



SAPIENZA  
UNIVERSITÀ DI ROMA

# ANTIBODY MODELLING AND STRUCTURE ANALYSIS

Application to biomedical problems



**Tutor**  
**Dr. Marcatili P.**  
**Supervisor**  
**Prof. Tramontano A.**  
**Coordinator**  
**Prof. Tripodi M.**

**Student**  
**Chailyan A.**

| XXV cycle | Pasteurian Sciences PhD School



SAPIENZA  
UNIVERSITÀ DI ROMA

“ANTIBODY MODELLING AND STRUCTURE  
ANALYSIS”

“Application to Biomedical Problems”

by

Anna Chailyan

---

**Tutor**

**Dr. Marcatili P.**

**Supervisor**

**Prof. Tramontano A.**

**Coordinator**

**Prof. Tripodi M.**

\_\_XXV cycle\_Pasteurian Sciences PhD School\_\_2013\_Rome\_\_

I dedicate this work to all who dare the blank page and the bare stage!

# Table of contents

---

<b>TABLE OF CONTENTS</b>	<b>4</b>
<b>ACKNOWLEDGEMENTS</b>	<b>7</b>
<b><u>ABSTRACT</u></b>	
BACKGROUND	8
AIM	9
RESULTS	9
REFERENCE:	11
<b>CHAPTER 1</b>	
<b><u>1.1 GENERAL INTRODUCTION</u></b>	<b><u>12</u></b>
<b><u>1.2 ANTIBODY SEQUENCE AND STRUCTURE</u></b>	
1.2.1 ANTIBODY STRUCTURE	14
1.2.2 ANTIBODY HYPERVARIABLE REGIONS	16
1.2.3 ANTIBODY NUMBERING	18
1.2.4 ANTIBODY GENETICS	20
<b><u>1.3 METHODS FOR IMMUNOGLOBULIN STRUCTURE PREDICTION</u></b>	<b><u>22</u></b>
1.3.1 CAVEATS AND OPEN PROBLEMS IN ANTIBODY MODELLING	27
<b><u>1.4 AIM AND CONTRIBUTIONS OF THE STUDY</u></b>	<b><u>30</u></b>

## **CHAPTER 2**

### **2.1 STRUCTURAL REPERTOIRE OF IMMUNOGLOBULIN $\lambda$ LIGHT CHAINS 32**

### **2.2 THE ASSOCIATION OF HEAVY AND LIGHT CHAIN VARIABLE DOMAINS IN ANTIBODIES: IMPLICATIONS FOR ANTIGEN SPECIFICITY 37**

## **CHAPTER 3**

### **3.1 A DATABASE OF IMMUNOGLOBULINS WITH INTEGRATED TOOLS: DIGIT 41**

## **CHAPTER 4**

### **4.1 HIERARCHICAL CLUSTERING OF B-CELL RECEPTOR STRUCTURES IN SPLENIC MARGINAL ZONE LYMPHOMA 45**

4.1.1 BACKGROUND 46

4.1.2 RESULTS 47

4.1.3 DISCUSSION 51

### **4.2 STRUCTURE VARIABILITY OF IMMUNOGLOBULINS EXPRESSED BY B CELLS IN CHRONIC LYMPHOCYTIC LEUKAEMIA 53**

4.2.1 BACKGROUND 53

4.2.2 MATERIALS AND METHODS 54

4.2.3 RESULTS 56

4.2.4 DISCUSSION 57

## **CHAPTER 5**

### **5.1 ANTIBODY CONFORMATIONAL CHANGE UPON ANTIGEN BINDING 60**

5.1.1 BACKGROUND 60

5.1.2 MATERIAL AND METHODS 62

5.1.3 RESULTS 68

## **CHAPTER 6**

### **6.1 CONCLUSIONS AND OUTLOOK 74**

### **REFERENCE LIST 76**

### **PAPER I 83**

### **PAPER II 84**

### **PAPER III 85**

## Acknowledgements

---

**It was a pleasure for me to work with all the wonderful people in the Biocomputing Group here in Rome.**

First of all, I wish to express sincerest gratitude to my supervisor, Prof. Tramontano for her on-going inspiration, guidance, immense knowledge and support encompassing both scientific and emotional. It's funny to reckon my random walk in the University campus and you taking me to work with you, without any skills. I learned a lot during this time and I am convinced that this knowledge will help me in the future. For everything you've done for me, dear Anna, I thank you. I have learned more from you than you realize.

I would also like to thank my tutor Dr. Marcatili, for encouraging me from the very first days my own process of changes – a period of personal growth that has been terrifically exciting, rewarding and completely chaotic. Paolo's ideas and tremendous support had a major influence on this thesis and he put a lot of efforts in assisting and guiding me in getting my scientific career started on the right foot. He provided me with the foundation for becoming a Scientist one day.

My thanks to my friends and colleagues for the great time I had in Biocomputing group. I enjoyed the atmosphere, their friendship, and their support. Many thanks to PierPaolo, Allegra, Mimmo, Luigi, and Daniel. Special thanks goes to Lori and Tiger for their friendship and for providing some much needed humour and entertainment. You had inspired me in research and life through our interactions during the long hours in the lab. Grazie.

A special thanks to Prof. Lesk, for inspiring collaboration and all the help. It was an honour to work with you and I hope we will be continuing this collaboration in the future.

I would especially like to thank my amazing family for the love, support and confidence in me throughout my life. Mamma, Papa, Gor and Tata Շնորհակալություն, Սիրուն եմ ու Շատ Կարողում!!! This thesis was definitely a challenge and did not eventuate without many sacrifices and I would not have completed this task without you.

This thesis has partly been supported by Research fellowship of Anna Tramontano (King Abdullah University of Science and Technology (KAUST) (Award number KUK-I1-012-43)). The support is gratefully acknowledged.

# Abstract

## Background

*The usefulness of antibodies and antibody derived artificial constructs in various medical and biochemical applications has made them a prime target for protein engineering, modelling, and structure analysis. The huge number of known antibody sequences, that far outpaces the number of solved structures, raises the need for reliable automatic methods of antibody structure prediction.*

*Antibodies have a very characteristic molecular structure that is reflected in their modelling technique. Currently, the most accurate models are produced using a quite peculiar modelling strategy, developed among others by our group: the framework regions are modelled with a standard comparative modelling approach, whereas the hypervariable loops are predicted using the ad-hoc “canonical structure method”, historically based on expert analysis of the available antibody solved structures. More than thirty years passed since this modelling method was initially developed, nonetheless there is still a huge effort in the academic and pharmaceutical communities to improve its accuracy. The reason for this lies in several error sources in the current modelling process. First of all, given the large amount of available structures, it was impossible to manually update “canonical structure” classes and rules. Moreover, the lack of specific studies on the packing between the VL and the VH domains and on possible conformational changes occurring upon antigen binding was impairing the integration in the modelling techniques of such factors.*



## **Aim**

*The general aim of this study is to carry out an extensive characterization and annotation of immunoglobulin molecules i.e. to deepen our understanding of the molecular basis of their specificity using a combination of bioinformatics sequence- and structure-based analysis. I carried out improvements to the antibody modelling protocols by revising the canonical structure definitions and by minimizing the errors arising from VL and the VH domain packing at the same time by taking care of the conformational changes occurring upon antigen binding.*

## **Results**

*During the past years, we successfully improved the description of the structural repertoire of immunoglobulins with lambda light chains, which has both practical (design, engineering and humanization) and theoretical applications (improvement of the antibody modelling)[1]. Our large-scale analysis of the association of heavy and light chain variable domains in antibodies showed that there are essentially two different modes of interaction that can be identified by the presence of key amino acids in specific positions of the antibody sequences [2]. Interestingly, we also found that the different packing modes are related to the volume and type of recognized antigen. These findings are clearly relevant for the design of antibodies and of antibody libraries. The investigation of the antibody conformational changes upon antigen binding allowed us to identify sections on variable and constant regions that show significant flexibility when comparing the antigen bound/unbound forms of immunoglobulins. The results of all the above-mentioned analyses have been implemented in our in-house immunoglobulin structure prediction server (PIGS, automatic Prediction of ImmunoGlobulin*

*Structure*), thus helping to minimize the sources of errors in the current modelling process. Consequent to our results, we were asked to write a chapter in *Encyclopaedia of Biophysics on antibody modelling* [3].

*A further step in the direction of improving the understanding of antibody recognition mechanisms was to put together all the annotations of immunoglobulins in a publicly available database. To this aim, we constructed a database of immunoglobulin sequences and integrated tools (DIGIT) [4], which is becoming an extensively used resource by the community. DIGIT stores sequences of annotated immunoglobulin variable domains and offers to the user several tools for searching and analysing them.*

*Our experience in antibody modelling allowed us to approach two biomedical problems in collaboration with Prof. Arcaini (University of Pavia) and Prof. Fabio Ghiotto (University of Genova). More specifically, by applying the tools we developed and all our theoretical knowledge we successfully analysed the immunoglobulin repertoires of SMZL (splenic marginal zone lymphoma) and CLL (chronic lymphocytic leukaemia) patient data. Both the CLL and SMZL patients are known to have a biased usage of immunoglobulin (IG) heavy variable (IGHV) genes and stereotyped B-cell receptors (BCRs), used as a marker in disease prognosis. We extended these analyses by taking into account VL germlines, VL-VH pairing and structural information, thus giving a more detailed view of the immunoglobulin repertoire in terms of sequence, structure and function. Analysing the immunoglobulins of patients with CLL, we discovered statistically significant differences among immunoglobulins in patients with favourable and unfavourable prognosis. A paper describing this work has been submitted [5]. The poster describing the results of SMZL repertoire analysis was accepted at the 2012 American Society of Haematology (ASH) meeting and published as an abstract [6].*

## Reference:

1. Chailyan, A., P. Marcatili, et al. (2011). "Structural repertoire of immunoglobulin lambda light chains." *Proteins* 79(5): 1513-1524.
2. Chailyan, A., P. Marcatili, et al. (2011). "The association of heavy and light chain variable domains in antibodies: implications for antigen specificity." *FEBS J* 278(16): 2858-2866.
3. Marcatili P., A. Chailyan, D. Cirillo and A. Tramontano. *Modelling of antibody structures. Encyclopaedia of Biophysics. Springer (2012).*
4. Chailyan, A., A. Tramontano, et al. (2012). "A database of immunoglobulins with integrated tools: DIGIT." *Nucleic Acids Res.* doi:10.1093/nar/gkr806.
5. Marcatili P., F. Ghiotto, C. Tenca, A. Chailyan, A. N. Mazzarello, X. Yan, M. Colombo, E. Albesiano, D. Bagnara, G. Cutrona, F. Morabito, S. Bruno, M. Ferrarini, N. Chiorazzi, A. Tramontano, F. Fais. "Immunoglobulins produced by chronic lymphocytic leukaemia B cells show limited binding site structure variability." *submitted*
6. Marcatili P., S. Zibellini, S. Rattotti, A. Chailyan, M. Varettoni, L. Morello, E. Boveri, M. Lucioni, M. Bonfichi, M. Gotti, V. Fiaccadori, M. Paulli, A. Tramontano, L. Arcaini. "Hierarchical Clustering of B-Cell Receptor Structures in Splenic Marginal Zone Lymphoma", abstract, *American Society of Haematology (ASH) meeting.*

## Chapter 1

---

### 1.1 General Introduction

**W**e live in extraordinary times. Advances in the life sciences day by day shed light on protein sequence, structure, and function relationship. We are more and more aware of confirmations and, even more interestingly, exceptions of the central tenet of molecular biology for which the function of a protein is completely determined by its three dimensional structure, that is in turn dictated by its sequence.

