during backbone generation, while the *One mobile side-chain* strategy refers to the new strategy, allowing one surrounding side-chain to be removed during backbone generation.

Results show an improvement of RMSD_{min} to the *holo* structures using the new sampling strategy, while the RMSD_{min} to the *apo* conformation remains unchanged. The RMSD_{min} to the *holo* conformation lies between 1.22 and 1.68 Å, depending on the target crystal structure (against 1.57 to 1.79 Å using the basic sampling strategy). This remains a higher value than the RMSD_{min} to the *apo* conformation. As mentioned in Section 5.3.1, the tyrosine that prevents the sampling of the *holo* conformation becomes obstructive following the displacement of the backbone of L2. Moving the dihedrals of the tyrosine side-chain allows to approximate the *holo* conformation more closely, but a more accurate approximation would require a treatment of backbone flexibility in the L2 loop, which constitutes a much more challenging problem.

Concerning top-scoring states, and the ranks of the closest states to a known conformation, AMBER yields much better results using the ensemble generated

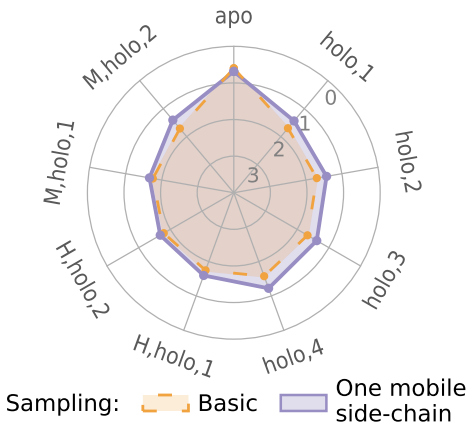


Figure 5.9. Lowest RMSD to the different crystal structures of the H3 loop, among sampled states from the *apo* scaffold, for the two different sampling strategies. RMSDs are calculated on the heavy atoms of the backbone.
The different crystal structures are distributed around the disk and their names indicated outside the disk.
The outer circle corresponds to a RMSD of 0 Å, while the center corresponds to RMSDs of 4 Å and above: a larger colored area indicates that the corresponding structure is found with greater accuracy in the sampled ensemble.

using the new sampling strategy (Figures 5.10 and 5.11). AMBER ranks the five closest states to the *apo* conformation as the 21^{st}, 24^{th}, 1^{st}, 64^{th} and 121^{st} best states, respectively. Concerning the closest states to the *holo* conformation, results are also greatly improved, with the top five closest states ranked 98^{th}, 9^{th}, 3252^{nd}, 174^{th} and 456^{th}, respectively. Overall, results are also improved for DFIRE2 using this sampling strategy, although to a lower extent. Other scoring methods demonstrate similar performance for the two sampling strategies. There is however one exception: ROSETTA places two states above 4 Å away from any known conformation as 2^{nd} and 5^{th} top scoring, while it yielded better results using the ensemble generated using the basic sampling strategy.

The improvement in the results of AMBER and DFIRE2 may be explained by a better placement of the side-chains among the closest states to the two known conformations. As mentioned in Section 4.4, side-chain placement is an important component of the sampling process, but the current side-chain placement method returns the first placement found to satisfy the collision constraints, without further optimization. If the first side-chain placement that is accepted is not energetically favorable, it may lower the quality of a state with an otherwise probable backbone. This is particularly true for DFIRE2, for which no relaxation is performed on the side-chain states. AMBER (just like ROSETTA) includes a relaxation step which may optimize side-chain placement. However, this optimization remains very local and these methods may also be impacted by a very bad side-chain placement.

Finally, concerning the landscapes (reproduced in Figure 5.12), it is interesting to observe that the new sampling strategy unlocks the sampling of intermediary

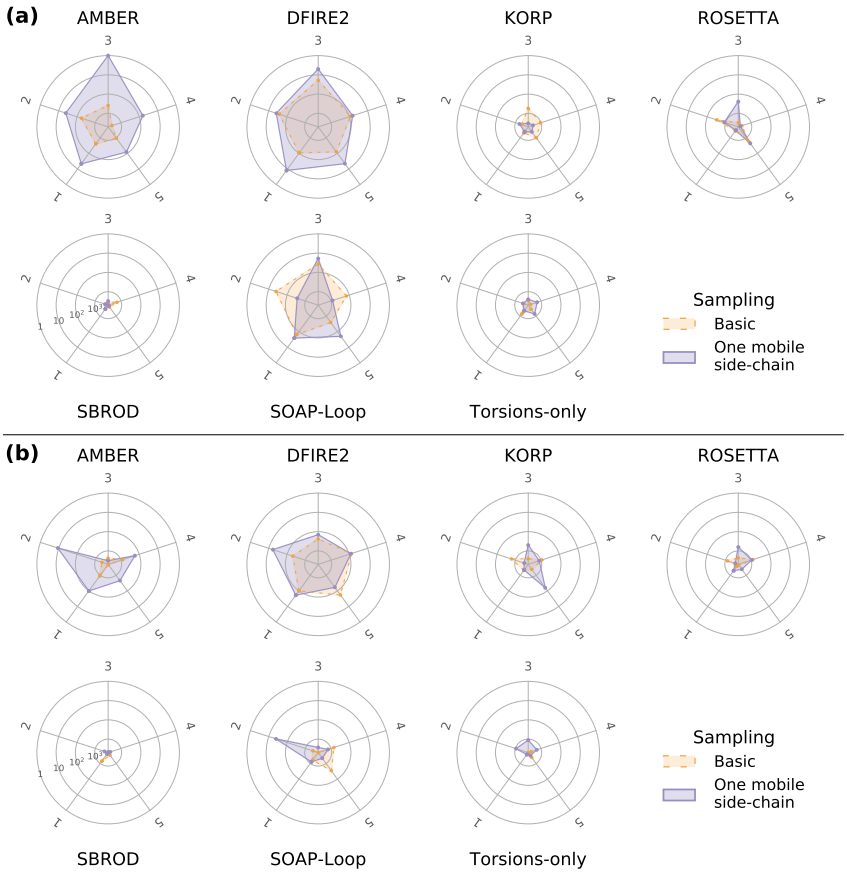


Figure 5.10. Ranks of the 5 closest states to the *apo* (a) and *holo* (b) conformations of the H3 loop, among the states sampled from the *apo* scaffold using 2 different sampling strategies. The ranks are represented by the proximity to the outer circle or to the center. The outer circle corresponds to the best scoring ranks, while the center corresponds to the worst rank. Note that the radial axis has a logarithmic scale.
With such a representation, points far from the center of the circle correspond to well-scored sampled states.

states between the two known conformations. Although the *holo* conformation is still not identified as meta-stable by any of the scoring methods, the basin containing the *apo* conformation extends towards that conformation in the landscapes generated using AMBER, DFIRE2, ROSETTA and SOAP-Loop. This was not the case in the landscapes generated from the ensemble sampled using the basic strategy (Figures 5.7 and 5.8).

5.4 Discussion

Overall, results confirm that DFIRE2 is a very reliable scoring method for the loop modeling process. SOAP-Loop, KORP, ROSETTA and AMBER also perform a highly satisfying scoring of the different sampled states. Note that the structural

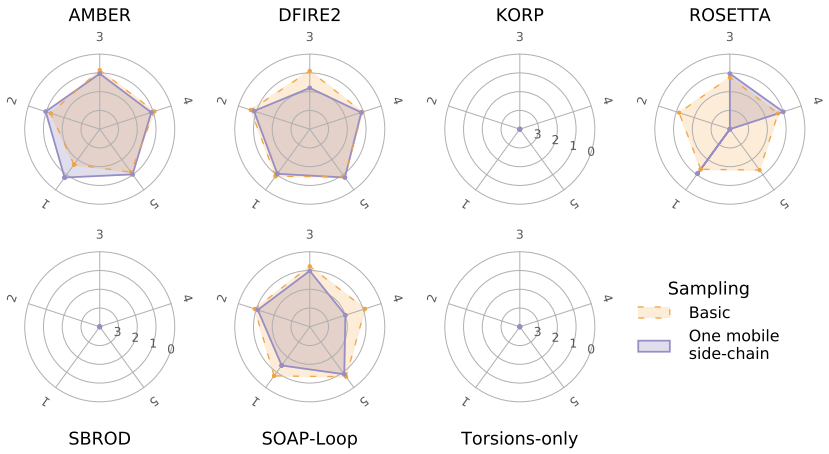


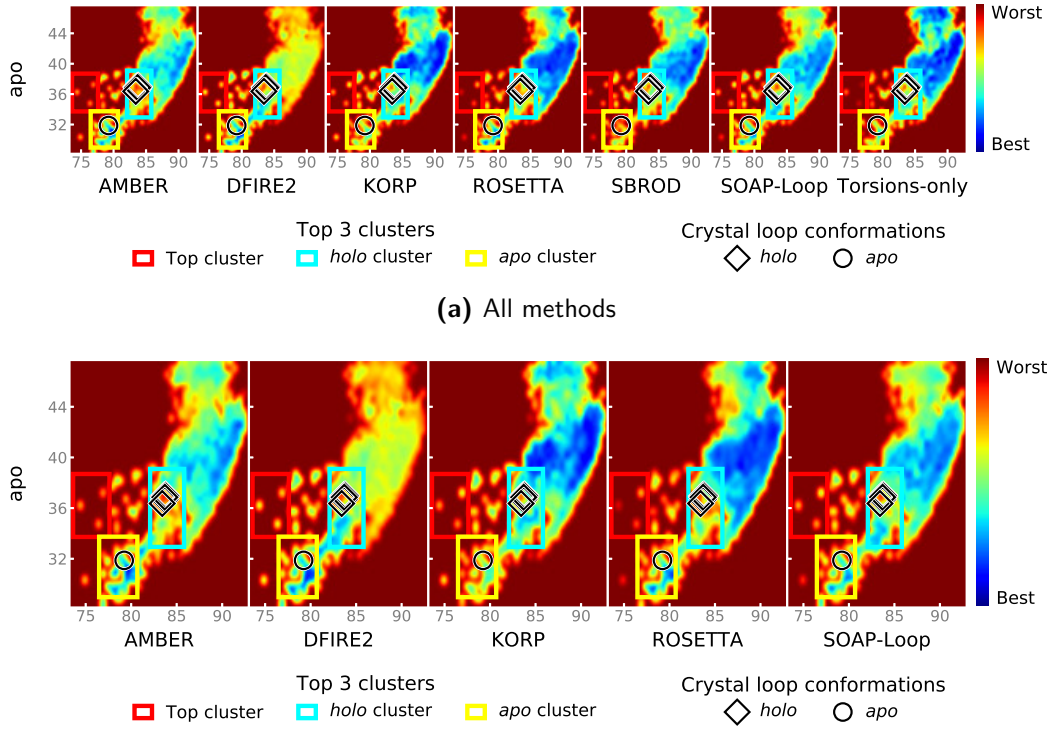
Figure 5.11. Backbone RMSD to the closest known conformation (*apo* or *holo*) of the five top-scoring states among those sampled from the *apo* scaffold, for the two sampling strategies. The outer circle corresponds to a RMSD of 0 Å, while the center corresponds to RMSDs of 4 Å and above: a larger colored area indicates that the top scoring states are close to a known stable conformation.

data employed in this chapter is not public yet, guaranteeing that it has not been used in the tripeptide state database from MoMA-LoopSampler, or in the data from which the different knowledge-based scoring methods were derived.

The importance of the starting scaffold is once again highlighted. The example of the *apo* scaffold showed that a careful treatment of the flexibility in the scaffold can greatly improve the results. Indeed, sampling the backbone using flexible surrounding side-chains generated more exhaustive ensembles, that in turn produced informative landscapes showing the intermediary states between the two stable conformations. Ideally, flexibility should also be considered in the scaffold backbone, although that makes the sampling considerably more complex.

The *apo* scaffold originated from Crystal structure 1 (Table 5.1), in which the CDRs were found to form multiple crystal contacts. In particular, H3 forms two hydrogen bonds with surrounding units in the crystal. Yet, in our results, the *apo* conformation is found in a basin despite the absence of neighboring Fabs for the H3 loop. This indicates that the *apo* conformation of the loop is stable even in the absence of these favorable crystal contacts. This suggests an interesting application for our methods: verifying that the structure in the crystal does correspond to a stable structure for the loop, even in the absence of binding partner or of artifactual crystal contacts.

It is likely that the mutations between the different scaffolds have an impact on the produced landscapes, and on the results in general. However, given that the mutations concern residues away from the H3 loop, we could not directly establish a connection between the differences observed in the results and specific mutations. More generally, we cannot conclude on the effects of humanization and maturation on the conformation or flexibility of the H3 loop. The differences in the results



(a) All methods

(b) 5 methods with the most reliable results

Figure 5.12. Energy landscapes of H3 obtained by the different scoring methods from the *apo* scaffold, using the sampling strategy allowing a side-chain to be replaced in the surroundings of the loop.
Each column correspond to a single scoring method. The x- and y- axes correspond to the first and second principal components of the PCA, respectively.

may have other sources such as a difference in the sampled ensembles, or minor conformational changes in the scaffolds.

Clustering the top states identified by the various methods led us to identify a conformation, $\mathcal{C}$, different from the *apo* and *holo* conformation, and adopted by many of the top-scoring states identified by the different scoring methods (except Torsions-only). This conformation is projected in a deep narrow basin that is topologically separated from the main sampled area of the conformational space. The limited size of this basin suggests a very low conformational entropy, which may explain why the conformation, although highly stable, is not likely to be observed.

Although data is insufficient to conclude on the existence of this conformation *in vivo*, this observation remains intriguing. On the one hand, if this conformation actually exists *in vivo*, this underlines once again the importance of using other experimental techniques than X-ray crystallography, such as nuclear magnetic resonance (NMR), to obtain loop structures. However, even with such methods, if the conformational entropy is too low, this conformation might still be missed. This highlights the possibilities offered by modeling methods. The fact that such a con-

formation may exist, but that its conformational entropy is too low to make it likely to be adopted by the loop, is an important information for loop design. If stabilizing the loop in this conformation constitutes a desirable objective, one may attempt to increase the conformational entropy of the loop in this conformation.

On the other hand, if this conformation is an artifact of modeling methods that is never actually adopted by the loop, this suggests a major flaw of the scoring methods. The fact that the most reliable scoring methods agree to identify the same conformation, but that this conformation is wrong, would remain a surprising result, given that these scoring methods are based on relatively different principles. This would indicate that using consensus on different scoring methods is not guaranteed to provide more reliable results.

