Computational design of peptide ligands based on antibody-antigen interface properties

By

Benjamin Thomas VIART

Thesis Advisor
Prof. Dra. Liza Figueiredo Felicori Vilela

Thesis Co-advisor
Dr. Franck Molina

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Bioinformática
Avenida Presidente Antônio Carlos, 6627
Pampulha 31270-901
Belo Horizonte - MG

To my parents,

Acknowledgments

- Firstly, I would like to express my sincere gratitude to my advisor Dr. Prof. Liza Figueiredo Felicori Vilela for the continuous support of my Ph.D study, for her patience, motivation and knowledge. Her advices and guidance helped me better understand the scientific gait, the rules of research, the importance of the details. I would like to thank her as well for making feel welcome in Brazil, helping me with the transition to this new culture that is now part of me as the French culture is part of her.

- Secondly, I would like to deeply thank my co-advisor Dr Franck Molina, without whom I would never have done this PhD. I am forever grateful to him for believing in me and helping me achieve this goal.

- From the department, I would like to thanks Prof. Jader, Prof. Vasco, Prof. Miguel, Prof. Lucas, Prof. Gloria as well as Sheila.

- This thesis is dedicated to my parents, Frédéric and Françoise Viart, who helped me throughout my life to become a good person. I thank them for their unconditional love and support.

- I would like to thank all my family, especially my sisters, Sophie and Lisa, for their smiles and grimaces; my grand-parents, Bob, Nanou and Mémé for their support and love and also my uncle, Bruno, for all his help when I was in first year of medical school.

- I would like to thank my girlfriend, Irene Benevides, for all she has done for me, all her advices, all her kindness, all her Portuguese lessons and so much more but ultimately

for her love that makes me happy. I love you.

- I am thankful to all of the Benevides Dutra Murta family and especially Ianara for welcoming me with so much respect and attention, for taking care of me and helping me throughout those years.

- I would like to thank all my friends from here, Edgar: for his advices and his partnership at Starcraft II, Loic: for our debates and discussion about so much things, Flavio (aka Pivete): for being so open minded and accessible, Graziele: for the climbing and Melissa for our conversations about the administrative nightmare. Thanks also to Juliana, Leidiane, Maina, Medhi, Jessica and Chedy among others.

- I would like to think all my friends from abroad, with whom I have kept contact despite the distance and the time-shift, Raphael, Thierry, Christophe, Elise, Aude, Sebastien, Emeline and JP.

Computational design of peptide ligands based on antibody-antigen interface properties

Benjamin VIART

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Bioinformática
Avenida Presidente Antônio Carlos, 6627
Pampulha 31270-901
Belo Horizonte â MG

ABSTRACT

Antibodies are an important class of biological drugs, but suffer from various limitations, such as inadequate pharmacokinetics, humanization protocol, adverse immunogenicity and high production costs. Synthetic peptides with high affinity and specificity for the desired target represent an important alternative to antibodies but the design of such peptides is limited by our knowledge of the antibody-antigen interface complementarity mechanisms.

The rapid expansion of the available antibody-antigen complex structure have made the study of those complex a major way to gain insight into the interface properties. To identify the interacting residues in a given antibody-antigen interface we used Interface Interacting Residue (I2R), a selection method based on computed molecular interactions. This new selection allowed us to assess other interface selection techniques and compare them. To store all the data such as the structure, epitope and paratope computed from the different selection methods and their properties we created the Epitope-Paratope Interface DataBase (EPI-DB) specially dedicated to study complementarity of Ab-Ag interface using structural and physicochemical properties. Using ensemble of linear model prediction based on physicochemical properties we were able to assess if a pair of epitope-paratope (both sequence coming from the same complex structure) was mismatched or not and achieved an area under the curve of 0.7633

showing the capacity of an epitope properties to predict paratope characteristics

The aggregation of all the molecular interactions between epitope and paratope residues allowed us to transform the 3D antibody-antigen complex structures into interface graphs. Based on these data and the probability of molecular interaction we developed EPI-Peptide Designer tool that uses predicted paratope residues for an epitope of interest to generate targeted peptide ligand libraries. EPI-Peptide Designer successfully predicted 301 peptides able to bind to LiD1 target protein (65% of the experimentally tested peptides). This tool should enable the development of a new generation of synthetic interacting peptides that could be very useful in the biosensor, diagnostic and therapeutic fields.

In addition, to further understand the complementarity mechanisms from Ab-Ag, we investigated the differences between Human antibodies and mice antibodies. Using the data available in EPI-DB we compared epitope and paratope according to the organism source and found that paratope of mice antibodies contains 5% more tyrosine than the Human one. Using linear model we found possible to predict if a protein epitope is complexed with Human Ab of mice Ab only using its secondary structure.

All of those results helps us better understand the complementarity mechanisms of the antibody antigen interface and will help improve peptide binders design and overpass some drug antibodies limitations.

Table of Contents

List of Figures

List of Tables

Glossary

- **%Occ** Percentage of occurrence
- **Å** Ångström
- **∆SAS** Difference in Solvent Accessible Surface
- **3D** Three Dimension
- **68Ga** Gallium 68
- **111In** Indium 111
- **Ab** Antibody
- **ABR** Antibody Binding Regions
- **Ag** Antigen
- **AUC** Area Under the Curve
- **CDR** Complementary Determining Region
- **DB** DataBase
- **DNA** DeoxyriboNucleic Acid
- **DBS** Distance Base Selection
- **ECS** Epitope Containing Sequence
- **EPI** Epitope Paratope Interface
- **FDA** Food and Drug Administration
- **FEP** Free energy perturbation
- **FPR** False Positive Rate
- **HER2** Human Epidermal growth factor Receptor 2
- **I2R** Interface Interacting Residue
- **IgG** Immunoglobulin G
- **IGG** Interface Graph Generators
- **IMGT** ImMunoGeneTics
- **IRA** Interface Research Algorithm
- **MAb** Monoclonal Antibody
- **mRNA** Ribonucleic acid Messenger
- **PCP** Physicochemical Properties
- **PCS** Paratope Containing Sequence
- **PET** Position Emission Tomographic
- **pM** Pico molar
- **PPI** Protein Protein Interface
- **RAFT** Regioselectively Addressable Functionalized Templates
- **ROC** Receiver Operating Characteristic
- **Sc** Shape Complementarity
- **ScVf** Single-chain Fragment variable

- **SQL** Structured Query Language
- **SS** Secondary Structure
- **SyAM** Synthetic Antibody Mimics
- **TPR** True Positive Rate

Amino acid

Amino acid	3-letter code	1-letter code	Group
Alanine	Ala	A	Small
Arginine	Arg	R	Positivelycharged
Asparagine	Asn	N	Polar
Asparticacid	Asp	D	Negativelycharged
Cysteine	Cys	C	Hydrophobic
Glutamicacid	Glu	E	Negativelycharged
Glutamine	Gln	Q	Polar
Glycine	Gly	G	Small
Histidine	His	H	Positivelycharged
Isoleucine	Ile	I	Hydrophobic
Leucine	Leu	L	Hydrophobic
Lysine	Lys	K	Positivelycharged
Methionine	Met	M	Hydrophobic
Phenylalanine	Phe	F	Aromatic
Proline	Pro	P	Hydrophobic
Serine	Ser	S	Alcohol
Threonine	Thr	T	Polar
Tryptophan	Trp	W	Aromatic
Tyrosine	Tyr	Y	Aromatic
Valine	Val	V	Hydrophobic

Table 1: All natural amino acids name, three and one letter code and group

Databases, webtools and other computational resources

Type: Databank	**Protein Databank, PDB**
Description:	The Protein Data Bank is the single worldwide archive of structural data of biological macromolecules. Berman et al. (2000)

Type: Database	**3D interacting domain, 3did**
Description:	The database of three-dimensional interacting domains is a collection of high-resolution three-dimensional structural templates for domain-domain interactions. It contains templates for interactions between two globular domains as well as novel domain-peptide interactions. Stein et al. (2011)

Type: Database	**Immunoglobulin Database, IMGT**
Description:	IMGT®, the international ImMunoGeneTics information system is the global reference in immunogenetics and immunoinformatics. IMGT® consists of sequence databases, genome database, structure database, and monoclonal antibodies database, Web resources and interactive tools. Lefranc (1998)

Type: Database	**The Structural Antibody Database, SAbDab**
Description:	Collecting, curating and presenting the antibody structural data from the PDB. Dunbar et al. (2014)

Type: Database	**Structural Epitope Database, SEDB**
Description:	A Web-based Database for the Epitope, and its Intermolecular Interaction Along with the Tertiary Structure Information Om Prakash et al. (2012)

Type: Database	**Python-based immunoglobulin classification, PyIgClassify**
Description	PyIgClassify that provides access to assignments of all CDR structures in the PDB to our classification system. The database includes assignments to the IMGT germline V regions for heavy and light chains for several species Adolf-Bryfogle et al. (2015)

Type: Database	**Antibody Antigen Interaction Database, AgAbDb**
Description	AgAbDb is a derived knowledge base that archives molecular interactions of protein and peptide antigens characterized by co-crystal structures. The interactions are characterized using AAIF (Antigen-Antibody Interaction Finder) developed by the same authors. (Kulkarni-Kale et al., 2014)

Type: Web Tool	**BLUE STAR STING**
Description:	A multiplatform environment for protein structure analysis Neshich et al. (2006)

Type: Web Tool	**Paratome**
Description:	The Paratome web server predicts the ABRs of an antibody, given its amino acid sequence or 3D structure Based on a set of consensus regions derived from a structural alignment of a non-redundant set of all known antibody-antigen complexes. Kunik et al. (2012a)

Type: Web Tool	**automatic Prediction of ImmunoGlobulin Structures, PIGS**
Description:	Pigs is a web server for the automatic modeling of immunoglobulin variable domains based on the canonical structure method. It has a user-friendly and flexible interface, that allows the user to choose templates (for the frameworks and the loops) and modeling strategies in an automatic or manual fashion. Its final output is a complete three-dimensional model of the target antibody that can be downloaded or displayed on-line. Marcatili et al. (2008)

Type: Web Tool	**Protein-peptide complexes, PepX**
Description:	PepX is a database containing unique protein-peptide interface clusters from the PDB, representing the diversity of structural information on protein-peptide complexes available in the Protein Data Bank. Vanhee et al. (2010)

Type: Web Tool	**Weblogo**
Description:	WebLogo is a web based application designed to make the generation of sequence logos. Crooks et al. (2004)

Type: Software	**CD-HIT**
Description	CD-HIT is a very widely used program for clustering and comparing protein or nucleotide sequences. Fu et al. (2012a)

CHAPTER 1
Introduction

1.1 Protein-Protein Interface

Protein-Protein interactions (PPI) are at the core of biological processes. They are involved in all steps of a living organism biochemistry and are crucial to the understanding of all *in vivo* functions, cellular regulation, biosynthesis, degradation pathways, signal transduction, initiation of DNA replication, transcription, translation, multi-molecular associations, packaging, oligomer formation and the immune response (Keskin et al., 2005). The heart of immune response rely on the Antibodies (Ab)-Antigen (Ag) recognition, making the Ab-Ag complex a specific type of PPI of great interest. Determining which parts of the Ab are essential for Ag recognition and vice-versa is necessary for understanding B cell-mediated immunity. Moreover Abs are commonly used in molecular biology and are a potent tool for biotechnology and biomedicine (Maynard and Georgiou, 2000). The region of the antibodies that recognize antigens, called paratope, is included in the Complementarity Determining Regions (CDR) and the region of the antigen recognized by antibodies is named epitope.

1.2 Complementarity Determining Regions

Complementarity Determining Regions are formed of six variable loops, three from the light chain (CDRL1-3) and three from the heavy chain (CDRH1-3) (Chothia et al., 1989; Mian et al., 1991). Defining the limits of the CDRs can nowadays be done with different methods.

Firstly Kabat and co-workers (Wu and Kabat, 1970; Kabat et al., 1983) used the high variability of the CDR compared to the framework region of the Abs to identify the boundaries in a systematic way. Using alignment, they developed a numbering technique to automatically mark positions for all new Abs sequences. Chothia and Janin used sequence of a small number of reference Abs to identify the CDRs. The method is based on the observation that CDR loop's amino-acids composition is highly variable compared to the antibody framework. Using an alignment of sequence was established a numbering system used to identify the conserved residues that delimitates the CDRs. Lefranc and co-worker developed the IMGT database (Lefranc, 1998) containing nowadays more than 176.000 immunoglobulins (IG) and T-cell receptors curated genes as well as more than 4000 annotated structure of Abs. The IMGT developed a uniform numbering system based on previous techniques and is homogeneous for various IG and TR including antibody heavy or light chain of different species. More recently Ofran and co-workers (Kunik et al., 2012b) developed Paratome which used a structural consensus of a set Ab-Ag complexes to determine the Ag Binding Regions (ABRs). They use this consensus to predict from other sequences or structures the limits of the ABRs. The identification of CDRs from sequence or structure defines one side of the interface. The amino-acids from the antigen in contact with the CDRs forming the other side, called the epitope.

1.3 Epitope prediction

Epitope can be divided into two categories: continuous (also called linear) in which all residues are consecutive in the sequence and discontinuous (or conformational) where the epitope is formed of multiple distant parts of the Ag sequence. Interface size and shape can take different forms and have different degrees of complementarity (Figure 1.1) leading to the fact that the majority of the epitopes are conformational (Pellequer et al., 1991).

Epitopes are at the center of the humoral immune response (Silverstein, 1990). Antigen recognition depends on the affinity and specificity of the antibody (Abbas and Lichtman, 2005), physicochemical properties and structure of the epitope (Greenbaum et al., 2007). Predicting an epitope limits for a given antigen remain problematic (Hopp and Woods, 1981), at least without structural information. Identified correctly, an epitope sequence can synthesize and replace the parent antigen, allowing antibody production, through immunization (Emini et al., 1985; Moyle and Toth, 2013), purification of interface specific monoclonal antibody (Murray et al., 2001) or antibody detection from patient serum using complement fixation test used to diagnose infection and other diseases (Rao, 2005) . Prediction of the epitope from protein sequence was firstly attempted in the 1980s and was based on on amino acid properties such as flexibility (Karplus and McCammon, 1986), hydropathy (Parker et al., 1986), antigenicity (Greenbaum et al., 2007) or structural properties such as beta turns (Pellequer and Westhof, 1993a) and accessibility (Davydov and Tonevitski, 2009). Later on, with the increase of available data, researchers improved the prediction method using multiple parameters such as solvent accessibility, flexibility, and secondary structure propensities (Pellequer et al., 1991; Pellequer and Westhof, 1993b; Alix, 1999).

Later on, a new generation of methods would combine many of those properties (Pellequer and Westhof, 1993a; Alix, 1999; Odorico and Pellequer, 2003). Then in 2005, Blythe and Flower showed that almost 500 properties did not perform sufficiently good. Since then epitope prediction has slided from simple propensity analysis to multiple parameters using more complex data mining and knowledge-based methods (Gao and Kurgan, 2014) such as neural network (ABCPred)(Saha and Raghava, 2006) or support vector machine (COBEpro)(Sweredoski and Baldi, 2009) or even graph model (BeTOP)(Zhao et al., 2012). Nowadays prediction methods encounter different difficulties like data quality (Greenbaum et al., 2007; Denh et al., 2011), quantity of positive elements or proper negative data (Subramanian and Chinnappan, 2013).

With the rapid expansion of crystallography techniques allowing the resolution of protein-Ab complexes, interfaces can be studied directly from the structures and give a reliable reference for the epitope prediction, but raised the problematic of the interface limits selection.

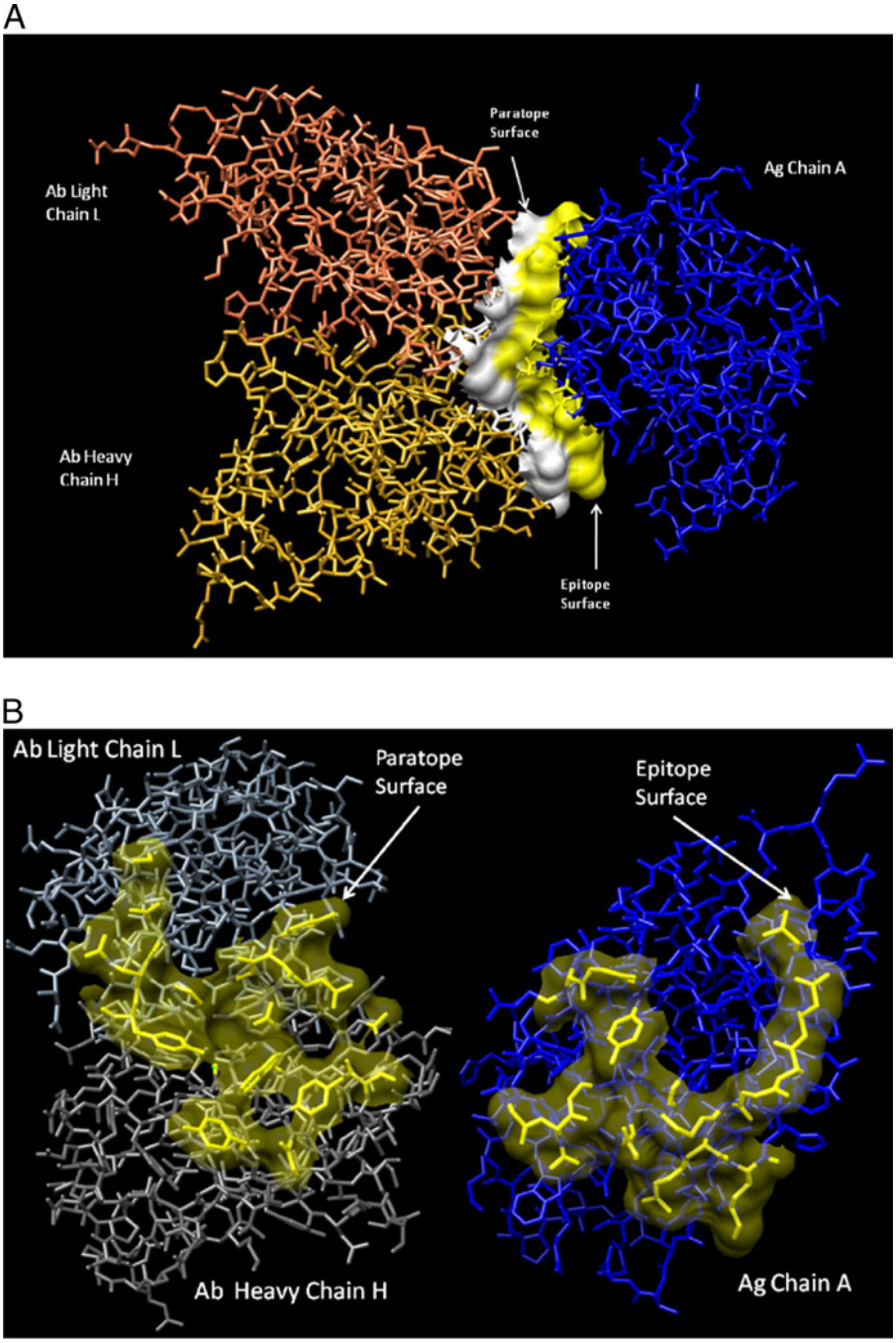


Fig. 1.1: Pheasant egg white lysozyme antigen-antibody complex structure (1JHL) showing (A) a side view of interaction region surfaces and (B) the interfaces turned 90° so that the epitope-containing and paratope-containing surfaces (with 80% transparency) can be better visualized. Images from: Ramaraj et al. (2012).

1.4 Interface limits selection

Three methods exist nowadays to determine interface boundaries: The first use a cutoff on the euclidean distance between the atoms of the antibody and antigen (Chothia and Janin, 1975; Lo Conte et al., 1999). The second is based on the Difference of Accessible Solvent Area (ΔASA) between the split and complexed proteins (Janin and Chothia, 1990; Chakrabarti and Janin, 2002). These two methods can be adapted using different cutoff values or even combined leading to the Epitope Containing Sequence and Paratope Containing Sequence as defined by Ramaraj and co-workers 2012. Finally, the last approach defines interfaces through computational geometry using Voronoi diagrams and the alpha shapes theory (Pontius et al., 1996) used by Goncalves-Almeida et al.2012. These selection methods have evolved and gained complexity with the increase in antibody-antigen complex structures available.

1.5 Antigen-Antibody interface

In 1986, Novotny et al. demonstrated using crystallographic structures that the accessible surface area (ASA) is a better parameter to assess the antigenicity than hydrophilicity used by the first epitope prediction methods. In 1990, Davies et al. used 6 Ab-Ag complexes including 2 different protein antigens (lysozyme and neuraminidase) to describe the interface in surface area, number of residues buried and in contact. The interface area was observed to be from 137 to 879 Ångström square ($Å^2$), the number of residues for the antigen was found to be 24 to 32 and 14 to 21 for buried and in contact respectively. For the antibody the number of buried amino-acid observed was from 22 to 32 while the contact residue number was 14 to 21. They also noticed that the Abs had a conserved structure, in which only the Complementarity Determining Regions (CDR) presented a high structural variability. The same year Laver et al. , using 5 structures of 2 different complexes to describe the protein-ab interface. They

found that the epitope was composed of 15 to 22 amino-acids compressed in a narrow span of 650 to 900 $Å^2$ and forming with the Ab 75 to 120 Hydrogen Bond (HB) as well as other molecular interactions. Analysis of the number of residues from the Ab in the interface was similar to the epitope (Laver et al., 1990). In the work from Mian et al., in 1990, six structures were analyzed among those only one was a Protein-Ab complex and confirmed the solvent exposure as primordial factor of the epitope prediction. For the paratope the most important residues were Ser, Thr, Trp and especially Tyr representing alone 25% of the residues. In 1993, Lawrence and Colman introduced the shape complementarity index (Sc) allowing to measure the nesting of the antibody with the antigen . Their work showed that the Ab-Ag interface has a significantly poorer Sc than other protein-protein complex. Later on, part of a PPI study by Jones and Thornton in 1996, 86 structures of protein interaction were described including 6 Ag-Ab. The ΔSAS of those complexes was found to have a mean of 777$Å^2$ with a Standard Deviation (SD) of 135.33, both measures being lower than any other types of PPI. The planarity of the interface was also reduced compared to other PPI. In 1996 MacCallum et al. gathered 26 Ab-Ag complexes and compared the interface shape in function of the Ag size. They concluded that the interface is concave for small Ag and planar for the big ones. The same year Cohen and Davies focus on 6 anti-idiotic structures of protein-Ab and observed the predominance of the heavy chain over the light one in term of ΔSAS as well as interaction count (Cohen et al., 1996). Three years later Lo Conte et al. in 1999 studied 19 interfaces including seven lysozymes-Ab. The average interface span was found to be 1680$Å^2$ with a SD of 260. They also defined the 3 groups of interface residues depending on the ΔSAS. They showed that for the epitope the most predominant residue in the surrounding group is Ser followed by Lys while the Tyr is highly enriched in the most central group.