The immune system is an extensive machinery that by orchestrating a complex network of regulatory mechanisms achieves an ample yet precise recognition of pathogens and foreign molecules, avoiding any potentially harmful response over self-molecules. Immunoglobulins (Igs, also known as antibodies, abbreviated Abs) are some of the molecules used by the immune system to adapt to huge number of different antigenic challenges: at any one time the average human body contains more than  $10^6$  different antibodies. Since their discovery in the first half of the

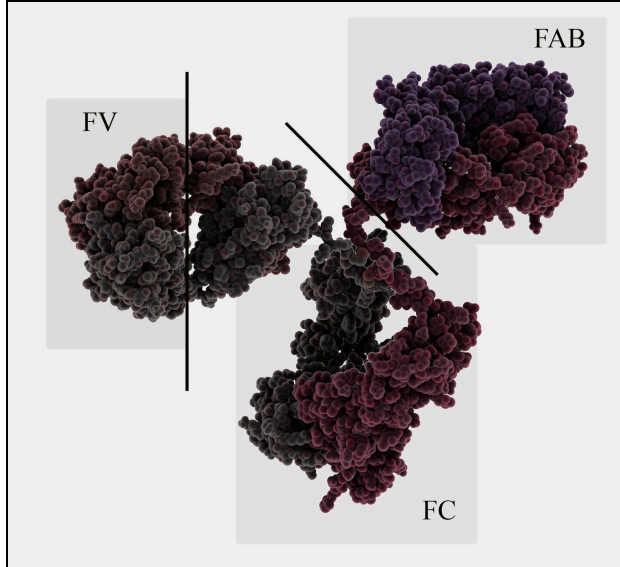
twentieth century, antibodies have occupied a prominent role in several research areas as well as in a large number of bio-technological and clinical applications.

The three dimensional structures of immunoglobulins and of proteins in general are determined mainly by two experimental techniques: X-ray crystallography and Nuclear Magnetic Resonance (NMR). The experimentally determined atomic coordinates are deposited in a public repository, the Protein Data Bank (PDB) (Berman et al., 2000). As of September 2012, there are approximately 2000 immunoglobulin structures, whereas the number of antibody sequences currently available in public databases is two orders of magnitude larger. The reason for this increasingly huge gap between the number of structures and that of sequences (Fox et al., 2008) is that it can take months or even years to solve one structure, which can cost hundreds of thousands dollars, whereas sequencing techniques are much cheaper and faster. This increase in the number of antibody sequences with unsolved structure produced by various high-throughput techniques is bridged by an alternative that allows obtaining an accurate antibody structural model through computational modelling and prediction in a relatively automatic fashion.

## **1.2 Antibody sequence and structure**

### **1.2.1 Antibody structure**

Antibodies generally assume one of these two roles: i) plasma membrane bound antigen receptor on the surface of a B-cell (B-cell receptor) or ii) free molecules in cellular fluids functioning to intercept and eliminate antigenic determinants. In either role, antibody function is intimately related to its structure. In 1959, a joint Nobel Prize was awarded to Edelman and Porter for their contributions towards the immunoglobulin properties by elucidating the basic primary structure of the immunoglobulin that can be written as  $H_2 L_2$  (Figure 1). The molecule consists of four polypeptide chains, two heavy and two light, joined by disulphide bonds so that each heavy chain is linked to a light and the two heavy chains are linked together.



**Figure 1.** Immunoglobulin structure. Heavy chains (grey, burgundy) and light chains (violet, brown) are shown in space-fill mode using Qutemol (<http://qutemol.sourceforge.net>). Fab, Fc and Fv regions are highlighted.

Both the chains are usually divided in a Variable region (Fv, fragment variable, VL light and VH heavy) and a Constant region (C<sub>H</sub> and C<sub>L</sub>). The variable region is so named because its sequence varies from one immunoglobulin to another. In contrast, the constant region consists of an amino acid sequence that is conserved for immunoglobulins of the same isotype. There are four different light chain isotypes,

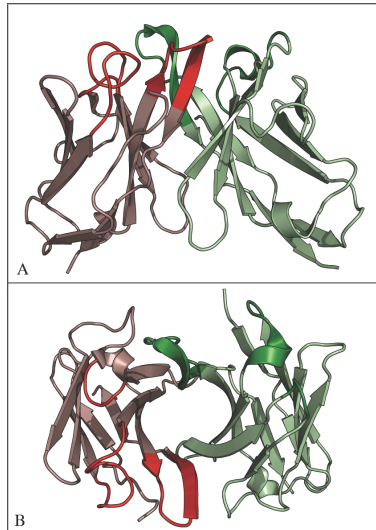
among which only two are found in mammals, namely the  $\kappa$ -kappa and  $\lambda$ -lambda type light chains. The heavy chains are of five types, designated as  $\alpha$ ,  $\gamma$ ,  $\delta$ ,  $\mu$  and  $\epsilon$  chains and they define the class and the function of the antibody. The soluble form of an antibody and its membrane-bound counterpart are identical with the exception of a portion of the C-terminus of the heavy chain constant region. The heavy chains of membrane-bound antibody molecules have a hydrophobic C-terminus, which anchors them in the lipid bilayer of the B cell's plasma membrane. The heavy chains of secreted antibody molecules, by contrast, have instead a hydrophilic C-terminus, which allows them to escape from the cell.

### **1.2.2 Antibody hypervariable regions**

By analysing the Fv fragment, Wu and Kabat, (i.e. the dimer of  $V_H$  and  $V_L$  domains, Figure 2) identified three sequence portions in the variable part of each chain, the so-called hypervariable regions, with an extremely variable amino acid composition in comparison with the other less variable parts (Wu and Kabat, 1970). It has been correctly predicted by the authors that these hypervariable regions assume a loop conformation arising from a relatively conserved framework and are responsible for the selective binding of the antigen. They therefore named them



“complementary determining regions” (CDRs) in contrast to the surrounding framework regions (FRs).



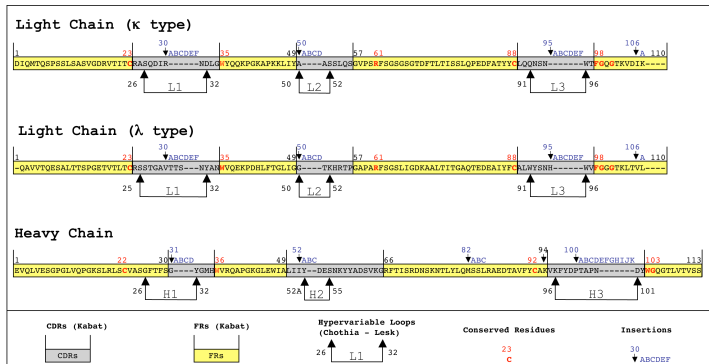
**Figure 2.** Two views of the Ig variable domain. Light chain in red (light red: framework, full red: hypervariable loops), heavy chain in green (light green: framework, full green: hypervariable loops). Panel A reports a lateral view (antigen should be located on the topmost part of the Ig molecule), panel B depicts the antigen view of the same structure.

### **1.2.3 Antibody numbering**

In an attempt to standardize and eventually automate the comparative analysis of a large number of related immunoglobulin sequences and structures, the unambiguous identification of structurally equivalent residues is of crucial importance. Fortunately, the architecture of antibodies, in which variable regions are interspersed between regions that are extremely conserved both in sequence and in structure, allowed the development of several unified numbering schemes for their sequences. Such numberings unambiguously define portions or specific residues of immunoglobulins that have a similar position in the three-dimensional structure. Currently there are several such numbering schemes (Abhinandan and Martin, 2008; Al-Lazikani et al., 1997; Chothia and Lesk, 1987; Honegger and Pluckthun, 2001; Kabat and Wu, 1991; Lefranc, 1999).

We describe here the Kabat-Chothia (KC) numbering scheme that is commonly adopted in antibody modelling. The Kabat numbering scheme was developed when none or little structural information was available. Chothia et al. analysed a small number of Ab structures and determined the relationship between the sequences of the Abs and the structures of their CDRs. Thus, the Chothia numbering scheme is almost identical to the Kabat scheme (they basically differ in the position of two

insertions), but based on structural considerations. In the KC scheme each residue of the variable domains of both heavy and light chain is identified by a different sequential number with the only exception of insertions that may occur in the CDRs or in some positions in the FRs, which are identified by the number of the position in which the insertion occurs followed by consecutive letters. Figure 3 summarizes this scheme. It is worth mentioning that some conserved FR residues are present in almost all immunoglobulins. Such residues are extremely useful in identifying the various regions of the immunoglobulins and are reported in red in the figure.



**Figure 3.** Kabat-Chothia numbering of VK, VL, and VH. The numbers above the sequences represent the KC numbering of specific residues, the remaining residues are numbered consecutively. Letters in blue correspond to insertions. Kabat definition of FRs and CDRs are highlighted in yellow and grey respectively, the Chothia and Lesk definition of hypervariable loops is indicated by arrows. Conserved residues are reported in red.

### 1.2.4 Antibody genetics

To give a more complete picture of immunoglobulin sequence-structure-function relationship, we will briefly describe some of the sophisticated mechanisms that the immune system developed to overcome the problem of recognizing a huge number of different antigens. The first source of variability is achieved in the Pre-B cell phase, when random assembling of gene segments generates the mature antibody genes. For the heavy chain in humans, selection and recombination takes place by the union of one from each of the 51 V<sub>H</sub> ("variable"), 27 D ("diversity") and 6 J<sub>H</sub> ("joining") genes, followed by a series of 5 C<sub>H</sub> ("constant") gene segments. The light chain loci ( $\lambda$  and  $\kappa$  in human) are formed by numerous V and J genes, but do not have D genes. The  $\lambda$  multigene family is formed by 30 V $\lambda$  gene segments, 4 J $\lambda$  gene segments, 4 C $\lambda$  gene segments, whereas the  $\kappa$  multigene family has 40 V $\kappa$  gene, 5 J $\kappa$  segments, and a single C $\kappa$  segment. The procedure of gene recombination is not precise and extra nucleotides are inserted during this process thus further increasing in a very effective way the number of possible antibody V region sequences. Another mechanism that augments the variety and the specificity of interactions with antigens and expands the potency of the immunoglobulin molecule is achieved through a process of random somatic

mutation. When a B cell recognizes an antigen, it is stimulated to proliferate. During proliferation, the rate of somatic mutation in the variable regions is extremely high, at least  $10^5$ - $10^6$  folds greater than the normal rate of mutation across the genome. Considering the variability produced by all these events, together with the combinatorial effect on light and heavy chain selection, one can have an approximate picture of the mechanisms that create the huge structural variety of immunoglobulins and help to overcome the problem of recognizing an almost infinite number of antigens.

## 1.3 Methods for immunoglobulin structure prediction

Antibody design, engineering and humanization partially or completely rely on the knowledge of the immunoglobulin structure. In the absence of an experimental structure, it is often possible to generate predictions of 3D structures by comparative modelling, (also known as homology modelling), as well as by *de novo* computational approaches. These modelling techniques often are reliable enough to be used in a very large number of applications, such as protein engineering (Morea et al., 1997a; Morea et al., 1997b) and docking (Pedotti et al., 2011).

*De novo* modelling methods calculate 3D coordinates of the target protein, starting from its amino acid sequence alone, by means of knowledge-based methods and/or physics-based energy functions. Their major limitation is that, due to our limited comprehension of the physicochemical principles governing protein structures, the energy functions used to evaluate the different conformations often do not distinguish a correct prediction from an incorrect one, even when additional criteria are used to filter the results. Approaches combining knowledge-based and *de novo* methods have also been described (Martin et al., 1989). Comparative modelling is

based on identifying a protein of known structure homologous to the one to be modelled and using it as template to predict the structure of the target protein.

Antibodies have very peculiar molecular structure and so is their modelling technique. The birth of antibody modelling techniques goes back to 1972, when Kabat and Wu carried out an analysis to determine the relationship between the amino acid sequences and the three-dimensional structures of the antigen-binding sites. They identified three *hypervariable* regions on both heavy and light chains, and correctly predicted such portions to assume a loop conformation arising from a relatively conserved framework and to be responsible for the selective binding of the antigen. The work of Chothia and Lesk extended this analysis by showing that five out of six regions (L1, L2, L3, H1 and H2) usually adopt just a limited set of discrete main-chain conformations. Moreover, they identified relatively few residues found within the hypervariable regions as well as in the conserved  $\beta$ -sheet framework that, through their packing, hydrogen bonding or ability to assume unusual  $\phi$ ,  $\psi$  or  $\omega$  angles, are primarily responsible for the backbone conformations of the hypervariable loops. These classes of commonly occurring backbone conformations of the hypervariable regions, identified by the length of the loop and by specific key residues, were consequently named “canonical structures.” The sixth

hypervariable region (H3) is a specific case: it is the most variable loop in terms of sequence and structure and is found to contact the antigen in the majority of solved antibody-antigen structures. A sequence-structure relationship can only be found for the main chain conformation of the ten residues (four residues from the N-terminus and six residues from the C-terminus) of the loop proximal to the framework (the “torso” of the loop (Morea et al., 1998)).