5.5 Conclusion

The methods from Chapters 3 and 4 were successfully applied to the flexible H3 loop from Ab, producing informative and consistent landscapes. Although general, the loop modeling methods developed in this thesis are suitable for antibody H3 loop modeling, which was the initial motivation for these methodological developments. The generality of these methods makes them more likely to succeed on antibody CDRs that cannot be assigned a canonical class, or with designed CDRs that do not resemble existing hypervariable loops.

This work confirmed the conclusions drawn in the previous chapter (Chapter 4) concerning: the most reliable scoring methods (DFIRE2, but also KORP, AMBER), the importance of the starting scaffold, (and, indirectly, of modeling flexibility in the loop surroundings), and the need of accurate side-chain placement methods. In addition, it revealed another possible application of the methods developed in this thesis: verifying that loop conformations provided by crystal structures are not artifacts of the crystallization process.

An interesting observation emerged from the modeled landscapes and the clustering of top results: a conformation predicted to be highly stable by most top-scoring methods was identified. However, the topography of the landscapes suggests that conformational entropy is very low in that region, and therefore, that this conformation may not by highly statistically likely. This may explain why it is not adopted by the loop in any of the crystal structures. Despite this fact, this observation constitutes an interesting information for loop design. An interesting direction for future work would be the validation of this conformation. Molecular dynamics simulations are currently being run to gain further insight into this conformation.

As mentioned in Section 1.3.2.4, loop modeling for antibodies may benefit from using antibody fragments for generating loop samples. One way of making MoMA-LoopSampler more antibody-specific would thus be to use a database of tripeptide states extracted from public antibody structures, although potential lack of data may prevent an exhaustive sampling of the loop conformational space. This constitutes an interesting direction for future work.

Conclusion

This thesis addressed the problem of better representing flexibility in antibodies through the development and the assessment of efficient methods for protein loop modeling. Firstly, an analysis of the sources of failure of docking algorithms applied to antibody-antigen complexes was conducted. This study started with the analysis of conformational changes occurring upon binding in 27 antibodies. Multiple metrics were employed, which enabled a detailed classification of the different classes of conformational changes and of their combination. The performance of four docking algorithms applied to these antibody-antigen systems were then analyzed in light of these conformational changes. This analysis concluded that docking algorithms were mainly impacted by large conformational changes in CDR loops located in the antigen-combining site, although other sources of docking algorithm failures were also hypothesized. This thesis work was therefore oriented towards the development of protein loop modeling methods that better represent and consider conformational flexibility.

In this context, a general method to sample the conformational space of protein loops in an exhaustive way was then developed. This loop sampling method employs a database of loop fragments, and enforces strict steric clash constraints. These two features contribute to the generation of high-quality ensembles, by filtering many improbable loop conformations. Moreover, besides the quality of the generated ensembles, diversity constitutes an essential element of this method. Sampling diversity is indeed crucial for an accurate representation of loop flexibility. Therefore, carefully investigating the exhaustiveness of the generated ensembles constituted a large part of the method validation work. Overall, this method was found to produce conformational ensembles that are better filtered than the ensembles generated by other state-of-the-art methods, while maintaining sufficient coverage of the conformational space. A variant employing reinforcement learning was also proposed, in order to speed up sampling. This variant produces larger ensembles with a higher resolution and without loss of diversity, provided a careful parameterization is adopted.

The third part of this thesis described a workflow that can be followed to model insightful energy landscapes of flexible protein loops. After performing an exhaustive sampling using the previously described method, this process requires an accurate scoring component for which multiple options were compared in this work. Several state-of-the-art loop scoring methods were integrated to this process, in order to produce 2D visualizations of the energy landscapes of eight flexible protein loops. These energy landscapes were then compared, based on their usability and consistency with experimental data. Results indicated that useful information can be extracted from the landscapes thus produced, about e.g.: loop flexibility, (meta-)stable conformations, ... In addition, this analysis provided guidelines as to what scoring methods are appropriate depending on the desired application.

Finally, the general methods developed in this thesis were used to model a flexible H3 loop. Results validated the application of these new methods to hypervariable antibody loops, our initial system of interest. Modeled energy landscapes identified both the *apo* and *holo* conformations of the loop as stable conformations. Such accurate landscapes can be valuable in the context of antibody design.

Future research

This work revealed multiple directions for future research. To begin with, the sampling method may be improved in several ways. For example, by including structural data from similar sequences for tripeptide types for which data is scarce, or through the development of other learning strategies with lower memory requirements.

Another way to improve the sampling method would be to divert the learning process from its initial use, by adjusting the scores of the leaves in the learning tree in order to influence the sampling distribution. For instance, scores could be regulated so that the tripeptide states that are employed in the generated ensemble follow the distribution encoded in the database. However, this solely makes sense if the distribution in the database is meaningful: for example, if it represents the implicit propensity of the different tripeptides to adopt the corresponding states in a coil fragment. If a meaningful distribution can be guaranteed in the database, and if the sampling strategy is designed to follow that distribution, the generated ensemble might ultimately be employed to estimate thermodynamic properties for the loop system. However, that would require regulating the frequencies of the so-called 'synthetic' states as well, meaning that the strategy consisting in concatenating tripeptide states may have to be revised. Instead, a strategy consisting in overlapping tripeptides states (using a sliding window of three residues) could be adopted.

Solutions should also be sought for a better treatment of flexibility in the sampling scaffold. The sampling method developed during this thesis allows for flexibility in surrounding side-chains during the placement of the loop side-chains. However, the need to allow flexibility in the loop surroundings earlier in the sampling process, i.e. during backbone generation, was repeatedly exposed in this work. Ideally, flexibility modeling should also be extended to the surrounding backbone, although that constitutes a substantially harder challenge.

The development of fast and accurate side-chain placement methods constitutes another direction for future work. State-of-the-art methods like SCWRL4 perform fast side-chain placements, but are prone to providing very poor results due to steric clashes in highly constrained environments. Conversely, our method provides clash-free solutions, but is highly time-consuming. Moreover, it does not include any energy consideration in the process and returns the first clash-free solution. Integrating a short optimization using a fast scoring methods like DFIRE2 constitutes an interesting option to improve the quality of the placement. Moreover, instead of randomly modifying dihedral angles to correct observed collisions, more

sophisticated geometric methods may be employed, saving considerable time and increasing the success rate of the side-chain placement step.

Analyzing energy landscapes in higher dimension might also provide valuable insight into loop structures and dynamics. Automatic detection of energy basins (e.g. employing clustering methods integrating energy evaluations) or of transition regions may be more accurate than projecting those landscapes in 2D, although not being as straightforward to interpret. Applying existing or designing new algorithms for loop motion modeling, using either the results of the combined sampling and scoring steps, or the reconstructed energy landscapes, also constitutes an interesting direction for future work. Finally, interesting applications of the process proposed in Chapter 4 could be investigated. First, it could be used to investigate the effects of mutations on the loop conformation, in order, for example, to explain the structural basis of a genetic disease. Other possible applications of the process include the validation of sequences proposed in the context of loop design (including negative design).

Appendix A

French Summary

Contents

Introduction

Les anticorps sont des protéines essentielles du système immunitaire. Ils sont capables d'identifier des molécules spécifiques exposées par les pathogènes (les antigènes) en s'y liant. La neutralisation du pathogène est engendrée soit directement par la liaison de l'anticorps, soit par la réponse immunitaire déclenchée par cette liaison.

Cette thèse traite de la modélisation d'anticorps, en particulier de celle des boucles hypervariables situées à l'interface avec l'antigène. Ces boucles protéiques, appelées CDRs, sont responsables de la reconnaissance spécifique de l'antigène, de la formation du complexe anticorps-antigène, et de l'affinité de l'anticorps pour l'antigène. La spécificité et l'affinité de cette interaction sont possibles grâce à une grande variabilité de séquence, ainsi qu'à la plasticité de ces fragments protéiques. En effet, contrairement à d'autres éléments structurels plus stables, comme les

hélices α ou les feuillets β, les boucles protéiques font preuve d'une grande flexibilité qui joue un rôle majeur dans de nombreux processus biologiques.

Ces régions protéiques flexibles constituent un défi considérable pour la biologie structurale. La plupart des données expérimentales liées aux structures protéiques sont obtenues par l'intermédiaire de cristallographie aux rayons X. Bien que cette technique soit capable de déterminer de manière précise la structure des éléments les plus stables, elle ne permet pas de représenter correctement les parties plus flexibles. En effet, elle fournit une structure unique, ce qui n'est pas adapté aux boucles protéiques, qui adoptent un ensemble de conformations différentes avec diverses probabilités associées. Comme le montrent plusieurs travaux récents, et en raison du biais de représentation dû à la nature des méthodes expérimentales employées et au manque de données liées à la flexibilité conformationnelle, les outils actuels de bioinformatique structurale ne sont pas capables de correctement modéliser les boucles protéiques.

La modélisation d'une boucle protéique s'effectue généralement en deux étapes. La première, appelée *échantillonnage*, consiste à générer un ensemble conformationnel le plus exhaustif possible, afin de représenter de manière globale ce fragment protéique. La seconde, appelée *évaluation*, consiste à attribuer un score à chacune des conformations générées lors de la phase précédente. Ce score est censé représenté les différences d'énergie entre les différents modèles échantillonnés. L'échantillonnage et l'évaluation restent des problèmes ouverts. En effet, les méthodes développées jusqu'à présent dans ce domaine se concentrent principalement sur la prédiction d'une unique conformation stable, ce qui n'est pas suffisamment représentatif.

Cette thèse a pour but principal de développer de nouvelles méthodes au delà de l'état de l'art pour la modélisation de boucles protéiques, et de démontrer l'intérêt dans le cas particulier de la modélisation des CDRs chez les anticorps.

Le chapitre 1 présente le contexte dans lequel s'inscrit le travail de thèse. Il commence par présenter les notions générales liées à la fonction et à la structure des protéines, avant de se concentrer sur les anticorps, notamment leur propriétés structurales, le problème de la prédiction de leur structure et de l'amarrage anticorps-antigène et celui du design d'anticorps. Le problème de la modélisation de boucles protéiques est ensuite présenté, et un certain nombre de méthodes récentes dans ce domaines sont décrites. Enfin, ce chapitre présente les paysages énergétiques, et diverses méthodes pour les explorer.

Le chapitre 2 présente ensuite une analyse des changements structuraux survenant chez les anticorps au moment de leur liaison avec l'antigène. Cette étude, basée sur l'analyse de données structurales expérimentales, montre que les changements conformationnels chez les anticorps (qui concernent principalement les boucles), peuvent être substantiels et ne sont pas suffisamment pris en compte. En particulier, les algorithmes d'amarrage voient leurs résultats se dégrader lorsqu'ils ont affaire à des boucles hypervariables particulièrement flexibles au niveau du site de liaison.

Suite à ce constat, les deux chapitres suivants sont orientés vers des méthodes générales pour la modélisation de boucles protéiques. Le chapitre 3 présente une

méthode pour l'échantillonnage exhaustif de l'espace conformationnel des boucles. Celle-ci inclut un composant inspiré de l'apprentissage par renforcement pour accélérer la génération de modèles de boucles. Cette méthode inspirée de la robotique emploie une représentation géométrique interdisant les collisions stériques les plus importantes entre les atomes. La procédure d'échantillonnage consiste à concaténer des fragments protéiques à partir d'une base de données conçue pour cette application.

Le chapitre 4 présente ensuite une analyse en profondeur des performances de plusieurs méthodes d'évaluation appliquées à plusieurs boucles flexibles pour lesquelles on dispose de données expérimentales. Un protocole combinant échantillonnage, évaluation et projection d'échantillons de boucles est employé afin de visualiser les paysages énergétiques implicitement modélisés par ces méthodes. L'analyse de ces paysages énergétiques permet d'identifier de manière précise à la fois les défauts d'échantillonnage et les limites des méthodes d'évaluation.

Enfin, ces méthodes furent appliquées à un anticorps dont une boucle présente un changement conformationnel au moment de la liaison avec l'antigène. Cette étude est décrite dans le chapitre 5. Les résultats montrent que les méthodes précédemment analysées et développées dans les chapitres 3 et 4 permettent de modéliser un paysage énergétique cohérent pour cette boucle flexible, identifiant à la fois la conformation adoptée par l'anticorps libre, et celle adoptée par l'anticorps lié à l'antigène, ainsi qu'une autre conformation potentiellement très rigide. Cela suggère que ces méthodes pourraient être appliquées dans le cadre du design d'anticorps, par exemple. La visualisation des paysages énergétiques modélisés pourrait en effet permettre de prédire la stabilité d'une boucle dans une certaine position (ou dans une autre), éliminant ainsi les séquences produisant des boucles insuffisamment stables ou qui adoptent des conformations indésirables.

Une conclusion générale résume le travail effectué durant la thèse et indique des directions possibles pour la recherche future.

Contributions de la thèse

Dans l'objectif de mieux modéliser les structures des anticorps, cette thèse apporte des contributions à différents niveaux. Ces contributions s'étendent d'une étude pour établir plus précisément les besoins méthodologiques au développement de techniques pour répondre à ces besoins, et terminent par la validation de ces méthodes sur notre système d'intérêt, les anticorps. Les contributions sont détaillées ci-après.

- *Analyse des changements conformationnels chez les anticorps survenant lors de la liaison avec l'antigène* : la thèse commence par mesurer et décrire les changements conformationnels de 27 systèmes anticorps-antigènes pour lesquels à la fois les structures libres et liées existent. En analysant séparément différentes sous-parties de la structure de l'anticorps, une analyse qualitative et quantitative des différents types de mouvement fut conduite. Combiner ces

observations avec les résultats rapportés de quatre algorithmes d'amarrage sur les systèmes étudiés permit d'identifier les importants mouvements de boucles CDR comme la source probable de l'échec de ces algorithmes. Ce travail fit l'objet d'une publication dans *Immunology Letters* [Barozet 2018].

- *Développement et tests approfondis d'une nouvelle méthode d'échantillonnage de boucles* : cette méthode constitue la première partie des contributions méthodologiques proposées pour répondre au problème de la mauvaise modélisation de la flexibilité des boucles protéiques. Cette méthode, appelée MoMA-LoopSampler, adopte une représentation géométrique avec stricte détection des collisions et concatène des états tripeptidiques extraits d'une base de données dédiée. Une nouvelle approche inspirée de l'apprentissage par renforcement fut intégrée pour accélérer l'échantillonnage. Étant donné que cette méthode fut principalement développée pour modéliser les boucles flexibles, il nous parut important d'assurer une diversité suffisante des conformations échantillonnées, tout en maintenant des ensembles conformationnels de bonne qualité, et c'est donc ce sur quoi l'accent fut mis lors du développement et de l'évaluation de MoMA-LoopSampler. La validité des méthodes employées en elles-mêmes, ainsi que la qualité des résultats obtenus et des ensembles générés furent étudiés en détail. Ce travail fit l'objet d'une publication dans *Bioinformatics* [Barozet 2019b].