In the 2000s the rapid progression of crystallographic techniques led to a fast progress concerning structures resolution including Ab-Ag complexes. In 2002, Chakrabarti and Janin used the notion of interface patches defined by Jones and Thornton (1996) to describe 18

Ab-Ag structures. A patch is a group of interface's residues clustered by an average linkage method (Johnson RA, 1996) using a threshold distance of 15Å. Most of the Ab-Ag interface only contained one patch but some were formed of 2 or 3 patches. Each of the patch was organized as described by Lo Conte et al.. In 2003 Sundberg et al. analyzed 30 Ab-Ag interfaces detailing energy and interactions. The authors analyzed the 'hot spots', defined as a residue having a $\Delta\Delta$G superior to 2.5 kcal/mol and noticed a higher concentration of hot spot residues in the center of the interface. Sundberg and co-workers also observed that the partial hydration increases the Sc and form water mediated HB. In 2006, Haste Andersen et al. selected 76 structures and focused on the epitope properties in order to improve B-cell epitope prediction. As previously shown, the majority of the epitope were conformational and contained 9 to 22 residues. More than 45% of the segments were composed of only one amino-acid. Rubinstein et al. selected only 53 structures that were curated and non-redundant in 2007. The results of their statistical analysis confirmed the significant enrichment of Tyr and Trp in the epitope as well as charged and polar residues. The epitopic surface was found to be preferentially planar and exposed with a high percentage of unorganized secondary structure. Rubinstein and co-worker also noticed that epitope's structure undergoes a compression when bound to the Ab leading to important shape changes. In the work of Chen et al. in 2009, 192 structures were selected. They divided this set into two, based on the Ag size, leading to a split analysis of the Protein-Ab and Peptide-Ab. Comparing the shape of bound peptide with its native protein structure showed very important variations since none of the superimposition could be done without clashes or bumps. The peptide epitope secondary structure, computed with STRIDE (Heinig and Frishman, 2004), showed a remarkable increase of Coil compared to the protein's ones. In 2012 Kringelum et al. gathered 107 non-similar Ab-Ag interfaces. They described the epitope surface as flat, oblong, oval shaped containing a majority of hydrophobic residue in the center and surrounded by charged amino-acids. Their statistical study of epitope composition concluded that none of the residue's propensity was significantly different compared

to the Ag surfaces. Unlike the epitope, the paratope amino-acid composition showed significant variations for 14 out of the 20 residues. Tryptophan and Tyrosine showed the major enrichment while Pro, Lys and Gly were significantly impoverished. Also in 2012, Ramaraj et al. selected 53 non-redundant interfaces and , as previously, observed enrichment of aromatics residues in the paratope. Based on an euclidean distance and ΔSAS they computed the interaction (based on distance) frequencies for all epitope-paratope amino-acid couple (Figure 1.2). Containing Sequence (ECS) and Paratope Containg Sequence (PCS) have very specific pattern of interaction in term of residues. Hydrophobic residues such as Iso and Leu interact with themselves to form hydrophobic bond. The Arg from the ECS is found to interact with the Trp while the PCS one interacts mainly with Tyr. The epitope Met having a low representation is found to interact with high specific interaction frequency with Phe and Met in the paratope

The same year Sela-Culang et al. and co-workers compared 49 Ab's free and bound structures. They concluded that concerning the variable domain of the Ab, only the CDR-H3 undergoes significant binding-related conformational changes in about one third of the antibodies structures. Meanwhile a loop from the H chain implicated in the inter-chains interaction present superior conformation modification than the variable ones. The constant regions surrounding the CDRs also present structure modification in a proportion related to the size of the Ag, the bigger the Ag, the bigger the change. Still in 2012, Dario et al used 28 structures of free and bound fragment Ab (Fab) to investigate dynamic coordination and intra-molecular interactions changes upon binding. Their study concluded that Fab internal dynamic, coordination pattern and molecular interactions are modified when binded to the Ag and not only in variable domains. In 2013 by Stave and Lindpaintner selected 111 Proteins-Ab crystallography from the PDB with different Abs. Using a 4Å selection they found that the size of the Ab and Ag interface was sensibly equivalent. The number of residues selected from the heavy chain was superior to the light chain in 92 out of the 111 structures, confirming precedent results

Table 1.1: Epitope and paratope properties

Epitope Residue Number	For Protein Ag : 9 to 22 Amino Acids.[5]
Epitope Residue Propensities	High occurrence of hydrophobic and aromatic in the center and charged in the surrounding area.[4,6]
Epitope sequence	Mostly conformational epitope, segment from 1 to 12 residues.[5]
Paratope Residue Number	14 to 20 Amino-acids, usually bigger than the epitope.[8]
Paratope Residue Propensities	Until 25% of Tyr, Trp and charged also enriched.[1,4,8]
Interface Dimension	From 600 to 2000Å depending on the Ag size.[4]
Epitope Surface	Oval, exposed and unorganized secondary structure.[3,7,8]
Epitope 3D shape	Planar for the big Ag and concave for the small ones.[3,6]
Shape complementarity	Significantly poorer than Protein-Protein Interaction.[2]
Epitope conformation changes	Compression upon binding.[6,7]
Ab conformation changes	Compression upon binding proportional to the Ag size, modification in the L-H chain contacts[6,7,9]

[1]Mian et al. (1991) [2]Lawrence and Colman (1993) [3]MacCallum et al. (1996) [4]Lo Conte et al. (1999) [5]Haste Andersen et al. (2006) [6]Rubinstein et al. (2008) [7]Chen et al. (2009) [8]Kringelum et al. (2012) [9]Sela-Culang et al. (2012)

about the superior importance of the heavy chain in the interface. To our knowledge the most recent study using a set of crystallographic structures to extract pattern of the interface was made by Robin et al. in 2014. Using a set of 227 antibody-antigen structures and, by analyzing free binding energy, they demonstrated that for the paratope as few as 8 residues out of 30 important positions are enough to explain 80% of the binding energy.

From the beginning of the 1990s to nowadays, the studies of Ag-Ab crystallographic complexes have become one of the major way to gain insight into the interface's mechanism (Table.1.1). The evolution was directly related to the improvement of the X-ray resolutions techniques that have rapidly evolve in the last 25 years. The knowledge obtained from such studies have directly benefited to the techniques such as epitope prediction, antibody engineering and design of mimetic peptides.

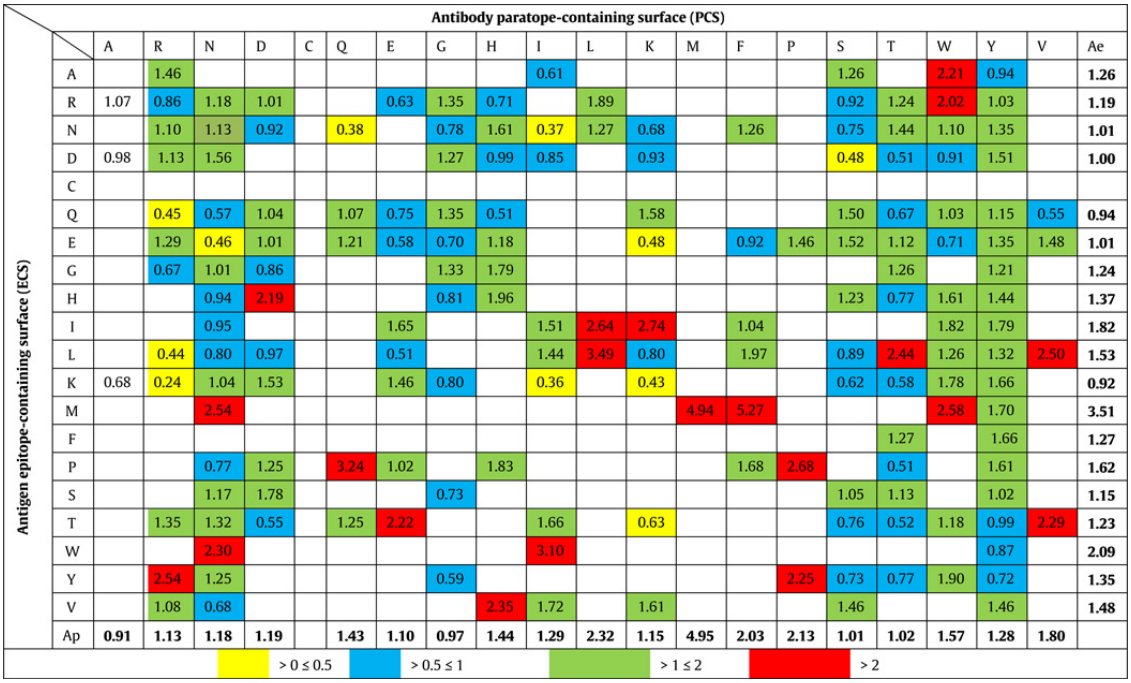


Fig. 1.2: Specific interaction frequency. Ae = Average over each row; Ap = Average over each column. Epitope Containing Sequence (ECS) and Paratope Containg Sequence (PCS). Highest to lowest interactions interaction between residues from the PCS (columns) are ECS (rows). Interaction ranking is graded as the highest being red > green > blue > yellow. Figure and legend from Ramaraj et al. (2012)

1.6 The importance of antibody

Antibodies are very versatile molecules used as molecular probes in research and also the most important biological drug with nowadays more than 30 Abs and derivative approved by the Food and Drug Administration (Beck et al., 2010). Even though full Immunoglobulin Gamma (IgG) are mostly used, smaller Ab fragments retaining binding capacity have been expending. The most used antibody fragment is Fragment Ab (Fab), produced by Papain enzymatic digestion, followed by the Single-chain Fragment variable (scFv), composed only by the variable fragment of both Light chain (C_L) and Heavy chain (C_H) joined by a linker (Porter, 1959; Bird et al., 1988). The figure 1.3 shows the different fragments and their construction. Both of those fragments maintain binding capacity but loose their immune system inducing function (Jain et al., 2007). The minibody is made of 2 ScFv joined by 2 C_H3 portion of the heavy chain.

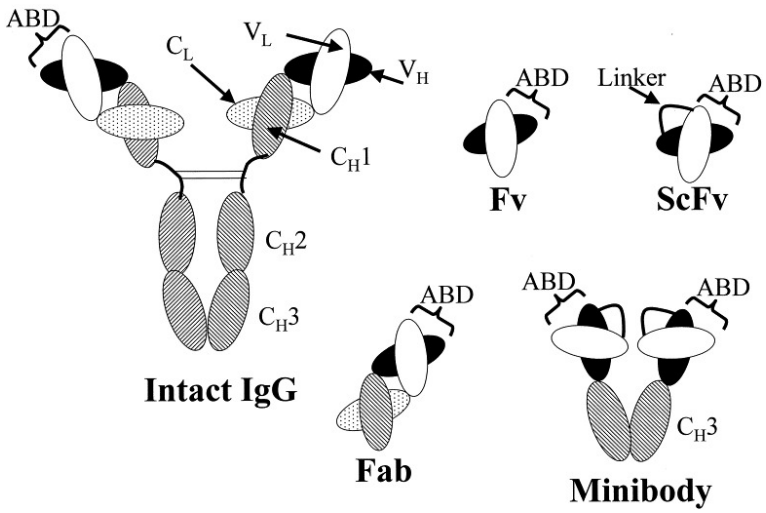


Fig. 1.3: Structure of intact immunoglobulin G (IgG) molecule, a minibody, and monomeric antibody fragments: Fab (antibody fragment), Fv (variable fragment), and scFv (single chain variable fragment). Variable domains of heavy (VH) and light (VL) chains are represented by black and white ovals, respectively. Constant regions of heavy (CH1-3) and light (CL) chains are represented by shaded and dotted ovals, respectively. ABD = antigen-binding domain (paratope). Figure and legend from Azzazy and Highsmith (2002)

1.7 Limitation of drug's antibody

The most common method to produce Monoclonal Antibodies (MAb) is through injection of purified protein antigen into animals and isolation of the Abs binding to the native Ag after spleen cells extraction. Selected cells are then fused with immortal myeloma cells obtaining an almost immortal cell producing antibody binding to the desired antigen (Figure 1.4). This methodology present various limitations and disadvantages. Obtaining a satisfactory immune response from a living organism require multiple injections of purified antigen (Leenaars and Hendriksen, 2005). Therefore immunization protocol suffers limitations from the cost and difficulties to obtain sufficient amount of pure antigen as well as usage of live animals. Immunization are also very problematic when dealing with lethal or toxic protein. Concerning the monoclonal antibody production both *in vivo* or *in vitro* techniques suffers limitations. When selection and production are successful the MAb needs to be purified and

stored which, in the case of large molecules, can be difficult requiring very low temperature in order to preserve their structure. In the case of therapeutic usage, monoclonal antibody lack of administration's route. Most therapeutic Abs require intravenous administration however some have been approved for subcutaneous or intramuscular but never orally (Wang et al., 2008). Drug antibody also suffer from poor pharmacokinetic due to poor tissue penetration and rapid proteolysis (Pimm, 1988). Being produced by animal organism (mice) MAb can also cause undesired allergic reaction to the patient. With the rapid expansion of antibody use in the year 2000, patents on antibody engineering have prompted the biotechnology companies to develop proprietary antibody humanization (Lugovskoy et al., 2010), adding legal restriction to the MAb usage (Hanf et al., 2014).

Improvement of the crystallographic techniques and structure prediction have made *de novo* protein design an important part of the biotechnology and gives hope for synthetic antibodies. Computationally predicting peptide or protein with specific binding affinities is nowadays possible with a certain accuracy (London and Ambroggio, 2013).Stranges and Kuhlman reviewed success and failure of *de novo* protein design and showed that predicted interfaces are smaller than real PPI and also that the predicted hydrogen bonds are not carefully computed. The biggest limitation of *de novo* protein design is the limited accuracy of structure prediction from protein sequence as described in the work of Pantazes et al.. Applying this methodology to Abs in order to create synthetic antibody is a complex task, partly due to the uncommon structure of the Ab framework. In order to overpass the antibody limitations many efforts have been made to design and create small biological molecules with binding capacity.

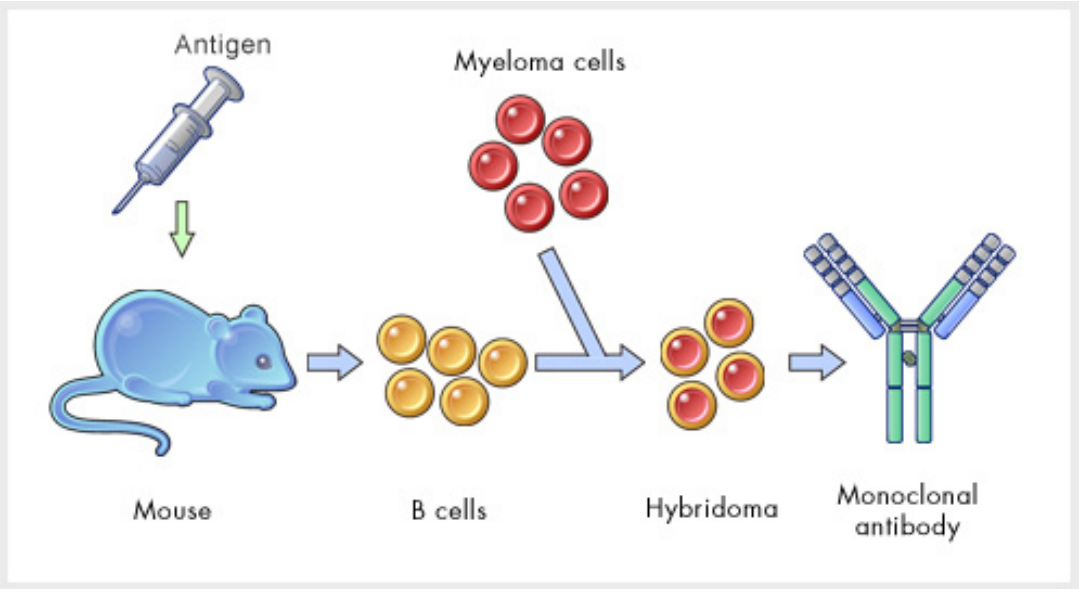


Fig. 1.4: Step by step protocol of monoclonal antibody production. Mouse are injected with the specific antigen to force the immune system to react and produce specific antibody against it. B-cells are extracted from the spleen and fused with immortal myeloma cells. Hybridoma cells produce antibody which are extracted and purified. Source: *http://www.kyowa-kirin.co.jp/antibody*

1.8 Peptide binder

1.8.1 Combinatorial peptide libraries using phage displayed

The design of peptide with affinity for a specific target would offer a viable replacement of drug antibody. To guide the design and increase the affinity and specificity of these peptide drugs, different methodologies exist such as directed evolution, high-throughput protein screening or rational design based on protein-peptide interactions (Pei and Wavreille, 2007; Yin et al., 2007; Vanhee et al., 2011). One way to produce peptide binders *in vitro* is through phage display using Combinatorial Peptide Libraries (CPLs) that allows discovery of new peptides able to bind to desired target protein such as receptors, enzymes, virus, materials or even small molecules. (Cwirla et al., 1997; Su et al., 2005; Hyde-DeRuyscher et al., 2000; Welch et al., 2007; Matsubara et al., 2010; Whaley et al., 2000; Wang et al., 2003; Rodi et al.,

1999). Peptides selected through phage display CPLs are often very closely related in term of sequence and structure to the natural ligand they aim to mimic. Peptide binders selected through CPLs can be used for a broad range of applications and can be stored in solution or solid state for long period without loss of effectiveness but are often limited in term of affinity to the target. Those issues can be explained by different reasons like the cell toxicity of some peptides (Saar et al., 2005) and the use of peptide libraries from small sizes $10^{7\sim9}$ (Smith and Petrenko, 1997; Hoess, 2001) To avoid issues related to living cell and bacteriophage handling cell free methodologies were developed such as ribosomes display, mRNA display and CIS display have been developed (Zahnd et al., 2007; Cotten et al., 2011; Odegrip et al., 2004a). Ribosome display has been used to select from CPLs peptide with affinity for antibody (Mattheakis et al., 1994) or streptavidin (Lamla and Erdmann, 2003). Concerning mRNA display, a covalent link peptide-mRNA is required but it allows larger size library $10^{12\sim14}$ (Cho et al., 2000). CIS display, works in similar way than mRNA display but uses DNA and replication protein A instead of puromycin (Ingmer et al., 2001; Odegrip et al., 2004b).

1.8.2 Rational design

Rational design of peptide ligand uses the increase in knowledge about peptide binding (Vanhee et al., 2009; London et al., 2010) to create and select possible binding sequences. Using *in silico* mutagenesis Clackson and Wells (1995) discovered that peptide-protein interfaces possess 'hotspot' similar to all protein-protein interface. The number of 'hotspot' residues depending on the length of the peptide ranging from two for a peptide length from 6-8 residues and three for peptides of size 9-11 (London et al., 2010). In case the structural information is available for a drug target, more complex computational methods can be used such as neural network (Honeyman et al., 1998), genetic algorithm, hidden Markov models or motif discovery algorithm (Lin et al., 2008). Peptide ligand can also be derived from a crystallographic

structure of a PPI. The first work achieving a such result was reported by Wild et al. (1994). The authors designed a 36 residues anti-HIV peptide blocking an early step in the virus life cycle before to reverse transcription. It became the first fusion inhibitor for HIV-1 therapy named enfuvirtide (Fuzeon®) that was approved by the United States Food and Drug Administration (Naider and Anglister, 2009). Nevertheless successful design of peptide binder from protein-protein interface is limited. Most of the working examples had to ensure that only a few residues are important for the interface.

Peptide ligands bring solutions to some of the most central limitations of antibodies as a drug such as production, storage and in a certain limit patent. Nevertheless designing specific active peptides is not an easy task especially due to flexibility and environment dependent conformation. In most of the successful approaches, modifications were done to, at least partially, limit the peptides possible conformations. In order to force a small protein sequence into a certain conformation scaffold can be used. Such a molecule can be of different natures and various sizes depending on the objectives, sequence properties and nature of the target. Most of the scaffolds used are derived from existing proteins or chemical constructs.

1.9 Proteic Scaffold

1.9.1 Fibronectin

The tenth type III domain of the human fibronectin is a β sandwich structured protein composed of 94 amino acids resembling the immonoglobulin domain (Figure 1.5a). Through randomization of one of the three exposed loops new binding activity was successfully created (Koide et al., 1998; Lipovsek, 2011; Chen et al., 2013). Using yeast surface display library Hackel et al. obtained a fibronectin domain with 3 pM affinity for lysozyme. Later on, the same authors (Hackel et al., 2012) engineered an fibronectin domain binding to the

epidermal growth factor receptor (EGFR). The fibronectin domain was labeled with copper 64 and successfully used to observe EGFR over-expressing xenografted tumors in mice using the positron emission tomographic (PET) technique.

1.9.2 Affibodies

The affibody (or z protein) is a protein composed of 58 residues forming three helical bundles (Figure 1.5b), one of the smallest known cooperatively folding structural domain (Wickstrom et al., 2006). Binding is usually obtained through randomization of 13 residues spatially closed located on the first and second helix. Most known affibodies were engineered to target Human Epidermal growth factor Receptor 2 (HER2) that achieved a 22 pM affinity and is used for cancer HER2-expressing imagery (Orlova et al., 2006). One derivative form of affibody was labeled with 111In and 68Ga for single-photon emission computed tomography imaging of metastatic breast cancer patient (Baum et al., 2010). Affibodies have also been able to target epidermal growth factor receptor (Tolmachev et al., 2010, 2009; Nordberg et al., 2008) and insulin-like growth factor (Tolmachev et al., 2012).

1.9.3 Two helix affibodies

Two-helix affibodies are a smaller version of the affibodies which is downsized to 36 amino acids. The loss of structural stability due to the third helix removal is partially compensated by adding disulfide bond between the two helices (Honarvar et al., 2013; Webster et al., 2009). Two-helix affibodies were also used for HER2 imaging achieving a lower uptake than the classical affibody at the price of lower affinity (Honarvar et al., 2013; Rosik et al., 2012)

1.9.4 Ankyrin

Ankyrin of Designed Ankyrin Repeat Protein (DARPin) is composed of a β turn followed by two α-helices repeated 4 to 6 times (Figure 1.5c). DARPins reached affinities of 270nM to 90nM for HER2 expressing tumor and was successfully used for imaging in mice using Single-photon emission computed tomography (Tamaskovic et al., 2012)

1.9.5 Knottins

Knottins are a group of protein with a size ranging from 30 to 50 amino acids containing three disulfide bonds forming a knotted structure shape (Moore and Cochran, 2012). Binding has been obtained by peptide grafting into one loop followed by a process of affinity maturation through randomization. Kimura et al. report an integrin-binding knottin . One of the best affinity obtained using a knottin scaffold was achieved against a tripsin II inhibitor from *Momordica cochinchinensis* with affinities of 3-6nM (Kimura et al., 2012).

1.9.6 Peptide aptamers

Peptide aptamers, also called thioredoxin-insert proteins, are the display of peptide ligands onto the thioredoxin scaffold (Figure 1.5e) (Borghouts et al., 2008). Thioredoxin is a 105 amino acids long protein involved in the redox signaling pathways. Most common form of thioredoxin used is the TrxA (Colas et al., 1996) from E. Coli (Li et al., 2011). The peptide sequence are commonly generated using CPLs and selected using the yeast two-hybrid system (Bickle et al., 2006). The peptide is inserted within a loop of the biological active center. The folding of this scaffold likely limits the possible conformations of the peptide. This importance of the scaffold influence on the peptide conformation was explored in the work of Klevenz et al.

1.9.7 Other protein scaffolds

A large variaty of other scaffold exists such as PDZ domain, neocarzinostatin or ribose-binding protein. Those scaffolds can be with different technique to achieve new binding such as error prone PCR, loop grafting or rational design (Figure 1.5f,g and h).

1.9.8 RAFT

Another possibility to proteic scaffold is the human-made Regioselectively Addressable Functionalized Templates (RAFT) scaffold consisting of a cyclic peptide composed of 8 Lys divided in two group of 4, each group separated by a prolylglycine. This modified residue acts as β-type II turn inducers. Forcing the 10 residues cyclic peptide to adopt two turns induce an antiparallel β sheet organization, locking the conformation of the structure. The Lys side chain are then modified to display desired chemical molecules or peptides (Dumy et al., 1996). Boturyn et al. used the RAFT platform to link the c[-RGDfK-] (molecule with a high affinity for $\alpha_v\beta_3$ integrin receptor) to different reporter group (including Biotin and fluorescein). This construct was successfully used to mark $\alpha_v\beta_3$ integrin receptor expressing endocytosis cells.