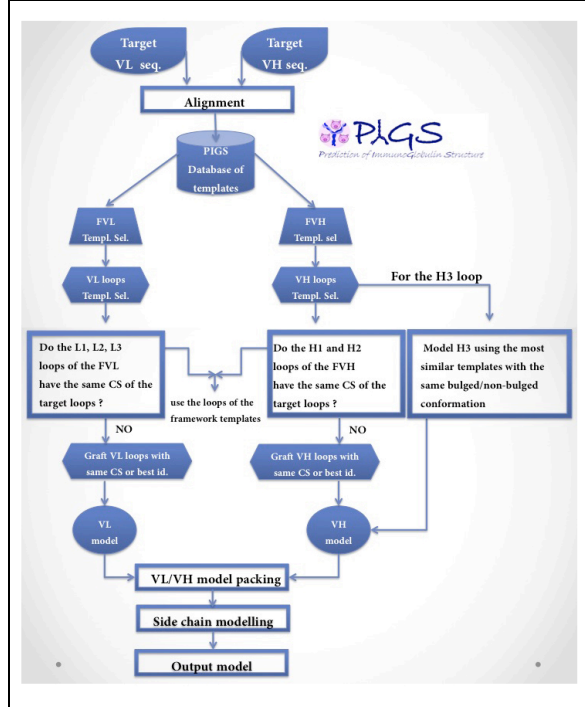
Thus, the peculiarity of the antibody molecules gives rise to an interesting modelling technique: the framework regions can be modelled using comparative modelling, whereas the hypervariable loops can be predicted based on the canonical structure method. The rules defining the canonical structures are being periodically revised and updated as more diverse antibody structures are solved (Al-Lazikani et al., 1997; Chailyan et al., 2011a; Morea et al., 1998; North et al., 2011).

Recently a blind study has been carried out to assess the state of the art in three-dimensional structure modelling of the antibody variable region (Almagro et al., 2011). The benchmark test set was composed by nine unpublished high-resolution X-ray Fab crystal structures. The Fv models generated by the four most successful immunoglobulin structure prediction methodologies were compared: two were generated by the homology modelling strategies independently developed by CCG



(Chemical Computer Group) and Accerlys Inc, and two were built using fully automated antibody modelling servers PIGS (Prediction of ImmunoGlobulin Structure) (Marcatili et al., 2008) and Rosetta Antibody (Sircar et al., 2009). The two web servers: PIGS (<http://arianna.bio.uniroma1.it/pigs>) and Rosetta Antibody Modelling Server (<http://antibody.graylab.jhu.edu/>) were developed in 2008.

The PIGS is a tool developed by our group that predicts the conformation of the antigen-binding site based on the canonical structure method (Figure 4). The framework of Rosetta Antibody homology modelling server is the Rosetta structure prediction suite, comprising knowledge-based techniques for template selection, grafting for non-H3 loops, *de novo* loop modelling for creating H3, and docking to optimize the relative orientation of  $V_H$  and  $V_L$  domains.



**Figure 4.** Schematic view of the PIGS web server pipeline. The template structures for each of the user uploaded target sequences can be selected manually or according to a predefined strategy.

Interestingly, when the models H1 and H2 were compared to the x-ray structures average RMSD calculated over the backbone yielded the best values for PIGS, 1.1 Å. The RMSD values of ACC, CCG, and Rosetta models were slightly higher and remarkable similar: 1.3 Å, regardless of the method used to model the Fv region

and/or the modeller expertise. Average RMSD values for the FRs and hypervariable loops with canonical structures (L1, L2, L3, H1, and H2) were close to 1.0 Å for all methods. The major source of error in the process are the modelling of the H3 loop, the possible changes of conformation occurring upon antigen binding and the reconstruction of the correct packing between the VL and the VH domains.

### **1.3.1 Caveats and open problems in antibody modelling**

As mentioned in the previous section, the errors in the modelling process arise mainly due to the H3 loop, the possible changes of conformation occurring upon antigen binding and the reconstruction of the correct packing between the VL and the VH domains.

The first of the aforementioned problems is still being fronted in many different ways but the accuracy of H3 modelling is far below that of the rest of the model. This problem is crucial for example in modelling camelid VHH antibodies that lack the light chain and usually display extremely long H3 loops with a peculiar

conformational repertoire. Accurate prediction of H3 is of great importance also because of its direct involvement in antigen binding.

Another open issue is the effect of antigen binding on the CDR loops conformation. Even though most of the loops display just a minor distortion upon antigen binding, the H3 loops (especially longer ones) can assume quite different conformations in the apo (unbound) and holo (bound) forms. It is therefore important to include this kind of information in the modelling process, and currently it cannot be achieved using the automated procedures described above.

The last point regards the correct packing between VL and VH domains. Recently, we (Chailyan et al., 2011c) described the existence of at least two different packing modalities that have a strong impact on the shape of the antigen-binding site and on the antigen specificity. This becomes relevant if two different solved structures are used as templates for the framework of the two chains; in this case the regions modelled for each of the two templates need to be packed together in order to obtain the final model. This process may introduce deviations between the model and the real structure that can have a relevant impact on the structure of the ABS, located at the tip of the domains. In order to avoid this, the user can choose where possible, the same template (even though with a lower sequence similarity) for several loops

and/or frameworks, thus minimizing the number of superposition needed to build the final model. Such a choice is not trivial and depends on the existence of a suitable template with a good sequence identity. In a typical scenario, if the templates are selected using the highest similarity criteria they may come from different antibodies, therefore they have to be packed together and this may introduce errors. At the same time, if the same template is used for different regions, it may have a low sequence identity to the target, and therefore a lower expected similarity to the target.

## **1.4 Aim and contributions of the study**

The usefulness of antibodies and of antibody-derived artificial constructs in various medical and biochemical applications has made them a prime target for protein engineering, modelling and structure analysis. The purpose of the study described here is to carry out an extensive characterization and annotation of immunoglobulin molecules i.e. to deepen our understanding of the molecular basis of their specificity. Original results are described in three independent papers, which are included in the order in which they appear in the thesis. A short summary before each of the articles outlines the main rationale and results. The thesis also contains unpublished investigations on antibody conformational changes upon the antigen binding and applications of thesis concepts to biomedical problems, more specifically to the analysis of the immunoglobulin repertoires of SMZL and CLL patients.

Chapters 2, 3, 4 and 5 outline the results of my studies. The description of the structural repertoire of immunoglobulins with lambda light chains, which has both practical (design, engineering and humanization) and theoretical applications, is given in Chapter 2. Besides, this chapter contains also the investigation on the association of heavy and light chain variable domains in antibodies, showing that

there are essentially two different modes of interaction (relevant for the design of antibodies and of antibody libraries). Chapter 3 comprises the description of the database of immunoglobulin sequences and integrated tools (DIGIT), a much-needed resource that stores sequences of annotated immunoglobulin variable domains enriched with tools for searching and analysing them. Finally, the chapter 4 contains two applications of these tools and methods to biomedical problems. The first of such analysis focuses on the immunoglobulin repertoire of Splenic Marginal Zone Lymphoma (SMZL) patients and investigates the sequence-structure-antigen relationship in this disease. The second biomedical application concerns the analysis of the immunoglobulins of patients affected by Chronic Lymphocytic Leukaemia (CLL). Our findings on the structural differences in the immunoglobulins expressed by patients with favourable and unfavourable prognosis are described. Such results have been included in a paper, prepared in collaboration with Prof. Fabio Ghiotto that has already been submitted for publication. Finally, the investigation of immunoglobulin conformational change upon antigen binding is discussed in chapter 5. A paper describing this work, carried out in collaboration with Prof. Arthur Lesk, is in preparation. A general discussion is given in Chapter 6.

## Chapter 2

---

### **2.1 Structural repertoire of immunoglobulin $\lambda$ light chains**

Improvements in immunoglobulin stability, affinity and suitability as therapeutics are often necessary for clinical and practical application, and they are in turn directly dependent to the knowledge of antibody three-dimensional structure. To this aim, bioinformatics methods come in hand when experimental data are not available. Standard homology modelling techniques are suitable to build models of the well-conserved light and heavy chain framework only. The complementarity-determining regions, which are mostly composed by loops, would have been impossible to properly model in a standard scenario. Luckily, thanks to a large effort of the structural bioinformatics community, it was possible to develop the “canonical structures” method, a sort of dictionary of recurring CDR topologies to be used as a guide in the modelling of such regions. From a structural perspective, the classification of immunoglobulin hypervariable loops in a discrete set of conformation (the so called *canonical structures, CS*) (Al-Lazikani et al., 1997) became an indispensable ingredient for the majority of antibody structure prediction tools. A large amount of available structural data has permitted to define a discrete

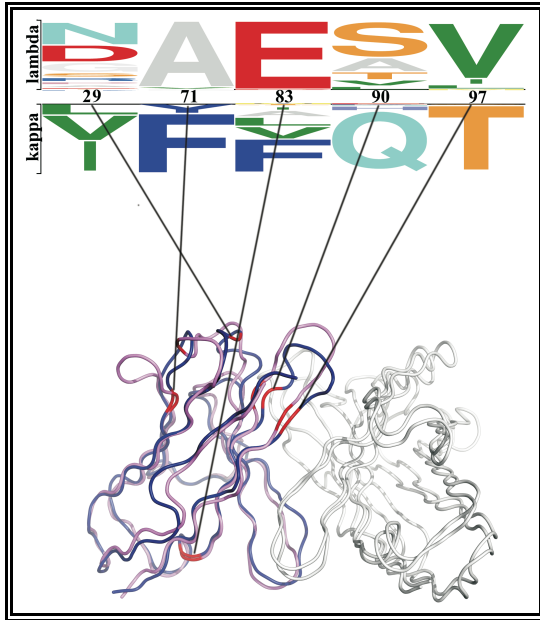


set of recurring topologies adopted by the main chain of hypervariable loops and, in the large majority of cases, sequence based rules to infer such CSs for two out of three heavy chain loops and for all the  $\kappa$  type<sup>1</sup> light chain loops. Instead, a comparably detailed analysis has not been possible for  $\lambda$  chains, mainly because of the limited set of solved structures. Gaining an understanding of the sequence/structure relationship of the hypervariable loops of  $\lambda$  chains is of extreme relevance for both practical and theoretical reasons. To this aim, by means of a novel integrative approach of bioinformatics (structural superposition, sequence alignments, physiochemical analysis of structures), machine learning (clustering, random forests) and statistical (correlation tests, validation of predictions) techniques, we performed a large-scale analysis of the so far observed CSs in  $\lambda$  type light chains. We first defined the main sequence differences with respect to  $\kappa$  light chains by analysing more than 5000 Ig sequences. At this step,  $\kappa/\lambda$  aligned sequences were compared and, using feature selection methods, the residues whose identity best discriminates between the two types of light chains were identified (Figure 5). The majority of such isotype specific positions had already been described as important to maintain the structure of the antigen binding-site (ABS),

---

<sup>1</sup> There are four isotypes of light chains encoded by different loci in vertebrates, only two of which, namely  $\kappa$  and  $\lambda$  types, are present in mammals.

suggesting that the structural diversity between the different isotypes reflects the evolutionary pressure related to their potential different role in the overall immunological response.



**Figure 5.** The logo diagram in the upper panel shows isotype-specific positions, which are mapped on the structure of two antibodies with  $\lambda$  (PDB code 1A6U, in pink) and  $\kappa$  (PDB code 25C8, in blue) light chains after superposition of their heavy chains (in grey) in the lower panel. Residues at position 29, 71, 90, and 97 influence the conformation of the antigen-binding site. The sequence logo was generated using the two-sample logo web server (Vacic et al., 2006).