- *Comparaison de méthodes récentes d'évaluation de boucles* : cette contribution correspond à la seconde partie des développements méthodologiques menés pendant la thèse. Dans ce travail, un procédé pour visualiser et facilement interpréter les paysages énergétiques des boucles flexibles modélisés par différentes méthodes d'évaluation fut développé. Des ensembles conformationnels furent échantillonnés avec MoMA-LoopSampler pour plusieurs systèmes protéiques contenant une boucle flexible. Des méthodes de pointe furent ensuite utilisées pour attribuer un score aux différents échantillons, et une projection 2D utilisant des descripteurs explicites fut employée pour dessiner les paysages énergétiques. Cette étude permit d'identifier les meilleures méthodes à utiliser pour évaluer les conformations de boucles flexibles de manière fiable, et indiqua les potentielles sources d'imprécisions dans les paysages produits. Un manuscrit décrivant ce travail est en phase de révision par le journal *Bioinformatics*.

- *Modélisation du paysage énergétique d'une boucle CDR flexible* : ce travail illustre la manière selon laquelle les méthodes développées lors de cette thèse peuvent être appliquées à une boucle hypervariable d'anticorps. En modélisant de manière correcte le paysage énergétique d'une boucle CDR, nous montrons comment cette thèse répond au problème initialement formulé, à savoir comment correctement modéliser la flexibilité dans les boucles hypervariables chez les anticorps. Un manuscrit est actuellement en préparation sur ce travail, et nous espérons publier les structures associées.

Bien que les anticorps constituent la motivation initiale pour les développements méthodologiques effectués lors de cette thèse, les méthodes présentées dans ce travail ne sont pas spécifiques aux anticorps et peuvent très bien être généralisées à l'analyse des boucles protéiques dans d'autres systèmes.

A.1 Contexte

Le Chapitre 1 introduit les concepts permettant de comprendre le travail présenté dans cette thèse. La Section 1.1 commence par décrire brièvement les notions de fonction et de structure des protéines. Les anticorps, qui sont des exemples de systèmes protéiques, sont ensuite abordés (Section 1.2). Leur structure est décrite de manière détaillée, notamment les boucles responsables de l'interaction avec l'antigène. Plusieurs applications nécessitant une modélisation précise de ces boucles sont également présentées. La section suivante (Section 1.3) introduit le problème de la modélisation des boucles protéiques, notamment l'échantillonnage et l'évaluation de boucles, qui sont respectivement au cœur des Chapitres 3 et 4. Finalement, la Section 1.4 présente les notions d'exploration de paysage énergétique nécessaires pour comprendre la conformation et la dynamique de protéines. L'intérêt porté à ces paysages énergétiques est double. Tout d'abord, la capacité à les modéliser de manière à la fois correcte et précise permettrait d'avancer considérablement dans la compréhension du mode d'action des protéines ou de leur boucles. Ils constituent donc un objectif pour les méthodes de modélisation de boucles. Ensuite, les paysages énergétiques sont utilisés tout au long de la thèse pour valider la cohérence des résultats fournis par les méthodes de modélisation de boucles protéiques.

A.2 Flexibilité des boucles d'anticorps

A.2.1 Présentation du Chapitre 2

Le Chapitre 2 analyse les boucles flexibles assurant la reconnaissance de l'antigène par l'anticorps d'un point de vue structural. Ces boucles CDR, décrites dans le Chapitre 1 (Section 1.2.3) sont déterminantes à la fois pour la spécificité et l'affinité de l'anticorps pour l'antigène. Plus précisément, ces deux caractéristiques cruciales de la reconnaissance de l'antigène par l'anticorps sont permises par la complexe flexibilité des fragments CDRs (Section 1.2.4). Cette variabilité conformationnelle constitue une difficulté considérable pour la prédiction de la pose d'amarrage anticorps-antigène, qui ignore *a priori* les conformations que les boucles adoptent dans la conformation liée de l'anticorps (Section 1.2.7).

Les changements conformationnels sont la plupart du temps analysés au cas par cas, quand les structures à la fois de l'anticorps en complexe et de l'anticorps libre deviennent disponibles. Bien que Sela-Culang *et al.* [Sela-Culang 2012] aient analysé les changements conformationnels survenant lors de la formation du complexe anticorps-antigène sur un large ensemble de 49 anticorps différents, leur anal-

yse se concentra sur les changements conformationnels significatifs sur l'ensemble des anticorps considérés. Le but principal de notre analyse est quant à lui de rassembler des données informations quantitatives sur les changements conformationnels pour un ensemble divers de cas, de manière à fournir une base permettant d'estimer l'amplitude des réarrangements possibles et de définir l'intervalle dans lequel peuvent varier ces changements conformationnels.

Les structures liées et libres des anticorps présents dans le protein-protein benchmark version 5.0 [Vreven 2015] furent comparées au moyen de mesures de la déviation de la racine de la moyenne des carrés (*root-mean-square deviations*, RMSD) des positions atomiques et des valeurs des angles dièdres, sur différentes sous-parties des fragments Fab. L'objectif de cette analyse des mesures de déviation est de mieux caractériser les changements conformationnels subis par les structures d'anticorps lors de la liaison avec l'antigène. Vreven et collègues [Vreven 2015] testèrent également quatre algorithmes d'amarrage sur les nouveaux cas ajoutés à leur benchmark. Leur résultats sont analysés ici en parallèle avec nos mesures, afin de mieux comprendre les effets des différents types de changements conformationnels sur la performance des algorithmes de prédiction d'amarrage. Le travail présenté dans le Chapitre 2 fut publié dans *Immunology Letters* [Barozet 2018].

A.2.2 Conclusions du Chapitre 2

Dans ce travail, les changements conformationnels survenant lors de la liaison de 27 anticorps avec leur antigène respectif furent analysés. Les résultats montrent que les régions cadres (*Framework Regions, FR*), sont stables d'un point de vue structural, malgré quelques mouvements des chaînes latérales en surface. Une observation plus importante est que la variabilité de ces régions est d'une amplitude similaire chez les différents anticorps, avec très peu de valeurs très divergentes. Les boucles hypervariables sont quant à elles bien plus flexibles dans l'ensemble, avec des variations beaucoup plus hétérogènes d'un système à l'autre. Certaines sont hautement stable structurellement et très rigides, quand d'autres subissent d'importants changements conformationnels au moment de la liaison de l'antigène. Les changements conformationnels sont de nature variée : grand mouvement de squelette, ou encore importants réarrangements de chaînes latérales. Certains mouvements sont uniquement locaux: la boucle reste globalement au même emplacement, mais sa conformation interne change, alors que certains mouvements concernent la boucle dans sa globalité. La classification des mouvements de boucles montra qu'il y a une diversité encore plus importante si l'on considère le site de liaison de l'antigène dans son ensemble: chaque anticorps montre une profile unique, différent de celui des autres cas étudiés. La taille limitée du jeu de données ne permet pas de tirer de conclusions plus poussées, mais l'extension de cette classification à un nombre plus conséquent de cas constitue une direction intéressante pour la recherche future. L'orientation entre le fragment variable (Fv) et le domaine constant du fragment Fab varie également beaucoup. Un nombre important d'anticorps présentent des variations de l'angle de coude (*elbow angle*) supérieures à 25°.

Les changements conformationnels expliquent partiellement la difficulté rencontrée par les algorithmes d'amarrage dans la plupart des cas. Les anticorps présentant d'importants mouvements de boucles CDR ou des réarrangements de chaînes latérales conséquents au moment de la liaison représentent des cas plus difficiles pour ces algorithmes, ce qui suggère que les changements dans la topologie du site de liaison sont une obstruction majeure à la réussite de l'amarrage informatique. Cependant, les changements conformationnels à eux seuls ne peuvent pas expliquer l'échec de la prédiction dans certains cas, ce qui suggère que l'évaluation correcte des différentes poses reste un problème important dans la prédiction de l'amarrage anticorps-antigène. Certains anticorps avec des sites de liaison rigides mais d'importantes variations de l'angle de coude font l'objet de mauvais résultats avec les algorithmes d'amarrage, ce qui suggère que les mouvements importants des domaines constants des anticorps peuvent perturber la prédiction de l'amarrage, peut-être en raison de l'imprécision des méthodes d'évaluation des poses. Il a aussi été établi que la nature de l'interface anticorps-antigène joue un rôle dans la performance des algorithmes d'amarrage. Les interfaces avec un nombre de contacts électrostatiques favorables largement supérieurs au nombre de contacts électrostatiques défavorables sont mieux prédits (en considérant les interfaces sans changements conformationnels), ce qui suggère que les fonctions d'évaluation donnent beaucoup d'importance aux contacts électrostatiques pour estimer la qualité d'une pose. L'amélioration de la prédiction d'amarrage anticorps-antigène passera donc probablement par des fonctions d'évaluations de meilleure qualité, qui donnent de meilleurs résultats pour les interfaces dont la nature n'est pas principalement électrostatique.

En mettant en évidence les changements conformationnels dans les boucles CDR et les mauvais résultats dus à l'incapacité des méthodes de prédiction de pose à gérer (ou modéliser) de tels changements, ces conclusions soulignent l'importance de bien modéliser les boucles flexibles. La modélisation de boucles s'opère en deux grandes étapes: l'échantillonnage et l'évaluation. La première est au centre du Chapitre 3 qui présente une méthode d'échantillonnage exhaustif pour explorer l'espace conformationnel des boucles protéiques. La seconde est étudiée dans le Chapitre 4, avec une estimation des performances des méthodes récentes d'évaluation de boucles et de leur capacité à modéliser le paysage énergétique des boucles protéiques flexibles.

A.3 Échantillonnage de boucles protéiques

A.3.1 Présentation du Chapitre 3

Le Chapitre 2 souligne l'importance de modéliser les boucles d'anticorps de manière précise et fiable, tout particulièrement quand ces boucles démontrent de la flexibilité. Comme cela est mentionné dans la Section 1.3.1, ce problème n'est pas spécifique aux anticorps et peut se généraliser aux boucles protéiques d'autres systèmes. Les méthodes de modélisation existantes se concentrent principalement sur la prédiction d'une unique conformation pour une boucle manquante. Par conséquent,

l'évaluation de la performance des méthodes d'échantillonnage de boucles consistent généralement à les associer à une méthode d'évaluation et à chercher des conformations proches de la boucles native dans les échantillons les mieux notés. Cependant, le défaut majeur de ce processus est qu'il fait implicitement l'hypothèse que la boucle ne peut adopter qu'une unique conformation, ignorant ainsi sa flexibilité potentielle. Pourtant, le manque de données structurales dans certaines portions des protéines dont la structure fut résolue par cristallographie aux rayons X indique en général que les régions concernées sont trop flexibles pour être observées à l'aide de cette méthode communément utilisée. Ces considérations suggèrent que le fait de représenter les boucles manquantes dans les structures cristallines par une unique conformation modélisée est intrinsèquement contradictoire.

Le Chapitre 3 présente une nouvelle méthode d'échantillonnage de boucles protéiques, appelée MoMA-LoopSampler, qui ne se concentre par sur la prédiction d'une unique conformation, mais qui a pour but une exploration plus exhaustive de l'espace conformationnel de la boucle. La méthode proposée n'est pas spécifique aux anticorps et peut être appliquées à n'importe quelle boucle protéique. Elle emploie une approche hybride de la modélisation de boucle qui construit des conformations en utilisant une base de données structurale de petits fragments protéique et un solveur de cinématique inverse (*inverse kinematics*, IK). La boucle est divisée en fragments consécutifs de trois résidus. Tous les fragments sauf un sont échantillonnés de manière itérative à partir de la base de donnée et légèrement perturbé pour augmenter la taille de l'espace conformationnel exploré. L'état du dernier fragment est résolu par cinématique inverse pour fermer la boucle. MoMA-LoopSampler varie le choix du fragment résolue par cinématique inverse pour augmenter davantage l'espace échantillonné. Lors du placement de chaque fragment dans la boucle en construction, les collisions sont détectées et la vérification de la distance restant à construire est effectuée afin de limiter l'échantillonnage à l'espace conformationnel possible. Le Chapitre 3 étudie également une version plus avancée de la méthode d'échantillonnage, qui incorpore une stratégie employant une heuristique inspirée de l'apprentissage par renforcement. Pour chaque fragment à construire dans la boucle, les échantillons présents dans la base de données sont groupés après une projection dans une espace de basse dimension. Les groupes d'échantillons permettant de former des conformations acceptables de boucles sont alors choisis plus fréquemment.

La validation de la méthode ainsi que le test de ses performances constitue une part conséquente de ce travail. Un certain nombre de caractéristiques de MoMA-LoopSampler furent soigneusement étudiés, notamment la base de données de fragments ainsi que la projection employées pour l'heuristique d'apprentissage par renforcement. Quant à la performance, elle fut évaluée sur plusieurs aspects. Tout d'abord, la capacité de la méthode à générer des conformations proches des conformations natives fut testée. Bien que ce ne soit pas l'objectif principal de la méthode, comme cela a déjà été précisé, échantillonner les conformations connues est une obligation de la méthode. De plus, cet aspect en particulier fut comparé à celui d'autres méthodes récentes alors que les autres aspects de la performance

ne sont pas aussi facilement comparables. Ensuite, en utilisant l'exemple d'une boucle flexible de la protéine streptavidine, la capacité de MoMA-LoopSampler à échantillonner de multiples états stables, ainsi que les états intermédiaires, fut étudiée. Les effets de l'activation de l'apprentissage par renforcement furent aussi minutieusement analysés et testés, avec l'objectif clair de maintenir la diversité conformationnelle dans les ensembles générés.

Le Chapitre 3 est organisé comme suit. La Section 3.2 décrit les détails de la méthode MoMA-LoopSampler et s'attarde sur la validation de certains de ses composants. La Section 3.3 démontre la capacité des approches basiques et utilisant l'heuristique d'apprentissage à générer des ensembles de conformations variés, et présente l'application de MoMA-LoopSampler à différents datasets de test de prédiction de boucles. Un résumé de ces résultats, ainsi que les perspectives de recherche future, sont discutés dans la Section 3.4.

Ce travail sera prochainement publié dans *Bioinformatics* [Barozet 2019b].