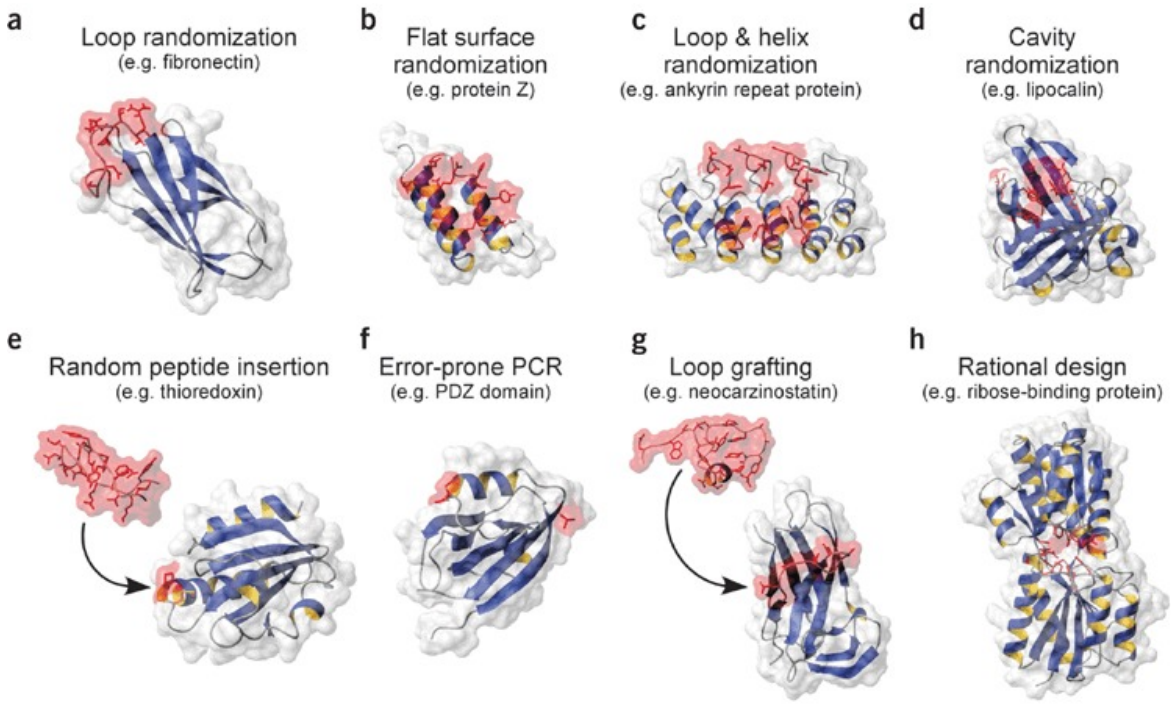


Fig. 1.5: (**a–h**) In combinatorial engineering approaches, sequences of a scaffold can be diversified at specified positions by means of defined randomized codons (e.g., in loops (**a**), flat surfaces (**b**), combinations of loops and helices (**c**), or cavities (**d**)), or a random peptide sequence is inserted into the scaffold (**e**), usually at a loop, or the scaffold sequence is randomized at undefined positions by error-prone PCR (**f**). Target-binding variants of the resulting libraries are subsequently isolated using selection or screening technologies. In rational engineering approaches, preexisting binding sequences (e.g. loops) have been grafted onto a novel scaffold (**g**), or binding sites have been engineered *de novo* into a suitable scaffold (**h**). The different engineering possibilities are illustrated by alternative binding molecules where the engineering in question has been applied: loop randomization (fibronectin)(Koide et al., 1998), flat surface randomization (protein Z)(Nord et al., 1997), loop and helix randomization (ankyrin repeat protein)41, cavity randomization (lipocalin)(Beste et al., 1999), random peptide insertion (thioredoxin)(Colas et al., 1996), error-prone PCR (PDZ domain)(Schneider et al., 1999), loop grafting (neocarzinostatin)(Nicaise et al., 2004) and rational design (ribose-binding protein)(Looger et al., 2003). Many other permutations of randomization strategies and scaffolds are conceivable; this figure illustrates each strategy with one published example. Figure and legend from Binz et al. (2005)

1.10 Chemical Scaffold

The chemical scaffold makes one more step in the direction of reducing the antibody limitation. By using a synthetic structure, limitations due to cell culture and protein expression are removed. Peptides binders derived from CDR or phage display for example, are structured using a chemical scaffold such as peptoid nanosheet, T2 and T3 platforms and synthetic antibody mimics.

1.10.1 Peptoid nanosheet

Peptoid or poly-N-substituted glycines, are a class of peptidomimetics molecules. When synthesized with a periodic amphiphilicity sequence and dissolved into aqueous solution the peptoids self assemble into a nanometer scale thin sheet (Nam et al., 2010). Taking advantage of the protein-like folding, Olivier et al. used to peptoid nanosheet structure presenting peptide sequence in a similar arrangement than encountered in the Ab (Figure 1.6A). Peptoid nanosheet are made from protease-resistant molecules that are capable of self-assembly into a stable sheet form. Peptides located on the sheet were shown to be accessible by enzymes and when the gold-binding peptide (Kulp et al., 2004) was used the peptoid nanosheet was shown to bind to the surface of the sheet using atom force microscopy.

1.10.2 T2 and T3 platforms

Timmerman et al. used two synthetic platforms they call T2 (α, α-dibromoxylene) and T3 (2,4,6-tris(bromo-methyl)mesitylene) on which peptide can be displayed. The T2 and T3 are linked to the peptide with cystein (Figure 1.6B). Using the specific Ab against gastrin17 the authors observed binding of the construct through screening of microarrays. Most of the successful sequences were derived from the CDRs sequence and obtained a binding with a

Kd in the micromolar range (100 μM).

1.10.3 SyAM, Synthetic Antibody Mimics

In 2013, McEnaney et al. designed and produced a Synthetic Antibody Mimics Protate cancer specific (SyAM-Ps) able to target prostate cancer cell (Figure 1.6C). This scaffold possess four arms, two on which are displayed the prostate cells binding region and other two at the end of which is found the immunoglobulin G receptor type I binding domain. While most of the binding molecules we saw that far lack the capacity to bind to Fc gamma receptor I (FcγRi), the SyAM possess two regions displaying FcγRi binding region allowing initiation of pro-inflammatory responses. SyAM-Ps offers various possibilities for future cancer treatment capable of recognizing specific cells and marking them for destruction.

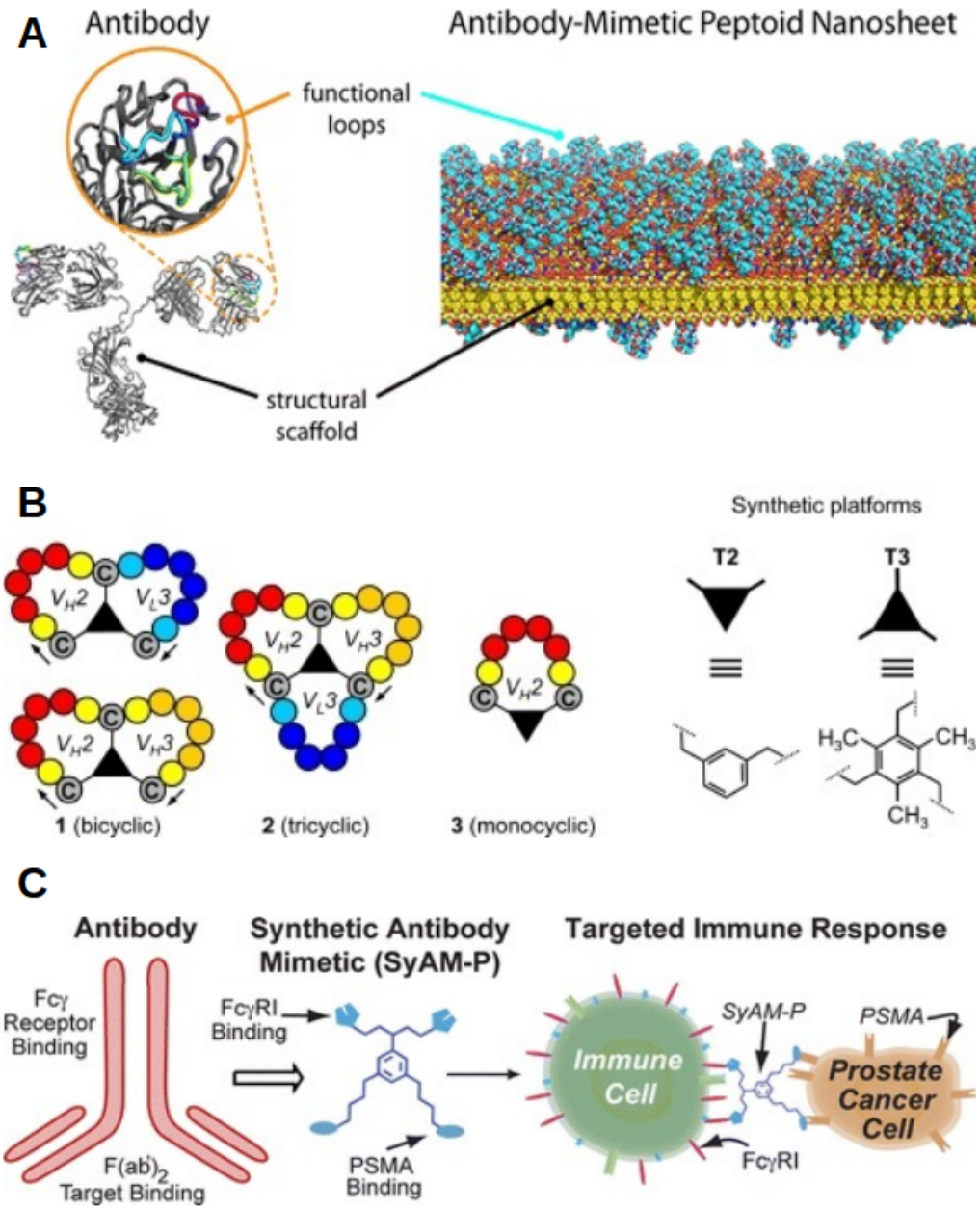


Fig. 1.6: (**A**) Architecture of an anti-canine lymphoma IgG2 antibody (PDB code: 1IGT). Antibody-mimetic loops displayed on a peptoid nanosheet scaffold (Olivier et al., 2013). (**B**) monocyclic, bicyclic, and tricyclic peptidomimetics are depicted together with molecular details for the scaffolds (T2 and T3) used in chemical linkage of peptides onto Scaffolds methodology (Timmerman et al., 2009). (**C**) Schematic illustration of the evolution of SyAM-Ps design from a monoclonal antibody template. Schematic depiction of SyAM-Ps proposed mechanism of action. Docking of SyAM-P into prostate-specific membrane antigen binding pocket and Fc gamma receptor I binding surface to determine the linker length needed to template a ternary complex (McEnaney et al., 2014).

CHAPTER 2
Objectives

General Objectives

The objective of this work is to gain insight about Antibody-Antigen interface properties in order to computationally generate peptide ligands.

Specific Objectives

1. To obtain a set of curated antigen-antibody structures.
2. To extract automatically epitope and paratope from interfaces using different selection method.
3. To design a database able to store efficiently all the data from the antibody-antigen interfaces along with a web interface.
4. To develop a methodology able to predict if a pair of epitope-paratope is mismatched or not.
5. To use computed molecular interaction between epitope and paratope to compare different interface selection methods.
6. To extract pattern from Paratope-Epitope using graphs.
7. To develop a program to computationally generate peptide ligands libraries.

8. To compare the interface properties recognized by specie-specific antibodies.

9. To predict the propensity of an epitope to be recognized by different antibody species and assess which properties give the best prediction.

CHAPTER 3

Epitope-Paratope Interface Database and webserver

The rapid increase in Antibody-Antigen(Ab-Ag) complexes available in the Protein Data-Bank (PDB, Berman et al. 2000) has led to the emergence of related databases helping to efficiently retrieve and analyze crystal Ab-Ag complex structures. The ImmunoGlobulin database (IMGT, Lefranc 1998) contains more than 4000 annotated structures of Antibodies. More recently, the Structural Antibody Database (SAbDab, Dunbar et al. 2014) focus on Complementary Determining Regions (CDR) sequence as well as maintaining a clustering of the different antibodies and contains more than 2000 references. In order gain insight into the Ab-Ag interface and help overpass the antibody's production limitation, we conducted a series of analysis based on physicochemical and structural properties of epitope-paratope interfaces. For this purpose we created several bioinformatics tools and a database with a web interface named Epitope-Paratope Interface DataBase (EPI-DB).

3.1 Interface Research Algorithm

In order to extract structures of antibody complexed with proteic antigen from the PDB we first used the dataset from Ramaraj et al. and Kunik et al. and selected Light and Heavy chain from the Ab to be used as reference sequences. Following redundancy removal from those two reference sets using CD-Hit (Fu et al., 2012a) with a cutoff of 0.9, we used a BioJava program we developed called Interface Research Algorithm (IRA, Figure 3.1). IRA

automatically computes a Smith and Waterman (Smith et al., 1981) local alignment of each of the reference sequences against each of the chains of all PDB files containing at least three chains. Using a threshold determined by aligning the reference set against itself, IRA labeled each chain as Ab Light, Ab Heavy or antigen. IRA selected PDB files that contain at least one antigen, one light chain and one heavy chain spatially close (using 5 Angstrom distance cutoff contacts). From those, were only selected the structures with a X-Ray resolution inferior or equal to 3Å . The files were checked to make sure that if the CDR were synthetic constructs they were done from the respective specie library. The antigen with a length inferior to 30 amino-acids was considered as peptide, bigger Ag as protein.

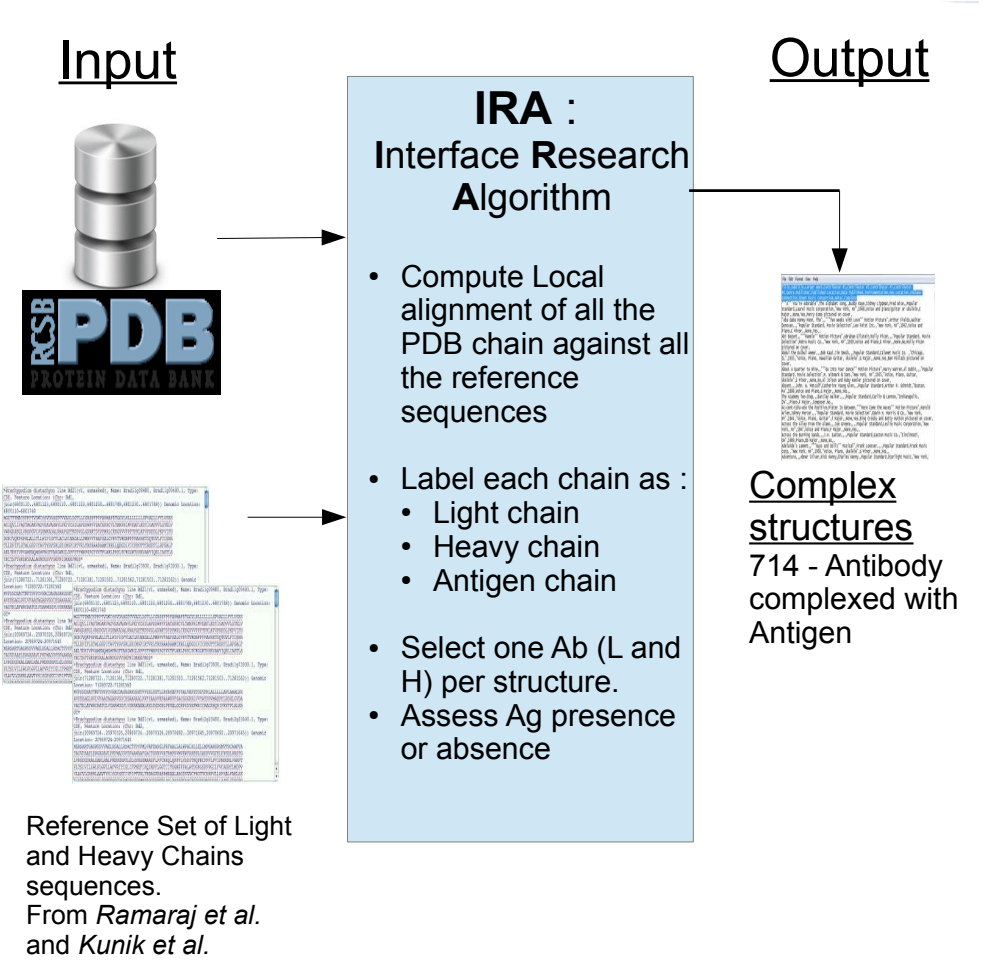


Fig. 3.1: Interface Research Algorithm methodology to extract antibody antigen complexed structures. IRA takes as input the Protein Data Bank and two sets of references sequences, one for light chain and one for heavy. One by one IRA determines if the structure contains heavy, light and non-antibody chain. The output gives the list of PDB files that contain at least one heavy, one light and one non-antibody chain.

3.2 Interface selection

To analyze the interface of Ab-Ag complexes, we used three different interface selection methods. First a selection based on the distance between atoms of the antigen and the antibody (Distance-Based Selection, DBS, (Figure 3.2**A**) as used by Chothia and Janin (1975); Lo Conte et al. (1999). An amino acid of the antigen is considered to be part of the distance selected epitope, if at least one of its atom is at a distance below a chosen cutoff. The paratope selection is done in the same manner. We computed DBS epitope and paratope with the following cutoff: 3.0, 3.5, 3.8, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0, 7.5 and 8.0 Ångström (Å). Second, we used an approach based on the difference of Solvent Accessible Surface (ΔSAS, Figure 3.2**B**), interfaces are selected based on the loss of solvent accessibility between the split and the complexed protein (Lo Conte et al., 1999). For this selection were computed the following cutoffs: 70, 60, 50, 40, 30, 20, 15, 10, 5 and 0^{+}Å^{2} (0^{+} meaning SAS loss is not null). Third, we developed a selection method in which the interface computed molecular interactions are extracted from STING RDB (Neshich et al., 2006). In this method, the interface is defined by all the amino acids that are involved in the molecular interactions between the antigen and the antibody chains and that are called, therefore, Interface Interacting Residues (I2R, Figure 3.2**C**). The selected antibody residues form the I2R Paratope and the selected antigen amino acids constitute the I2R Epitope. This methodology has no direct cutoff since it relies on computed interactions. We took as example the pheasant egg white lysozyme antigen-antibody complex structure (1JHL) to image the differences in residue selection. Using the DBS with 5Å cutoff for the epitope, 15 amino acids are selected while only 9 residues compose the I2R epitope. The biggest epitope is obtained using the ΔSAS with 0^{+}Å^{2} (0^{+} cutoff constituted of 18 residues. The biggest difference is observed for the paratope where the ΔSAS selects all of the center of the antibody (55 residues) while I2R and DBS only select 17 and 12 residues respectively.

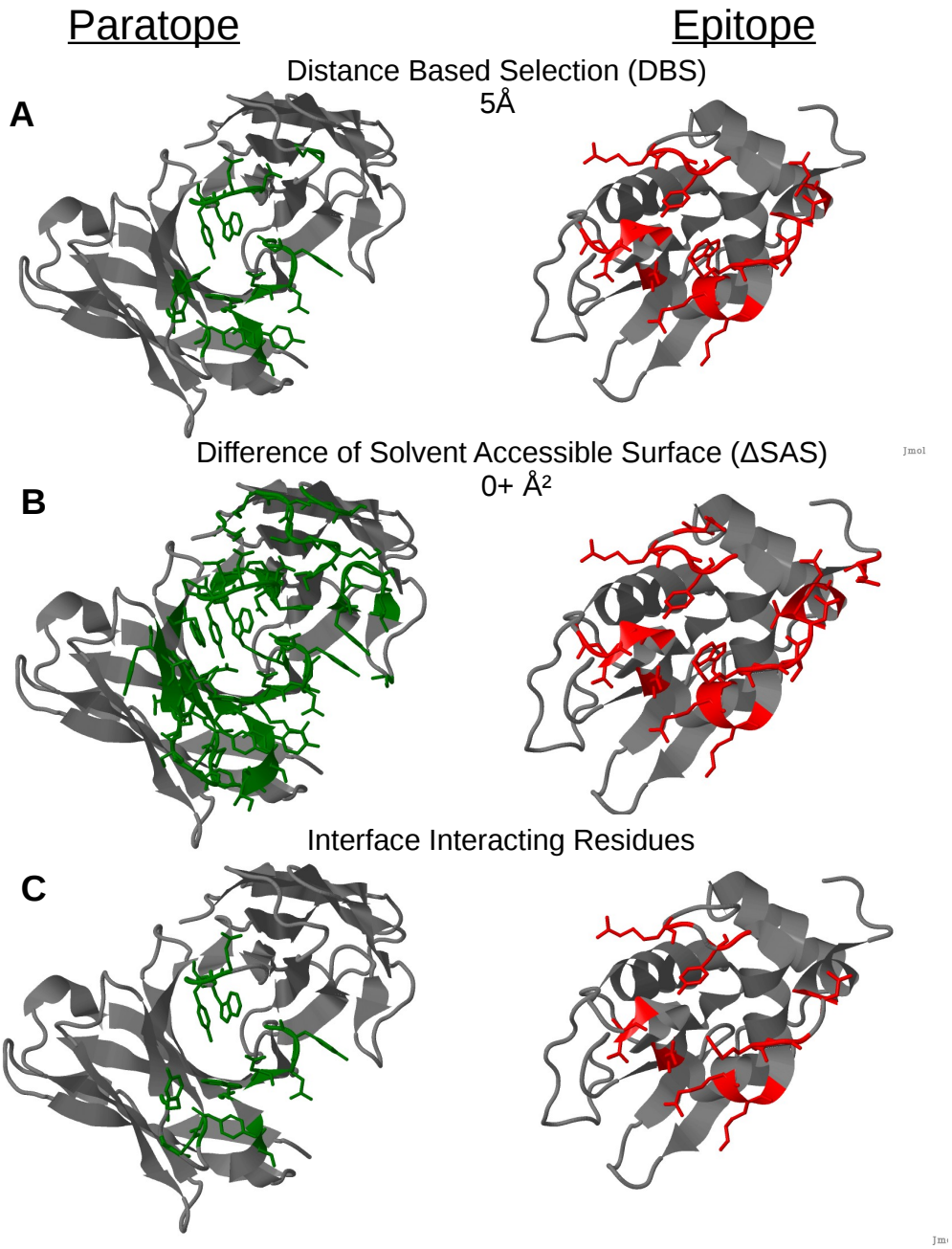


Fig. 3.2: Pheasant egg white lysozyme antigen-antibody complex structure (1JHL) paratope (left side) and epitope (right side) using different selection techniques. Residues selected for the antibody appear in green while residues selected from the antigen appear in red. **A** Distance based selection with a 5Å cutoff. **B** Selection based on loss of solvent accessible surface with cutoff of 0^{+}Å^{2} (loss is not null). **C** Interface Interacting Residue selection based on computed molecular interactions

3.3 Dataset description and redundancy removal

Using IRA we extracted 714 Antibody-Antigen (Ab-Ag) complex structures from the PDB (Berman et al., 2000). After removing the files that didn't correspond to our criteria of resolution and origin, we manually curated the remaining PDB files to make sure of their quality leading to the base redundant dataset that was inserted into our database, EPI-DB, currently composed by 543 PDBs. Due to different methods of selection and cutoffs we have 12025 pairs of paratope-epitope. We defined a 'pair' as the epitope and paratope obtained from a given structure using the same selection method and the same cutoff. From these, 6504 were obtained by Distance Based Selection (DBS), 5420 from difference of solvent accessibility surface (ΔSAS). Interface Interacting Residues were only added for the non redundant dataset resulting in 101 pairs of epitope-paratope. Some pairs were removed because the selection method and cutoff did not select any residues for one or the other side of the interface. Each epitope has 44 properties as does the paratope. We can see in the table 3.2 that most antibodies are from mice (316) and Humans(203).

To extract meaningful information from the interface dataset, we removed redundancies by selecting only the DSE and DSP sequences from the complex with a cutoff of 6Å. Using the CD-Hit global sequence identity score (Fu et al., 2012b), we only selected interfaces with a score lower than 0.90 for both interface sides. Global sequence identity score is defined as the number of identical amino acids in alignment divided by the length of the shorter sequence. The selected files were manually curated to confirm their quality. This provided us with a non-redundant dataset composed of 101 PDB structures (corresponding to the I2R) (Tables 3.1).