On the next step, our in-house semi-automatic pipeline allowed us to define for each hypervariable loop of the light chain a set of canonical structures (eight for loop L1, two for L2 and five for L3) and to identify the amino acid positions that are most informative for predicting them from their sequence alone. Of course this analysis has a direct practical application to antibody modelling, but it also pointed at a very interesting observation: the conformational landscape of  $\lambda$  CSs is wider than that of  $\kappa$  chains, i.e. the hypervariable loops adopt a larger and more varied set of conformations. We believe the reason for this lies in the B cell ontogeny. It is thought that  $\lambda$  gene recombination follows, as a sort of “rescue” mechanism, that of  $\kappa$  gene. The different role of the two isotypes is also supported by a bias in the  $\kappa$  chain codon usage towards more volatile codons (i.e. more subject to nonsynonymous mutations) both in the CDR and in the framework regions, favouring the insurgence of more drastic mutations. This suggests that the relatively lower probability of somatic diversity of  $\lambda$  chains is compensated by a larger genetically encoded structural variability that might reflect a different strategy in balancing genetic and somatic diversity adopted by the two light chain types to achieve ABS (antigen-binding site) differentiation.

Interestingly, this diversity is related to subtype evolution: the majority of subtype-

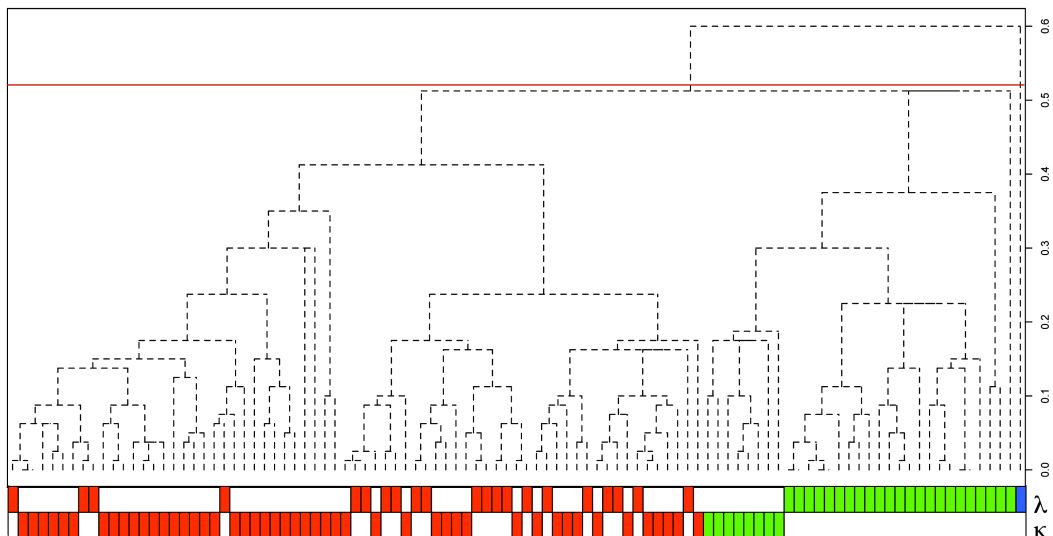
specific positions determined in this analysis play a role in shaping the conformation of the L1 and L3 loops and therefore of the ABS. Their conservation in the subtype germlines strongly suggests that structural diversity is a fundamental important feature driving the evolution of light chains. Moreover, the  $\lambda$  repertoire in human is larger than the corresponding murine one in terms of number of CSs. At last, surprisingly, unlike what is observed for  $\kappa$  and heavy chains, there is a very limited structural similarity between human and mouse CSs, which has obvious implications in selecting the appropriate strategies for humanization of therapeutically relevant antibodies.

## **2.2 The association of heavy and light chain variable domains in antibodies: implications for antigen specificity**

The success in clinical trials using antibodies with low toxicity and high efficiency has raised expectations for the development of next-generation protein therapeutics. Nevertheless, the process of obtaining therapeutic antibodies remains time consuming and empirical, and computational tools play a fundamental role in it. Reliable antibody models are indispensable in order to be able to design and engineer antibodies with improved affinities and physicochemical properties. The known pitfalls in this sense are to be found in these two tasks: (i) modelling the antigen-binding site (discussed in the previous chapter) and (ii) predicting the relative orientations of the variable heavy ( $V_H$ ) and light ( $V_L$ ) chains. At present, when building an immunoglobulin structural model, the models of the two chains, heavy and light, are built independently and next placed with each other. Previous studies (Davies and Metzger, 1983; Mariuzza et al., 1987; Narayanan et al., 2009; Novotny et al., 1983) have shown that the mode of interaction between the heavy and light chain variable domains can modify the relative positions of the hypervariable loops, which in turn, can alter the general shape of antigen-binding

site. In this study, we carried out a comprehensive analysis of the VH-VL interface of several currently available experimentally solved structures of immunoglobulins. This large-scale analysis showed that there are essentially two different modes of interaction of the domains that can be identified by the presence of key amino acids in specific positions of the antibody sequences (Chailyan et al., 2011b). Interestingly, we also found that the different packing modes are related to the size of recognized antigen.

A complex pipeline developed by us allowed analysing the variability of VH-VL domain packing using cluster analysis based on a distance by LGA software. The number of antibody crystal structures with  $\kappa$  light chain in the PDB is much larger than that of antibodies with  $\lambda$  light chain, and thus we used a ‘balanced light chain set’, which included similar numbers of  $V\kappa$  and  $V\lambda$  antibody sequences, respectively. We found that VH-VL pairings can be divided into two major clusters where one cluster is only observed in mouse antibodies and the other is observed both in mouse and in human, indicating the importance of template selection in humanization procedure (Figure 6). We pointed out that these clusters could be assigned from the sequences by looking at some specific positions (L8, L28, L36, L41, L42, L43, L44, L66) in the sequences.



**Figure 6.** Results of the cluster analysis. Dendrogram based on the difference between the positions of residues at the interface in the light and heavy chain variable domains. The red line indicates the clustering with the highest silhouette value (0.47). In the bottom panel, red, green and blue indicate the A, B and C cluster, respectively. The type of light chain is shown in the bottom panel.

An instructive example of the relevance of this study for antibody design can be found in the work by Worn and colleagues (Worn et al., 2000). These authors produced two single-chain Fv humanized intrabody versions of a murine anti-GCN4

immunoglobulin molecule (with a  $\kappa$  chain) using, as recipient, two human antibodies that differed in the type of light chain ( $\kappa$  in one case and  $\lambda$  in the other) and in only seven residues (including residues L36, L43 and L44). The  $\kappa$ -graft variant had an activity comparable with the wild-type antibody, whereas the  $\lambda$ -graft variant, although extraordinarily stable in vitro, had a five order of magnitude decreased antigen affinity, presumably, as the authors suggest, caused by differences in the mutual orientation of the two domains.

Finally, we would like to mention that the ability of type B antibodies to bind smaller antigens, and the presence of the pocket described, might open up the possibility of using them as potential drug delivery vectors. Indeed, this has been proposed already in the case of the 1IND antibody (Love et al., 1993), a type B immunoglobulin with an exceptionally high affinity binding for an indium-chelate hapten.

The ability to use sequence data to predict the mode of association of the variable domains of antibodies also has implications for methods to predict their structure. Indeed, the information obtained through the analysis described here is being used to implement a better prediction protocol in our PIGS immunoglobulin structure prediction server (Marcatili et al., 2008).



## Chapter 3

---

### **3.1 A database of immunoglobulins with integrated tools: DIGIT**

As the amount of biomedical information available in the literature continues to increase, databases that aggregate this information continue to grow in significance and scope. Given the importance of studying the antibody sequence/structure function relationship, the need to have organized data is even more evident because of an overgrowing urgency to gain an overall view of the immune repertoire and of its dynamics. This is of primary importance in many practical applications (e.g. vaccine development) and theoretical breakthroughs (e.g. to find out similarities and differences between B-cell receptors in pathogenic diseases, autoimmune states and lymphoproliferative disorders). This so-called “-omic” perspective of course necessitates all the publicly available data on immunoglobulins to be annotated and organized in a database and ready to be integrated and analysed.

Unfortunately, all the annotations that include information on the type of antigen that an antibody binds, its germline sequences and the pairing between light and

heavy chains were scattered in literature and thus couldn't altogether be retrieved from any supported database. To this aim, taking advantage of our experience and suggestions from our experimental collaborators we constructed a database of immunoglobulin sequences and integrated tools (DIGIT) (Chailyan et al., 2011d), a resource that stores sequences of annotated immunoglobulin variable domains enriched with tools for searching and analysing them. In less than one year it became an extensively used resource by the community (more than 1,300 queries and 800 unique users).

At the current state the database houses 145,759 heavy chain sequences and 71,404 light chain sequences (47168 kappa type and 24236 lambda type) retrieved using isotype-specific HMM profiles developed by us, with assigned canonical structures for the hypervariable loops.

The user can query the database using the antigen type, source organism, accession number, chain type (heavy, lambda, kappa, lambda+kappa) or free text (disease, process etc.) with the option of selecting only complete immunoglobulins (VL+VH). Other annotations are computed on the fly (and therefore can also be obtained for user submitted sequences), for example:

- 1) Numbering of the sequence according to the Kabat numbering scheme
- 2) Identifications of the Complementary Determining Regions (CDRs) in the sequence and of the Framework Regions (FRs)
- 3) Assignment of the canonical structures for the CDRs
- 4) Identification of mutations with respect to the germline
- 5) Automatic link to our 3D modelling tool for immunoglobulin variable domains.
- 6) Sequence searching that, given the input immunoglobulin sequence of interest (amino acid or nucleotide sequence of heavy chain variable domain sequence; light chain variable domain sequence or both), retrieves the closest sequences (sorted according to e-value or % id).

Human annotated complete antibody sequences of the database are approximately 2,500 (i.e. with both light and heavy chains). Among those, more than 1,000 have some information on the antigen, more than 250 are annotated as autoimmune or autoantibody, more than 400 have been sequenced in lymphomas or leukaemias, making possible aforementioned analyses. We are convinced that DIGIT might be extremely useful to researchers interested in immunology as well as to scientists

performing experiments such as antibody humanization, stabilization and functionalization.

## Chapter 4

---

### **4.1 Hierarchical Clustering of B-Cell Receptor Structures in Splenic Marginal Zone Lymphoma**

Our long lasting experience in antibody modelling allowed us to approach two biomedical problems raised in collaborations with Prof. Arcaini (University of Pavia) and Prof. Fabio Ghiotto (University of Genova). More specifically, by applying the developed tools and all our theoretical knowledge we carried out a comprehensive annotation of the repertoires and the characteristics of immunoglobulins from patients with splenic marginal zone lymphoma (SMZL) and chronic lymphocytic leukaemia (CLL) providing novel insights into the molecular mechanisms and the etiology of two the diseases. The following chapter describes in detail the multilayer characterization of the sequence and structure properties of SMZL patient samples and points to both “normal” antigenic and superantigenic simulation.

### **4.1.1 Background**

Splenic marginal zone lymphoma (SMZL) is recognized by the WHO classification as an individual entity, it is frequently associated with HCV infection and autoimmune disorders. SMZL is a disease with currently unknown molecular pathogenesis. Several studies pointed that it may be associated with a chronic antigenic stimulation, in fact previous studies by other groups and us demonstrated that SMZL presents a biased usage of immunoglobulin (IG) heavy variable (IGHV) genes and stereotyped B-cell receptors (BCRs), confirming the likelihood of involvement of antigens and/or superantigens in lymphomagenesis. This characterization, however, is mainly based on the heavy chain alone, even if strong evidences are emerging on the role of light chain (Bikos et al., 2012). The aim of this study was to extend the current analysis of SMZL samples by taking into account Ig light variable genes (IGLV), VL-VH pairing and structural information, thus giving a more detailed view of the immunoglobulin repertoire in terms of sequence, structure and function, and to investigate the sequence-structure-antigen relationship.