A.3.2 Conclusions du Chapitre 3

La méthode présentée dans le Chapitre 3 emploie de la connaissance séquence-dépendante sur les structures locales et des techniques géométriques, ainsi que de l'apprentissage par renforcement pour échantillonner l'espace conformationnel des boucles protéiques de manière efficace et exhaustive. Les résultats montrent que cette nouvelle méthode a des performances comparables à (voire meilleures que) les méthodes existantes du domaine en terme de temps de calcul et que l'ensemble des conformations de boucles échantillonnées contient bien les structures déterminées expérimentalement (l'état dit 'natif' de la boucle). L'approche basée sur l'apprentissage par renforcement qui a été implémentée permet à MoMA-LoopSampler d'accélérer l'échantillonnage tout en maintenant la diversité conformationnelle (évitant ainsi le surapprentissage), et peut être appliquée à des boucles plus longues (15 résidus). Ce travail montre aussi que MoMA-LoopSampler est capable d'échantillonner correctement les boucles pouvant adopter plusieurs conformations présentes dans des bassins de basse énergie distincts. Cette méthode constitue donc un outil intéressant pour étudier les paysages énergétiques et les transitions conformationnelles.

Les améliorations possibles de la méthode incluent l'amélioration du composant d'apprentissage pour limiter son coût en mémoire. Une autre piste à envisager est l'ajustement du score des feuilles de la structure d'apprentissage de manière à ce que la distribution des conformations échantillonnées corresponde à la distribution des états tripeptidiques présents dans la base de données. Des ajustements à la base de données pourraient également améliorer la qualité des résultats. Par exemple, filtrer la base de données pour ne garder qu'un représentant parmi des tripeptides très similaires pourrait accélérer l'échantillonnage, et ajouter les états des séquences similaires aux états correspondants à des séquences tripeptidiques rares pourrait permettre l'échantillonnage des conformations pour l'instant inaccessibles en raison d'un potentiel manque de données.

Enfin, construire des ensembles de boucles corrects nécessite à la fois un composant d'échantillonnage et un composant d'évaluation. Alors que MoMA-LoopSampler a pour but de fournir un ensemble divers de conformations possibles, cette méthode n'évalue pas les conformations échantillonnées ou leur probabilité d'existence. Le développement d'une méthode d'évaluation appropriée, ou l'intégration de telles méthodes dans MoMA-LoopSamplerconstitue une direction intéressante de recherche.

Cette dernière observation a motivé une étude comparative de plusieurs méthodes récentes d'évaluation de boucles. Cette analyse avait pour but de déterminer une méthode d'évaluation qui complémenterait bien MoMA-LoopSampler tout en fournissant une évaluation précise de la qualité des conformations échantillonnées. L'analyse détaillée fait l'objet du Chapitre 4.

A.4 Évaluation d'états de boucles, paysages énergétiques

A.4.1 Présentation du Chapitre 4

La modélisation de boucles protéiques s'opère en deux étapes: une première étape d'échantillonnage conformationnel, pour générer un ensemble exhaustif de conformations de boucles, suivie d'une étape d'évaluation dont le but est d'éliminer les conformations impossibles et de sélectionner les plus probables. Le Chapitre 3 introduit une nouvelle méthode qui effectue un échantillonnage de boucle exhaustif efficace d'un point de vue computationnel. Le Chapitre 4 se concentre quant à lui sur la second étape, *l'évaluation*. Plus précisément, il compare les méthodes récentes existantes sur leur capacité à évaluer correctement divers états conformationnels de boucles obtenus par échantillonnage, et à modéliser un paysage énergétique cohérent.

Tout comme les méthodes d'échantillonnage de boucles, les méthodes d'évaluation de boucles sont principalement évaluées sur leur capacité à identifier la conformation "native" parmi un ensemble de conformations. Cependant, plusieurs applications, à l'instar du design de boucles [Kundert 2019], nécessite de pouvoir accomplir des tâches plus complexes comme: déterminer l'espace conformationnel effectif d'une boucle, identifier non pas un seul mais tous ses états conformationnels méta-stables, ou encore comprendre ses mouvements. Plus qu'une unique conformation stable, ces tâches requièrent une description précise et correcte du paysage énergétique de la boucle. Dans ce contexte, il apparait qu'une méthode d'évaluation doit être capable d'attribuer à chaque conformation de boucle un score qui soit en accord avec son énergie associée.

Les méthodes d'évaluation basées sur la physique (Section 1.3.3) estiment directement l'énergie potentielle de la boucle. Cependant, ces méthodes sont très coûteuses d'un point de vue computationnel, et on tendance à modéliser des paysages énergétiques trop rugueux pour être correctement interprétés. À l'inverse, les méth-

odes basées sur les statistiques (Section 1.3.3) sont plus rapides et modélisent en général des paysages plus lisses. Cependant, elles utilisent des données issues de cristallographie aux rayons X, qui ne fournit qu'une unique conformation pour une boucle protéique. Dans ce contexte, la capacité des potentiels statistiques à correctement évaluer les boucles flexibles et à identifier les conformations alternatives est incertaine.

Le travail présenté dans le Chapitre 4 compare la performance de plusieurs méthodes d'évaluation sur de multiples systèmes protéiques comprenant une boucle dont la flexibilité est connue. Plusieurs méthodes d'évaluation sont employées pour attribuer des scores aux différents états conformationnels de boucles issus d'ensemble exhaustifs échantillonnés avec MoMA-LoopSampler. Les résultats sont ensuite utilisés pour évaluer la capacité de ces méthodes à identifier une ou plusieurs des conformations connues. En s'appuyant sur l'idée mise en place dans la Section 3.3.1.4, des projections en 2D sont employées pour visualiser et analyser les paysages énergétiques implicitement modélisés. La cohérence de ces paysages avec les conformations de boucles connues est vérifiée et l'influence de la conformation de l'environnement protéique est détaillée.

En analysant les paysages énergétiques produits ainsi que la concordance entre les structures connues et les états conformationnels ayant les meilleurs scores, ce travail se propose t'identifier les différences qualitatives entre les résultats obtenus par les différentes méthodes d'évaluation. Ensuite, en examinant ces différences, nous espérons pouvoir fournir des indications quant aux méthodes à employer en fonction du problème à résoudre, et aux conditions qui doivent être respectées pour pouvoir espérer des résultats corrects. En particulier, on se propose en faisant cette comparaison de vérifier si le compromis entre coût computationnel et précision offert par les méthodes basées statistiques reste attractif.

Section 4.2 décrit le protocole *in silico* employé. Section 4.3 détaille les résultats obtenus par les différentes méthodes sur les différents systèmes protéiques, tandis que la Section 4.4 tente de résumer et de rassembler les différents résultats et ainsi de déterminer des tendances plus générales sur le fonctionnement des différentes méthodes d'évaluation.

À travers le Chapitre 4, la distinction est faite entre la *boucle*, ensemble physique d'atomes formant un fragment de protéine flexible; un *état* conformationnel, entièrement déterminé par les valeurs prises par ses degrés de liberté internes; et une *conformation* de boucle, définie comme un état consensus, ou comme un ensemble d'états similaires.

Au moment de la rédaction de ce manuscrit, ce travail est en révision par le journal *Proteins: Structure, Function and Bioinformatics* [Barozet 2019a].

A.4.2 Conclusions du Chapitre 4

Dans ce travail, nous avons testé la capacité des méthodes de pointe d'échantillonnage et d'évaluation de boucles à modéliser les boucles protéiques flexibles, pouvant adopter plusieurs conformations (méta)-stables. Cette anal-

yse montre que malgré les résultats assez encourageants obtenus lors des phases d'échantillonnage et d'évaluation, d'importantes avancées méthodologiques sont encore nécessaires pour pouvoir identifier les conformations les plus probables de manière fiable et précise.

Tout d'abord, le succès des méthodes d'échantillonnage est limité par leur difficulté à prendre efficacement en compte la flexibilité conformationnelle autour de la boucle et à correctement placer les chaînes latérales. En effet, quelle que soit la méthode employée pour la phase d'évaluation, les échafaudages structuraux employé pour modéliser la boucle se sont avérés déterminants dans la topographie des paysages implicites. L'intégration d'un composant gérant la flexibilité dans les méthodes d'échantillonnage de boucle constitue donc une piste intéressante pour le travail futur. Concernant les chaînes latérales, DFIRE2, qui prend en compte la position de tous les atomes et qui est parmi les méthodes d'évaluation les plus rapides, pourrait être employée pour optimiser le placement des chaînes latérales avant l'évaluation. Bien que la relaxation structurale ne soit pas nécessaire en théorie en amont de l'utilisation de cette méthode, l'optimisation locale des chaînes latérales générées par une stratégie de recherche globale (comme c'est le cas dans ce travail) pourrait améliorer les résultats de manière substantielle.

En ce qui concerne les méthodes d'évaluation, certaines peuvent identifier les états impossibles de manière fiable, et sont capables d'apporter des informations importantes concernant la topographie globale du paysage énergétique de la boucle. Cependant, les paysages modélisés restent souvent trop flous pour permettre de modéliser précisément l'espace conformationnel de la boucle. En outre, la plupart des méthodes d'évaluations donnent des résultats irréguliers d'une boucle à l'autre, si bien que leur performance sur un système inconnu reste trop imprévisible. En pratique, toutes ces observations suggèrent que les méthodes d'évaluation peuvent être employées pour des applications nécessitant un filtrage grossier et rapide des états conformationnels, mais que leurs résultats ne sont pas assez précis pour des applications comme le design de protéines. Plus précisément, la comparaison qualitative des méthodes d'évaluation pour la modélisation de boucles flexible présentée dans le Chapitre 4 valide l'utilisation des potentiels statistiques rapides tels que KORP ou DFIRE2 comme premiers filtres ou comme méthodes d'évaluation globale de la qualité pour de larges ensembles conformationnels de boucles. En effet, ces méthodes peuvent identifier les états proches des conformations statistiquement probables, en dépit d'une mauvaise géométrie locale ou de faibles collisions internes. Cependant, leur faible sensibilité aux changements conformationnels mineurs les empêche de fournir une évaluation plus précise. Pour de tels cas, les méthodes hybrides ou basées sur la physique semblent plus appropriées, pourvu que les relaxations structurales nécessaires soient correctement menées en amont.

Ensemble, les Chapitres 3 et 4 fournissent un protocole général pour analyser le paysage énergétique d'une boucle protéique flexible et des indications relatives à la fiabilité des différentes méthodes d'évaluation. Cependant, les tests ne furent appliqués qu'à des systèmes de référence, et pas à des anticorps, nos protéines d'intérêt. C'est pour cette raison que le Chapitre 5 se concentre sur l'application

du protocole complet à un anticorps de Sanofi présentant une boucle H3 flexible. Cette étude a pour but de vérifier que les conclusions précédentes s'appliquent également à un tel système. Les méthodes d'évaluation testées dans le Chapitre 4 seront également appliquées à ce système, afin de (1) vérifier leur performance sur des données non publiques et (2) d'analyser les différents paysages énergétiques à la lumière des résultats présentés dans le Chapitre 4.

A.5 Modélisation de la boucle H3 d'un anticorps de Sanofi

A.5.1 Présentation du Chapitre 5

Comme le Chapitre 2 le révèle, la flexibilité des boucles CDR s'avère un élément crucial dans la compréhension des mécanismes liés aux anticorps. Cependant, comme cela est souligné dans ce chapitre, la modélisation précise le la plasticité des boucles CDR reste un problème ouvert, pour lequel des méthodes efficaces sont toujours recherchées. Afin de combler ce manque, le Chapitre 3 présente une méthode pour échantillonner des ensembles conformationnels de boucle variés et de qualité, tandis que le Chapitre 4 se concentre sur l'évaluation précise des conformations générées, afin de modéliser des paysages énergétiques intéressants. Cependant, les méthodes presentées dans les Chapitres 3 et 4 ne sont pas spécifiques aux anticorps et leur performance sur les boucles CDR n'ont pas encore été démontrées. Le Chapitre 5 se propose donc de tester ces méthodes sur un anticorps dont la boucle H3 est flexible.

Dans le cadre d'un projet d'humanisation et de maturation artificielle d'un anticorps à Sanofi, plusieurs structures du fragment Fab furent obtenues sous différentes conditions (libre, lié, humanisé, maturé, ...). Ces structures révèlent de la flexibilité dans le boucle CDR H3, qui adopte deux conformations distinctes: une conformation *apo* et une conformation *holo* (identiques pour toutes les structures liées de l'anticorps).

Les différentes structures nous fournissent un cadre idéal pour tester les méthodes présentées dans cette thèse. Bien que l'humanisation et la maturation ne se soient pas concentrées sur la boucle H3, les conditions nécessaires pour suivre un protocole similaire à celui du Chapitre 4 sont rassemblées:

- Plusieurs conformations de la boucle sont connues
- Plusieurs structures de Fab pouvant être employées comme échafaudage pour échantillonner la boucle H3 existent.

La Section 5.2 détaille les structures employées dans ce travail, et résume les méthodes employées pour modéliser le paysage énergétique. La Section 5.3 décrit les résultats des différentes méthodes d'évaluation, et les paysages énergétiques modélisés, de manière analogue à la Section 4.3. La Section 5.4 discute ces résultats.

Enfin, la Section 5.5 conclut sur l'applicabilité des méthodes générales de modélisation de boucles développées au cours de cette thèse au cas particulier des boucles CDR chez les anticorps.

Un manuscrit décrivant ce travail est en cours de rédaction au moment de la rédaction de ce manuscrit.

A.5.2 Conclusions du Chapitre 5

Les méthodes des Chapitres 3 et 4 furent appliquées avec succès à la boucle H3 flexible de l'anticorps d'intérêt du Chapitre 5. Les paysages produits sont à la fois informatifs et cohérents avec les données expérimentales. Bien que générales, les méthodes de modélisation de boucles développées ou étudiées dans cette thèse sont adaptées à la modélisation de boucle H3 d'anticorps, qui constituait la motivation initiale de ces développements méthodologiques. La généralité de ces méthodes les rend encore plus susceptibles de réussir à modéliser les boucles CDRs pour lesquelles aucune classe canonique ne peut être attribuée, ou les boucles CDRs issues de design et qui ne ressemblent à aucune boucle hypervariable existante.

Ce travail a confirmé les conclusions tirées dans le Chapitre 4 concernant les méthodes d'évaluation les plus fiables (DFIRE2, mais aussi KORP et AMBER), l'importance de l'échafaudage de départ, (et indirectement, de la nécessité de la prise en compte de la flexibilité lors de l'échantillonnage), et le besoin en méthodes fiables et rapides de placement des chaînes latérales. En outre, ce travail a révélé une autre application possible des méthodes développées dans cette thèse: vérifier que les conformations de boucles fournies par les structures cristallographiques ne sont pas des artefacts du processus de cristallisation.