Table 3.1: Non-redundant dataset detail, Part 1

PDB	Light Chain	Heavy Chain	Antigen Chain	Resolution	Ab Specie
1H0D	A	B	C	2	Human
1N0X	L	H	P	1.8	Human
1RZJ	L	H	G	2.2	Human
1TJI	L	H	P	2.2	Human
1W72	L	H	ACD	2.15	Human
2B0S	L	H	P	2.3	Human
2CMR	L	H	A	2	Human
2DD8	L	H	S	2.3	Human
2FX7	L	H	P	1.76	Human
2H9G	A	B	R	2.32	Human
2NY7	L	H	G	2.3	Human
2QQN	L	H	A	2.2	Human
2R0L	L	H	A	2.2	Human
2UZI	L	H	R	2	Human
2VXQ	L	H	A	1.9	Human
2XRA	L	H	A	2.3	Human
2XWT	B	A	C	1.9	Human
3BN9	C	D	B	2.17	Human
3D85	A	B	C	1.9	Human
3GBN	L	H	AB	2.2	Human
3GJF	L	H	ACE	1.9	Human
3GRW	L	H	A	2.1	Human
3H0T	A	B	C	1.89	Human
3H42	L	H	AB	2.3	Human
3HI6	L	H	A	2.3	Human
3IDX	L	H	G	2.5	Human
3K2U	L	H	A	2.35	Human
3KR3	L	H	D	2.2	Human
3L95	A	B	X	2.19	Human
3LEV	L	H	A	2.5	Human
3MA9	L	H	A	2.05	Human
3MAC	L	H	A	2.5	Human
3MLR	L	H	P	1.8	Human
3MLT	L	H	P	2.49	Human
3MLX	L	H	P	1.9	Human
3MLY	L	H	P	1.7	Human
3MXW	L	H	A	1.83	Human
3NPS	C	B	A	1.5	Human
3P0Y	L	H	A	1.8	Human
3PGF	L	H	A	2.1	Human

PDB	Light Chain	Heavy Chain	Antigen Chain	Resolution	Ab Specie
3Q1S	L	H	I	2.15	Human
3RU8	L	H	X	2.07	Human
3SE8	L	H	G	1.9	Human
3SE9	L	H	G	2	Human
3SKJ	L	H	E	2.5	Human
3SOB	L	H	B	1.9	Human
3THM	L	H	F	2.1	Human
3TJE	L	H	F	1.93	Human
3U30	B	C	A	2.43	Human
3U7Y	L	H	G	2.45	Human
3UJI	L	H	P	1.6	Human
3UJJ	L	H	P	2	Human
4AL8	L	H	C	1.66	Human
4D9R	L	H	A	2.42	Human
4DGV	L	H	A	1.8	Human
4DTG	L	H	K	1.8	Human
1A3R	L	H	P	2.1	Mouse
1BGX	L	H	T	2.3	Mouse
1EJO	L	H	P	2.3	Mouse
1FE8	L	H	AC	2.03	Mouse
1FNS	L	H	A	2	Mouse
1JHL	L	H	A	2.4	Mouse
1KB5	L	H	AB	2.5	Mouse
1N64	L	H	P	2.34	Mouse
1NBY	A	B	C	1.8	Mouse
1NCA	L	H	N	2.5	Mouse
1NDG	A	B	C	1.9	Mouse
1ORS	A	B	C	1.9	Mouse
1OSP	L	H	O	1.95	Mouse
1P2C	A	B	CF	2	Mouse
1QKZ	L	H	AP	1.95	Mouse
1SY6	L	H	A	2.1	Mouse
1TET	L	H	P	2.3	Mouse
1UWX	K	M	BQ	2.2	Mouse
1VFB	A	B	C	1.8	Mouse
1WEJ	L	H	F	1.8	Mouse
1YQV	L	H	Y	1.7	Mouse
1ZTX	L	H	E	2.5	Mouse
2ADF	L	H	A	1.9	Mouse
2AEP	L	H	A	2.1	Mouse

PDB	Light Chain	Heavy Chain	Antigen Chain	Resolution	Ab Specie
2B2X	L	H	A	2.2	Mouse
2CK0	L	H	P	2.2	Mouse
2DQF	A	B	CF	2.5	Mouse
2J4W	L	H	D	2.5	Mouse
2JEL	L	H	P	2.5	Mouse
2QHR	L	H	P	2	Mouse
2VXT	L	H	I	1.49	Mouse
2XQY	K	J	E	2.05	Mouse
2Y5T	B	A	EFG	2.2	Mouse
3FFD	B	A	P	2	Mouse
3G5Y	A	B	E	1.59	Mouse
3GI9	L	H	C	2.48	Mouse
3HB3	D	C	B	2.25	Mouse
3LIZ	L	H	A	1.8	Mouse
3O2D	L	H	A	2.19	Mouse
3O6L	L	H	C	2.1	Mouse
3QWO	L	H	P	1.9	Mouse
3RKD	L	H	A	1.9	Mouse
3RVV	C	D	A	1.9	Mouse
4AEI	L	H	AB	2.3	Mouse
4ETQ	L	H	C	2.1	Mouse

Table 3.2: EPI-DB general statistics

PDB	543	paratope	epitope
Selection method	DBS	6504	6504
	ΔSAS	5420	5420
	I2R	101	101
	Total	12025	12025
Organism	Mouse	316	
	Human	203	
	Others	24	

3.4 Epitope Paratope Database, EPI-DB

Epitope Paratope Database (EPI-DB) was implemented in MySQL and specifically designed to store antibody-antigen structures extracted from the Protein DataBank (Berman et al., 2000), antigen and antibody description and sequence, paratope and epitope selected using various methodologies and cutoffs, physicochemical and structural properties (Figure 3.4). Seven tables compose the core of the EPI-DB (Figure 3.3). The PDB table stores information related to the PDB file, such as the PDB identifier, experiment type, resolution, number of chains, antibody and antigens chain-ids and date of the submission. More detailed information about each antigens and antibodies chain can be found in the proteins table. This includes number and sequence of amino acids, protein-name, Swissprot-id, expression system... Antigens and antibodies tables are aggregations of the protein table defining for each PDB the antigen(s) chain and the antibody(ies) chain. Epitopes and Paratopes tables store the information about the interface, including the sequences and position of the residues, the selection method and cutoff used to determine them. Physicochemical and structural properties for each of the epitope and paratope are stored in the Properties tables and detailed in the table 3.3. Physicochemical descriptors were computed using a perl script combining tools from the ExPASy platform as ProtParam and pI/Mw (Gasteiger et al., 2005) while structural ones were obtained using Stride (Heinig and Frishman, 2004).

This data is used to better understand the relationship between pairs, defined as a paratope and the corresponding epitope both selected from the same PDB using the same selection method and cutoff. In total, there are 12.025 pairs available and therefore 1.058.200 properties values calculated from sequences of epitope and paratope. The EPI-DB also offers expansion to accommodate data from literature associated to Ab-Ag complexes and epitopes derived from other proteins as well as the experimental data to determine epitope.

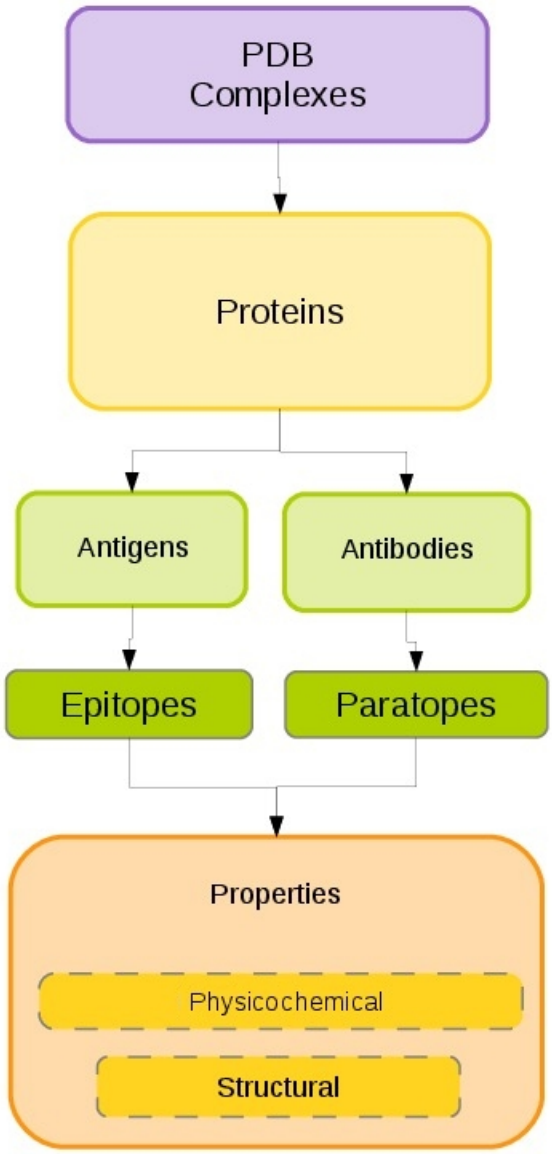


Fig. 3.3: Epitope-Paratope Database core design. Each full line square represents one of the central table of EPI-DB. PDB complexes contains information about the PDB file that references antibody-antigen structure. Protein table list all the chains of the PDB and store details like sequence, name and other databases references like Swissprot. Antigen and antibody table are aggregation of the protein table defining for each PDB the antigen(s) chain and the antibody(ies) chain. Epitope and paratope correspond to the sequence obtained using a given selection method and cutoff on an antigen and an antibody. Properties store the results of the 44 descriptors computed for each epitope and paratope. The properties can be divided into two groups, structural and physicochemical.

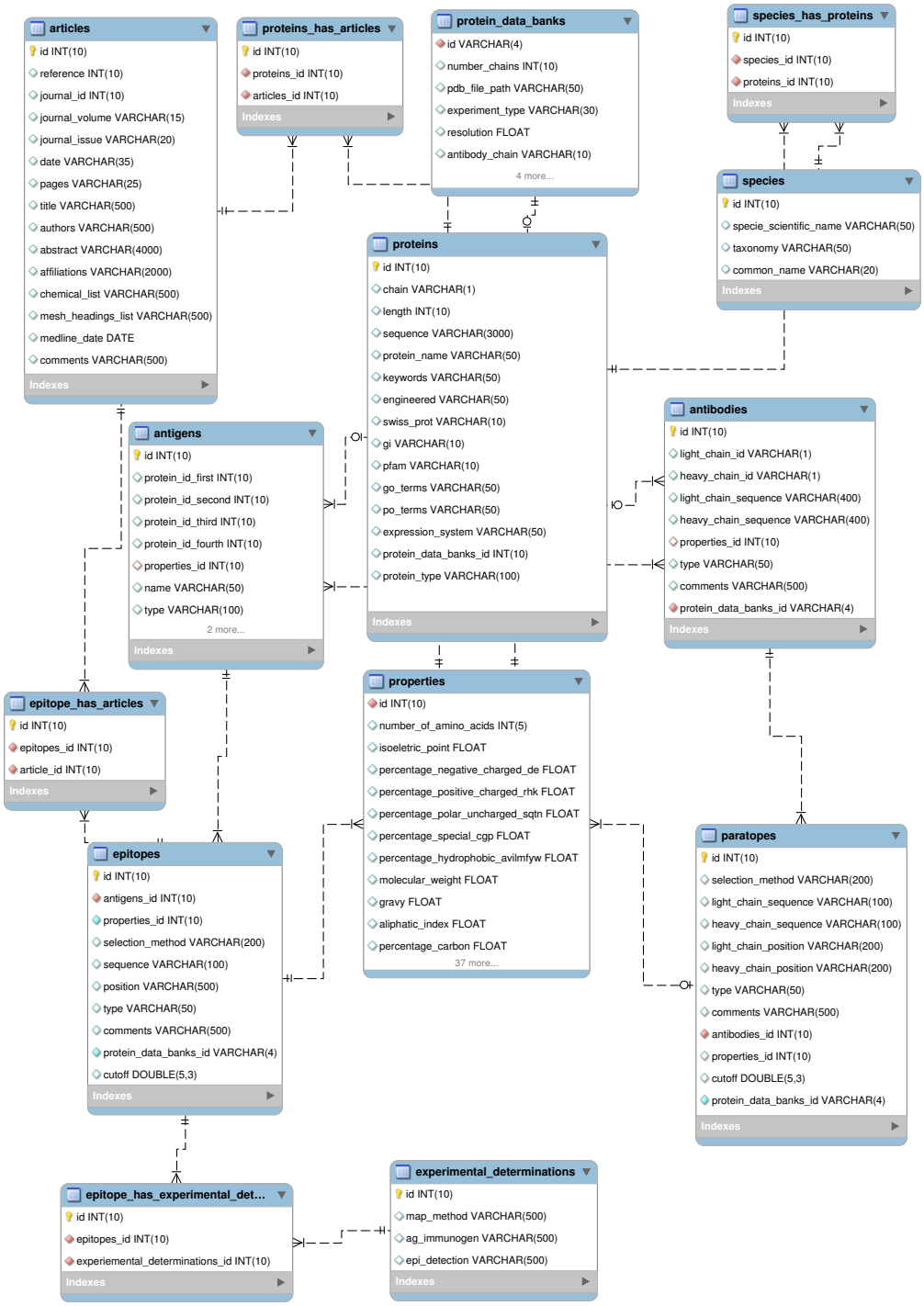


Fig. 3.4: Epitope-Paratope interface database layout.

Table 3.3: Physicochemical properties detail

Property	**Description**
Number of amino acids	Amino acids count
Isoeletric point	Estimative value for Isoelectric Point.*
GRAVY	The grand average of hydropathy is the sum of hydropathy values of all the amino acids, divided by the number of residues in the sequence. *
Aliphatic index	Relative volume occupied by aliphatic side chains.*
Molecular weight	Molecular Weight computed pI/Mw *
Accessibility	Solvent accessible surface when complexed **
Accessibility split	Solvent accessible surface when not complexed **
α-helix	Percentage of Alpha Helix **
3-10 helix	Percentage of 3-10 Helix **
π-helix	Percentage of π-Helix **
Extended conformation	Percentage of Extended Conformation **
Isolated bridge	Percentage of Isolated Bridge **
Turn	Percentage of Turn **
Coil	Percentage of Coil **
Negative charged	Percentage of all negative charged residues (D, E)
Positive charged	Percentage of all positive charged residues (H,K,R)
Polar Uncharged	Percentage of all polar and uncharged residues (G, S, T, C, Y, N, Q)
Special(CGP)	Percentage of Special residues (C,G and P)
Hydrophobics	Percentage of Hydrophobic Amino Acids (G, A, V, L, I, P, F, M and W)
Atomic proportion	Percentage of Carbon, Oxigen, Nitrogen, Hydrogen and Sulfur(Contains 5 properties).*
Percentage of each amino acids	Amino acid composition of a protein sequence (Contains 20 Properties).*

* ExPASy platform tools ProtParam and pI/Mw (Gasteiger et al., 2005).

** Structural properties were computed using Stride (Heinig and Frishman, 2004).

3.5 Web interface for EPI-DB

Once the EPI-DB was filled with the data previously described, a web interface was implemented in order to deal with data visualization and retrieval (Figure 3.5) and was created using a framework based in HTML and PHP. The interface allows to download the full database in MySQL for the users willing to implement it locally. The Data tab allows the user to parse directly from the browser the data but only table by table for now. The site also contains basic statistics of the database to keep track of its evolution.

3.6 Epitope and paratope properties hierarchical clustering analysis

As a first analysis, we investigated the relations between the properties of the paratope and the epitope using clustering methodology. This analysis assessed the capacity of the properties to become potential predictors. From our database we extracted the 101 pairs of paratope-epitope properties corresponding to the I2R selection method. We made a hierarchical clustering of the epitope-paratope properties. Zero-variance properties were filtered out, removing 7 properties and leaving only 37 for this analysis. With these properties we created a square matrix with the absolute Pearson correlations of all paratope properties against all epitope properties. This computation was performed with “cor“ build-in function of R (Becker et al., 1988). This correlation matrix was clustered with the function ”pvclust“ of the package ”pvclust“ (Suzuki and Shimodaira, 2006). We performed 2000 bootstraps and used distance based on absolute correlation. We highlighted clusters with more than 95% unbiased probability and 0.7 of absolute correlation.

The cross correlation analysis between the epitope and paratope parameters forms 6 strongly related cluster of parameters, with an R-value above 0.7 (Figure 3.6). Most of those clusters

are expected in term of parameters, such as Met with Sulfur percentage or aliphatic index with GRAVY. We can also observe the two very close clusters (number 3 and 4) representing negative and positive charged residues respectively. It is important to note that the Lys and His are not part of the positive charged cluster. A surprising aggregation is the presence of the Iso-electric Point (IP) parameter within the positive charge cluster showing that the IP is more related to the positive charged than the negative ones. Concerning the secondary structures parameters we can see the combination of the extended conformation with the turn percentage proving that the structure of the epitope influences the structure of the paratope and vice versa. Considering the cluster Tyr, very important residue for the paratope, we can note the high correlation with the percentage of Coil, therefore linking the most important amino acid of the paratope with secondary structure.

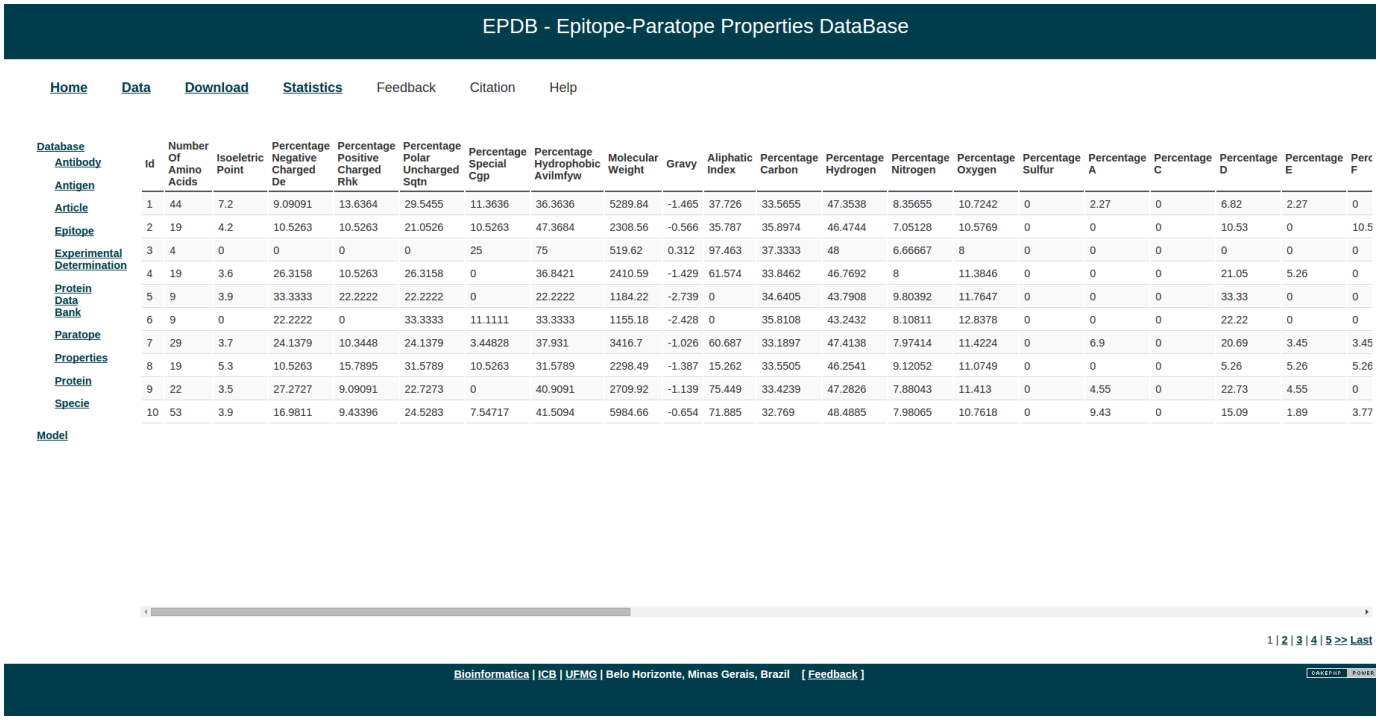


Fig. 3.5: EPI-DB web interface displaying the ten first line of the Properties table. The interface contains four implemented tabs. Home contains the presentation of the EPI-DB and interface. Data, currently displayed tab, allows the user to parse the data directly from the browser, here for example the properties table. Download allow full EPI-DB MySQL file download and Statistics contains generals statistics about the EPI-DB.

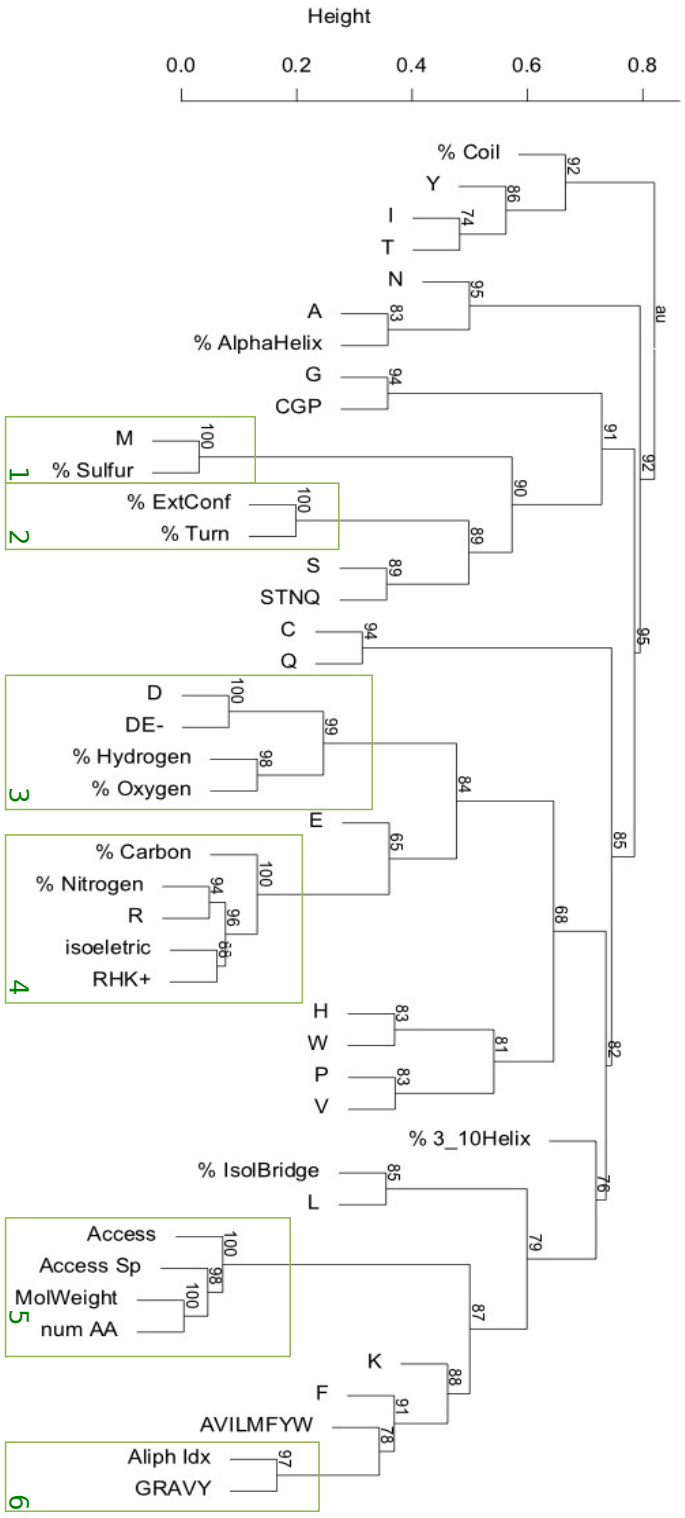


Fig. 3.6: Clustering of epitope and paratope based on absolute correlation. We highlighted clusters with more then 95% unbiased probability and 0.7 of absolute correlation (green square).