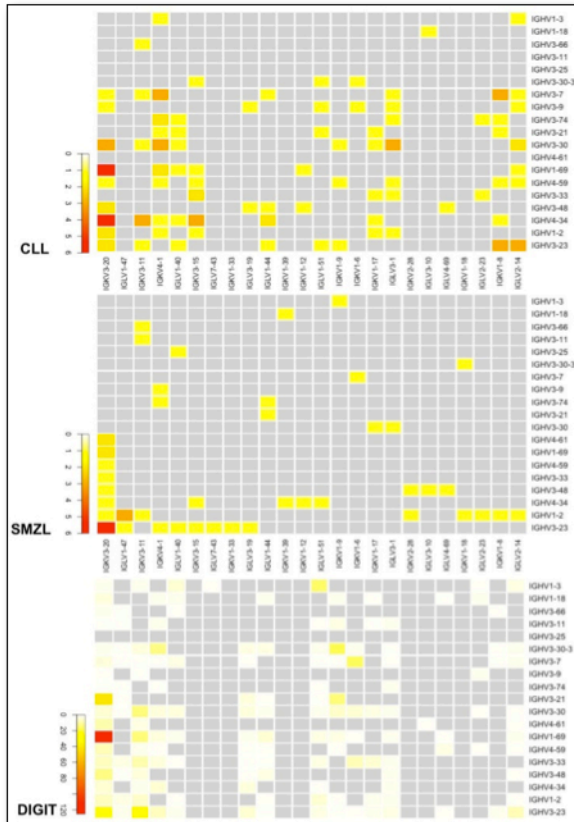
## 4.1.2 Results

In this study for the first time the VL-VH paired sequences of immunoglobulins from SMZL patients diagnosed according to Matutes criteria (Matutes et al., 2008) have been extensively analysed. Sources for analysis were bone marrow in 38 cases and peripheral blood in 14 cases. Ig rearrangements have been characterized by amplification and direct sequencing. Sequences were analysed using the IMGT/DBs and the IMGT/V-QUEST tool. The PIGS web server was used to build 3D models of all antibodies (Abs). The Ab structures were compared using LGA and clustered together according to a score accounting for structure and sequence similarity. Using the DIGIT DB and tools, Igs of all the clusters were analysed and compared to other Igs.

Based on the IGHV nucleotide sequence identity to the germline, 7 sequences (13%) were considered 'truly unmutated' (100% sequence identity), 20 (39%) were 'minimally or borderline mutated' (97-99.9%) whereas 25 (48%) were 'significantly mutated' (<97%). IGHV families were used as follows: IGHV3 (58%), IGHV1 (27%) and IGHV4 (15%). The majority of patients carried kappa light chain (69%). The most frequently used IGKV families were IGKV3 (58%) and IGKV1 (28%), the most frequent IGLV family was IGLV1 (56%). Considering the VL-VH paired

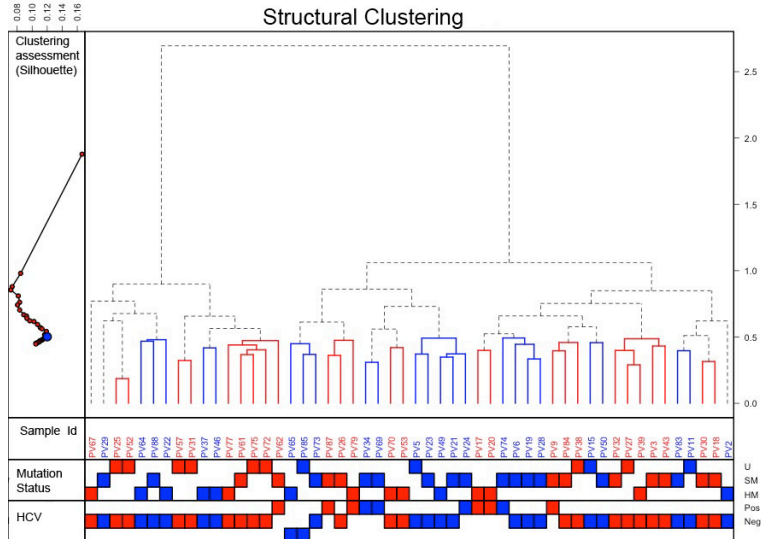
sequences, the two pairings IGHV3-23/IGKV3-20 (n=6) and IGHV1-02/IGLV1-47 (n=3) were significantly over-represented when compared to CLL and DIGIT DB sequences (Figure 7), indicating that the pairing between VL-VH chains was non-random. The IGHV1-02/IGLV1-47 paired sequences showed a high number of somatic mutations (>3%), whereas samples using the IGHV1-02 gene (n=10) but a VL gene other than from IGLV1-47 displayed a low number of mutations, suggesting a significant role for the light chain.





**Figure 7.** Biased usage of VL/VH pairing in SMZL. The closest light chain germline gene is on the X-axes, the closest heavy chain germline gene is on the Y-axes. Colour code: the redder the dot on the cross is the bigger the number of immunoglobulins with that specific VL/VH pairing.

In order to reveal the possible functional role of light chain, we analysed the structural similarity of Ag binding sites (ABSs), performing hierarchical clustering on the similarity obtained by an all-against-all structural superposition of each ABS. Twenty structural clusters were identified (8 with  $\geq 3$  samples) (Figure 8). Igs in the same major groups showed a similar mutation rate, pointing out a likely common Ag selection at least in a fraction of patients. In most cases, Igs in the same clusters display ABSs with similar physicochemical characteristics: positively charged binding sites (2 clusters), hydrophobic patches (3 clusters) or small pockets in the middle of the ABS (3 clusters) that might be clue for different Ags specific for each cluster. HCV infection was found in 1 major and 2 minor clusters (Figure 8), mainly associated with unmutated clones, indicating a likely common antigenic stimulation. In the other major clusters, the role for an Ag-driven selection different from HCV in SMZL lymphomagenesis can be postulated. In particular, 3 clusters, containing both mutated and unmutated samples, displayed a statistically significant similarity to CLL clones ( $p < 0.05$ ), and 1 cluster was structurally similar to autoimmune clones (Kawasaki disease) ( $p = 0.05$ ). Of note, other clusters showed a degree of similarity with samples connected to diseases that involve an Ag-independent or superantigenic stimulation (EBV, Rabies virus, Rotavirus).



**Figure 8.** Hierarchical Clustering of B-Cell Receptor Structures in Splenic Marginal Zone Lymphoma. The bottom panel shows the status of the mutation and HCV prognosis. The number of clusters is defined based on silhouette value.

### 4.1.3 Discussion

The multi-layered characterization of the sequence and structure properties of paired VL-VH in SMZL identified a non-random pairing between heavy and light chains. Structural cluster analysis identified Abs with similar physicochemical properties, similar mutation rate and similar HCV status in a fraction of our dataset.

Comparing Abs of our cases to a large dataset of human annotated Abs derived from the DIGIT DB, a subset resulted similar to CLL or autoimmune clones, whereas other Abs appeared more similar to polyreactive Abs and to Abs possibly targeted by superantigens (Figure 7). These findings could explain the large diversity observed in the Igs expressed in SMZL and provide new insights in SMZL pathogenesis. Future large-scale analysis is anticipated.

## **4.2 Structure variability of immunoglobulins expressed by B cells in chronic lymphocytic leukaemia**

### **4.2.1 Background**

Chronic lymphocytic leukaemia (CLL) is a cancer of the blood cells that affects B-lymphocytes. CLL pathogenesis is still unclear, but several factors contribute to the evolution and the expansion of the mature B-lymphocytes (Chiorazzi et al., 2005). The disease is heterogeneous and some patients progress rapidly having a short survival, whereas others have a more stable clinical course that may not need treatment for years. Interestingly, in ~ 50% of clones, the IGHV and IGL/KV regions of Igs expressed on the surface of leukemic cells undergo somatic hypermutation (SHM) (Fais et al., 1998; Hadzidimitriou et al., 2009), and patients expressing such clones usually have a more indolent clinical courses with respect to those with unmutated ones (Hamblin et al., 1999). Last but not least, CLL clones of different patients may rearrange Igs with remarkable similarities (Ghiotto et al., 2004; Tobin et al., 2003). These so called “stereotyped Igs” are mainly characterized by their HCDR3 amino acid composition and length and recently a large-scale confirmation study of thousands of CLL H chain IGVs showed that the

30% of CLL cases fall within one of more than 300 described subgroups of stereotyped Igs (Agathangelidis et al., 2012; Murray et al., 2008).

All the previously exposed experimental evidences, i.e. the presence of somatic mutations, the biased use of certain *IGHV* genes and the presence of "stereotyped" immunoglobulin complementarity-determining region-3 (*HCDR3*) sequences, indicate that antigenic exposure of B cells might play an important role in the etiology and progression of CLL. If this is the case, even though the HCDR3 plays a crucial role in the antigen-Ig interaction, a comprehensive analysis of the whole ABS structure is expected to give a more complete view of the molecular characteristics of receptors and possibly on the antigen pressure involved in CLL.

Therefore, we analysed here for the first time the structures of Igs from CLL patients, in order to assess whether the structural information on the ABS can provide novel insights into the antigen-driving role in CLL pathogenesis.

#### **4.2.2 Materials and Methods**

We analysed the paired (VL+VH) sequences of 366 patients affected by CLL. For the analysis of Ig structural features, the PIGS server was used to derive the

sequence alignments of the Ig frameworks (using the Kabat-Chothia numbering scheme as reported in (Al-Lazikani et al., 1997)) and to build three-dimensional models of 342 among 366 Igs in our dataset. An all-against-all structural alignment of the antigen-binding sites was performed using the LGA package (Zemla, 2003) in order to group together structurally similar receptors. To this aim, we used several distance measures (RMSD, Global Distance Test (GDT) and Template Modelling (TM) score) and clustering methods (agglomerative hierarchical clustering and divisive hierarchical clustering with a number of clusters ranging from 10 to 50) and selected, by means of the silhouette analysis (PJ, 1987), the one that held to the best cluster topology, i.e. that was able to produce tight and well-separated clusters. The best average silhouette value was obtained with TM-score, divisive clustering method and 21 clusters. The 15 out of 21 clusters containing more than 5 samples were used for all subsequent analyses.

We then performed a comparative analysis of the information of 3D ABS structural model versus that of the linear VH sequence. In order to do this, we built a clustering of the CLL dataset samples (342 patient Igs) based on sequence similarity alone. For both the structure-based and the sequence-based clustering results we generated statistical models of their members in the form of HMMs. All the HMMs

were then used to score a large dataset of 2441 human Igs included in the DIGIT database (Chailyan et al., 2012).

### **4.2.3 Results**

Structural clustering of the 342 CLL Igs according to their ABS structural similarity displays a strong correlation with the Ig mutation status, which is one of the most common prognostic factors in CLL. We used a 2-class partition with a 2% cut-off for defining M (mutated) and U (unmutated) groups to identify the mutation level of Igs. For each cluster we computed the probability that an equal or higher number of Igs belonging to the same class could be found by chance in a randomly extracted subset of the same size (hypergeometric distribution). The Bonferroni-Holm (BH) method was used to correct for multiple testing. We found that M- and U-CLL Igs segregated with a significant over-representation of either M or U Igs in 5 out of 15 clusters thus showing that the structural features of 180 out of the 342 clustered Igs correlate with their mutation status, supporting the hypothesis of an active role of antigen selection.

Apart from the structural similarity, most members of the clusters also shared properties such as type of light chain, stereotypes and in some cases homogeneity



also in terms of the IGHV-IGK/LV usage and H/L chains pairing. Interestingly enough, such structural clusters were CLL specific. If we build statistical models of each cluster and use them to assign a score to samples not present in our dataset we obtain scores significantly higher for CLL samples with respect to non-CLL ones. Additionally, only samples annotated as autoimmune or autoreactive have scores comparable to the CLL ones, in accordance to previous literature (Herve et al., 2005). Such results depend on the structural similarity adopted to cluster the ABSs and not to sequence similarity alone: if the same procedures described above are repeated on a clustering obtained by mere sequence similarity no discriminative power between CLL and non-CLL samples is observed. This result strongly suggests that CLL specific features of the Igs are intrinsic to the atomic structure of the binding site and hints to the involvement of antigenic structures in CLL pathogenesis.

#### **4.2.4 Discussion**

No definite answer has to be found to the question whether the B lymphocyte clones found in CLL patients display a selective specificity for a restricted number of antigens. Support to this hypothesis is provided by the existence of stereotyped IgV rearrangements, even though they are present in a

minority of CLL cases. This is likely due to the definition of stereotyped BCRs, mostly based on the HCDR3 amino acid sequence alone.