Comme cela est mentionné dans la Section 1.3.2.4, la modélisation de boucles pour les anticorps pourrait voir ses résultats améliorés par l'emploi de fragments d'anticorps pour l'échantillonnage. Une manière de rendre MoMA-LoopSampler plus spécifique au cas des anticorps serait donc d'utiliser une base de données de tripeptides extraits de structures d'anticorps publiques, bien que le manque potentiel de données soit susceptibles d'empêcher un échantillonnage réellement exhaustif de l'espace conformationnel. Cela constitue une piste de travail intéressante.

Conclusion

Cette thèse aborde le problème de mieux représenter la flexibilité chez les anticorps à travers le développement et l'évaluation de méthodes efficaces pour la modélisation des boucles protéiques. Tout d'abord, une analyse des facteurs d'échec des algorithmes d'amarrage appliqués aux complexes anticorps-antigènes fut menée. Cette étude commença par l'analyse des changements conformationnels survenant lors de la liaison chez 27 anticorps. Diverses métriques furent employées, permettant une classification détaillée des différentes classes de changements conformationnels, ainsi que de leur combinaison. La performance de quatre algorithmes d'amarrage appliqués à ces systèmes anticorps-antigène fut ensuite analysée à la lumière de la

connaissance de ces changements conformationnels. Cette analyse conclut que les algorithmes d'amarrage était principalement négativement impactés par la survenue de larges changements conformationnels dans les boucles CDR situées au niveau du site de liaison de l'antigène, bien que d'autres sources d'échecs pour ces algorithmes furent aussi avancées. Le travail de thèse fut dont orienté vers le développement de méthodes de modélisation de boucles qui représenteraient et prendraient mieux en compte la flexibilité conformationnelle.

Par conséquent, une méthode générale pour échantillonner l'espace conformationnel des boucles protéiques d'une manière exhaustive fut ensuite développée. Cette méthode d'échantillonnage emploie une base de données de fragments de boucles, et respecte des contraintes strictes en terme de collisions stériques. Ces deux facteurs contribuent à la génération d'ensembles de qualité élevée, en éliminant un certain nombre de conformations de boucles improbables. En outre, en plus de la qualité des ensembles générés, la diversité constitue un élément essentiel de cette méthode. En effet, la diversité de l'échantillonnage est cruciale pour obtenir une représentation correcte de la flexibilité d'une boucle. Par conséquent, vérifier en profondeur l'exhaustivité des ensembles générés constitua une part importante du travail de validation méthodologique. Globalement, il fut prouvé que la méthode produit des ensembles conformationnels mieux filtrés que ceux générés par d'autres méthodes récentes, tout en maintenant une couverture suffisante de l'espace conformationnel. Une variante employant une heuristique d'apprentissage par renforcement fut également proposée, afin d'accélérer l'échantillonnage. Cette variante produit des ensembles plus grands, avec une plus grande résolution et sans perte de diversité conformationnelle, sous réserve qu'une paramétrisation adaptée soit employée.

La troisième partie de cette thèse décrit un protocole qui peut être utilisé pour modéliser des paysages pertinents de boucles protéiques flexibles. Après un échantillonnage exhaustif effectué à l'aide de la méthode précédemment décrite, cette procédure requiert un composant d'évaluation fiable, pour lequel plusieurs options furent comparées dans ce travail. Plusieurs méthodes récentes d'évaluation de boucles furent intégrées à ce protocole, afin de produire des visualisations 2D des paysages énergétiques de huit boucles protéiques flexibles. Ces paysages énergétiques furent ensuite comparés, en fonction de leur informativité et de leur cohérence avec les données expérimentales. Les résultats indiquèrent que des informations utiles peuvent être tirées des paysages ainsi produits, concernant par exemple la flexibilité de la boucle, les conformations (meta-)stables, ... De plus, cette analyse fournit des indications quant aux méthodes d'évaluations appropriées aux différentes applications possibles.

Enfin, les méthodes générales développées au cours de cette thèse furent utilisées pour modéliser une boucle H3 flexible. Les résultats validèrent l'application de cette nouvelles méthodes aux boucles hypervariables d'anticorps, notre système d'intérêt de départ. Les paysages énergétiques modélisés identifièrent à la fois les conformations *apo* et *holo* de la boucle comme des conformations stables. Des

paysages énergétiques aussi précis peuvent s'avérer précieux dans le contexte du design d'anticorps.

Recherche future

Ce travail révèle de multiples directions pour la recherche future. Pour commencer, la méthode d'échantillonnage pourrait être améliorée de multiples façons. Par exemple, en incluant des données structurales issues de séquences similaires pour les tripeptides dont la séquence est rare, ou à travers le développement d'autres stratégies d'apprentissage par renforcement avec un coût moindre en mémoire.

Une autre manière d'améliorer la méthode d'échantillonnage serait de détourner le processus d'apprentissage de son utilisation initiale, en ajustant les scores des feuilles dans l'arbre d'apprentissage afin d'influencer la distribution de l'échantillonnage. Par exemple, les scores pourraient être réglés de manière à ce que les tripeptides employés pour former l'ensemble conformationnel généré suivent la distribution de la base de donnée. Cependant, cela n'a du sens que si la distribution dans la base de données a un sens en elle-même: par exemple, si elle représente la propension implicite des tripeptides à adopter les états correspondants lorsqu'ils se trouvent dans une boucle. Si une distribution statistique pertinente peut être garantie dans la base de données, et si la stratégie d'échantillonnage est adaptée pour suivre cette distribution, l'ensemble générer pourrait alors être employé pour estimer des propriétés thermodynamiques de la boucle. Cependant, cela nécessiterait de régler les fréquences des tripeptides 'synthétiques' (ceux obtenus après concaténation et pas directement extraits de la base de données), ce qui signifierait que la stratégie consistant à concaténer les tripeptides devrait être modifiée. À la place, une stratégie consistant à superposer les tripeptides (en utilisant une fenêtre glissante de trois résidus) pourrait être adoptée.

Des solutions pourraient aussi être cherchées pour un meilleur traitement de la flexibilité dans la structure 'échafaudage' employée pour l'échantillonnage. La méthode d'échantillonnage développée dans cette thèse ne permet la flexibilité dans les chaînes latérales entourant la boucle qu'au moment du placement des chaînes latérales de la boucle. Cependant, la nécessité de prendre en compte la flexibilité plus tôt dans le processus d'échantillonnage fut mis en évidence à plusieurs reprises dans ce travail. Idéalement, la modélisation de la flexibilité devrait être étendue au squelette environnant, bien que cela constitue un défi autrement plus difficile.

Le développement de méthodes de placement de chaînes latérales à la fois justes et rapides constitue une autre direction pour le travail futur. Des méthodes récentes comme SCWRL4 sont rapides, mais ont tendance à renvoyer des résultats de piètre qualité plein de collisions dans les environnements très contraints. A l'inverse, notre méthode fournit des résultats sans collision, mais est hautement chronophage. En outre, elle n'inclut aucune considération énergétique et retourne la première solution sans collision. Intégrer une courte optimisation utilisant une méthode d'évaluation rapide comme DFIRE2 constitue une option intéressante pour améliorer la qualité du placement. De plus, au lieu de modifier aléatoirement les angles dièdres pour

corriger les collisions observées, des méthodes géométriques plus sophistiquées pourraient être envisagées, représentant un gain de temps considérable et augmentant le taux de succès des cette étape de placement des chaînes latérales.

L'analyse des paysages énergétiques en plus haute dimension pourrait également fournir des informations intéressantes sur la structure et la dynamique des boucles protéiques. La détection automatisée des bassins énergétiques (en employant par exemple des méthodes de partitionnement intégrant des considérations énergétiques) ou des régions de transitions pourrait être plus fiable que la projection de ces paysages énergétiques en 2D, bien que l'interprétation puisse s'avérer plus ardue. L'application d'algorithmes existants ou le développement de nouvelles méthodes pour la modélisation de mouvements de boucles, en utilisant soit les résultats combinés des étapes d'échantillonnage et d'évaluation, soit les paysages énergétiques reconstitués, constitue également une direction intéressante pour la recherche future. Finalement, des applications intéressantes du processus proposé dans le Chapitre 4 pourraient être étudiées. D'abord, ce protocole pourrait être utilisé pour étudier les effets de mutations sur la conformation d'une boucle, afin par exemple d'exposer la base structurale d'une maladie génétique. D'autres applications possibles du processus existent, comme la validation de séquences dans le contexte du design de boucles (en particulier dans le cas du design négatif).

Bibliography

[Al-Bluwi 2012] Ibrahim Al-Bluwi, Thierry Siméon and Juan Cortés. *Motion planning algorithms for molecular simulations: A survey.* Comput. Sci. Rev., vol. 6, no. 4, pages 125–143, 2012. (Cited in page 29.)

[Al-Lazikani 1997] Bissan Al-Lazikani, Arthur M Lesk and Cyrus Chothia. *Standard conformations for the canonical structures of immunoglobulins.* J. Mol. Biol., vol. 273, no. 4, pages 927–948, 1997. (Cited in pages 12 and 46.)

[Alford 2017] Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme and Jeffrey J. Gray. *The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design.* J. Chem. Theory Comput., vol. 13, no. 6, pages 3031–3048, 2017. (Cited in pages 25 and 111.)

[Almagro 2004] Juan C. Almagro. *Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires.* J. Mol. Recognit., vol. 17, no. 2, pages 132–143, 2004. (Cited in pages 16 and 17.)

[Almagro 2014] Juan C. Almagro, Alexey Teplyakov, Jinquan Luo, Raymond W. Sweet, Sreekumar Kodangattil, Francisco Hernandez-Guzman and Gary L. Gilliland. *Second antibody modeling assessment (AMA-II).* Proteins, vol. 82, no. 8, pages 1553–1562, 2014. (Cited in page 15.)

[Altis 2007] A. Altis, P. H. Nguyen, R. Hegger and G. Stock. *Dihedral angle principal component analysis of molecular dynamics simulations.* J. Chem. Phys., vol. 126, no. 24, page 244111, 2007. (Cited in page 29.)

[Amaro 2018] Rommie E. Amaro, Jerome Baudry, John Chodera, Özlem Demir, J. Andrew McCammon, Yinglong Miao and Jeremy C. Smith. *Ensemble Docking in Drug Discovery.* Biophys. J., vol. 114, no. 10, pages 2271–2278, 2018. (Cited in page 25.)

[Anfinsen 1973] C. B. Anfinsen. *Principles That Govern the Folding of Protein Chains.* Science, vol. 181, no. 4096, pages 223–230, 1973. (Cited in page 26.)

[Babor 2011] Mariana Babor, Daniel J. Mandell and Tanja Kortemme. *Assessment of flexible backbone protein design methods for sequence library prediction in the therapeutic antibody Herceptin-HER2 interface: Flexible Backbone Protein Design Methods.* Protein Sci., vol. 20, no. 6, pages 1082–1089, 2011. (Cited in page 19.)

[Barozet 2018] Amélie Barozet, Marc Bianciotto, Thierry Siméon, Hervé Minoux and Juan Cortés. *Conformational changes in antibody Fab fragments upon binding and their consequences on the performance of docking algorithms.* Immunol. Lett., vol. 200, pages 5–15, 2018. (Cited in pages 3, 32, 166, and 168.)

[Barozet 2019a] Amélie Barozet, Marc Bianciotto, Marc Vaisset, Siméon Thierry, Hervé Minoux and Juan Cortés. *Protein loops with multiple meta-stable conformations: a challenge for sampling and scoring methods.* 2019. Submitted. (Cited in pages 3, 110, and 173.)

[Barozet 2019b] Amélie Barozet, Kevin Molloy, Marc Vaisset, Thierry Siméon and Juan Cortés. *A reinforcement-learning-based approach to enhance exhaustive protein loop sampling.* Bioinformatics, 2019. In press. (Cited in pages 3, 62, 166, and 171.)

[Becker 1997] Oren M Becker and Martin Karplus. *The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics.* J. Chem. Phys., vol. 106, no. 4, pages 1495–1517, 1997. (Cited in page 29.)

[Berman 2000] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov and Philip E. Bourne. *The Protein Data Bank.* Nucleic Acids Res., vol. 28, no. 1, pages 235–242, 2000. (Cited in page 7.)

[Bondi 1964] A. Bondi. *van der Waals Volumes and Radii.* J. Phys. Chem., vol. 68, no. 3, pages 441–451, 1964. (Cited in page 87.)

[Bonvin 2006] Alexandre MJJ Bonvin. *Flexible protein-protein docking.* Curr. Opin. Struct. Biol., vol. 16, no. 2, pages 194–200, 2006. (Cited in page 19.)

[Brändén 1999] Carl-Ivar Brändén and John Tooze. Introduction to Protein Structure. Taylor & Francis, 1999. (Cited in pages 7 and 36.)

[Brandt 2008] Bernd W. Brandt, Jaap Heringa and Jack A. M. Leunissen. *SEQATOMS: a Web Tool for Identifying Missing Regions in PDB in Sequence Context.* Nucleic Acids Res., vol. 36, no. suppl_2, pages W255–W259, 2008. (Cited in page 20.)

[Brooks 1983] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan and Martin Karplus. *CHARMM: A program for macromolecular energy, minimization, and dynamics calculations.* J. Comput. Chem., vol. 4, no. 2, pages 187–217, 1983. (Cited in page 28.)

[Cahill 2003] Sean Cahill, Michael Cahill and Kevin Cahill. *On the Kinematics of Protein Folding.* J. Comput. Chem., vol. 24, no. 11, pages 1364–1370, 2003. (Cited in page 89.)

[Canutescu 2003] Adrian A. Canutescu and Roland L. Dunbrack. *Cyclic coordinate descent: A robotics algorithm for protein loop closure.* Protein Sci., vol. 12, no. 5, pages 963–972, 2003. (Cited in pages 21 and 89.)

[Case 2005] David A. Case, Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang and Robert J. Woods. *The Amber biomolecular simulation programs.* J. Comput. Chem., vol. 26, no. 16, pages 1668–1688, 2005. (Cited in pages 75, 95, 116, 117, and 118.)

[Case 2016] D.A. Case, R.M. Betz, D.S. Cerutti, III T.E. Cheatham T.A. Darden, R.E. Duke, T.J. Giese, H. Gohlke, A.W. Goetz, N. Homeyer, S. Izadi, P. Janowski, J. Kaus, T.S. Lee A. Kovalenko, S. LeGrand, P. Li, C.Lin, T. Luchko, R. Luo, B. Madej, D. Mermelstein, K.M. Merz, G. Monard, H. Nguyen, H.T. Nguyen, I.Omelyan, A. Onufriev, D.R. Roe, A. Roitberg, C. Sagui, C.L. Simmerling, W.M. Botello-Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, L. Xiao and P.A. Kollman. *AMBER 2016, University of California, San Francisco*, 2016. (Cited in pages 75, 116, 117, and 118.)