3.7 Assessing rightful and mismatched pairs of paratope-epitope based on epitope properties

As shown by the previous results some properties from paratope are correlated with epitope's properties. To analyze more deeply the relation between the properties we used linear model for all the epitope properties to predict a specific paratope property and then combined those linear models to form 'ensemble' able to predict all paratope properties. The objective of this computational methodology is to be able to discriminate a rightful pair of paratope-epitope from a mismatched one. A pair of epitope-paratope being the sequences from the antigen and antibody respectively obtained from an Ab-Ag complex structure using a given selection method and cutoff.

Prediction of a single property using linear model

Linear models were created with the "GLM" R-package using Gaussian family (Dobson, 1990; Hastie and Pregibon, 1992; McCullagh and Nelder, 1989; Venables and Ripley, 2002). Each model was build adding properties one by one and evaluating the change in correlation between predictions and real values. If the new property increased the correlation by at least 0.02, the attribute was accepted or removed from the model. Properties were added in decreasing order of correlation.

Creation of ensemble of linear models for multiple properties prediction

Ensembles of models were created the following way: each linear model made a prediction, this prediction was compared with real value and a z-score was calculated by taking the absolute difference between predicted and the real values and dividing by the y-standard deviation of the prediction. After calculating the z-score for every model we summed them result-

ing in the final prediction value. Only linear models with a certain threshold of correlation were added to the ensemble. In order to test the ranking ability of the ensemble of models we created a set of correct paratope-epitope pairs along with a set of mismatched pairs. We attempted to predict if each epitope was paired to it's corresponding paratope based on the sum-z-score. Pairs (correct and mismatched) were sorted by this score and Area Under the roc-Curve (AUC) was calculated with the function "roc.area" from the R-package "verification" (Mason and Graham, 1982).

Evaluation of the models and ensembles prediction using double cross-validation

In order to have realistic, not over fitted measures of quality for the models we performed a double cross-validation in which the dataset was first divided in 10 k-folds (outer) cross-validation and in each k-fold was performed a leave-one-out (inner) cross-validation. The inner cross-validation was used to build and evaluate the linear models. For each model was repeated the leave-one-out 100 times using the bootstrapping method. The outer cross-validation was used to evaluate the ranking ability of the ensemble, it's training set was composed of 91 instances (rows in the data-set) while the test sets contained 10. The 10 test instances were taken from correct pair (positive data-set), which we shuffled to create another 10 instances of mismatched pairs (negative data-set). Area Under the Curve (AUC) evaluation was performed for the 202 predictions. The minimum threshold of absolute correlation required to be part of an ensemble was tested from 0.3 up to 0.6 with a step of 0.05 in order to maximize AUC the of the prediction. The best ensemble was obtained using absolute correlation cutoff of 0.5 (Figure 3.7) and reached an AUC of 0.6420.

The table 3.4 contains the detail of the properties used by the different models that constitute the best ensemble (cutoff 0.5). The prediction is done by an aggregation of five models that predict the following paratope properties, number of residues, percentage of Trp, percentage of hydrogen, percentage of Asp and the molecular weight based on epitope properties.

The models with the highest correlation correspond to the number of residues and molecular weight which are two related properties and have low discrimination capacity. The percentage of Trp and Asp are more discriminating properties but their correlation are lower.

Those results show the relative prediction capacity of sets of properties and allowed to quantify the quality of prediction for each one of them. This type of methodology could be applied to assess the binding likelihood of a Complementary Determining Regions to an epitope of interest.

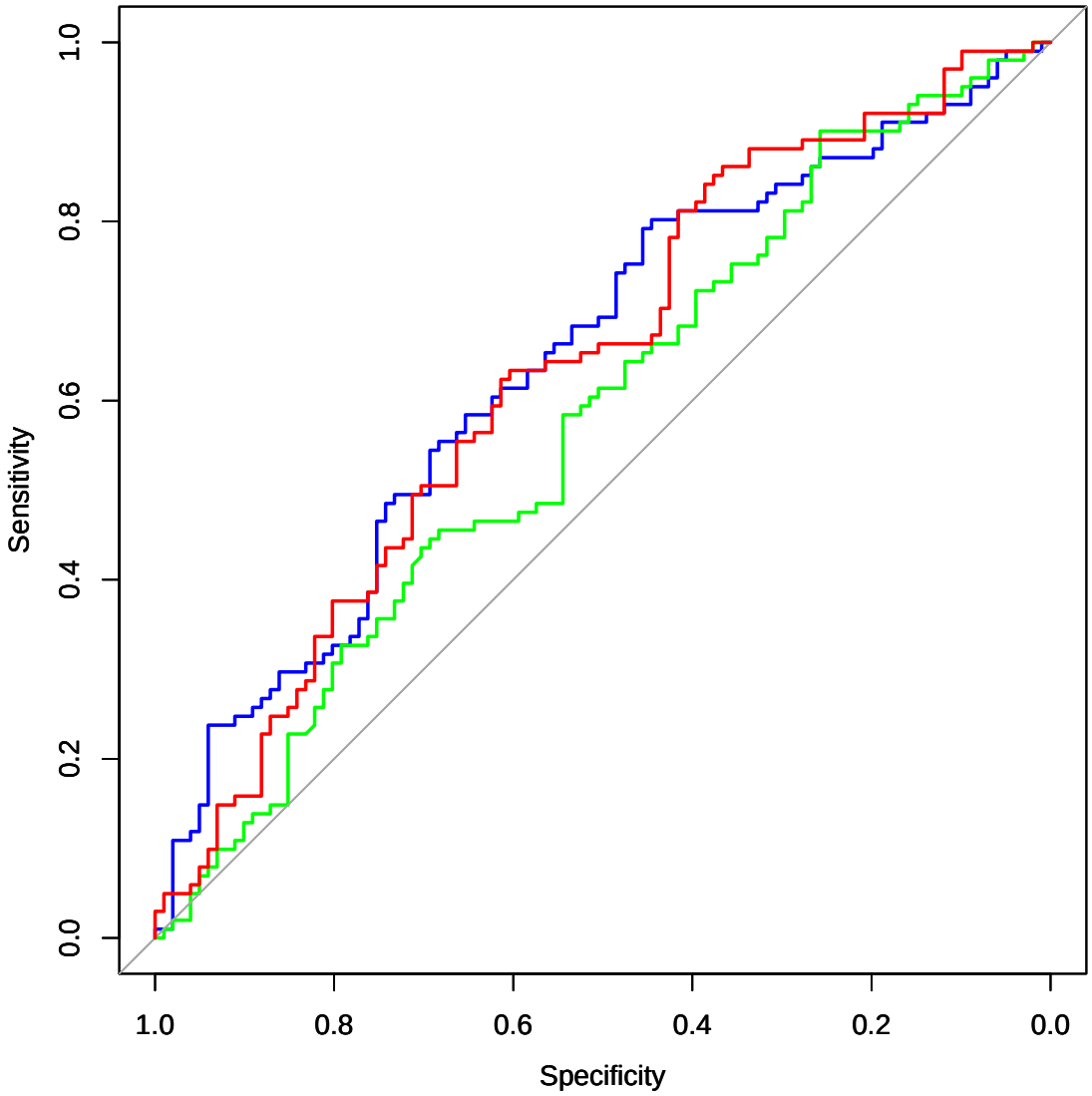


Fig. 3.7: ROC curve of the three best ensemble of models. Each curve represent an ensemble of model with a different minimum threshold for linear model to be part of it. The green curve correspond to a 0.55 of absolute correlation threshold and reach and AUC of 0.5761. The red curve correspond to a 0.45 absolute correlation cutoff and obtain an AUC of 0.6332. The blue curve correspond to a cutoff of 0.5 and reached the highest AUC with 0.6420.

Table 3.4: Best ensemble of models properties detail

	Predicted Paratope Properties				
Epitope Properties	**N°Residues**	**%Trp**	**%Hydrogen**	**% Asp**	**MW**
Molecular Weight	×				×
%Polar& Uncharged	×				
% negative charged		×			
% Iso		×	×	×	
% Trp		×			
% His		×	×		
% Ser		×			
% Cys					
Isoeletric Point			×		
% Leu			×		×
% Phe			×		
% Tyr				×	
Gravy				×	
% Met				×	
% Positive Charged				×	
% Asp				×	
% Val				×	
% Hydrophobics				×	
Aliphatic Index					×
correlation	0.58955	0.5220628	0.5138915	0.511735	0.5993296

CHAPTER 4

EPI-Peptide Designer

Antibodies play an increasingly important role in both basic research and the pharmaceutical industry. Fully understanding the complementarity of epitope and paratope is of great interest but still remains a challenge. Different bioinformatics methodologies are used to gain insight into the molecular mechanisms such as *in silico* alanine scanning (Robin et al., 2014), specialised antibody-antigen docking such as SnugDock (Sircar and Gray, 2010), *ab initio* antibody contact residue prediction such as Antibody i-Patch (Krawczyk et al., 2013) or statistics analysis from set of complex Ab-Ag crystal structure. Crystal structures analysis have been very helpful to understand basics principles of complementarity but lack more complex approaches. Using the previously constructed non-redundant dataset of Ab-Ag structure and computed molecular interaction extracted from the BLUE STAR STING server (Neshich et al., 2006) we developped a tool to represent the interface using graph format. From this representation, using graph extraction and Bayesian probabilities, we implemented Epitope-Paratope Interface (EPI) Peptide Designer, a new tool to generate peptide binder libraries biased based on a target epitope sequence and the patterns extracted from the Ab-Ag interfaces. In order to prove predicted EPI-Peptide capacity to bind, we conducted an experimental validation using LiD1 (GI: 33348850) (Felicori et al., 2006) sequence epitope as target. EPI-PeptideDesigner successfully predicted 301 peptides able to bind to LiD1 protein (65% of the experimentally tested peptides). The detailed methodology is described in the following manuscript submitted to the Oxford Bioinformatics journal.

BIOINFORMATICS

Vol. 00 no. 00 2015
Pages 1–9

EPI-Peptide Designer : a tool for designing specific peptide ligand libraries based on Epitope-Paratope Interactions

Viart B [1], Gonzalez E [1], Dias-Lopes C [1], Oliveira C F B [1], Nguyen C [3], Neshich G [2] , Chávez-Olórtegui C [1] , Molina F [3] , and Felicori L [1*]

[1] Universidade Federal do Minas Gerais, Brazil
[2] Embrapa Informática Agropecuária, Campinas, SP, Brazil
[3] Sys2Diag, FRE3690-CNRS/ALCEDIAG, Montpellier, France

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation:

Antibodies are an important class of biological drugs, but with limitations, such as inadequate pharmacokinetics, adverse immunogenicity and high production costs. Synthetic peptides with high affinity and specificity for the desired target represent an important alternative to antibodies. However, no computational tool exists to guide the design of these peptides.

Results:

To identify the interacting residues in a given antibody-antigen interface we used Interface Interacting Residue (I2R), a selection method based on computed molecular interactions. The aggregation of all the molecular interactions between epitope and paratope residues allowed us to transform the 3D antibody-antigen complex structures into interface graphs. Based on these data and the probability of molecular interaction we developed EPI-Peptide Designer tool that uses predicted paratope residues for an epitope of interest to generate targeted peptide ligand libraries. EPI-Peptide Designer successefully predicted 301 peptides able to bind to LiD1 target protein (65% of the experimentally tested peptides). This tool should enable the development of a new generation of synthetic interacting peptides that could be very useful in the biosensor, diagnostic and therapeutic fields.

Availability:

All software developed in this work are available at http://www.biocomp.icb.ufmg.br/biocomp/

Contact: liza@icb.ufmg.br

1 INTRODUCTION

Protein-protein interactions are at the heart of biological processes and protein functions are highly related to their binding properties (Chakrabarti and Janin, 2002). For instance, the immune response relies on antigen recognition by a specific antibody and the Antibody-Antigen (Ab-Ag) complex represents a specific type of protein-protein interaction characterized by high affinity and specificity. Identifying the key residues and interaction patterns on the Ab-Ag interface could help improving antibody humanization as well as the design of new antibodies (Morea et al., 2000) and peptide ligands based on the antibody properties.

The use of peptides for therapeutic purpose instead of antibodies has plenty of advantages such as lower manufacturing costs, less immunogenic profile, greater stability and better organ/tumor penetration. Several chemical approaches have been generated to overcome therapeutic peptides limitations such as low oral bioavailability and biodistribution (Vlieghe et al., 2010). Indeed, much research effort is focused on the use of peptide ligands as a viable alternative to antibodies in targeted therapies (Wada, 2013). For instance, mimetic peptides derived from the anti-HER2/ERBB antibody can inhibit the tyrosine kinase activity of this receptor and consequently impair tumour growth (Park et al., 2000; Ponde et al., 2011). Presently, over 50 peptide drugs are approved for clinical use (Reichert J., 2010). To guide the design and increase the affinity and specificity of these peptide drugs, different tools, based on various methodologies (e.g., directed evolution, high-throughput protein screening or rational design based on protein-peptide interactions) have emerged (Pei and Wavreille, 2007; Yin et al., 2007; Vanhee et al., 2011). In silico rational design of peptides based on molecular interactions is also a fundamental proof-of-concept for the current understanding of the physical-chemical basis of molecular recognition. Moreover, this approach could become a powerful complement to the current library-based screening methods because it allows targeting specific patches on the surface of a protein (Fleishman et al., 2011). Computational design also gives the opportunity to program protein-protein interactions for specific applications. However, currently no computational methodology to design this kind of peptides is available.

In this work, we propose a computational method to generate libraries of peptide ligands or paratope mimetics based on the Epitope-Paratope Interaction (EPI) patterns and on a target epitope input sequence. This software, called EPI-Peptide Designer, uses a set of Ab-Ag complex structures from the Protein Data Bank (PDB) (Berman et al., 2000) and the BlueStar STING server and STING_DB (Neshich et al., 2006) containing hundreds of interaction descriptors reported in residue by residue fashion

*to whom correspondence should be addressed

to compute the Bayesian probabilities of molecular interactions between epitope and paratope. EPI-Peptide Designer generates peptide binder sequences based on the epitope sequence entered by the user and the patterns extracted from the Ab-Ag interfaces. The method was experimentally validated using as target a dermonecrotic protein LiD1 from the brown spider venom. We have synthesized a library of 460 peptides and 65% of them were able to bind to LiD1. This is, to our knowledge, the first generator of peptide ligand libraries based on EPI.

2 METHODS

Dataset extraction

To extract structures of Ab-Ag complexes from the PDB (Berman et al., 2000), we first used the datasets from Ramaraj et al. and Kunik et al. to select the antibody light and heavy chains to be used as reference sequences. After redundancy removal using CD-Hit (Fu et al., 2012), we processed the two reference sequence datasets with Interface Research Algorithm (IRA), a BioJava program we developed. IRA automatically computed the Smith and Waterman local alignment (Smith et al., 1981) of each sequence against each chain of all the PDB files that contain at least three protein chains. Using a threshold determined by aligning the reference dataset against itself, IRA labelled each chain as Antibody Light, Antibody Heavy or Antigen. IRA selected structures that contain at least one antigen, one light chain and one heavy chain spatially close (i.e., presenting inter-atomic contacts using the 5 Ångström (Å) distance cutoff). From these, the PDB files with X-ray resolution lower or equal to 2.5Å and present in STING RDB were extracted (Neshich et al., 2006).

Interface selection

To analyse the interface of Ab-Ag complexes, we used three different interface selection methods. First, in the selection based on the distance between atoms of the antigen and the antibody (distance-based selection, DBS) (Chothia and Janin, 1975; Lo Conte et al., 1999), an amino acid of the antigen is considered to be part of the Distance Selected Epitope (DSE), if one or more of its atoms are at a distance below a chosen cutoff (in our study, from 3 to 8 Ångström). The Distance Selected Paratope (DSP) is selected in the same manner. Second, in the approach based on the difference of Solvent Accessible Surface (ΔSAS), interfaces are selected based on the loss of solvent accessibility between the separated and the complexed protein (Lo Conte et al., 1999). Third, we developed a selection method in which the interface computed molecular interactions are extracted from STING RDB (Neshich et al., 2006). In this method, the interface is defined by all the amino acids that are involved in the molecular interactions between the antigen and the antibody chains and that are called, therefore, Interface Interacting Residues (I2R). The selected antibody residues form the I2R Paratope and the selected antigen amino acids constitute the I2R Epitope.

Computation of the interface molecular interactions

Molecular interactions (salt bridges, hydrogen bonds, aromatic stacking and hydrophobic interactions) were taken from STING RDB IFR (Mancini et al., 2004). This tool identifies all potential intra- and inter-protein chain contacts stored in STING RDB (Neshich et al., 2006) by (1) classifying the atoms in groups according to their electrostatic behaviour and position in the amino acid (main or side chain) and (2) by then selecting atoms based on the type of contacts they potentially can make and on the experimentally defined distance restrictions (Harris and Mildvan, 1999; Sobolev et al., 1999; Swindells, 1995).

Redundancy removal

To extract meaningful information from the interface dataset, we removed redundancies by selecting only the DSE and DSP sequences from the complex (with a cutoff of 6Å). Using the CD-Hit global sequence identity score (Fu et al., 2012), we only selected interfaces with a score lower than 0.90 for both interface sides. Global sequence identity score is define as the number of identical amino acids in alignment divided by the length of the shorter sequence. The selected files were manually curated to confirm their quality. This provided us with a non-redundant dataset composed of 101 PDB structures, 21 antibody-peptide complexes (here, peptides are defined as molecules smaller than 30 amino acids) and 80 antibody-protein complex.

Interface statistical analysis

To compute the percentage of occurrence (%Occ) of the epitopes and paratopes selected by I2R we used :

$$\%Occ_n = \frac{Occ_n}{Occ_{total}} \times 100,$$

where n is an amino acids, $\%Occ_n$ is the percentage of occurrence of n, Occ_n is the occurrence of n and Occ_{total} is the occurrence of all the residues. The results were compared to all STING RDB protein-protein interaction (Neshich et al., 2006) occurrence values after exclusion of our 101 PDB Files. The statistical comparison of the amino acids was done using a t-test of differential distribution and was considered significant when the p-value was lower than 0.01.

Comparison of the interface selection methods

To compare the interface residue selection by the three methods we computed the Receiver Operating Prime Curve (ROC') of the performance of the distance-based selection and ΔSAS, using various cutoffs, against I2R. As the aim was the comparison of selected interface residues, the true negatives were not considered. We computed the ROC' curve as follows. The True Positive Rate (TPR), also called recall, was computed as:

$$TPR = \frac{TP}{TP + FN}$$

and the False Discovery Rate (FDR) as:

$$FDR = \frac{FP}{FP + TP}$$

where TP is the True Positive, FP the False Positive and FN the False Negative.

Computation of the most frequent interface partners using graph analysis

To analyse the interface in a multi-level manner, we developed Interface to Graph Generator (IGG). IGG is a BioJava program that takes as input PDB codes and two sets of chains. Molecular interactions between those two sets are recovered from PDB structures using STING RDB (Neshich et al., 2006). The interface is

automatically transformed into a graph, where all I2Rs are vertices and all interactions are edges. The vertex label holds the information concerning the interface side and the amino acid type (Table 1). The edges are labelled according to the type of interaction, such as hydrogen bonds, salt bridges, hydrophobic interactions and aromatic stacking. Using GASTON (Nijssen and Kok, 2004), we extracted the most conserved sub-graphs from the complete set of interfaces containing two and three nodes. Subgraphs plot was done using R (R Development Core Team, 2008) and the "igraph" package (Csardi and Nepusz, 2006).

Table 1. Amino acids group used for graph and subgraphs analysis

Group	Residue
Small	A,G
Charged +	K,R,H
Charged -	D,E
Hydrophobic	V,I,L,C,M,P
Alcohol	S,T
Aromatic	Y,W,F
Polar	Q,N

Assessment of paratope residue prediction

Based on the Bayesian probabilities extracted from the epitope-paratope graphs, we predicted the amino acid sequence and the interaction of a given paratope using a given epitope sequence. To evaluate the prediction of residues and interactions, we used a leave-one-out cross validation of the 21 antibody-peptide PDB interfaces from our dataset. Antigens were considered as peptides if their size was equal or lower than 30 amino acids. The evaluation considered each residue from the input epitope and defined as True Positive (TP) a correct "interaction type and paratope residue" couple, as False Positive (FP) any interaction where the interaction type or the residue group was incorrect, as False Negative (FN) any existing couple not added by the program and as True Negative (TN) any possible not existing and not added interaction type-paratope residue couple.

EPI-Peptide Design tool

Using all the Ab-Ag interaction patterns and the residue occurrence data obtained in this study, we developed EPI-Peptide Designer in BioJava. EPI-Peptide Designer includes the IGG program described above. The program takes as input a real or putative epitope sequence (linear or conformational; gaps in the sequence can be represented by -), a cutoff score representing the importance of the epitope sequence in the design and the number and size of peptides needed by the user. To design peptide ligands, EPI-Peptide Designer uses the Base Residue Library (BRL) composed of all residues from all the paratopes in the input dataset. The computed probabilities include: probability of an epitope residue type to do an interaction and, for each type of interaction, the probability of the target paratope residue type and the influence of the epitope neighbour residues on the interaction. Using these probabilities and the input sequence, EPI-Peptide Designer ranks the predicted paratope residues in decreasing order of likelihood. The paratope residues are then added according to the decreasing order of likelihood to the BRL until the defined cutoff score is reached (i.e., for a BRL of 100 residues and a cutoff score of 10%, EPI-Peptide Designer will add 10 residues to the BRL). The thus obtained biased amino acid library (i.e., modified to become specific for a given epitope sequence) is then used to generate random EPI-peptide sequences of the length and in the number defined by the user.

EPI-peptide design, peptide synthesis on cellulose membranes and binding assay

In order to test the effectiveness of the method, we generated 800 EPI-Peptides using the protein LiD1 (GI: 33348850, Felicori et al., 2006) catalytic sequence epitope ($_{37}$FDDNANPEYTYHGIP$_{51}$) and default parameter of EPI-Peptide Designer (Ab-peptide dataset, length of 15 amino acid and a score of 50). To ensure solubility, only sequences which contained less than 50% hydrophobic residues; at least 25% of charged residues and less than 75% of D, E, H, K, N, Q, R, S, T and Y were selected and synthesized (Following recommandations from Life technologies peptide solubility website, http://www.lifetechnologies.com). Four hundred and sixty peptides were synthetized on a cellulose membrane as previously described by Laune et al. The membrane was blocked by incubation with 3% BSA and 5% sacarose at room temperature overnight, and then membranes were probed LiD1 covalently linked to biotin at a concentration of 20μg/ml in blocking buffer at room temperature for 90 min. Biotinalytion of LiD1 was conducted using commercial available Biotinylation kit (Sigma-Aldrich, BK101). Protein binding was revealed by incubation (at room temperature for 90 min) with alkaline phosphatase-conjugated avidin (1:10,000) and 5-bromo-4-chloro-3-indolyl phosphate (BCIP) plus 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) as substrate. To remove molecules and precipitated blue die attached, membranes were sequentially treated with dimethylformamide, 1% SDS, 0.1% 2-mercaptoethanol in 8 M urea, ethanol/water/acetic acid (50:40:10, vol/vol/vol) and, finally, methanol and further employed in other assays. Peptide reactivity was assessed based on manual reading and consensus of triplicate assays. Positive sequences were analysed by GibbsCluster (Andreatta et al., 2013) and Weblogo (Crooks et al., 2004) tools.