In this work we modelled and analysed the structure of CLL Ig ABSs, and we found a limited repertoire of conformations that is CLL specific. When compared to non-CLL samples, this repertoire presented a degree of similarity to autoreactive samples. Clusters of similar receptor often shared a similar mutation rate, suggesting the presence of some functional constraint that is likely to be related to the structure of the binding site.

This is the first time that immunoglobulins from CLL patients are analysed from a structural point of view, and we believe that our results point to the relevance of using this approach on a larger scale, which can now be easily handled by current methodologies for modelling and structural analysis.

The correlation between antigen-binding site structure and clinical outcome, if confirmed, may provide novel tools for a more robust prognostic stratification of CLL also thanks to the fact that sequencing of the IgVL can easily become a standard laboratory test, as it is already the case for IgVH sequencing and the modelling and clustering protocols are very well defined and available.

At present it is essentially impossible to identify the antigen given the structure of the cognate antibody-binding site, but this might change in the future and hopefully we might also be able to gain insight in the nature of the antigens associated with CLL pathogenesis, which would obviously have important applications for therapy.

## Chapter 5

---

# **5.1 Antibody conformational change upon antigen binding**

### **5.1.1 Background**

The immune system is a major area for the studies on molecular recognition, given its ability to produce a finite number of antibodies that must bind a virtually infinite range of foreign molecules. Several experimental and structural studies evidenced that antibodies show flexibility upon antigen binding, which might play an important role in antigen recognition and in immunoglobulin function in general. Conformational change in Abs has been widely recognized, albeit its role in immunoglobulin function and adaptive immunity in general has not been satisfactorily elucidated yet. The limitation created by the small amount of experimentally solved antibody structures has led to several controversial reports. Some of the studies highlight small overall changes, movements of side-chains and changes in the VH-VL relative orientation, whereas others point at large conformational changes and rearrangement of the CDRs. Thus the topic is well discussed in the literature and at the same time is full of open issues.

Basing on their sequence variability, immunoglobulin polypeptide chains can be divided in two functionally distinct regions: the “variable region”, that is responsible for the binding of antigens, and the conserved “constant region”, that is known to participate in signalling events, such as the activation of the complement cascade. The central dogma of immunology claims that the specificity of the antibody is solely a result of the interaction between the immunoglobulin variable region with the antigen, but recent studies highlighted the importance of the constant region in antibody affinity and specificity. For example, Pritsch et al. (Pritsch et al., 1996) found that two human mAbs that share identical variable domains but use different heavy chain isotypes bind with significantly different affinities the same antigen (tubulin). Since the sequence differences observed by the investigators were found at the CH1 domain level, the authors suggested an active role of this domain in the affinity. A similar work by Casadevall (Torres and Casadevall, 2008) showed that four mAb isotypes that share identical variable domains undergo different structural changes upon the binding to common Ag, hinting at structural cross-talk between constant and variable domains.

Thus, all these evidences raise the fundamental question of how the binding of an antigen to an antibody Fab site can trigger reactions in the distant FC domains and

what are the crucial residues that are responsible for the transmission of structural changes from the constant region to the variable region and vice versa.

In this study, we describe a general large-scale analysis of the antibody binding characteristics. By comparing all antibodies with experimentally determined structures both in free and antigen bound form we will provide information about statistically significant structural modifications at a residue level in both variable and constant domains (side-chains movements, site-specific changes, change in the relative orientation of variable domains, changes in the relative orientation of variable and constant domains).

### **5.1.2 Material and Methods**

#### ***Dataset “AB”***

We scanned the sequences of proteins with solved structure present in the PDB database (Dutta et al., 2009) with in-house HMM profiles (for kappa, lambda and heavy chains) to identify immunoglobulin structures (obtaining 1294 molecules). Next, all the structures with unresolved residues in the VH-VL domains have been removed and the resulting dataset has been renumbered according to Kabat-Chothia

(i.e. structurally-derived) numbering scheme. We also removed Bence Jones proteins, heavy chain dimers and immunoglobulins with missing residues indispensable to correctly identify the ABS (Cys 23\_L, Cys 22\_H, Phe 98\_L, Trp 103\_H). We used an in house routine to define whether the immunoglobulin is in apo (unbound) or holo (bound by ligand) state. To this aim we checked if any atom not belonging to the immunoglobulin light or heavy chain is in a 10Å radius from the ABS barycenter. When such atoms were present, the antibody was labelled as “bound” and all the chains to which these atoms belong were considered as “antigens”, with the only exception of small molecules (for which the chain label in the PDB file usually coincides with either the light or heavy chain label) where only the atoms belonging to the molecule itself were added. This procedure allowed us to carry out the study without regard of the type of antigen, thus selecting also haptens that have been recently found to induce conformational change in antibodies. For all the cases in which the antigen identified with the aforementioned routine was an antibody itself we performed a manual checking of the corresponding article and removed the Igs that were not described as anti-idiotypic, considering the contact as an artefact of the crystal structure. We excluded from the analysis all the Igs (28 structures) for which the RMSD calculated on the Cαs of multiple copies after superimposing the core of their variable regions (light chain=20\_L: 98\_L; heavy

chain=20\_H: 101\_H) was larger than 0.6 Å (this value being the average RMSD of multiple copies in our complete dataset plus one standard deviation). Next, the resulting dataset (864 Ig structures) has been culled by resolution (2Å, using the PISCES web server (Wang and Dunbrack, 2003)) and by B-factor, excluding the structures with a B-factor of any backbone atom (C, C $\alpha$ , O, N) more than 80. The final dataset contained 265 immunoglobulin structures. In order to collect all the structures of the same antibody in apo and holo states, we grouped the immunoglobulins according to the sequence similarity. Two structures were considered as being of the same antibody if both of their chains shared 100% sequence identity. This allowed us to obtain 37 groups containing at least 2 Ig molecules, among which 14 contained at least 1 apo and 1 holo solved structure of the same antibody (dataset AB\_VAH, AB variable apo-holo), 20 groups contained the Igs solved just in holo state (dataset AB\_VHH, AB variable holo-holo) and 3 groups were composed of immunoglobulins in apo state. Finally, we performed all the possible apo-holo and holo-holo comparisons inside each group (for example, if the group contains 1 Ig in apo state and 3 Igs in holo state, we perform 3 comparisons of Igs in apo-holo state and 3 comparison of the Igs in holo-holo state), obtaining 25 apo/holo pairs in the AB\_VAH dataset and 82 holo/holo pairs in the AB\_VHH dataset.



### ***Dataset “AB\_constant”***

Starting from the AB\_VAH and AB\_VHH datasets we derived two datasets composed of immunoglobulins with identical variable and constant domains (first constant domain of the heavy chain, CH1 and constant domain of the light chain, CL) sequences. First, we removed from the AB\_VAH and AB\_VHH datasets the Igs that have missing residues in the constant domains. Next, similarly to what we did for the variable domains, we grouped the immunoglobulins by considering two structures to be of the same antibody if both their chains shared 100% sequence identity. The final AB\_CAH (dataset AB constant apo-holo) dataset contained 17 apo-holo pairs derived from 10 groups of different experimental structures of the same antibody, whereas the AB\_CHH (dataset AB constant holo-holo) dataset contained 24 pairs from 8 groups. All the datasets with the corresponding number of immunoglobulins and pairs are presented in Table 1.

**Table 1.** The AB dataset content.

	<b>VAH</b>	<b>VHH</b>	<b>CAH</b>	<b>CHH</b>
	<i>Variable region</i>		<i>Constant region</i>	
<b>Number of groups</b>	14 Igs	20 Igs	10 Igs	8 Igs
<b>Number of pairs</b>	25 pairs	82 pairs	17 pairs	24 pairs

## **Structural superposition**

### ***Variable region “AB” dataset, Constant region “AB\_constant” dataset***

The analysis described below was performed separately for the variable and constant domains. All the immunoglobulins of the four datasets (AB\_VAH, AB\_VHH, AB\_CAH, AB\_CHH) were renumbered according to the Kabat-Chothia numbering scheme in order to identify the corresponding positions. For each pair of the AB\_VAH and AB\_VHH variable domain datasets we performed an all-against-all structural superposition of the whole variable domain comprising residues 1-110 for the VL region and residues 1-113 for the VH region. The structural superposition has been carried out by LGA package with the following parameters: sequence dependent analysis and cut-off threshold of 10 Å. For each structural

alignment we calculated the C $\alpha$  root-mean-square deviation (RMSD) per each position of the variable domain sequence.

Similarly to what has been done for variable domain, we performed structural superposition of the whole constant domain (CL=> 110-210; CH1=> 113-213) independently from variable domains with subsequent C $\alpha$  RMSD calculation.

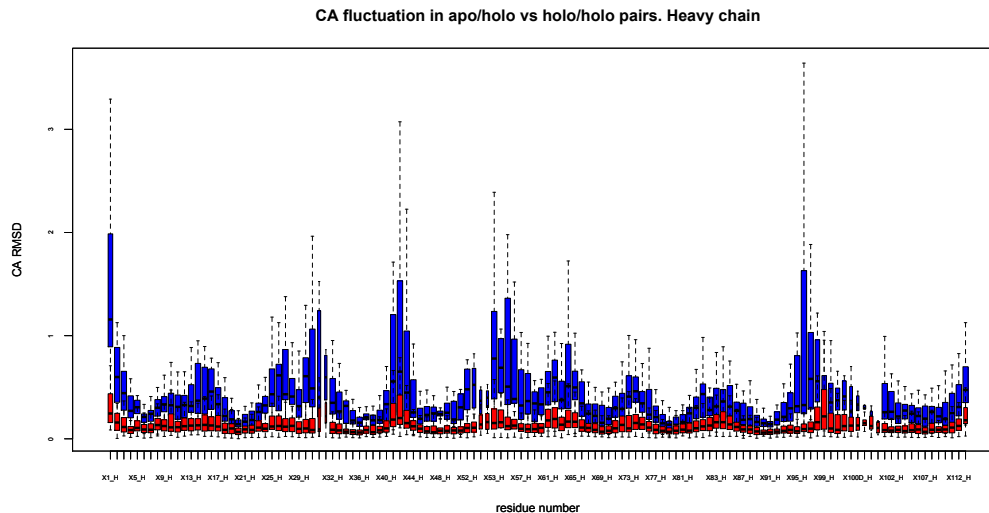
**Table 2.** The loop regions with corresponding residues according to Kabat-Chothia numbering scheme.

	Residues
L1 loop	24-34
L2 loop	48-54
L3 loop	89-98
H1 loop	24-34
H2 loop	51-57
H3 loop	94-103

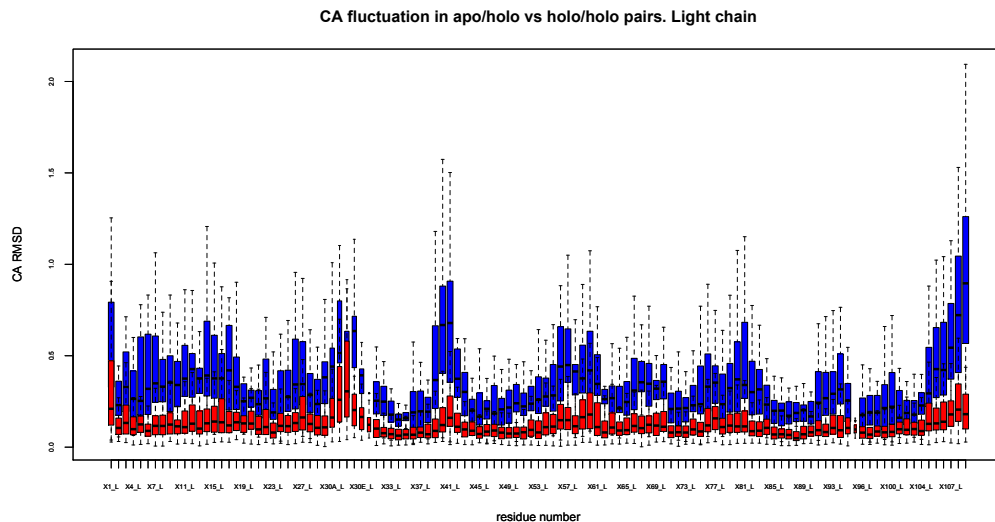
### 5.1.3 Results

An increased number of solved structures allowed us to investigate the hypothesis of conformational changes upon antigen binding by means of a comparative analysis of free and bound forms of the same antibody. In this analysis we used pairs of structures of the same antibody determined with or without its corresponding antigens. The dataset used for the analysis of variable domains consists of 100 PDB structures of 34 different antibodies. We performed not only bound-unbound comparisons, but also bound-bound comparisons of the same antibody to provide a control that allows determining whether the change is unequivocally determined by antigen binding. Overall this dataset consists of 25 bound-unbound and 82 bound-bound pairs of structures of the same antibody. We identified changes occurring at a residue level by performing a structural superposition of bound-unbound and bound-bound pairs and calculated the RMSD for  $C\alpha$  atoms of structurally equivalent residues. The results of this analysis averaged for all the immunoglobulins of the dataset are shown in Figures 9 (heavy chain) and 10 (light chain).  $C\alpha$  fluctuation of each residue of bound-unbound pairs is shown in blue, whereas of bound-bound pairs are depicted in red. In almost all positions, the RMSD calculated between free and bound structures is greater than

the RMSD between the two bound structures. Besides the hypervariable loops that showed flexibility as a result of direct interaction with antigen, we identified two loop regions far from ABS on the heavy (residues 40-44) and light chain (41-45) variable domains, that interact with each other and are part of heavy-light interface.