[Cavanagh 2006] J. Cavanagh, W. J. Fairbrother, A. G. Palmer and N. J. Skleton. Protein NMR spectroscopy: principles and practices. Royal Society of Chemistry, 2006. (Cited in page 20.)

[Cazals 2015] Frédéric Cazals, Tom Dreyfus, Dorian Mazauric, Christine-Andrea Roth and Charles H Robert. *Conformational ensembles and sampled energy landscapes: Analysis and comparison.* J. Comp. Chem., vol. 36, no. 16, pages 1213–1231, 2015. (Cited in page 29.)

[Chames 2009] Patrick Chames, Marc Van Regenmortel, Etienne Weiss and Daniel Baty. *Therapeutic antibodies: successes, limitations and hopes for the future.* Br. J. Pharmacol., vol. 157, no. 2, pages 220–233, 2009. (Cited in page 10.)

[Chang 2007] Chia-en A. Chang, Wei Chen and Michael K. Gilson. *Ligand configurational entropy and protein binding.* Proc. Natl. Acad. Sci. U.S.A., vol. 104, no. 5, pages 1534–1539, 2007. (Cited in page 28.)

[Chen 2003a] Rong Chen, Li Li and Zhiping Weng. *ZDOCK: An initial-stage protein-docking algorithm.* Proteins, vol. 52, no. 1, pages 80–87, 2003. (Cited in page 54.)

[Chen 2003b] Rong Chen and Zhiping Weng. *A novel shape complementarity scoring function for protein-protein docking.* Proteins, vol. 51, no. 3, pages 397–408, 2003. (Cited in page 54.)

[Cheng 2007] Tammy Man-Kuang Cheng, Tom L. Blundell and Juan Fernandez-Recio. *pyDock: electrostatics and desolvation for effective scoring of rigid-*

body protein-protein docking. Proteins, vol. 68, no. 2, pages 503–515, 2007. (Cited in page 54.)

[Choi 2010] Yoonjoo Choi and Charlotte M. Deane. *FREAD revisited: Accurate loop structure prediction using a database search algorithm.* Proteins, vol. 78, no. 6, pages 1431–1440, 2010. (Cited in page 23.)

[Choi 2011] Yoonjoo Choi and Charlotte M. Deane. *Predicting antibody complementarity determining region structures without classification.* Mol. Biosyst., vol. 7, no. 12, page 3327, 2011. (Cited in pages 23 and 24.)

[Chothia 1987] Cyrus Chothia and Arthur M. Lesk. *Canonical structures for the hypervariable regions of immunoglobulins.* J. Mol. Biol., vol. 196, no. 4, pages 901–917, 1987. (Cited in pages 11, 12, 35, and 46.)

[Chothia 1989] Cyrus Chothia, Arthur M. Lesk, Anna Tramontano, Michael Levitt, Sandra J. Smith-Gill, Gillian Air, Steven Sheriff, Eduardo A. Padlan, David Davies, William R. Tulip, Peter M. Colman, Silvia Spinelli, Pedro M. Alzari and Roberto J. Poljak. *Conformations of immunoglobulin hypervariable regions.* Nature, vol. 342, no. 6252, pages 877–883, 1989. (Cited in pages 12, 35, and 46.)

[Chothia 1992] C. Chothia, A. M. Lesk, E. Gherardi, I. M. Tomlinson, G. Walter, J. D. Marks, M. B. Llewelyn and G. Winter. *Structural repertoire of the human VH segments.* J. Mol. Biol., vol. 227, no. 3, pages 799–817, 1992. (Cited in page 13.)

[Chys 2013] Pieter Chys and Pablo Chacón. *Random Coordinate Descent with Spinor-matrices and Geometric Filters for Efficient Loop Closure.* J. Chem. Theory Comput., vol. 9, no. 3, pages 1821–1829, 2013. (Cited in pages 23 and 89.)

[Clementi 2008] Cecilia Clementi. *Coarse-grained models of protein folding: toy models or predictive tools?* Curr. Opin. Struct. Biol., vol. 18, no. 1, pages 10–15, 2008. (Cited in page 27.)

[Connolly 1983] M. L. Connolly. *Solvent-accessible surfaces of proteins and nucleic acids.* Science, vol. 221, no. 4612, pages 709–713, 1983. (Cited in page 18.)

[Cortés 2004] J. Cortés, T. Siméon, M. Remaud-Siméon and V. Tran. *Geometric algorithms for the conformational analysis of long protein loops.* J. Comput. Chem., vol. 25, no. 7, pages 956–967, 2004. (Cited in pages 22 and 66.)

[Cortés 2005] J. Cortés, T. Siméon, V. Ruiz de Angulo, D. Guieysse, M. Remaud-Siméon and V. Tran. *A path planning approach for computing large-amplitude motions of flexible molecules.* Bioinformatics, vol. 21 Suppl 1, pages i116–125, 2005. (Cited in page 29.)

[Coutsias 2004] Evangelos A. Coutsias, Chaok Seok, Matthew P. Jacobson and Ken A. Dill. *A kinematic view of loop closure.* J. Comput. Chem., vol. 25, no. 4, pages 510–528, 2004. (Cited in pages 22 and 67.)

[Creighton 1993] Thomas E Creighton. Proteins: structures and molecular properties. Macmillan, 1993. (Cited in page 6.)

[DePristo 2003] Mark A. DePristo, Paul I. W. de Bakker, Simon C. Lovell and Tom L. Blundell. *Ab Initio Construction of Polypeptide Fragments: Efficient Generation of Accurate, Representative Ensembles.* Proteins, vol. 51, no. 1, pages 41–55, 2003. (Cited in page 85.)

[Devaurs 2014] Didier Devaurs, Thierry Siméon and Juan Cortés. *Efficient sampling-based approaches to optimal path planning in complex cost spaces.* In International Workshop on the Algorithmic Foundations of Robotics (WAFR), page 16, 2014. (Cited in page 29.)

[Dill 1997] Ken A. Dill and Hue Sun Chan. *From Levinthal to pathways to funnels.* Nat. Struct. Biol., vol. 4, no. 1, pages 10–19, 1997. (Cited in page 26.)

[Dinner 2000] Aaron R. Dinner. *Local Deformations of Polymers With Nonplanar Rigid Main-Chain Internal Coordinates.* J. Comput. Chem., vol. 21, no. 13, pages 1132–1144, 2000. (Cited in pages 22 and 67.)

[Dominguez 2003] Cyril Dominguez, Rolf Boelens and Alexandre M. J. J. Bonvin. *HADDOCK: A Protein–Protein Docking Approach Based on Biochemical or Biophysical Information.* J. Am. Chem. Soc., vol. 125, no. 7, pages 1731–1737, 2003. (Cited in page 54.)

[Dong 2013] Guang Qiang Dong, Hao Fan, Dina Schneidman-Duhovny, Ben Webb and Andrej Sali. *Optimized atomic statistical potentials: assessment of protein interfaces and loops.* Bioinformatics, vol. 29, no. 24, pages 3158–3166, 2013. (Cited in pages 25 and 111.)

[Dr. Andrew C.R. Martin's Group at UCL 1995] Dr. Andrew C.R. Martin's Group at UCL. Canonicals - Chothia Canonical Assignment. `http://www.bioinf.org.uk/abs/chothia.html`, 1995. (Cited in page 35.)

[Dunbar 2016] James Dunbar and Charlotte M. Deane. *ANARCI: antigen receptor numbering and receptor classification.* Bioinformatics, vol. 32, no. 2, pages 298–300, 2016. (Cited in page 32.)

[Engh 1991] R. A. Engh and R. Huber. *Accurate Bond and Angle Parameters for X-Ray Protein Structure Refinement.* Acta Crystallogr., Sect. A.: Found. Adv., vol. 47, no. 4, pages 392–400, 1991. (Cited in page 63.)

[Fiser 2000] András Fiser, Richard Kinh Gian Do and Andrej Šali. *Modeling of loops in protein structures.* Protein Sci., vol. 9, no. 9, pages 1753–1773, 2000. (Cited in page 20.)

[Fox 2014] Naomi K. Fox, Steven E. Brenner and John-Marc Chandonia. *SCOPe: Structural Classification of Proteins—Extended, Integrating SCOP and ASTRAL Data and Classification of New Structures.* Nucleic Acids Res., vol. 42, no. D1, pages D304–D309, 2014. (Cited in page 64.)

[Frenkel 2002] D. Frenkel and B. Smit. Understanding Molecular Simulations: From Algorithms to Applications. Academic Press, 2002. (Cited in page 26.)

[Goding 1996] James W Goding. Monoclonal antibodies: principles and practice. Elsevier, 1996. (Cited in page 9.)

[Gray 2003] Jeffrey J. Gray, Stewart Moughon, Chu Wang, Ora Schueler-Furman, Brian Kuhlman, Carol A. Rohl and David Baker. *Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations.* J. Mol. Biol., vol. 331, no. 1, pages 281–299, 2003. (Cited in page 19.)

[Greally 1991] J. M. Greally. *The physiology of anti-idiotypic interactions: from clonal to paratopic selection.* Clin. Immunol. Immunopathol., vol. 60, no. 1, pages 1–12, 1991. (Cited in page 16.)

[Gu 2009] Jenny Gu and Philip E Bourne. Structural bioinformatics, volume 44. John Wiley & Sons, 2009. (Cited in page 6.)

[Guddat 1994] Luke W. Guddat, Lin Shan, Jerry M. Anchin, D. Scott Linthicum and Allen B. Edmundson. *Local and Transmitted Conformational Changes on Complexation of an Anti-sweetener Fab.* J. Mol. Biol., vol. 236, no. 1, pages 247–274, 1994. (Cited in page 10.)

[Halperin 2002] Inbal Halperin, Buyong Ma, Haim Wolfson and Ruth Nussinov. *Principles of docking: An overview of search algorithms and a guide to scoring functions.* Proteins, vol. 47, no. 4, pages 409–443, 2002. (Cited in page 17.)

[Harder 2010] Tim Harder, Wouter Boomsma, Martin Paluszewski, Jes Frellsen, Kristoffer E. Johansson and Thomas Hamelryck. *Beyond Rotamers: a Generative, Probabilistic Model of Side Chains in Proteins.* BMC Bioinf., vol. 11, no. 1, page 306, 2010. (Cited in pages 9, 94, and 117.)

[Haspel 2010] Nurit Haspel, Mark Moll, Matthew L. Baker, Wah Chiu and Lydia E. Kavraki. *Tracing conformational changes in proteins.* BMC Struct. Biol., vol. 10 Suppl 1, page S1, 2010. (Cited in page 29.)

[Hastings 1970] W. K. Hastings. *Monte Carlo sampling methods using Markov chains and their applications.* Biometrika, vol. 57, no. 1, pages 97–109, 1970. (Cited in page 28.)

[Herges 2004] T. Herges and W. Wenzel. *An All-Atom Force Field for Tertiary Structure Prediction of Helical Proteins.* Biophys. J., vol. 87, no. 5, pages 3100–3109, 2004. (Cited in page 27.)

[Herron 1991] J. N. Herron, X. M. He, D. W. Ballard, P. R. Blier, P. E. Pace, A. L. M. Bothwell, E. W. Voss and A. B. Edmundson. *An autoantibody to single-stranded DNA: Comparison of the three-dimensional structures of the unliganded fab and a deoxynucleotide–fab complex.* Proteins, vol. 11, no. 3, pages 159–175, 1991. (Cited in page 10.)

[Hu 2007] Xiaozhen Hu, Huanchen Wang, Hengming Ke and Brian Kuhlman. *High-resolution design of a protein loop.* Proc. Natl. Acad. Sci. U.S.A., vol. 104, no. 45, pages 17668–17673, 2007. (Cited in page 25.)

[Iannuzzi 2003] Marcella Iannuzzi, Alessandro Laio and Michele Parrinello. *Efficient exploration of reactive potential energy surfaces using Car-Parrinello molecular dynamics.* Phys. Rev. Lett., vol. 90, no. 23, page 238302, 2003. (Cited in page 28.)

[Jackson 1999] R. M. Jackson. *Comparison of protein-protein interactions in serine protease-inhibitor and antibody-antigen complexes: implications for the protein docking problem.* Protein Sci., vol. 8, no. 3, pages 603–613, 1999. (Cited in page 17.)

[Jacobsen 2007] N. E. Jacobsen. NMR spectroscopy explained: simplified theory, applications and examples for organic chemistry and structural biology. Wiley-Interscience, 2007. (Cited in page 20.)

[Jacobson 2004] Matthew P. Jacobson, David L. Pincus, Chaya S. Rapp, Tyler J. F. Day, Barry Honig, David E. Shaw and Richard A. Friesner. *A hierarchical approach to all-atom protein loop prediction.* Proteins, vol. 55, no. 2, pages 351–367, 2004. (Cited in pages 23, 85, and 89.)

[Jaillet 2011] Léonard Jaillet, Francesc J. Corcho, Juan-Jesús Pérez and Juan Cortés. *Randomized tree construction algorithm to explore energy landscapes.* J. Comput. Chem., vol. 32, no. 16, pages 3464–3474, 2011. (Cited in page 29.)

[James 2003] Leo C. James, Pietro Roversi and Dan S. Tawfik. *Antibody multispecificity mediated by conformational diversity.* Science, vol. 299, no. 5611, pages 1362–1367, 2003. (Cited in page 14.)

[Jin 2003] Wenzhen Jin, Ohki Kambara, Hiroaki Sasakawa, Atsuo Tamura and Shoji Takada. *De Novo Design of Foldable Proteins with Smooth Folding Funnel: Automated Negative Design and Experimental Verification.* Structure, vol. 11, no. 5, pages 581–590, 2003. (Cited in page 25.)

[Jolliffe 2002] I. T. Jolliffe. Principal component analysis. Springer Verlag, 2002. (Cited in page 29.)

[Kabat 1977] E. A. Kabat, T. T. Wu and H. Bilofsky. *Unusual distributions of amino acids in complementarity-determining (hypervariable) segments of heavy and light chains of immunoglobulins and their possible roles in specificity of antibody-combining sites.* J. Biol. Chem., vol. 252, no. 19, pages 6609–6616, 1977. (Cited in page 11.)

[Kabsch 1983] W. Kabsch and C. Sander. *Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features.* Biopolymers, vol. 22, no. 12, pages 2577–2637, 1983. (Cited in page 64.)

[Kaiser 2007] Chris A Kaiser, Monty Krieger, Harvey Lodish and Arnold Berk. Molecular cell biology. WH Freeman, 2007. (Cited in page 6.)