3 RESULTS

Analysis of the Interface Interacting Residues (I2R) allows evaluating the distance-based selection and the difference of solvent-accessible surface methods

To compare the three interface residue selection techniques, we selected interfaces from the 101 PDB structures by computing the Euclidean distance DBS, the ΔSAS and the interface molecular interactions (I2R). We then compared the selections made with the DBS and ΔSAS methods against the I2Rs by computing the ROC' curves (Fig.1). Comparison of the selection made based on the Euclidean distance with the extracted I2Rs showed that the maximum precision was obtained with a 3Å distance, while the maximum TPR (also called Recall) was reached with 8Å. The DBS had a higher surface under the curve and the highest value

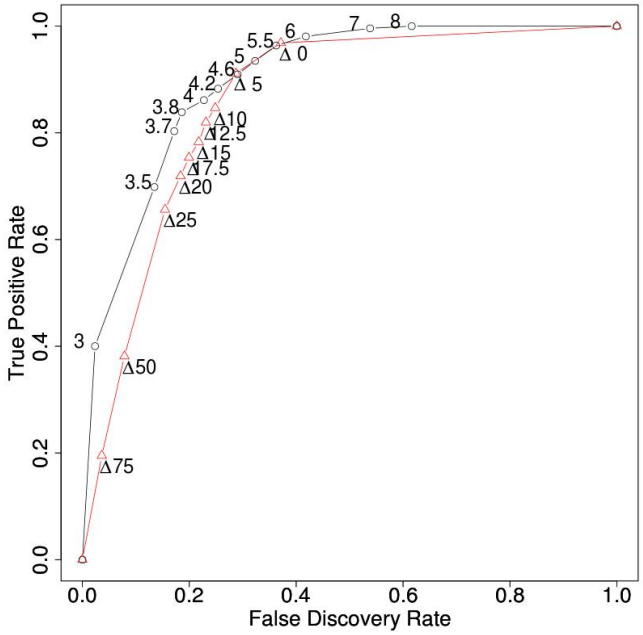


Fig. 1. Comparison of DBS (black circles) and ΔSAS (red triangles) residue selection using different cutoffs relative to the I2R method.

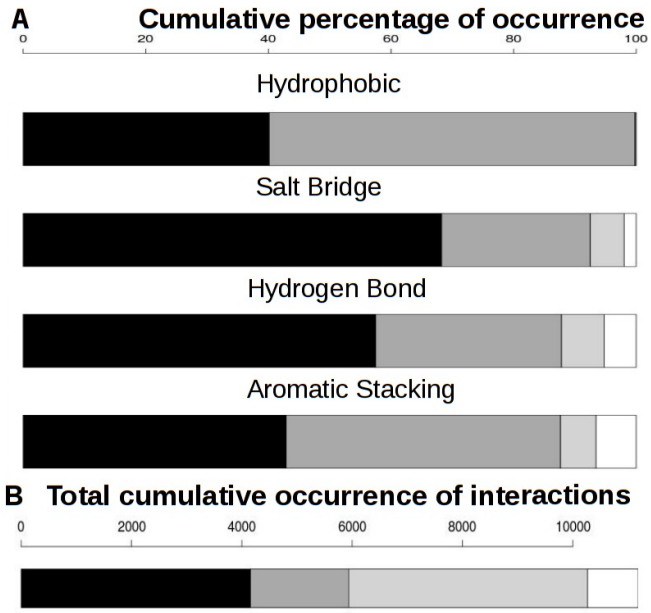


Fig. 2. **A:**Percentage of molecular interactions by type using DSB from 0 to 3Å (black), from 3 to 4Å (dark gray),from 4 to 5Å light gray) and from 5 to 8Å (white). **B:** Cumulative occurrence of hydrophobic interactions (black), salt bridges (dark grey), hydrogen bonds (light grey) and aromatic stacking (white) at the antigen-antibody interface.

of $TPR - FDR$ was reached for a distance of 3.8Å. Most DBS-based Ab-Ag structure studies use a cutoff between 4Å and 6Å. For a distance of 5Å, with this plot, 91.5% (TPR) of interacting residues were selected; however, 32% of the selected residues did not to do any kind of interaction. Surprisingly, to reach the maximum TPR, a distance cutoff of 8Å was needed. As most of the molecular interaction maximum distances are lower than 6Å, we further investigated the interaction repartition.

As all interface interactions are not selected by the 5Å cutoff, we were interested in the interaction repartition in function of the distance. The bar plots (Fig.2A) of the interactions relative to the chosen distance showed that the distance of 5Å, as expected based on the previous results, allowed the selection of most interactions, but still missed 8.5% of them, specifically 2% of all salt bridges, 5.2% of all hydrogen bonds and 6.5% of all aromatic stacking, but none of the hydrophobic interactions. The hydrogen bonds with a distance bigger than 5Å were all water-mediated, thus explaining the unusual long distance. The cumulative bar plot of the interactions (Fig.2B) showed that the hydrophobic interactions were quantitatively the most important, followed closely by hydrogen bonds. Conversely, salt bridges and aromatic stacking were less frequent on the antibody-antigen interface.

Amino acid occurrence in epitopes and paratopes selected with the Interface Interacting Residue (I2R) method

Compared to all interacting residues in STING RDB, I2R paratopes (grey columns in Fig.3) were significantly enriched in Tyr, Ser, Trp, Gly, Asn and Thr. I2R paratopes were depleted of most of the other amino acids, but for Ala, Asp and Phe the occurrence of which was not significantly different compared with all STING RDB interacting residues. I2R epitopes (black columns in Fig.3)

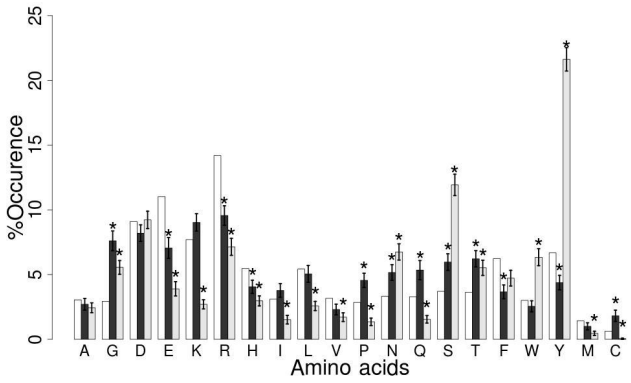


Fig. 3. Comparison of the occurrence (in percentage) of all interacting residues in STING RDB (white), I2R epitopes (black) and I2R paratopes (grey). Error bars are calculated as the standard deviation divided by the root square of the set size. Stars represent statistically significant differences compared to STING RDB, p value <0.01 using a standard t-test.

were enriched in Gly, Pro, Asn, Gln, Ser, Thr and Cys and depleted of Glu, Arg, His Phe and Tyr.

A bipartite graph representation of the paratope-epitope interactions indicated that the interacting residues had a very asymmetric distribution (Fig.4). In the paratope, Tyr, the most frequent residue, interacted with almost all the epitopic amino acids via different types of interactions. Tyr interacted most frequently with hydrophobic amino acids, particularly Pro, Gln, Gly, Phe, and with the charged Lys and Arg in the epitope. Indeed, paratopic

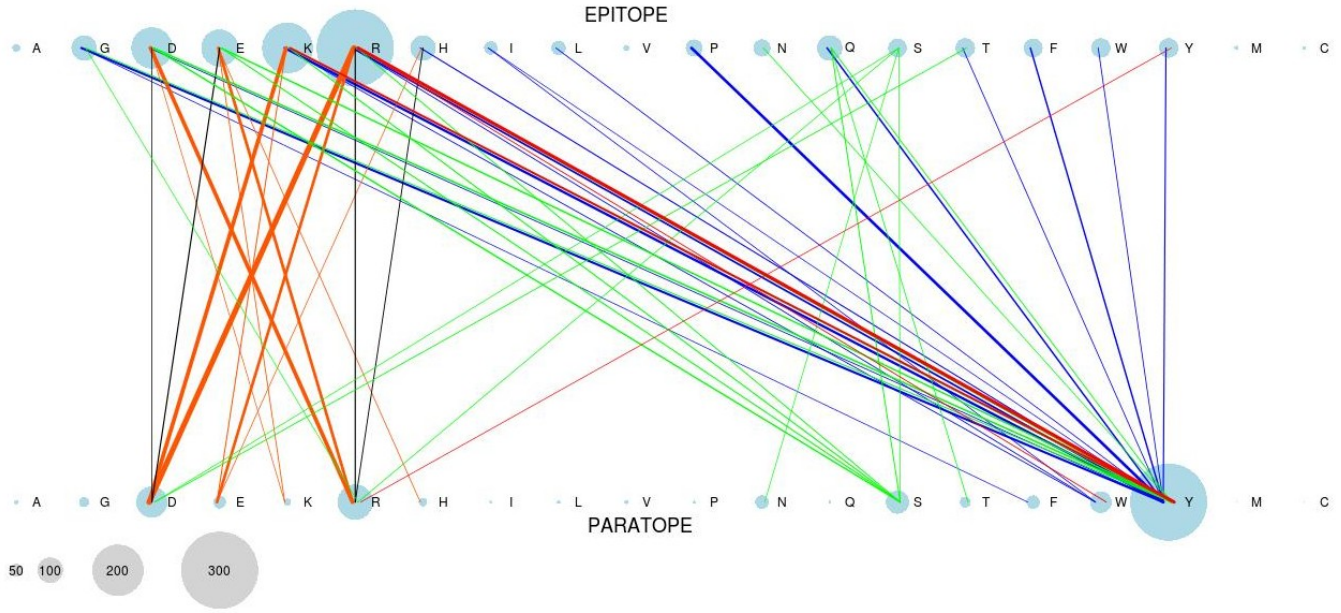


Fig. 4. The bipartite graph representation of the molecular interactions between I2R paratopes and I2R epitopes highlight the strong asymmetric pattern of epitope-paratope interactions. The sphere size of each residue is proportional to the amino acid occurrence in its respective side. The vertex width is proportional to the occurrence of the specific type of interaction; green, hydrogen bonds; blue, hydrophobic interactions; orange, attractive salt bridges; black, repulsive salt bridges; red, aromatic stacking. Only vertices with an occurrence higher than 25 are represented.

Tyr interacted with positively charged epitopic residues via cation-π interactions and with negatively charged epitopic residues via hydrogen bonds. The Ser in the paratope seemed important for establishing a network of hydrogen bonds with charged amino acids and also with Gln and Ser in the epitope. Among the charged amino acids in the paratope, a high prevalence of salt bridges done by Arg and Asp was observed. More heterogeneous interactions were observed among the epitope residues. Although Arg was less frequent than in other kinds of protein-protein interactions (Fig.4), it was the most frequent residue in epitopes and was involved in all kinds of interactions. Epitopic Arg interacted mostly with Tyr residues in the paratope via aromatic stacking, hydrogen bonds and hydrophobic interactions. It also formed salt bridges preferentially with Asp, but also with Glu, and repulsive salt bridges with Arg in the paratope. Lys in the epitope formed a similar network with Tyr in the paratope.

The most conserved subgraphs highlight the importance of cation-π interactions in the epitope-paratope interface

The extraction of the most conserved subgraphs from the complete dataset with two of the three nodes showed that paratopic aromatic residues (Tyr) predominantly interacted with positively charged residues in the epitope through an aromatic stacking interaction (cation-π interaction) (Fig.5A). Specifically, 84 of the 101 selected structures contained at least one cation-π interaction in which the positive charge was hold by the epitope. In addition 51 structures contained a double cation-π interaction (Fig.5B) composed of a positively charged residue in the epitope that interacted with two aromatic amino acids from the paratope. The subgraphs also showed that salt bridges often involved three residues: two negatively charged from the paratope with one positively charged from the epitope. Hydrogen bonds had a low score, although they were the second most frequent type of interaction observed in Ab-Ag interfaces. This can be explained by the variety of amino acid group couples that can form such interaction, thus reducing the frequency of same residue group - same interaction couples.

Assessment of the paratope residue prediction

Using these antibody-antigen graph patterns, we then developed a new methodology to design antibody mimetics using the antigen sequence Fig.6. First, we computed the Bayesian probability of all kinds of interactions to predict the residue-interaction couples. Then, to test the predictions, we used the 21 antibody-peptide interfaces from our dataset and a leave-one-out cross-validation method with all the interactions and the seven residue groups (Table 1). Using a cutoff of 5%, meaning that a paratope-residue interaction couple had to have a Bayesian probability of 0.05 to be added, we obtained a sensitivity of 23% and a specificity of 95%, with an accuracy of 92%.

EPI-Peptide Designer tool

From a set of user-defined Ab-Ag complexes (Fig.6A), the EPI-Peptide Designer computed the graph representation of the interfaces (Fig.6B). Then, from the set of graphs, the program computed the amino acid occurrence in the second side (in our study the paratope) and the interaction probability (Fig.6C and Fig.6D). To demonstrate how the EPI-Peptide Designer works, we used the epitope from the PDB structure 1TET that contains the choleric toxin complexed with an antibody. We chose this example because it was not in our dataset and is a small linear epitope composed of only one segment: VEVPGSQHIDSQKKA. We used our non-redundant 101 PDB structures as dataset input and as epitope input the 5Å epitope extracted from the 1TET structure. Using a score

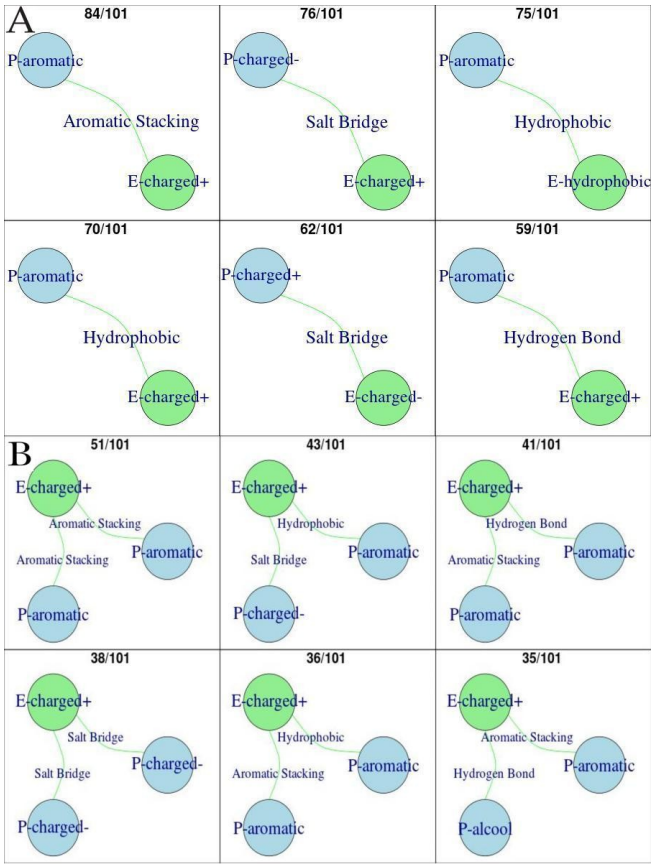


Fig. 5. Each cell contains one of the six most common subgraphs with two (**A**) or three nodes (**B**) from the interface graphs based on the 101 PDB structures dataset. The title indicates in how many interfaces the motif was observed at least once.

of 50% (representing the importance of the epitope sequence in the design), the percentage of occurrence of five amino acids from the BRL was modified by at least 2% (see Fig.6E).

Predicted EPI-peptides are able to bind to target LiD1 protein

To test the ability of EPI-peptide designer to successfully predict peptides able to bind to the epitope ($_{37}$FDDNANPEYTYHGIP$_{51}$) of LiD1 protein, 460 peptides were predicted, chemically synthesized and assayed against biotinylated LiD1. As a control for inespecific binding, the membrane was probed with AvidinAlkaline Phosphatase alone and no reactive peptides were observed (supplementary Figure 1). From 460 sequences synthetized, 218 were considered highly positive (47% of sequences, squares on Fig.7**A**) and 83 had a lower reactivity (18%). 159 peptides (35%) presented no reactivity. Highly positive peptides were clusterized in two groups and a graphical representation of the patterns from each multiple sequence alignment was computed (Fig.7**B**,**C**). Cluster 1 (supplementary table 1) contains 107 sequences and shows two conserved tyrosine at position 9 (59%) and 10 (41%) as well as two conserved arginines (positions 14 and 15). The second most frequent amino acid in those positions is also aromatic (Trp). Similarly, the cluster 2 (supplementary table 2) includes two conserved tyrosines at position 14 and 15 (53 and 54% respectively).

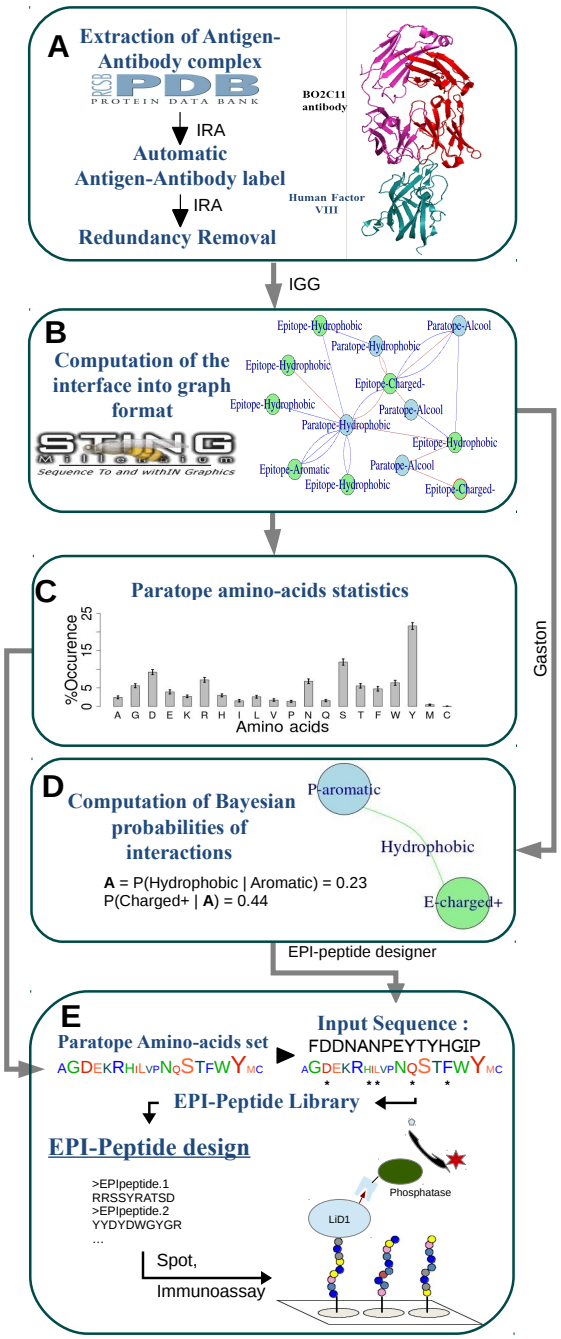


Fig. 6. Schematic of the EPI-peptide Design method to design targeted libraries of peptide ligands. **A**. The Interface Research Algorithm (IRA). **B**. Interface Graph Generator transform the epitope-paratope interface into a graph format using computed molecular interactions extracted the BlueStar STING database. **C**. Computation of the paratope amino acid occurrence using Gaston (Nijssen and Kok, 2004). **D**. Interaction probability. **E**. The epitope sequence modifications are entered in the Based Residue Library (BRL) using the previously computed probabilities. The size of the amino acid font represents the occurrence percentage in the libraries. The star represents amino acid frequencies that have been modified by at least 2% based on the epitope sequence specificity (biased library). EPI-Peptide are synthesized and binding is validated using immunoassay techniques

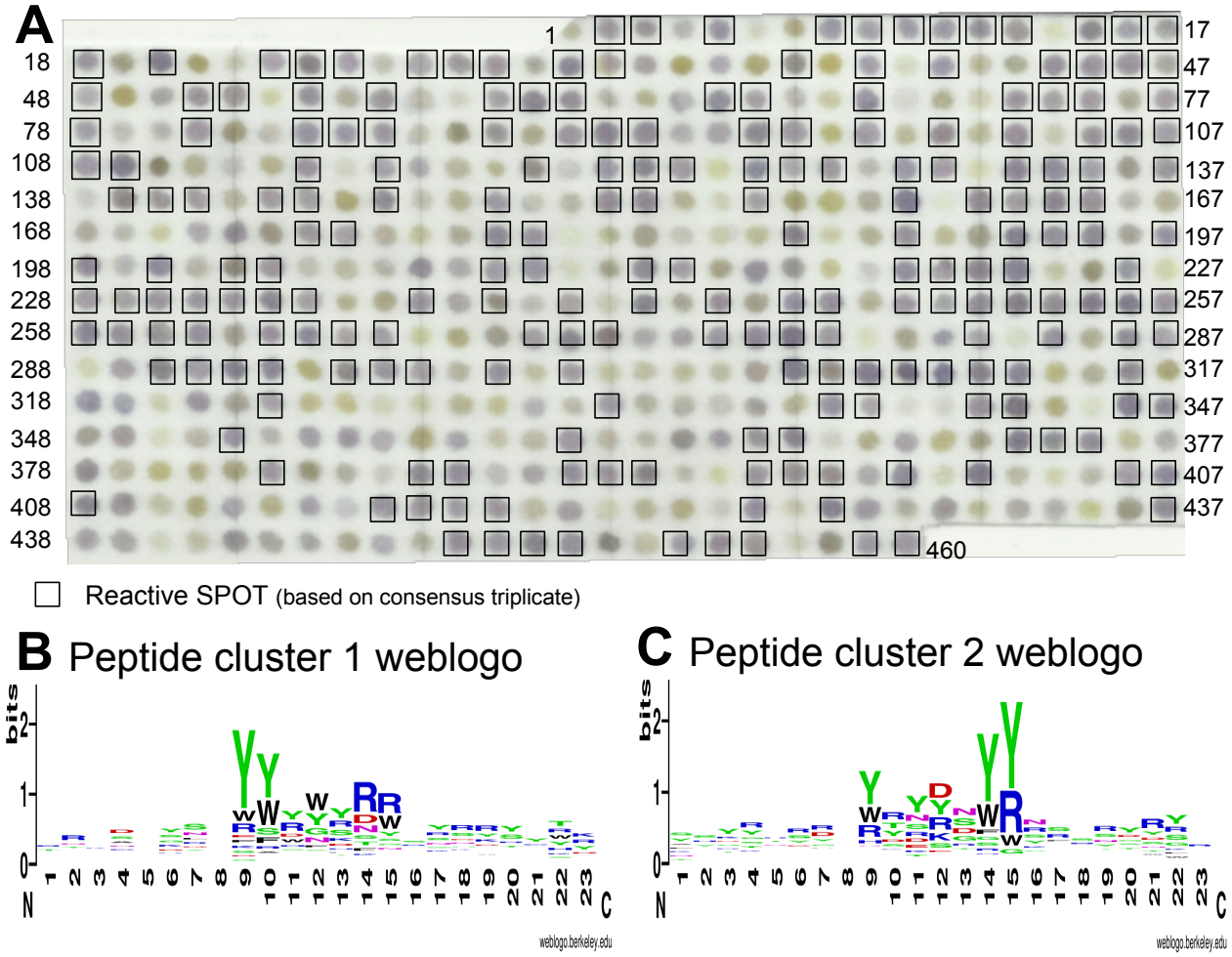


Fig. 7. Experimental validation and analysis of EPI-peptides prepared by SPOT method (**A**). 460 EPI-peptides predicted against LiD1 protein epitope was synthesized. 20 μ/ml of LiD1-biotin followed by alkaline phosphatase-conjugated avidin. 1:10,000 revealed binding peptides. Black boxes represent highly reactive peptides. Weblogo representation of the alignment obtained from reactive peptides grouped in 2 clusters: cluster 1 (**B**) and cluster 2 (**C**).

4 DISCUSSION

To overcome the many antibody limitations, such as their inadequate pharmacokinetics, poor tissue accessibility and adverse immunogenicity including high production costs, enormous efforts have been focused on finding alternative strategies (Yin and Hamilton, 2005), such as non-peptidic protein binders (Margulies and Hamilton, 2010), smaller antibody fragments that retain the original binding property (Hudson and Souriau, 2003; Holliger and Hudson, 2005; Nelson and ert, 2009) and even peptidomimetics inferred from the antibody Complementarity Determining Region (CDR) (Wada, 2013). The generation of CDR-derived peptidomimetics is challenging, but it would pave the way to ample biomedical (therapeutic and diagnostic) applications (Timmerman et al., 2010, 2009; Fontenot et al., 1998; Ponde et al., 2011; Park et al., 2000). However, it has been shown that some positions within the CDRs never participate in antigen binding and some off-CDR residues often contribute critically to the interaction with the antigen (Sela-Culang et al., 2012). For this reason, the present work proposes a new *in silico* methodology to design targeted libraries of ligand peptides that is not based on CDRs, but on the amino acids that are important for the interaction with the antigen. The design of these peptides is not arbitrary, but based on the antigen sequence.