**Figure 9.**  $C\alpha$  fluctuation in apo/holo (blue) vs holo/holo (red) immunoglobulin pairs. Heavy chain variable domain. X-axis residue number, Y-axis RMSD of  $C\alpha$  atoms between apo/holo and holo/holo Ig pairs.

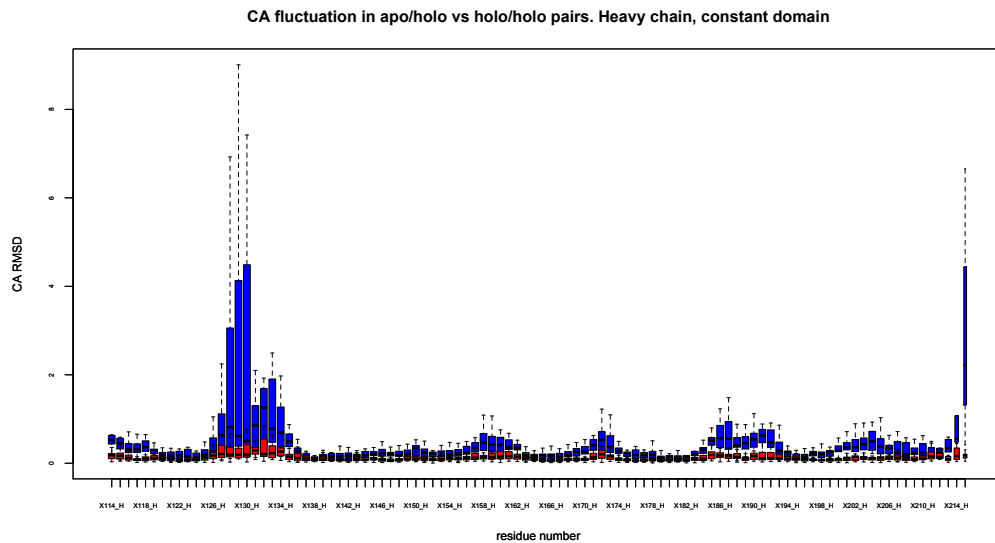


**Figure 10.**  $C\alpha$  fluctuation in apo/holo (blue) vs holo/holo (red) immunoglobulin pairs. Light chain variable domain. X-axis residue number, Y-axis RMSD of  $C\alpha$  atoms between apo/holo and holo/holo Ig pairs.

Taking into the account recent evidences of the constant region importance in Ag binding, we analysed also the first constant regions of immunoglobulin heavy and light chains. In this case the pairs of structures that share 100% sequence identity in both variable and constant domains of H and L chains (first constant heavy and constant of light chain) have been used. For this analysis 50 PDB structures

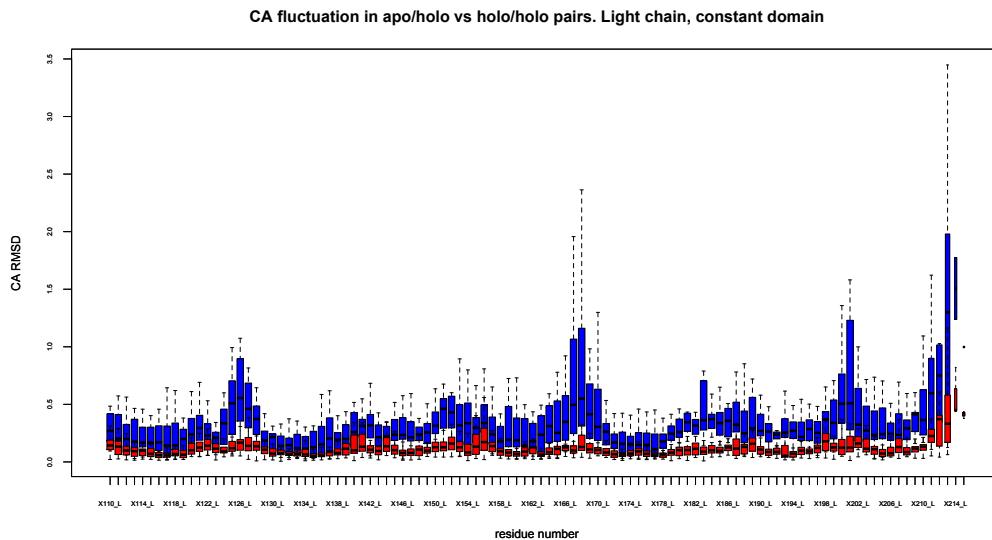
representing 18 different antibodies make up the dataset. As before, we compared the differences noticed in the bound-unbound superpositions to those observed in bound-bound superpositions of the same antibody to provide a control that allows determining whether the change is related to antigen binding. Thus, the overall number of bound-unbound pairs in this analysis was 17 whereas the number of bound-bound pairs was 24. As for the variable domain, we identified changes occurring at the residue level by performing a structural superposition of bound-unbound and bound-bound pairs and calculated the RMSD for  $C\alpha$  atoms of structurally equivalent residues. The results of this analysis averaged for all the immunoglobulins of the dataset are shown in Figures 11 (heavy chain) and 12 (light chain).  $C\alpha$  fluctuation of each residue of bound-unbound pairs is shown in blue and bound-bound pairs are depicted in red. The RMSD calculated between free and bound structures was greater than the RMSD between the two bound structures, indicating that these changes are related to Ag binding. This analysis allowed identifying a loop region in CH1 domain of the heavy chain (residues 126-138) that undergoes substantial changes upon Ag binding (Figure 11). Since this loop, implicated in the interaction between H and L chains, has been described as being intrinsically disordered in specific conditions and is involved in complement binding, we think it may have a role in Ab function. We are currently performing

further work on identifying the structural bases of the signal transduction from constant to variable region and the role of Ag type in it. Similar analysis on side chains, analysis of the heavy-light chain relative orientation and CDR comparisons are in progress.



**Figure 11.** C $\alpha$  fluctuation in apo/olo (blue) vs holo/olo (red) immunoglobulin pairs. Heavy chain constant domain. X-axis residue number, Y-axis RMSD of C $\alpha$  atoms between apo/olo and holo/olo Ig pairs.





**Figure 12.** C $\alpha$  fluctuation in apo/holo (blue) vs holo/holo (red) immunoglobulin pairs. Light chain constant domain. X-axis residue number, Y-axis RMSD of C $\alpha$  atoms between apo/holo and holo/holo Ig pairs.

## Chapter 6

---

### **6.1 Conclusions and outlook**

Nowhere is the importance of complex dynamics and architectures clearer than in biological systems. Novel experimental techniques and data give us the opportunity to explore this complex world, with the aid of different frameworks, models and tools since in biology, in most of the cases, complex questions require complex answers. Even when considering a single protein family, the immunoglobulins, such complexity cannot (or at least not yet) be reduced to a single framework. In order to gain useful information, one has to integrate interactions with expression data, structures, mutation experiments, medical information and sequences. Integrating all these data and methods in a biologically meaningful framework is an extremely hard though exciting task. In this work I present a number of these methods, developed by our group and published in the papers reported in the final section of this thesis, that share the common line of switching from the single-case analyses of Igs that has been carried out since now to a systematic analysis.

The first part of the thesis describes the theoretical tools developed to analyse the antibodies in a comprehensive way. These tools were used to help the antibody

modelling and to develop a database of immunoglobulins. Whereas, the second part demonstrates how these tools and theoretical knowledge let us give a comprehensive description of the repertoires and the characteristics of immunoglobulins in two diseases (CLL and SMZL) that confirmed the results of previous studies and revealed new features that can be useful as diagnostic markers in the prognosis of these diseases and hopefully shed a light on its causative agents.

The role of immunoglobulin molecules is quite well known, nevertheless some basic questions are still open and it is never quite straightforward to find the answers since its understanding is often a hard if not impossible task. We hope that, in this work, we were able to give an at least partial picture of this scenario and to answer some biologically relevant questions.

## Reference list

Abhinandan, K.R., and A.C. Martin. 2008. Analysis and improvements to Kabat and structurally correct numbering of antibody variable domains. *Mol Immunol.* 45:3832-3839.

Agathangelidis, A., N. Darzentas, A. Hadzidimitriou, X. Brochet, F. Murray, X.J. Yan, Z. Davis, E.J. van Gastel-Mol, C. Tresoldi, C.C. Chu, N. Cahill, V. Giudicelli, B. Tichy, L.B. Pedersen, L. Foroni, L. Bonello, A. Janus, K. Smedby, A. Anagnostopoulos, H. Merle-Beral, N. Laoutaris, G. Juliusson, P.F. di Celle, S. Pospisilova, J. Jurlander, C. Geisler, A. Tsaftaris, M.P. Lefranc, A.W. Langerak, D.G. Oscier, N. Chiorazzi, C. Belessi, F. Davi, R. Rosenquist, P. Ghia, and K. Stamatopoulos. 2012. Stereotyped B-cell receptors in one-third of chronic lymphocytic leukemia: a molecular classification with implications for targeted therapies. *Blood.* 119:4467-4475.

Al-Lazikani, B., A.M. Lesk, and C. Chothia. 1997. Standard conformations for the canonical structures of immunoglobulins. *J Mol Biol.* 273:927-948.

Almagro, J.C., M.P. Beavers, F. Hernandez-Guzman, J. Maier, J. Shaulsky, K. Butenhof, P. Labute, N. Thorsteinson, K. Kelly, A. Teplyakov, J. Luo, R. Sweet, and G.L. Gilliland. 2011. Antibody modeling assessment. *Proteins.* 79:3050-3066.

Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235-242.

Bikos, V., E. Stalika, P. Baliakas, N. Darzentas, Z. Davis, A. Traverse-Glehen, A. Dagklis, G. Kanellis, A. Anagnostopoulos, A. Tsaftaris, M. Ponzoni, F. Berger, P.

Felman, P. Ghia, T. Papadaki, D. Oscier, C. Belessi, and K. Stamatopoulos. 2012. Selection of antigen receptors in splenic marginal-zone lymphoma: further support from the analysis of the immunoglobulin light-chain gene repertoire. *Leukemia*. 26:2567-2569.

Chailyan, A., P. Marcatili, D. Cirillo, and A. Tramontano. 2011a. Structural repertoire of immunoglobulin lambda light chains. *Proteins*. 79:1513-1524.

Chailyan, A., P. Marcatili, and A. Tramontano. 2011b. The association of heavy and light chain variable domains in antibodies: implications for antigen specificity. *FEBS J*. 278:2858-2866.

Chailyan, A., P. Marcatili, and A. Tramontano. 2011c. The association of heavy and light chain variable domains in antibodies: implications for antigen specificity. *Febs Journal*. 278:2858-2866.