[Karami 2018] Yasaman Karami, Frédéric Guyon, Sjoerd De Vries and Pierre Tufféry. *DaReUS-Loop: Accurate Loop Modeling Using Fragments From Remote or Unrelated Proteins.* Sci. Rep., vol. 8, no. 1, page 13673, 2018. (Cited in page 23.)

[Karasikov 2018] Mikhail Karasikov, Guillaume Pagès and Sergei Grudinin. *Smooth orientation-dependent scoring function for coarse-grained protein quality assessment.* Bioinformatics, 2018. (Cited in pages 25 and 116.)

[Katchalski-Katzir 1992] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo and I. A. Vakser. *Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques.* Proc. Natl. Acad. Sci. U.S.A., vol. 89, no. 6, pages 2195–2199, 1992. (Cited in page 18.)

[Knapp 2017] Bernhard Knapp, James Dunbar, Marta Alcala and Charlotte M. Deane. *Variable Regions of Antibodies and T-Cell Receptors May Not Be Sufficient in Molecular Simulations Investigating Binding.* J. Chem. Theory Comput., vol. 13, no. 7, pages 3097–3105, 2017. (Cited in page 56.)

[Koga 2012] Nobuyasu Koga, Rie Tatsumi-Koga, Gaohua Liu, Rong Xiao, Thomas B. Acton, Gaetano T. Montelione and David Baker. *Principles for designing ideal protein structures.* Nature, vol. 491, no. 7423, pages 222–227, 2012. (Cited in page 25.)

[Krivov 2009] Georgii G. Krivov, Maxim V. Shapovalov and Roland L. Dunbrack. *Improved prediction of protein side-chain conformations with SCWRL4.* Proteins, vol. 77, no. 4, pages 778–795, 2009. (Cited in pages 9 and 137.)

[Krumrine 2003] Jennifer Krumrine, Florian Raubacher, Natasja Brooijmans and Irwin Kuntz. *Principles and Methods of Docking and Ligand Design.* In

Structural Bioinformatics, pages 441–476. John Wiley & Sons, Inc., 2003. (Cited in page 18.)

[Kundert 2019] Kale Kundert and Tanja Kortemme. *Computational design of structured loops for new protein functions.* Biol. Chem., vol. 400, no. 3, pages 275–288, 2019. (Cited in pages 25, 110, and 172.)

[Kunik 2012] Vered Kunik, Bjoern Peters and Yanay Ofran. *Structural consensus among antibodies defines the antigen binding site.* PLoS Comput. Biol., vol. 8, no. 2, page e1002388, 2012. (Cited in page 16.)

[Kunik 2013] Vered Kunik and Yanay Ofran. *The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops.* Protein Eng. Des. Sel., vol. 26, no. 10, pages 599–609, 2013. (Cited in pages 16 and 17.)

[Laio 2002] Alessandro Laio and Michele Parrinello. *Escaping free-energy minima.* Proc. Natl. Acad. Sci. U.S.A., vol. 99, no. 20, pages 12562–12566, 2002. (Cited in page 28.)

[Landau 2005] D. P. Landau and K. Binder. A guide to Monte Carlo simulations in statistical physics. Cambridge University Press, 2005. (Cited in page 28.)

[Le Trong 2006] Isolde Le Trong, Dimitri G. L. Aubert, Neil R. Thomas and Ronald E. Stenkamp. *The High-Resolution Structure of (+)-Epi-Biotin Bound to Streptavidin.* Acta Crystallogr., Sect. D: Biol. Crystallogr., vol. 62, no. Pt 6, pages 576–581, 2006. (Cited in page 94.)

[Le Trong 2011] Isolde Le Trong, Zhizhi Wang, David E. Hyre, Terry P. Lybrand, Patrick S. Stayton and Ronald E. Stenkamp. *Streptavidin and its Biotin Complex at Atomic Resolution.* Acta Crystallogr., Sect. D: Biol. Crystallogr., vol. 67, no. Pt 9, pages 813–821, 2011. (Cited in page 94.)

[Leach 2001] A. R. Leach. Molecular modelling: Principles and applications. Pearson Education, 2001. (Cited in page 26.)

[Lefranc 2003] Marie-Paule Lefranc, Christelle Pommié, Manuel Ruiz, Véronique Giudicelli, Elodie Foulquier, Lisa Truong, Valérie Thouvenin-Contet and Gérard Lefranc. *IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains.* Dev. Comp. Immunol., vol. 27, no. 1, pages 55–77, 2003. (Cited in page 11.)

[Levitt 1976] Michael Levitt. *A Simplified Representation of Protein Conformations for Rapid Simulation of Protein Folding.* J. Mol. Biol., vol. 104, no. 1, pages 59–107, 1976. (Cited in page 66.)

[Li 2000] Yili Li, Hongmin Li, Sandra J. Smith-Gill and Roy A. Mariuzza. *Three-Dimensional Structures of the Free and Antigen-Bound Fab from Monoclonal*

Antilysozyme Antibody HyHEL-63. Biochemistry, vol. 39, no. 21, pages 6296–6309, 2000. (Cited in page 10.)

[Li 2010] Xiaofan Li, Iain H. Moal and Paul A. Bates. *Detection and refinement of encounter complexes for protein–protein docking: Taking account of macromolecular crowding.* Proteins, vol. 78, no. 15, pages 3189–3196, 2010. (Cited in page 54.)

[Li 2013] Yaohang Li. *Conformational Sampling in Template-Free Protein Loop Structure Modeling: An Overview.* Comput. Struct. Biotechnol. J., vol. 5, 2013. (Cited in page 21.)

[Liu 2009] Pu Liu, Fangqiang Zhu, Dmitrii N. Rassokhin and Dimitris K. Agrafiotis. *A self-organizing algorithm for modeling protein loops.* PLoS Comput. Biol., vol. 5, no. 8, page e1000478, 2009. (Cited in page 22.)

[López-Blanco 2016] José Ramón López-Blanco, Alejandro Jesús Canosa-Valls, Yaohang Li and Pablo Chacón. *RCD+: Fast loop modeling server.* Nucleic Acids Res., vol. 44, no. W1, pages W395–400, 2016. (Cited in pages 23, 89, and 96.)

[López-Blanco 2019] José Ramón López-Blanco and Pablo Chacón. *KORP: knowledge-based 6D potential for fast protein and loop modeling.* Bioinformatics, 2019. (Cited in pages 25 and 116.)

[MacCallum 1996] R. M. MacCallum, A. C. Martin and J. M. Thornton. *Antibody-antigen interactions: contact analysis and binding site topography.* J. Mol. Biol., vol. 262, no. 5, pages 732–745, 1996. (Cited in pages 16 and 17.)

[Maier 2015] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser and Carlos Simmerling. *ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB.* J. Chem. Theory Comput., vol. 11, no. 8, pages 3696–3713, 2015. (Cited in pages 25, 75, 111, and 117.)

[Mandell 2009] Daniel J. Mandell, Evangelos A. Coutsias and Tanja Kortemme. *Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling.* Nat. Methods, vol. 6, no. 8, pages 551–552, 2009. (Cited in page 24.)

[Manivel 2000] V. Manivel, N. C. Sahoo, D. M. Salunke and K. V. Rao. *Maturation of an antibody response is governed by modulations in flexibility of the antigen-combining site.* Immunity, vol. 13, no. 5, pages 611–620, 2000. (Cited in page 14.)

[Manocha 1994] D. Manocha and J. F Canny. *Efficient inverse kinematics for general 6R manipulators.* IEEE Trans. Robot. Autom., 1994. (Cited in pages 22 and 67.)

[Marcatili 2008] Paolo Marcatili, Alessandra Rosi and Anna Tramontano. *PIGS: automatic prediction of antibody structures.* Bioinformatics, vol. 24, no. 17, pages 1953–1954, 2008. (Cited in page 15.)

[Marillet 2015] Simon Marillet, Marie-Paule Lefranc, Pierre Boudinot and Frédéric Cazals. *Dissecting Interfaces of Antibody -Antigen Complexes: from Ligand Specific Features to Binding Affinity Predictions.* Technical Report RR-8770, Inria Sophia Antipolis, 2015. (Cited in page 11.)

[Marks 2017] Claire Marks, Jaroslaw Nowak, Stefan Klostermann, Guy Georges, James Dunbar, Jiye Shi, Sebastian Kelm and Charlotte M Deane. *Sphinx: Merging Knowledge-Based and Ab Initio Approaches to Improve Protein Loop Prediction.* Bioinformatics, vol. 33, no. 9, pages 1346–1353, 2017. (Cited in page 24.)

[Marks 2018] Claire Marks, Jiye Shi, Charlotte M. Deane and Alfonso Valencia. *Predicting loop conformational ensembles.* Bioinformatics, vol. 34, no. 6, pages 949–956, 2018. (Cited in pages 20 and 111.)

[Martin 1996] A. C. Martin and J. M. Thornton. *Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies.* J. Mol. Biol., vol. 263, no. 5, pages 800–815, 1996. (Cited in page 35.)

[Maximova 2016] Tatiana Maximova, Ryan Moffatt, Buyong Ma, Ruth Nussinov and Amarda Shehu. *Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics.* PLoS Comput. Biol., vol. 12, no. 4, page e1004619, 2016. (Cited in page 27.)

[McCammon 1977] J. A. McCammon, B. R. Gelin and M. Karplus. *Dynamics of folded proteins.* Nature, vol. 267, no. 5612, pages 585–590, 1977. (Cited in page 28.)

[McQuarrie 1999] D. A. McQuarrie and J. D. Simon. Molecular thermodynamics. University Science Books, 1999. (Cited in page 26.)

[Messih 2015] Mario Abdel Messih, Rosalba Lepore and Anna Tramontano. *Loop-Ing: a Template-Based Tool for Predicting the Structure of Protein Loops.* Bioinformatics, vol. 31, no. 23, pages 3767–3772, 2015. (Cited in page 23.)

[Metropolis 1953] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller and Edward Teller. *Equation of State Calculations by Fast Computing Machines.* J. Chem. Phys., vol. 21, no. 6, pages 1087–1092, 1953. (Cited in page 28.)

[Moal 2010] Iain H. Moal and Paul A. Bates. *SwarmDock and the Use of Normal Modes in Protein-Protein Docking.* Int. J. Mol. Sci., vol. 11, no. 10, pages 3623–3648, 2010. (Cited in page 54.)

[Molloy 2013] Kevin Molloy and Amarda Shehu. *Elucidating the ensemble of functionally-relevant transitions in protein systems with a robotics-inspired method.* BMC Struct. Biol., vol. 13 Suppl 1, page S8, 2013. (Cited in page 29.)

[Morea 1997] V. Morea, A. Tramontano, M. Rustici, C. Chothia and A. M. Lesk. *Antibody structure, prediction and redesign.* Biophys. Chem., vol. 68, no. 1-3, pages 9–16, 1997. (Cited in page 12.)

[Mu 2005] Y. Mu, P. H. Nguyen and G. Stock. *Energy landscape of a small peptide revealed by dihedral angle principal component analysis.* Proteins, vol. 58, no. 1, pages 45–52, 2005. (Cited in page 29.)

[Nikoloudis 2014] Dimitris Nikoloudis, Jim E. Pitts and José W. Saldanha. *A complete, multi-level conformational clustering of antibody complementarity-determining regions.* PeerJ, vol. 2, page e456, 2014. (Cited in page 12.)

[Nilmeier 2009] Jerome Nilmeier and Matthew P. Jacobson. *Monte Carlo Sampling with Hierarchical Move Sets: POSH Monte Carlo.* J. Chem. Theory Comput., vol. 5, no. 8, pages 1968–1984, 2009. (Cited in pages 22 and 29.)

[Nilmeier 2011] Jerome Nilmeier, Lan Hua, Evangelos A. Coutsias and Matthew P. Jacobson. *Assessing protein loop flexibility by hierarchical Monte Carlo sampling.* J. Chem. Theory Comput., vol. 7, no. 5, pages 1564–1574, 2011. (Cited in page 22.)

[North 2011] Benjamin North, Andreas Lehmann and Roland L. Dunbrack. *A New Clustering of Antibody CDR Loop Conformations.* J. Mol. Biol., vol. 406, no. 2, pages 228–256, 2011. (Cited in pages 12 and 46.)

[Onuchic 1997] José Nelson Onuchic, Zaida Luthey-Schulten and Peter G. Wolynes. *Theory of protein folding: the energy landscape perspective.* Annu. Rev. Phys. Chem., vol. 48, no. 1, pages 545–600, 1997. (Cited in page 26.)

[Onuchic 2004] José Nelson Onuchic and Peter G Wolynes. *Theory of protein folding.* Curr. Opin. Struct. Biol., vol. 14, no. 1, pages 70–75, 2004. (Cited in page 26.)

[Padlan 1995] E. A. Padlan, C. Abergel and J. P. Tipper. *Identification of specificity-determining residues in antibodies.* FASEB J., vol. 9, no. 1, pages 133–139, 1995. (Cited in pages 16 and 20.)

[Paës 2012] Gabriel Paës, Juan Cortés, Thierry Siméon, Michael J. O'Donohue and Vinh Tran. *Thumb-loops up for catalysis: a structure/function investigation of a functional loop movement in a GH11 xylanase.* Comput. Struct. Biotechnol. J., vol. 1, 2012. (Cited in page 29.)

[Pantazes 2010] R. J. Pantazes and C. D. Maranas. *OptCDR: a general computational method for the design of antibody complementarity determining regions for targeted epitope binding.* Protein Eng. Des. Sel., vol. 23, no. 11, pages 849–858, 2010. (Cited in page 19.)

[Pedotti 2011] Mattia Pedotti, Luca Simonelli, Elsa Livoti and Luca Varani. *Computational docking of antibody-antigen complexes, opportunities and pitfalls illustrated by influenza hemagglutinin.* Int. J. Mol. Sci., vol. 12, no. 1, pages 226–251, 2011. (Cited in pages 15, 18, and 19.)

[Petoukhov 2002] Maxim V Petoukhov, Nigel A J Eady, Katherine A Brown and Dmitri I Svergun. *Addition of Missing Loops and Domains to Protein Models by X-Ray Solution Scattering.* Biophys. J., vol. 83, no. 6, pages 3113–3125, 2002. (Cited in page 20.)

[Pillardy 2001] Jarosław Pillardy, Cezary Czaplewski, Adam Liwo, Jooyoung Lee, Daniel R. Ripoll, Rajmund Kaźmierkiewicz, Stanisław Ołdziej, William J. Wedemeyer, Kenneth D. Gibson, Yelena A. Arnautova and others. *Recent improvements in prediction of protein structure by global optimization of a potential energy function.* Proc. Natl. Acad. Sci. U.S.A., vol. 98, no. 5, pages 2329–2333, 2001. (Cited in page 27.)