The first step to develop this methodology was to better understand the Ab-Ag interactions. Specifically, we identified the amino acids that are most frequently present in the epitope-paratope interactions, the most frequent physicochemical types of interactions and the most frequent partners in these interactions.

The amino acid frequency in the Ab-Ag interface was analysed in several previous works. However, different cutoffs and methodologies were used to determine the interface boundaries, such as the distance between atoms of the antigen and the antibody (DBS) and the difference of solvent-accessible surface (ΔSAS). Here, we developed a new method based on the interface molecular contact (I2R) to extract from the Ab-Ag interface only the amino

acids that make interactions, using the STING database (Neshich et al., 2006). By comparing the selections obtained using the I2R, DBS and ΔSAS methods, we show that DBS and ΔSAS missed part of the interacting residues that are important for the interface. Indeed, with a distance cutoff of 8Å, 60% of the amino acids that do not interact are selected in addition to the amino acids that do interactions. With a distance cutoff of 4Å, more than 10% of interacting residues are not selected and more than 20% of selected residues are not involved in interactions.

The I2R method also allowed studying the type of interactions and gave an approximation of the residue energetic contribution to the interface in a fast and easy way. Moreover, this selection method could be used to select targets for free-energy perturbation (FEP) (Xia et al., 2012), or to identify binding hot-spots to facilitate the humanization of mouse antibodies (Hanf et al., 2013). As previously noted with other selection techniques (Rubinstein 2008, Kringelum 2012, Ramaraj 2012), we found that the paratope was significantly enriched in Tyr, Ser and Trp residues. However, by comparing the occurrence of the I2R-selected amino acids and of all protein-protein interactions found in the STING database (Neshich et al., 2006), we found that the occurrence of most of the Ab-Ag interface residues was significantly different (but not for Ala, Glu and Phe), thus characterizing the antigen-antibody interface as a special kind of protein-protein interaction. Concerning the extraction of the most frequent partners, we highlighted the importance of the cation-π interaction. Dalkas and colleagues (Dalkas et al., 2014) previously reported that this type of interaction represents only 5% of the Ab-Ag interfaces, whereas in our study 84 of the 101 structures contained at least one cation-π interaction, where the positive charge is hold by the epitope. Moreover, 51 of them contained a double cation-π interaction composed of a positively charged residue in the epitope that interacted with two aromatic amino acids from the paratope. These results suggest that the cation-π interaction is highly conserved interaction in antigen-antibody interfaces but with low frequency as showed by Dalkas et al.

Besides gaining insights into the antigen-antibody interface characteristics, in this work we also describe a methodology to design peptide binders based on the epitope-paratope interface. In addition, this methodology was experimentaly validated showing that 65% of the predicted peptides are reactive. Those peptides contain two consecutive conserved Tyr, a key residue in paratopes. Moreover, those Tyr could interact with hydrophobic amino acids from LiD1 epitope sequence (Phe37, Pro 43, Gly 49, Pro 51) or positively charged residue (Hys 48) via cation-π or even negatively charged residues via hydrogen bond (Asp 38 and Asp 39). The computational design protocol is far from perfect because it does not take into account the antibody structural properties. However, strategies, such as cysteine-constrained peptides, could be employed to mimic antibody loops as shown by Burns et al. and thus force a constrained conformation of our predicted peptides. In conclusion, our study provides insights into the principles that guide Ab-Ag interactions and describes an original methodology (EPI-Peptide Designer) to design ligand peptide libraries, based on a given antigen sequence. These targeted peptide ligand libraries might be useful for proteomic and high-throughput analyses for antigen characterization because they minimize the work to produce antibodies *in vivo*. Finally, this methodology might guide the development of a new generation of biosensors as well as therapeutic and diagnostic molecules.

Funding

This research was supported by Coordenação de Aperfeioamento de Pessoal de Nível Superior, Brazil (CAPES), Fundação de Amparo a Pesquisa do Estado de Minas Gerais, Brazil (FAPEMIG) and by funds of the Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil (CNPq).

REFERENCES

Andreatta, M., Lund, O., and Nielsen, M. (2013). Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. Bioinformatics, **29**(1), 8–14.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Res., **28**(1), 235–242.

Burns, V. A., Bobay, B. G., Basso, A., Cavanagh, J., and Melander, C. (2008). Targeting RNA with cysteine-constrained peptides. Bioorg. Med. Chem. Lett., **18**(2), 565–567.

Chakrabarti, P. and Janin, J. (2002). Dissecting protein-protein recognition sites. Proteins, **47**(3), 334–343.

Chothia, C. and Janin, J. (1975). Principles of protein-protein recognition. Nature, **256**(5520), 705–708.

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. Genome Res., **14**(6), 1188–1190.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, **Complex Systems**, 1695.

Dalkas, G. A., Teheux, F., Kwasigroch, J. M., and Rooman, M. (2014). Cation-, amino-, -, and H-bond interactions stabilize antigen-antibody interfaces. Proteins, **82**(9), 1734–1746.

Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E. M., Wilson, I. A., and Baker, D. (2011). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science, **332**(6031), 816–821.

Fontenot, J. D., Tan, X., and Phillips, D. M. (1998). Structure-based design of peptides that recognize the CD4 binding domain of HIV-1 gp120. AIDS, **12**(12), 1413–1418.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, **28**(23), 3150–3152.

Hanf, K. J., Arndt, J. W., Chen, L. L., Jarpe, M., Boriack-Sjodin, P. A., Li, Y., van Vlijmen, H. W., Pepinsky, R. B., Simon, K. J., and Lugovskoy, A. (2013). Antibody humanization by redesign of complementarity-determining region residues proximate to the acceptor framework. Methods.

Harris, T. K. and Mildvan, A. S. (1999). High-precision measurement of hydrogen bond lengths in proteins by nuclear magnetic resonance methods. Proteins, **35**(3), 275–282.

Holliger, P. and Hudson, P. J. (2005). Engineered antibody fragments and the rise of single domains. Nat. Biotechnol., **23**(9), 1126–1136.

Hudson, P. J. and Souriau, C. (2003). Engineered antibodies. Nat. Med., **9**(1), 129–134.

Kunik, V., Peters, B., and Ofran, Y. (2012). Structural consensus among antibodies defines the antigen binding site. PLoS Comput. Biol., **8**(2), e1002388.

Laune, D., Molina, F., Ferrieres, G., Villard, S., Bes, C., Rieunier, F., Chardes, T., and Granier, C. (2002). Application of the Spot method to the identification of peptides and amino acids from the antibody paratope that contribute to antigen binding. J. Immunol. Methods, **267**(1), 53–70.

Lo Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. J. Mol. Biol., **285**(5), 2177–2198.

Mancini, A. L., Higa, R. H., Oliveira, A., Dominiquini, F., Kuser, P. R., Yamagishi, M. E., Togawa, R. C., and Neshich, G. (2004). STING Contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. Bioinformatics, **20**(13), 2145–2147.

Margulies, D. and Hamilton, A. D. (2010). Combinatorial protein recognition as an alternative approach to antibody-mimetics. Curr Opin Chem Biol, **14**(6), 705–712.

Morea, V., Lesk, A. M., and Tramontano, A. (2000). Antibody modeling: implications for engineering and design. Methods, **20**(3), 267–279.

Nelson, A. L. and ert, J. M. (2009). Development trends for therapeutic antibody fragments. Nat. Biotechnol., **27**(4), 331–337.

Neshich, G., Mazoni, I., Oliveira, S. R., Yamagishi, M. E., Kuser-Falcao, P. R., Borro, L. C., Morita, D. U., Souza, K. R., Almeida, G. V., Rodrigues, D. N., Jardine, J. G., Togawa, R. C., Mancini, A. L., Higa, R. H., Cruz, S. A., Vieira, F. D., Santos, E. H., Melo, R. C., and Santoro, M. M. (2006). The Star STING server: a multiplatform

CHAPTER 5

Epitope-Paratope interfaces shows differences depending on the antibody's organism source

Antibody as drugs present various limitations but one of the most important is the necessity of animal immunization therefore requiring humanization before they can be used. Antibody engineering depends on the knowledge of antibodies and their interactions. To help this process it would be useful to know more about the differences of interface properties between mice antibodies and human ones. To do so we used EPI-DB data and interface properties. We created two groups, the Mouse group composed of all the structures where the antibody had a murine origin and the Human group composed of all the Human antibodies. The Mouse group contained 316 structures while the Human one contained 203 as detailed in the table 3.2. The complete dataset presenting redundant structures we used the 101 non-redundant dataset containing 56 Human antibody complexed structures and 45 murine ones.

5.1 Mouse and human interface's shows different amino acids statistics

In order to determine if the Human's antibody interfaces might differ from the mice's Ab in terms of amino acids statistics we computed a statistical analysis of the residues in both groups. The I2RE percentage of occurrence (%Occ) from the Human and Mouse group

show remarkable differences (Figure 5.1**A**). The Mouse epitopes show an increase in negative charged residues, Lys, Asn and Ser while the Human I2RE are enriched in aromatics residues and Ile. Concerning the I2RP (Figure 5.1**B**) the Tyr shows a remarkable increase in the Mouse's paratope followed by Thr and Asn (5.7, 2.9 and 2.2% respectively). The Human paratope presents higher occurrences for the hydrophobic residues and also Phe. Those results show that it exists differences in the interfaces from the Mouse and Human groups.

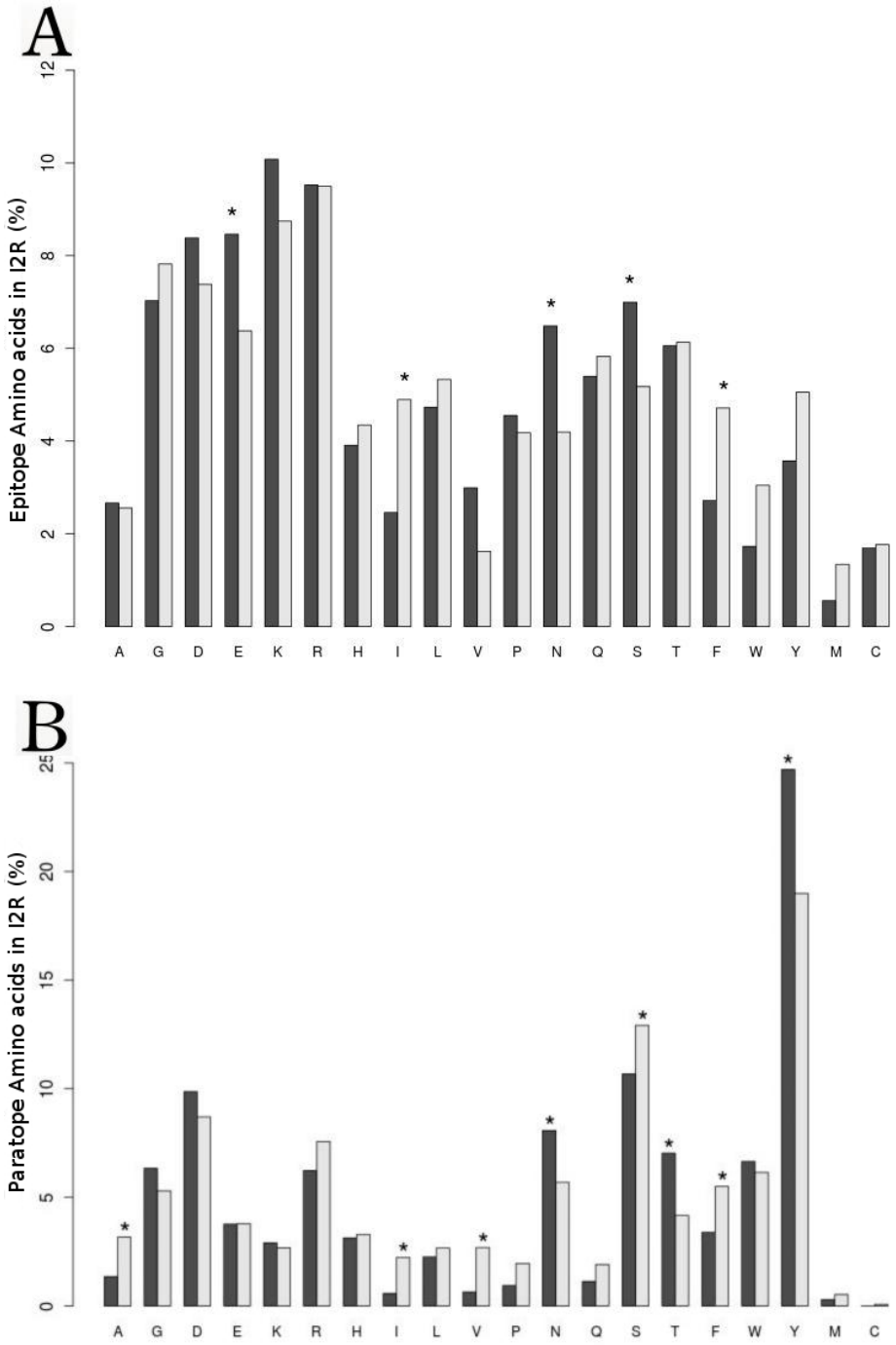


Fig. 5.1: Comparison of the Mouse and Human groups I2R. percentage of occurrence of the I2R epitope (**A**) and I2R paratope (**B**) from the Mouse (Dark Grey) and Human Group (Light Grey). Star indicate a variation of at least 1.5% between the two groups.

5.2 Interacting residues energy analysis

The statistic study of the interface leads to the conclusion that it exists important variations between the Mouse and Human group interface amino acids statistics. To further investigate those differences we realized a residue's percentage of occurrence in function of the average contact energy plot for the Human and Mouse groups. Molecular interactions were extracted from STING_RDB (Oliveira et al., 2007) and the energy was taken from the standard energy of contact defined in BLUE STAR STING web server(table 5.1).

Table 5.1: BLUE STAR STING values for the contacts energies

Contact Type	STING Contact Energy [Kcal/mol]
van der waals	0.08
Hydrophobic interaction	0.6
Aromatic stacking	1.5
H-Bond	2.6
salt-bridge	10.0

Concerning the paratope (Figure 5.2), we can observe the residues split in 3 different groups. The Charged residues with high occurrence and very high energy per residue, the very frequent with a low energy composed of Tyr and Ser and the rest of the amino-acids that have low occurrence and low interface energy. Comparing the Mouse and Human plots we can see that the Human I2RP uses more E (Glu), D(Asp) and K (Lys) even if their %Occ are very similar. As we previously saw the Ser is increased in Human group while the Tyr is impoverished but no energetic variation is observed within the two groups. The epitope analysis (Figure 5.3) shows a two groups organization with on one hand the high energy and high %Occ composed of the charged residues and on the other hand all the other residues. The main difference between the Mouse and the Human epitopes lies on the switch of E (Asp) and D (Glu). Mouse I2RE promotes the E (Glu) while R (Arg) is lowered in energy but not in %Occ. We can also note higher energy of Met and His in this group. Concerning the Human group the Asp is

enriched energetically while being less frequent in the Mouse group. Those two plots clearly show the most important residues for the interface as well as differences in the composition of the Mouse and Human groups.

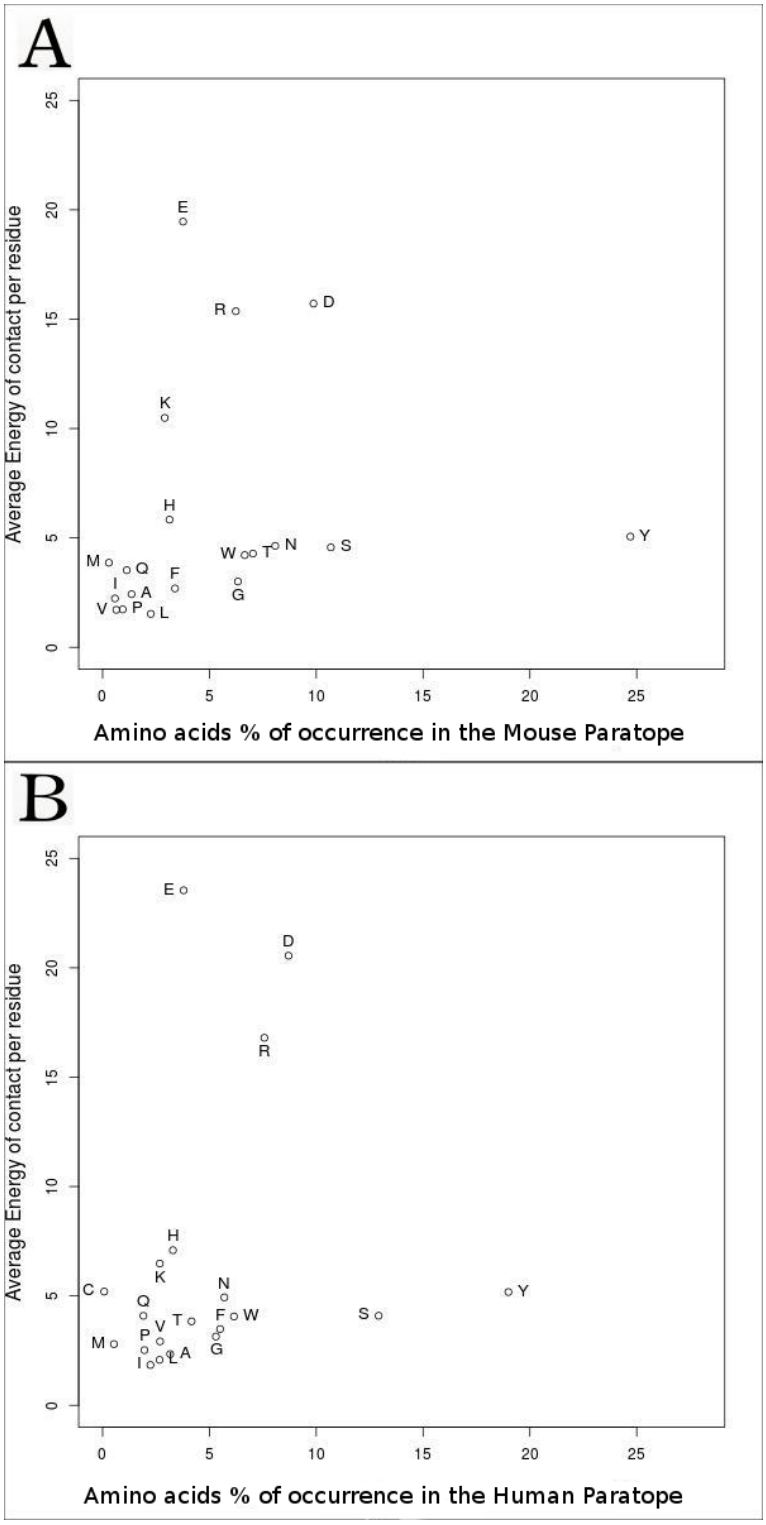


Fig. 5.2: **A** Plot of the I2RP Mouse residues in function of %Occ and average energy of contact. **B** Plot of the I2R Paratope Human residues in function of %Occ and average energy of contact.

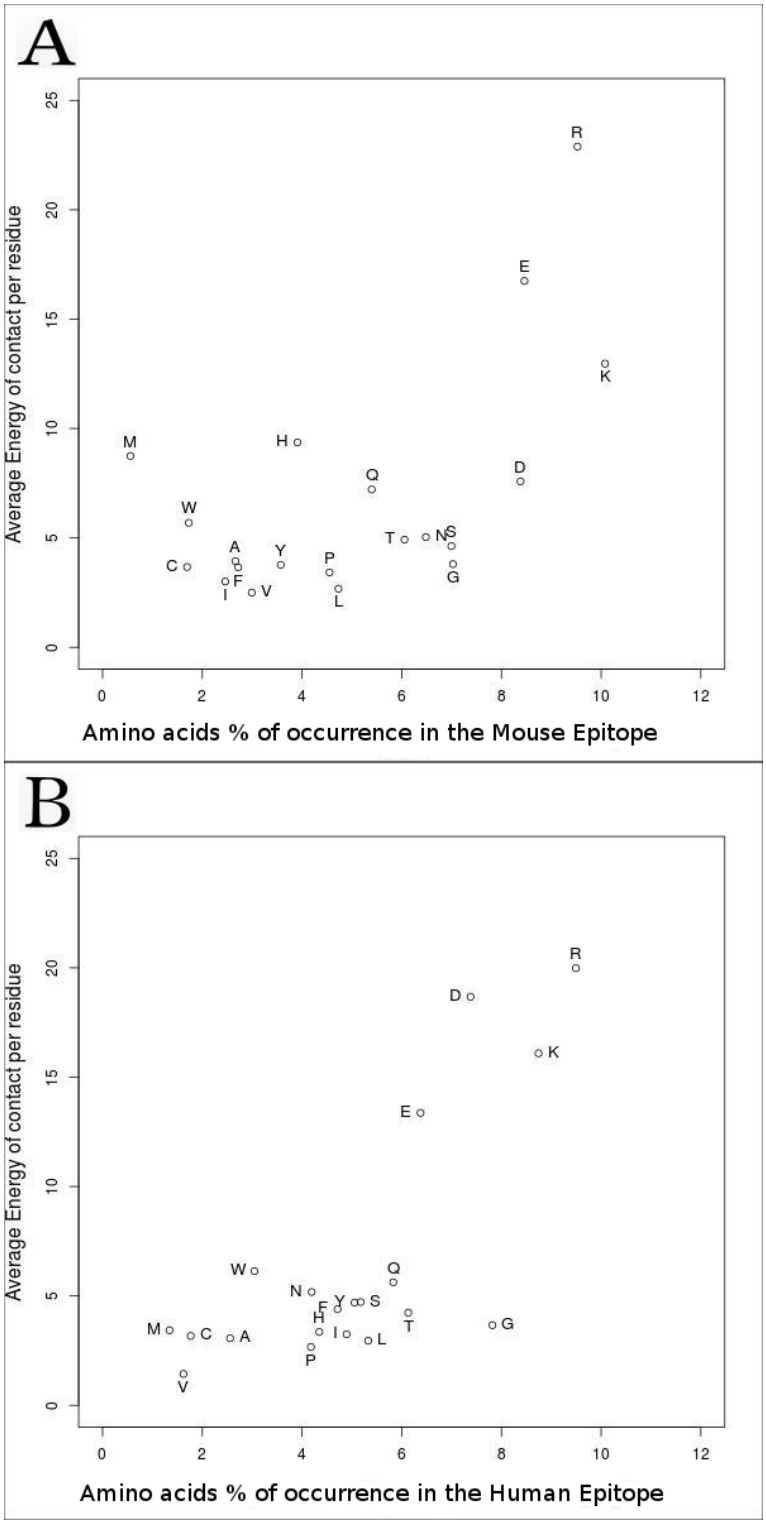


Fig. 5.3: **A** Plot of the I2RE Mouse residues in function of %Occ and average energy of contact. **B** Plot of the I2R Epitope Human residues in function of %Occ and average energy of contact.

5.3 Most common subgraphs analysis reveals differences in interaction patterns

The previous results confirmed noticeable differences between the Mouse and Human group interfaces in terms of amino acids statistics as well as molecular interactions. In order to better understand the differences of interaction patterns from one group to another we conducted the common subgraph ranking throughout both groups using the same methodology as described in Epi-Peptide Designer publication. Comparing the most common subgraphs we can observe that 4 out of 5 are identical between the two groups but their order changed (Figure 5.3). The most common subgraph is the same for both interface but the second most common in the Human group is ranked fifth in the Mouse group. The two subgraphs that aren't shared between the two groups, fifth from Human and fourth from Mouse, have partially the same structure. The positive charge is hold by epitope in both cases. The fifth subgraphs of the Human group correspond to the fourth most frequent subgraph with three nodes of the complete dataset (Fig.5, EPI-Peptide Designer publication), we can then deduce that this motif is very specific from the Human group. Moreover the fourth subgraph from the Mouse group does not appear in the six most frequent ones from the complete dataset, meaning that this subgraph is nearly exclusive to the Mouse interface.