Chailyan, A., A. Tramontano, and P. Marcatili. 2011d. A database of immunoglobulins with integrated tools: DIGIT. *Nucleic Acids Res*.

Chailyan, A., A. Tramontano, and P. Marcatili. 2012. A database of immunoglobulins with integrated tools: DIGIT. *Nucleic Acids Res*. 40:D1230-1234.

Chiorazzi, N., K.R. Rai, and M. Ferrarini. 2005. Chronic lymphocytic leukemia. *N Engl J Med*. 352:804-815.

Chothia, C., and A.M. Lesk. 1987. Canonical structures for the hypervariable regions of immunoglobulins. *J Mol Biol*. 196:901-917.

Davies, D.R., and H. Metzger. 1983. Structural basis of antibody function. *Annu Rev Immunol*. 1:87-117.

Dutta, S., K. Burkhardt, J. Young, G.J. Swaminathan, T. Matsuura, K. Henrick, H. Nakamura, and H.M. Berman. 2009. Data deposition and annotation at the worldwide protein data bank. *Mol Biotechnol.* 42:1-13.

Fais, F., F. Ghiotto, S. Hashimoto, B. Sellars, A. Valetto, S.L. Allen, P. Schulman, V.P. Vinciguerra, K. Rai, L.Z. Rassenti, T.J. Kipps, G. Dighiero, H.W. Schroeder, Jr., M. Ferrarini, and N. Chiorazzi. 1998. Chronic lymphocytic leukemia B cells express restricted sets of mutated and unmutated antigen receptors. *J Clin Invest.* 102:1515-1525.

Fox, B.G., C. Goulding, M.G. Malkowski, L. Stewart, and A. Deacon. 2008. Structural genomics: from genes to structures with valuable materials and many questions in between. *Nature Methods.* 5:129-132.

Ghiotto, F., F. Fais, A. Valetto, E. Albesiano, S. Hashimoto, M. Dono, H. Ikematsu, S.L. Allen, J. Kolitz, K.R. Rai, M. Nardini, A. Tramontano, M. Ferrarini, and N. Chiorazzi. 2004. Remarkably similar antigen receptors among a subset of patients with chronic lymphocytic leukemia. *J Clin Invest.* 113:1008-1016.

Hadzidimitriou, A., N. Darzentas, F. Murray, T. Smilevska, E. Arvaniti, C. Tresoldi, A. Tsaftaris, N. Laoutaris, A. Anagnostopoulos, F. Davi, P. Ghia, R. Rosenquist, K. Stamatopoulos, and C. Belessi. 2009. Evidence for the significant role of immunoglobulin light chains in antigen recognition and selection in chronic lymphocytic leukemia. *Blood.* 113:403-411.

Hamblin, T.J., Z. Davis, A. Gardiner, D.G. Oscier, and F.K. Stevenson. 1999. Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. *Blood.* 94:1848-1854.

Herve, M., K. Xu, Y.S. Ng, H. Wardemann, E. Albesiano, B.T. Messmer, N. Chiorazzi, and E. Meffre. 2005. Unmutated and mutated chronic lymphocytic leukemias derive from self-reactive B cell precursors despite expressing different antibody reactivity. *J Clin Invest.* 115:1636-1643.

Honegger, A., and A. Pluckthun. 2001. Yet another numbering scheme for immunoglobulin variable domains: An automatic modeling and analysis tool. *Journal of Molecular Biology.* 309:657-670.

Kabat, E.A., and T.T. Wu. 1991. Identical V region amino acid sequences and segments of sequences in antibodies of different specificities. Relative contributions of VH and VL genes, minigenes, and complementarity-determining regions to binding of antibody-combining sites. *J Immunol.* 147:1709-1719.

Lefranc, M.P. 1999. The IMGT unique numbering for immunoglobulins, T-cell receptors, and Ig-like domains. *Immunologist.* 7:132-136.

Love, R.A., J.E. Villafranca, R.M. Aust, K.K. Nakamura, R.A. Jue, J.G. Major, R. Radhakrishnan, and W.F. Butler. 1993. How the Anti-(Metal Chelate) Antibody Cha255 Is Specific for the Metal-Ion of Its Antigen - X-Ray Structures for 2 Fab' Hapten Complexes with Different Metals in the Chelate. *Biochemistry.* 32:10950-10959.

Marcatili, P., A. Rosi, and A. Tramontano. 2008. PIGS: automatic prediction of antibody structures. *Bioinformatics.* 24:1953-1954.

Mariuzza, R.A., S.E. Phillips, and R.J. Poljak. 1987. The structural basis of antigen-antibody recognition. *Annu Rev Biophys Biophys Chem.* 16:139-159.

Martin, A.C.R., J.C. Cheetham, and A.R. Rees. 1989. Modeling Antibody Hypervariable Loops - a Combined Algorithm. *Proceedings of the National Academy of Sciences of the United States of America*. 86:9268-9272.

Matutes, E., D. Oscier, C. Montalban, F. Berger, E. Callet-Bauchu, A. Dogan, P. Felman, V. Franco, E. Iannitto, M. Mollejo, T. Papadaki, E.D. Remstein, A. Salar, F. Sole, K. Stamatopoulos, C. Thieblemont, A. Traverse-Glehen, A. Wotherspoon, B. Coiffier, and M.A. Piris. 2008. Splenic marginal zone lymphoma proposals for a revision of diagnostic, staging and therapeutic criteria. *Leukemia*. 22:487-495.

Morea, V., A. Tramontano, M. Rustici, C. Chothia, and A.M. Lesk. 1997a. Antibody structure, prediction and redesign. *Biophysical Chemistry*. 68:9-16.

Morea, V., A. Tramontano, M. Rustici, C. Chothia, and A.M. Lesk. 1997b. Antibody structure, prediction and redesign. *Biophys Chem*. 68:9-16.

Morea, V., A. Tramontano, M. Rustici, C. Chothia, and A.M. Lesk. 1998. Conformations of the third hypervariable region in the VH domain of immunoglobulins. *J Mol Biol*. 275:269-294.

Murray, F., N. Darzentas, A. Hadzidimitriou, G. Tobin, M. Boudjogra, C. Scielzo, N. Laoutaris, K. Karlsson, F. Baran-Marzsak, A. Tsaftaris, C. Moreno, A. Anagnostopoulos, F. Caligaris-Cappio, D. Vaur, C. Ouzounis, C. Belessi, P. Ghia, F. Davi, R. Rosenquist, and K. Stamatopoulos. 2008. Stereotyped patterns of somatic hypermutation in subsets of patients with chronic lymphocytic leukemia: implications for the role of antigen selection in leukemogenesis. *Blood*. 111:1524-1533.



- Narayanan, A., B.D. Sellers, and M.P. Jacobson. 2009. Energy-based analysis and prediction of the orientation between light- and heavy-chain antibody variable domains. *J Mol Biol.* 388:941-953.
- North, B., A. Lehmann, and R.L. Dunbrack, Jr. 2011. A new clustering of antibody CDR loop conformations. *J Mol Biol.* 406:228-256.
- Novotny, J., R. Bruccoleri, J. Newell, D. Murphy, E. Haber, and M. Karplus. 1983. Molecular anatomy of the antibody binding site. *J Biol Chem.* 258:14433-14437.
- Pedotti, M., L. Simonelli, E. Livoti, and L. Varani. 2011. Computational Docking of Antibody-Antigen Complexes, Opportunities and Pitfalls Illustrated by Influenza Hemagglutinin. *International Journal of Molecular Sciences.* 12:226-251.
- PJ, R. 1987. Silhouettes – a graphical aid to the interpretation and validation of cluster-analysis. *J Comput Appl Math.* 20:53-65.
- Pritsch, O., G. Hudry-Clergeon, M. Buckle, Y. Petillot, J.P. Bouvet, J. Gagnon, and G. Dighiero. 1996. Can immunoglobulin C(H)1 constant region domain modulate antigen binding affinity of antibodies? *J Clin Invest.* 98:2235-2243.
- Sircar, A., E.T. Kim, and J.J. Gray. 2009. RosettaAntibody: antibody variable region homology modeling server. *Nucleic Acids Res.* 37:W474-479.
- Tobin, G., U. Thunberg, A. Johnson, I. Eriksson, O. Soderberg, K. Karlsson, M. Merup, G. Juliusson, J. Vilpo, G. Enblad, C. Sundstrom, G. Roos, and R. Rosenquist. 2003. Chronic lymphocytic leukemias utilizing the VH3-21 gene display highly restricted Vlambda2-14 gene use and homologous CDR3s: implicating recognition of a common antigen epitope. *Blood.* 101:4952-4957.

Torres, M., and A. Casadevall. 2008. The immunoglobulin constant region contributes to affinity and specificity. *Trends Immunol.* 29:91-97.

Vacic, V., L.M. Iakoucheva, and P. Radivojac. 2006. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics.* 22:1536-1537.


Wang, G., and R.L. Dunbrack, Jr. 2003. PISCES: a protein sequence culling server. *Bioinformatics.* 19:1589-1591.

Worn, A., A.A. der Maur, D. Escher, A. Honegger, A. Barberis, and A. Pluckthun. 2000. Correlation between in vitro stability and in vivo performance of anti-GCN4 intrabodies as cytoplasmic inhibitors. *Journal of Biological Chemistry.* 275:2795-2803.

Wu, T.T., and E.A. Kabat. 1970. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *The Journal of experimental medicine.* 132:211-250.

Zemla, A. 2003. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res.* 31:3370-3374.

## Paper I



**PROTEINS**  
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Research Article

**Structural repertoire of immunoglobulin  $\lambda$  light chains<sup>†</sup>**

Anna Chailyan<sup>1,‡</sup>, Paolo Marcatili<sup>1,‡,\*</sup>, Davide Cirillo<sup>1</sup>, Anna Tramontano<sup>1,2</sup> **Issue**

Article first published online: 1 MAR 2011  
DOI: 10.1002/prot.22979  
Copyright © 2011 Wiley-Liss, Inc.

Proteins: Structure, Function, and Bioinformatics  
Volume 79, Issue 5, pages 1513–1524, May 2011

**Link:** <http://onlinelibrary.wiley.com/doi/10.1002/prot.22979/full>

## Paper II



The screenshot shows a journal article page with a teal header containing the text "the FEBS Journal". The article title is "The association of heavy and light chain variable domains in antibodies: implications for antigen specificity", accompanied by a "FREE" icon. The authors listed are Anna Chailyan<sup>1,†</sup>, Paolo Marcatili<sup>1,†</sup>, and Anna Tramontano<sup>1,2</sup>. The article was first published online on 28 JUN 2011, with a DOI of 10.1111/j.1742-4658.2011.08207.x. The copyright notice states "© 2011 The Authors Journal compilation © 2011 FEBS". To the right of the text is a section labeled "Issue" with a horizontal line. Below this is a thumbnail image of the journal cover, which features a molecular structure and the text "the FEBS Journal". To the right of the thumbnail, the text reads "FEBS Journal Volume 278, Issue 16, pages 2858–2866, August 2011".

the FEBS Journal

**The association of heavy and light chain variable domains in antibodies: implications for antigen specificity** 

Anna Chailyan<sup>1,†</sup>, Paolo Marcatili<sup>1,†</sup>, Anna Tramontano<sup>1,2</sup>

Article first published online: 28 JUN 2011  
DOI: 10.1111/j.1742-4658.2011.08207.x  
© 2011 The Authors Journal compilation © 2011 FEBS

Issue



FEBS Journal  
Volume 278, Issue 16, pages 2858–2866, August 2011

**Link:** <http://onlinelibrary.wiley.com/doi/10.1111/j.1742-4658.2011.08207.x/full>

## Paper III

OXFORD JOURNALS

# Nucleic Acids Research

**A database of immunoglobulins with integrated tools: DIGIT**

Anna Chailyan<sup>1</sup>, Anna Tramontano<sup>1,2,\*</sup> and Paolo Marcatili<sup>1,\*</sup>

**This Article**

Nucl. Acids Res. (2011)  
doi: 10.1093/nar/gkr806  
First published online: November 10, 2011

**Link:** <http://nar.oxfordjournals.org/content/early/2011/11/10/nar.gkr806.full.pdf+html>