[Ponder 1987] J. W. Ponder and F. M. Richards. *Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes.* J. Mol. Biol., vol. 193, no. 4, pages 775–791, 1987. (Cited in page 19.)

[Ponder 2003] Jay W. Ponder and David A. Case. *Force fields for protein simulations.* Adv. Protein Chem., vol. 66, pages 27–85, 2003. (Cited in page 25.)

[Raghunathan 2012] Gopalan Raghunathan, Jason Smart, Joseph Williams and Juan Carlos Almagro. *Antigen-binding site anatomy and somatic mutations in antibodies that recognize different types of antigens.* J. Mol. Recognit., vol. 25, no. 3, pages 103–113, 2012. (Cited in pages 16 and 17.)

[Rapaport 2007] D. C. Rapaport. The art of molecular dynamics simulation. Academic Press, 2007. (Cited in page 28.)

[Rata 2010] Ionel A. Rata, Yaohang Li and Eric Jakobsson. *Backbone Statistical Potential from Local Sequence-Structure Interactions in Protein Loops.* J. Phys. Chem. B, vol. 114, no. 5, pages 1859–1869, 2010. (Cited in pages 25 and 116.)

[Rini 1992] J. M. Rini, U. Schulze-Gahmen and I. A. Wilson. *Structural evidence for induced fit as a mechanism for antibody-antigen recognition.* Science, vol. 255, no. 5047, pages 959–965, 1992. (Cited in page 10.)

[Russell 2009] Stuart Russell and Peter Norvig. Artificial intelligence: A modern approach. Prentice Hall Press, 3rd edition, 2009. (Cited in page 66.)

[Schmidt 2013] A. G. Schmidt, H. Xu, A. R. Khan, T. O'Donnell, S. Khurana, L. R. King, J. Manischewitz, H. Golding, P. Suphaphiphat, A. Carfi, E. C. Settembre, P. R. Dormitzer, T. B. Kepler, R. Zhang, M. A. Moody, B. F. Haynes, H.-X. Liao, D. E. Shaw and S. C. Harrison. *Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody.* Proc. Natl. Acad. Sci. U.S.A., vol. 110, no. 1, pages 264–269, 2013. (Cited in page 14.)

[Schrödinger, LLC 2015] Schrödinger, LLC. The PyMOL Molecular Graphics System. Version 1.8, 2015. (Cited in page 36.)

[Sela-Culang 2012] Inbal Sela-Culang, Shahar Alon and Yanay Ofran. *A Systematic Comparison of Free and Bound Antibodies Reveals Binding-Related Conformational Changes.* J. Immunol., vol. 189, no. 10, pages 4890–4899, 2012. (Cited in pages 32 and 167.)

[Sela-Culang 2013] Inbal Sela-Culang, Vered Kunik and Yanay Ofran. *The structural basis of antibody-antigen recognition.* Front. Immunol., vol. 4, page 302, 2013. (Cited in page 16.)

[Sellers 2010] Benjamin D. Sellers, Jerome P. Nilmeier and Matthew P. Jacobson. *Antibodies as a model system for comparative model refinement.* Proteins, vol. 78, no. 11, pages 2490–2505, 2010. (Cited in page 23.)

[Shehu 2006] Amarda Shehu, Cecilia Clementi and Lydia E. Kavraki. *Modeling protein conformational ensembles: from missing loops to equilibrium fluctuations.* Proteins, vol. 65, no. 1, pages 164–179, 2006. (Cited in pages 20 and 22.)

[Shehu 2012] Amarda Shehu and Lydia E. Kavraki. *Modeling Structures and Motions of Loops in Protein Molecules.* Entropy, vol. 14, no. 12, pages 252–290, 2012. (Cited in page 21.)

[Shenkin 1987] P. S. Shenkin, D. L. Yarmush, R. M. Fine, H. J. Wang and C. Levinthal. *Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures.* Biopolymers, vol. 26, no. 12, pages 2053–2085, 1987. (Cited in pages 21 and 89.)

[Shirai 1996] Hiroki Shirai, Akinori Kidera and Haruki Nakamura. *Structural classification of CDR-H3 in antibodies.* FEBS Lett., vol. 399, no. 1-2, pages 1–8, 1996. (Cited in page 12.)

[Simonelli 2013] Luca Simonelli, Mattia Pedotti, Martina Beltramello, Elsa Livoti, Luigi Calzolai, Federica Sallusto, Antonio Lanzavecchia and Luca Varani.

Rational engineering of a human anti-dengue antibody through experimentally validated computational docking. PLoS One, vol. 8, no. 2, page e55561, 2013. (Cited in page 20.)

[Soto 2008] Cinque S Soto, Marc Fasnacht, Jiang Zhu, Lucy Forrest and Barry Honig. *Loop Modeling: Sampling, Filtering, and Scoring.* Proteins, vol. 70, no. 3, pages 834–843, 2008. (Cited in pages 85 and 89.)

[Sotriffer 2000] C. A. Sotriffer, W. Flader, R. H. Winger, B. M. Rode, K. R. Liedl and J. M. Varga. *Automated docking of ligands to antibodies: methods and applications.* Methods, vol. 20, no. 3, pages 280–291, 2000. (Cited in page 18.)

[Spassov 2013] Velin Z. Spassov and Lisa Yan. *pH-selective mutagenesis of protein-protein interfaces: In silico design of therapeutic antibodies with prolonged half-life.* Proteins, vol. 81, no. 4, pages 704–714, 2013. (Cited in page 19.)

[Stanfield 2006] Robyn L. Stanfield, Adam Zemla, Ian A. Wilson and Bernhard Rupp. *Antibody Elbow Angles are Influenced by their Light Chain Class.* J. Mol. Biol., vol. 357, no. 5, pages 1566–1574, 2006. (Cited in page 39.)

[Stein 2013] Amelie Stein and Tanja Kortemme. *Improvements to Robotics-Inspired Conformational Sampling in Rosetta.* PLoS One, vol. 8, no. 5, page e63090, 2013. (Cited in page 24.)

[Sugita 1999] Y. Sugita and Y. Okamoto. *Replica-exchange molecular dynamics method for protein folding.* Chem. Phys. Lett., vol. 314, no. 1-2, pages 141–151, 1999. (Cited in page 28.)

[Tang 2014] Ke Tang, Jinfeng Zhang and Jie Liang. *Fast Protein Loop Sampling and Structure Prediction Using Distance-Guided Sequential Chain-Growth Monte Carlo Method.* PLoS Comput. Biol., vol. 10, no. 4, page e1003539, 2014. (Cited in pages 20, 24, 89, 90, and 92.)

[Tang 2015] Ke Tang, Samuel W.K. Wong, Jun S. Liu, Jinfeng Zhang and Jie Liang. *Conformational sampling and structure prediction of multiple interacting loops in soluble and β -barrel membrane proteins using multi-loop distance-guided chain-growth Monte Carlo method.* Bioinformatics, vol. 31, no. 16, pages 2646–2652, 2015. (Cited in page 24.)

[Teilum 2009] Kaare Teilum, Johan G. Olsen and Birthe B. Kragelund. *Functional aspects of protein flexibility.* Cell. Mol. Life Sci., vol. 66, no. 14, pages 2231–2247, 2009. (Cited in page 6.)

[Teplyakov 2014] Alexey Teplyakov, Jinquan Luo, Galina Obmolova, Thomas J. Malia, Raymond Sweet, Robyn L. Stanfield, Sreekumar Kodangattil, Juan Carlos Almagro and Gary L. Gilliland. *Antibody modeling assessment*

II. Structures and models. Proteins, vol. 82, no. 8, pages 1563–1582, 2014. (Cited in page 15.)

[Teplyakov 2016] Alexey Teplyakov, Galina Obmolova, Thomas J. Malia, Jinquan Luo, Salman Muzammil, Raymond Sweet, Juan Carlos Almagro and Gary L. Gilliland. *Structural diversity in a human antibody germline library.* mAbs, vol. 8, no. 6, pages 1045–1063, 2016. (Cited in pages 12 and 15.)

[Tomlinson 1992] I. M. Tomlinson, G. Walter, J. D. Marks, M. B. Llewelyn and G. Winter. *The repertoire of human germline VH sequences reveals about fifty groups of VH segments with different hypervariable loops.* J. Mol. Biol., vol. 227, no. 3, pages 776–798, 1992. (Cited in page 13.)

[Tsuchiya 2016] Yuko Tsuchiya and Kenji Mizuguchi. *The diversity of H3 loops determines the antigen-binding tendencies of antibody CDR loops.* Protein Sci., vol. 25, no. 4, pages 815–825, 2016. (Cited in pages 16 and 17.)

[Van Regenmortel 2014] Marc H. V. Van Regenmortel. *Specificity, polyspecificity, and heterospecificity of antibody-antigen recognition.* J. Mol. Recognit., vol. 27, no. 11, pages 627–639, 2014. (Cited in pages 10 and 16.)

[Vanommeslaeghe 2010] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov and A. D. Mackerell. *CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields.* J. Comput. Chem., vol. 31, no. 4, pages 671–690, 2010. (Cited in page 25.)

[Vargas-Madrazo 1995] E. Vargas-Madrazo, F. Lara-Ochoa and J. C. Almagro. *Canonical structure repertoire of the antigen-binding site of immunoglobulins suggests strong geometrical restrictions associated to the mechanism of immune recognition.* J. Mol. Biol., vol. 254, no. 3, pages 497–504, 1995. (Cited in pages 13 and 17.)

[Verma 2009] Abhinav Verma and Wolfgang Wenzel. *A Free-Energy Approach for All-Atom Protein Simulation.* Biophys. J., vol. 96, no. 9, pages 3483–3494, 2009. (Cited in page 27.)

[Vreven 2015] Thom Vreven, Iain H. Moal, Anna Vangone, Brian G. Pierce, Panagiotis L. Kastritis, Mieczyslaw Torchala, Raphael Chaleil, Brian Jiménez-García, Paul A. Bates, Juan Fernandez-Recio, Alexandre M. J. J. Bonvin and Zhiping Weng. *Updates to the integrated protein-protein interaction benchmarks: Docking benchmark version 5 and affinity benchmark version 2.* J. Mol. Biol., vol. 427, no. 19, pages 3031–3041, 2015. (Cited in pages 32, 54, and 168.)

[Wales 2003] David Wales. Energy landscapes: Applications to clusters, biomolecules and glasses. Cambridge University Press, 2003. (Cited in page 26.)

[Wang 2007] Chu Wang, Philip Bradley and David Baker. *Protein-protein docking with backbone flexibility.* J. Mol. Biol., vol. 373, no. 2, pages 503–519, 2007. (Cited in page 19.)

[Weitzner 2015] Brian D. Weitzner, Roland L. Dunbrack and Jeffrey J. Gray. *The origin of CDR H3 structural diversity.* Structure, vol. 23, no. 2, pages 302–311, 2015. (Cited in page 12.)

[Wilson 1993] Ian A. Wilson and Robyn L. Stanfield. *Antibody-antigen interactions.* Curr. Opin. Struct. Biol., vol. 3, no. 1, pages 113–118, 1993. (Cited in page 10.)

[Wilson 1994a] I. A. Wilson, J. B. Ghiara and R. L. Stanfield. *Structure of anti-peptide antibody complexes.* Res. Immunol., vol. 145, no. 1, pages 73–78, 1994. (Cited in page 14.)

[Wilson 1994b] Ian A. Wilson and Robyn L. Stanfield. *Antibody-antigen interactions: new structures and new conformational changes.* Curr. Opin. Struct. Biol., vol. 4, no. 6, pages 857–867, 1994. (Cited in page 10.)

[Wong 2011] Sergio E. Wong, Ben D. Sellers and Matthew P. Jacobson. *Effects of somatic mutations on CDR loop flexibility during affinity maturation.* Proteins, vol. 79, no. 3, pages 821–829, 2011. (Cited in pages 14 and 19.)

[Woolfson 1997] M. M. Woolfson. An introduction to X-ray crystallography. Cambridge University Press, 1997. (Cited in page 20.)

[Xiang 2002] Zhexin Xiang, Cinque S Soto and Barry Honig. *Evaluating Conformational Free Energies: the Colony Energy and its Application to the Problem of Loop Prediction.* Proc. Natl. Acad. Sci. U.S.A., vol. 99, no. 11, pages 7432–7437, 2002. (Cited in page 89.)

[Xiang 2006] Zhexin Xiang. *Advances in Homology Protein Structure Modeling.* Curr. Protein Pept. Sci., vol. 7, no. 3, pages 217–227, 2006. (Cited in page 89.)

[Yang 2008] Yuedong Yang and Yaoqi Zhou. *Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions.* Protein Sci., vol. 17, no. 7, pages 1212–1219, 2008. (Cited in pages 25 and 111.)

[Yao 2008] Peggy Yao, Ankur Dhanik, Nathan Marz, Ryan Propper, Charles Kou, Guanfeng Liu, Henry van den Bedem, Jean-Claude Latombe, Inbal Halperin-Landsberg and Russ Biagio Altman. *Efficient algorithms to explore conformation spaces of flexible protein loops.* IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 5, no. 4, pages 534–545, 2008. (Cited in page 22.)

[Zacharias 2005] Martin Zacharias. *ATTRACT: protein-protein docking in CAPRI using a reduced protein model*. Proteins, vol. 60, no. 2, pages 252–256, 2005. (Cited in pages 18 and 19.)

[Zaki 2008] M. Zaki and C. Bystroff. Protein structure prediction. Methods in Molecular Biology, 413. Humana Press / Springer, 2008. (Cited in page 15.)

[Zhao 2011] Suwen Zhao, Kai Zhu, Jianing Li and Richard A. Friesner. *Progress in Super Long Loop Prediction*. Proteins, vol. 79, no. 10, pages 2920–2935, 2011. (Cited in page 85.)

[Zhu 2006] Kai Zhu, David L. Pincus, Suwen Zhao and Richard A. Friesner. *Long loop prediction using the protein local optimization program*. Proteins, vol. 65, no. 2, pages 438–452, 2006. (Cited in page 23.)

[Zhu 2013] Kai Zhu and Tyler Day. *Ab initio structure prediction of the antibody hypervariable H3 loop*. Proteins, vol. 81, no. 6, pages 1081–1089, 2013. (Cited in page 23.)

[Zuckerman 2010] Daniel M Zuckerman. Statistical physics of biomolecules: an introduction. CRC Press, 2010. (Cited in page 26.)