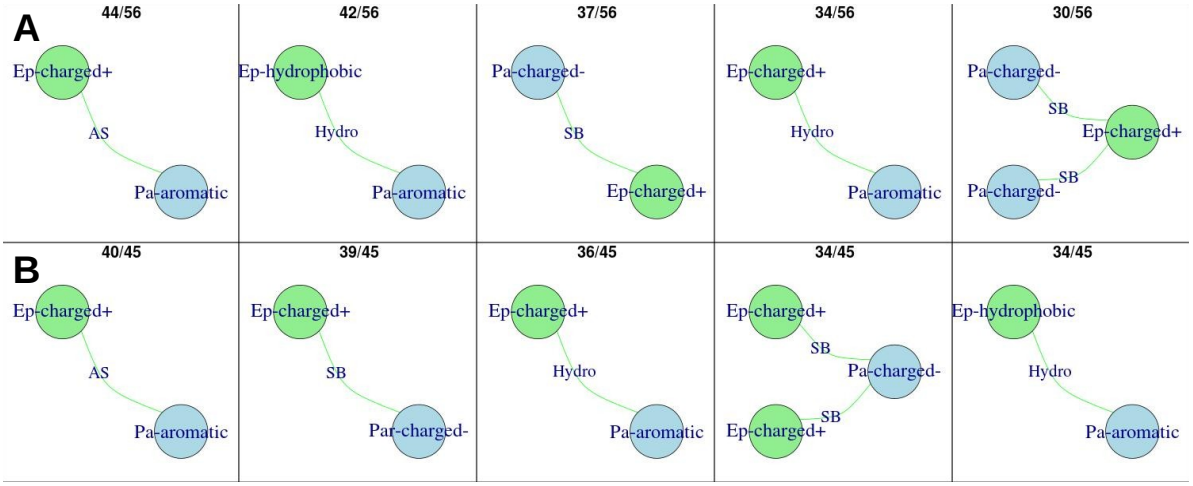


Fig. 5.4: Each cell represent one of the five most conserved subgraphs from the interface where the antibody comes from Human (**A**) or Mouse (**B**). The title indicates in how much interfaces was the subgraph found. 'Ep' stands for Epitope and 'Pa' for paratope. Molecular interactions used here are: Aromatic Stacking (AS), Hydrophobic (Hydro) and Salt Bridge (SB)

5.4 Epitope complexed with mice antibodies have higher coil and turn occurrence

To further analyze the differences between the Mouse and Human interfaces we computed the secondary structure using STRIDE (Heinig and Frishman, 2004) and assessed the representation of coil,helix,turn,α-helix and strand. For this part we only used protein antigen. Very small antigens have a very high flexibility and their structure contains a very high amount of unstructured parts. The Mouse protein group contains 34 structures while the human protein contains 44. We also used the DBS 5Å to select epitope since the previous results showed the I2R secondary structure lack prediction capacity. As we can see on the figure 5.5 the most represented secondary structure is the turn with more than 30%. The α-helix, coil and strand have representations of 19.2, 22.8 and 19.6% respectively. Those results are similar to Rubinstein's (Rubinstein et al., 2008), the epitope are mostly composed of unorganized structure (Turn and Coil). Comparing the Human and Mouse secondary structures, we can notice that

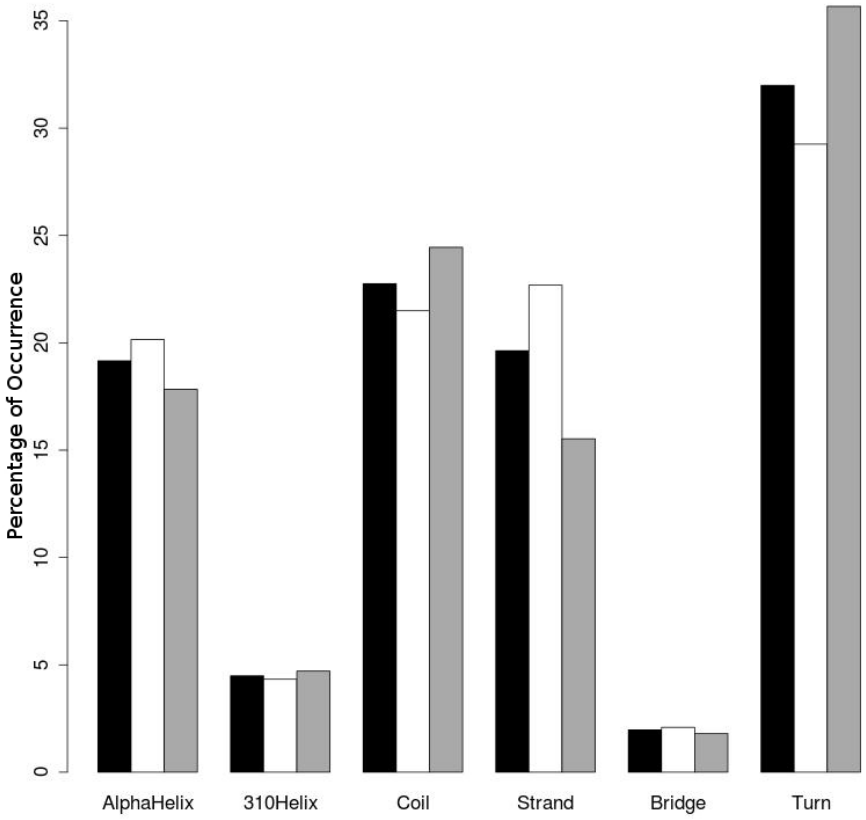


Fig. 5.5: Percentage of secondary structure of the DSE 5Å residues from the Protein Group(black), Human protein(white) and Mouse protein(Gray)

the strand representation is higher in the Human group while the coil and turn are enriched in the Mouse DSE. This observation reinforces the idea of different patterns of recognition of epitope for the antibody from Humans and Mouse even on a structural level.

5.5 Prediction of Antibody species from epitope sequence using linear regression analysis

As shown by the previous results epitope recognized by mice and humans antibodies have different patterns in terms of amino acids and molecular interactions. In order to understand better the differences between the Mouse and Human groups we tried to predict the group based only on the epitope properties. In this experiment we tried the group prediction with two different methods of selection, Distance based Selected Epitope (DBE) with 6Å cutoff and Interface Interacting Residues Epitope (I2RE). We also tested physicochemical and structural properties independantly. This led to four different datasets by crossing selection method and properties:

- Distance based Selected Epitope with PhysicoChemical Properties (DSE.PCP)
- Distance based Selected Epitope with Secondary Structure properties (DSE.SS)
- Interface Interacting Residues Epitope with PhysicoChemical Properties (I2RE.PCP)
- Interface Interacting Residues Epitope with Secondary Structure properties (I2RE.SS)

The regression analysis was applied to the matrices using as criteria the equation 1 as described by McDonald (2009). We assigned $b(i) = log(0.99/(1-0.99))$ when mouse and $b(i) = log(0.01/(1-0.01))$ when human for equation 2. By solving (2) we used (1) as a linear predictor and to minimize the mis-classification rate as described before in Elden (2007). Logistic regression equation 1:

$$log(\frac{p(x)}{1-p(x)})\beta_0 + x\beta$$

Multiple regression model, equation 2:

$$Y_1 = a + b_1 \times X_{1,1} + b_2 \times X_{1,2} + b_3 \times X_{1,3}...$$

$$Y_2 = a + b_1 \times X_{2,1} + b_2 \times X_{2,2} + b_3 \times X_{2,3}...$$

To compare the different linear models obtained we used a multiple ROC curve using False Positive Rate (FPR) and True Positive Rate (TPR) computed as follow :

$$FPR = \frac{FP}{FP + TN}$$

$$TPR = \frac{TP}{TP + FN}$$

with TP being True Positive, FP False Positive, FN False Negative and TN true negative.

The figure 5.5 shows the results of the different multiple regression predictions. For the PCP set of parameters, comparing the I2RE and DSE selection we can see that surprisingly the prediction is slightly better using distance selected epitope. The significant parameters p-value ≤ 0.05) for the DSE.PCP are presented in the table 5.2 with their respective p-values. The most important parameter for the prediction are the polar and alcohol amino acids. This suggest a difference of the hydrogen bond interactions patterns between the Mouse and Human group. Most notable differences are found using the structural parameters. The DSE secondary structure parameters allow a perfect prediction of the group using all six parameters while the I2RE structure give a worst prediction than the PCP. Those observations confirmed the previous idea that epitope residues, spatially close from the Ab but not making interactions, play a structural role that is important for the interface. The ability to predict affinity for a specific specie's antibody could improve immunization results.

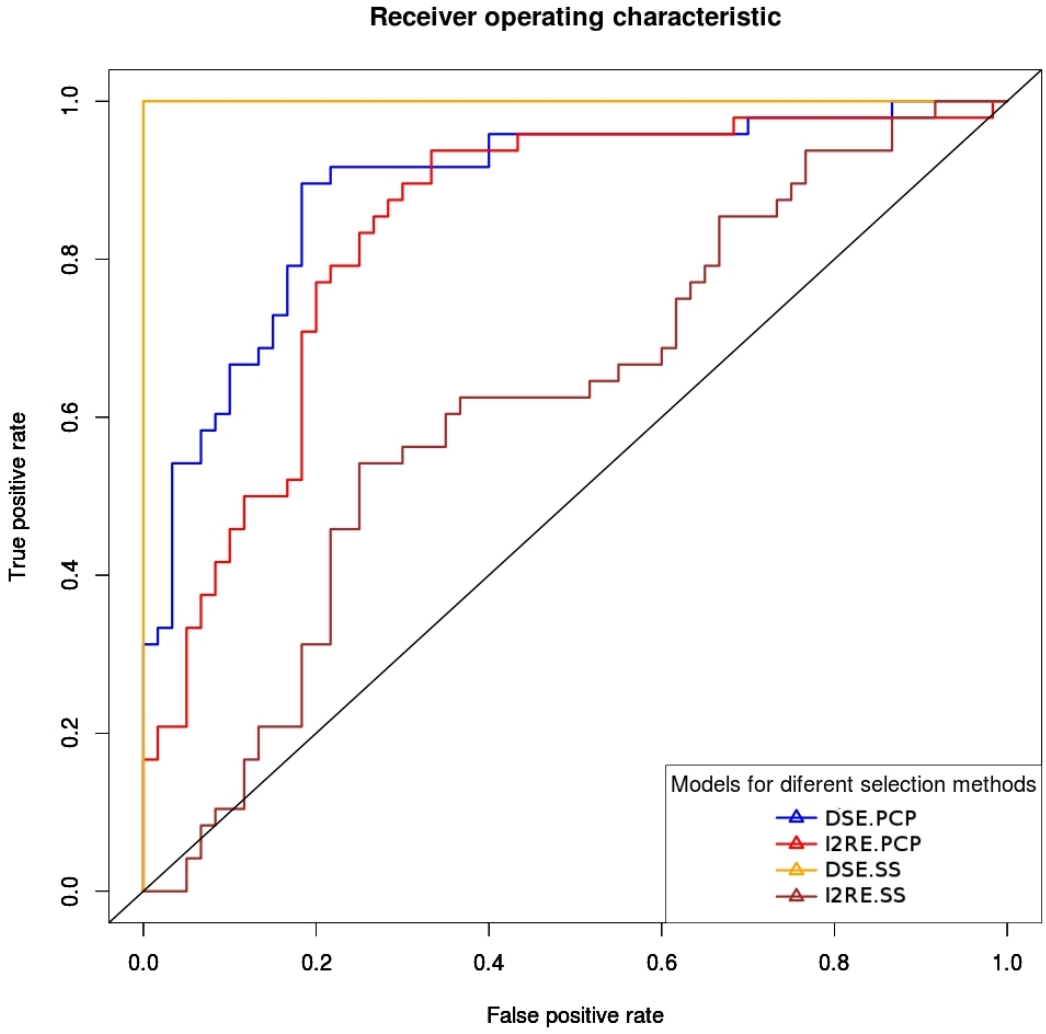


Fig. 5.6: ROC curve of the linear model prediction for the distance based selection epitope and PCP (DSE.PCP), DSE and secondary structure (DSE.SS), I2RE and PCP (I2RE.PCP) and I2RE secondary structure (I2RE.SS)

Table 5.2: Significant parameter for DSE.PCP

Parameter	P-value
Percentage N	0.00471
Percentage Q	0.00678
Percentage S	0.01839
Percentage T	0.01939
Percentage H	0.02027
Aliphatic Index	0.02243
Negative charged	0.02718
Percentage W	0.0274
Percentage H	0.02771
Percentage Y	0.02775
Percentage A	0.02796
Gravy	0.02913
Percentage R	0.02987
Percentage K	0.03159
Percentage L	0.03856

CHAPTER 6
Discussion

Antibodies are used in a broad spectrum of context in biology due to their specificity and affinity for their target protein. This large diversity of applications has triggered a lot of research in order to understand their structure, the mechanisms of their paratope diversity, how to engineer antibodies against desired target and produce large quantity of them. Some of this problems have been solved to some extent and led to three Nobel prizes in physiology during the last fifty years. In 1972 Gerald M. Edelman and Rodney R. Porter were awarded for their discovery on the structure of the antibodies, in 1984 Niels K. Jerne, Georges J.F. Köhler and César Milstein for the production's principle of monoclonal antibodies. Finally Susumu Tonegawa won the Nobel prize in 1987 for the discovery of genetic principle for the generation of antibody diversity.

Those discoveries enabled the apparition and development of the immunotherapy. With nowadays more than 30 antibodies approved by the Food and Drug Administration more limitations of antibodies appeared such as the poor tissue accessibility, fast clearance and high cost of production and storage. The first approach to overpass those limitations was a switch to smaller forms of antibody retaining binding capacity such as fragmented antibody. In a second time, researchers attempted to obtain binding using small protein sequences through large combinatorial libraries or by using existing interface called rational design and use an existing protein-protein interface to engineer in order to create a new binding molecule to a specific target. Those techniques are limited by our knowledge of the complementary mechanisms of PPI.

Antibody-antigen is a specific type of PPI possessing very high affinity and sensibility as well as a great source of inspiration for rational design. Existing research on the Ab-Ag interface lack deep analysis in terms of properties and complexed patterns extractions necessary to guide and help design of new protein binding domain. In order to understand better the mechanisms and characteristics of this interface we developed the Epitope-Paratope Interface DataBase (EPI-DB). EPI-DB was especially designed to store informations extracted from analysis of Ab-Ag complexed crystal structures that provide a more reliable source to investigate the complementary binding properties. EPI-DB was implemented in MySQL reliable and free SQL language and uses different techniques of selection with a broad range of threshold. Storing interfaces selected using distance based and difference of solvent accessible surface is especially interesting since it requires a substantial amount of computation. For each interface, we selected the sequences as well as the position in the chain and stored this information for easy retrieval and fasta format extraction but is also stored in another field the position in the chain of the selected residues. This information can be used to deduct interface limits from combination of DBS and ΔSAS like Ramaraj et al. used. Using 3D structures as source limits the quantity of data available since obtaining crystal structure isn't trivial especially for large, non-covalently bound complexed protein. Antibody structures are nowadays indexed in specialized database such as the IMGT/3D or SAbDab. The IMGT has more than 4000 references to antibody structure with annotated CDRs and numbered residues using the IMGT numbering system (Ruiz and Lefranc, 2002; Ehrenmann and Lefranc, 2011). SAbDab has more than 2000 structures and label structure were the antibody is in complex. Those two databases are specialized on antibodies and their main purpose are to identify CDRs and help improve antibody structure prediction. EPI-DB focus on the specificity of the Ab-Ag interface and therefore store epitope and paratope as a pair. EPI-DB only contains structures of reliable resolution, maximum 3Å, and complexed with protein antigen. This approach also fundamentally differs from epitope-only databases such as the IEDB (Vita et al., 2015), Bcipep

(Kumar et al., 2005) or SEDB (Om Prakash et al., 2012) containing respectively 142.175, 3031 and 614 epitopes. IEDB contains immune epitopes curated including from the published literature. Bcipep contains only curated epitopes from literature while SEDB contains only epitopes with structures from Tcell, Bcell or MHC. The Antigen Antibody DataBase (AgAbDB) is the most similar tool to EPI-DB focusing on Ab-Ag complexed structure only (Kulkarni-Kale et al., 2014). The authors develop the Antigen-Antibody Interaction Finder (AAIF) a program especially designed to compute molecular interactions. To the last update the AgAbDB has 505 annotated structure which is a bit inferior to the EPI-DB possessing 543 structures. They also added the solvent accessible surface to the residues as well as the boundaries of the CDR using the Wu-Kabat numbering but did not compute the interface properties. The list of properties we computed is intended to be as complete as possible with a focus on parameters that were used for epitope prediction from sequence. The list parameters includes, accessibility used by Hopp, who obtained an epitope's prediction performance of 60% , hydrophobicity from Manavalan and Ponnuswamy 1978 reaching 61% secondary structure used by Pellequer and Westhof reaching 70%. Since EPI-DB was designed for the biologists as a source of structural information on epitope and paratope from structure we implemented a web interface. The interface allows the user to understand what is EPI-DB, parse the data of the database table by table and allow download of the full SQL file to implement EPI-DB locally if needed. The interface is currently basic but provides a friendly way to retrieve data like epitope and paratope sequences or mice or Human antibodies complexed with antigen structures.

The objective of EPI-DB is to become a reference in term of epitope and paratope study using crystal structure. To this end we decided to develop Interface Research Algorithm (IRA) and we implemented it in Java (Biojava) allowing and automatic extraction of very specific structure from the Protein DataBase (Berman et al., 2000). Biojava offers a complete set of class and functions to handle the PDB file, including an automatic reader, chain-specific

sequence extraction as well as resolution techniques and quality. Using this tool and already implemented Smith and Waterman (Smith et al., 1981) protein alignment we developed a tool able to precisely extract and automatically label Ab-Ag structures from a PDB file. IRA is also multithreaded for speed optimization and can also be used to extract any other types of PPI from the PDB by changing its input.

Epitope prediction nowadays suffers from problem related to the quality of the data and low ratio of positive/negative used to develop prediction model (Greenbaum et al., 2007; Denh et al., 2011; Subramanian and Chinnappan, 2013). In order to avoid this issue we worked with crystal structure extracted interfaces. All along this work we also relied on the notion of pair of epitope and paratope. A pair is obtained from a single structure using a defined selection method and associated cutoff. This means that epitope and paratope are selected from the antigen and antibody respectively in a mirrored way. Using the computed properties based on the epitope and paratope we investigated their correlation and their predictive capacity. By doing the absolute correlation clustering (Figure 3.6) of the properties we were able to easily see the behaviour of the different variables. The presence of very strong clusters shows the prediction capacity of a good portion of the properties. To assess the quality of the prediction we came up with the system called rightful pair prediction (Figure 3.7). The objective of the model was to assess if an epitope and a paratope formed a rightful pair (coming from the same structure, selection method an cutoff) or not. From the dataset (101 interfaces) were taken a given number of pairs of epitope paratope forming the positive pairs. Those pairs were shuffled to form the same amount of mismatched pairs. This methodology of prediction validation ensure the data quality and the same number of negative and positive elements. The development of ensemble of models allowed us to test various combinations of properties in an automatic way. The objective was the maximization of the AUC from the rightful pair prediction. The best ensemble of models reached a prediction of 0.6420 of AUC. These results show that complex correlations exist between the paratope and the

epitope properties. Using the epitope properties it is possible to predict, with good accuracy, if a sequence corresponds to its cognate partner. This approach trained with set of epitopes and corresponding CDR could be an interesting addition to the methodology developed by DeKosky et al. which consist of antibody CDR high-throughput sequence. This methodology would potentially be able to predict which CDR sequence would bind to the antigen used for the immunazation.

Due to the difficulty of the epitope prediction from sequence has emerged a new trend of using supporting information. By aggregating additional information to the input data such as epitope prediction using protein family or for specific antibody as suggested by Sela-Culang et al.. The collaborative work done with other coworkers following this trend led to a publication in BMC bioinformatics (See annex 1).

Even if X-Ray crystallography is one of the best way to determine epitopes from antigen, the sequences obtained can vary considerably depending on the interface selection method. Most of selections used are first distance between atoms of the Ag and the Ab, difference of solvent accessible surface or combination of those two. All of those methods can be adapted using various cutoff values and therefore result in different epitopes selected from the same interface. Those methodologies lead to divergent results such as the work of Rubinstein et al. that found epitope significantly enriched in Tyr and Trp using a DBS with a 4Å cutoff and Kringelum et al. observing that epitope composition was not significantly different from the rest of the antigen surface using the Contacts of Structural Units server (Sobolev et al., 1999). In this work we developed a new method based on Interface Interacting Residues (I2R) to select epitope-paratope residues that allows to compare these different methodologies.

Comparing the I2R to the distance based and ΔSAS selection showed that the two latest miss a set of residues involved in interaction and are therefore important for the interface. Nevertheless the residues selected by the two selection techniques previously mentioned have a selection span that will take into account close amino acids which, even though they do not

make interaction, might be important for the epitope conformation.

The I2R selection method we developed allows an approximation of the residues energetic contribution of the interface in a fast and easy way. Moreover this selection could be used to select target for free-energy perturbation (FEP) (Xia et al., 2012) or to investigate the binding hot-spot to ease murine antibody humanization (Hanf et al., 2013).

Protein Protein interfaces involve complicated recognition mechanisms on which rely most of the cell interactions. Better understanding of the complementarity and mechanisms that lies within the interface would help improve field like drug design, protein *de novo* design and many others. To extract interaction pattern from the interface computational methods are required. Interface Graph Generator (IGG) allows the users to easily transform the interface into a graph format that can be used in computational pattern searches. IGG was developed in java, a cross platform language, that enabled us to implement an easy to use graphical user interface. The output of the graph can be obtained into two different formats, 'pajek' which is one of the most common format and the format corresponding to Gaston's input (Nijssen and Kok, 2004).

EPI-Peptide designer was implemented in combination with IGG since it relies on the graph representation of the interface and share similar input. The same interface was used for both programs. EPI-Peptides are the results of the fusion of the two main methodologies for peptide ligands selection being the randomization of peptide and the rational design.

In vitro display selection of peptides from libraries has been a successful methodology for the development of new peptide binders. From the phage display approach (Smith and Petrenko, 1997; Hoess, 2001) to the ribosomes display (Mattheakis et al., 1994), mRNA display (Roberts and Szostak, 1997; Cho et al., 2000) and CIS display (Odegrip et al., 2004a) various techniques have successfully produced peptide binders. The success of those methodologies relies on large combinatorial peptide libraries and a multi step process of affinity selection and

mutation. By creating biased libraries for a specific target EPI-Peptide designer may allow to reduce the size of the library and therefore simplify these methodologies.

The analysis conducted on the Mouse and Human showed that antibodies from mice and Humans have different amino acids propensities. In 2003, Zemlin et al. already investigated the differences of the CDR-H3 between Human and murine Abs. The results we obtained on full paratope go in the same direction than their residues statistics analysis for the Tyr but differ for Ser and Thr propensities. Both statistic analysis match with each other, showing the elevated level of small hydrophobic residue (Val, Leu, Iso) and Phe in the Human antibodies interfaces. The statistics differ for the Thr where we found its occurrence higher in the Mouse group and they found the Thr increased in the Human group. This is most likely due to the difference of selection method. The difference between interfaces from mice antibody and Human antibody but our analysis showed that the composition of the epitope they recognize is also different. The epitope in the Mouse group showed higher propensities of Ile and Phe while the Human epitopes where enriched in Glu, Asn and Ser. The results we obtained using the energetic analysis and conserved subgraph extraction show that Human antibodies and mouse antibodies recognize preferentially different antigen parts.

The prediction obtained using the distance based selected epitope and the secondary structure reaching a perfect prediction is probably due to over fitting and all those predictions should be done using a 5 cross fold validation. Nevertheless the results obtained show clearly that the properties from distance selected epitope give better results than the I2R selection. This is probably due to the larger selection of the DBS epitope and it rose the hypothesis that non interacting residues confer properties to an epitope that help improve the prediction.

CHAPTER 7

Conclusions and Perspectives

All along this work, we have gained insight into the antibody-antigen interfaces. We observed the differences of epitope and paratope sequences due to different interface selection methods, the capacity of the epitope physicochemicals properties to predict paratope's ones, the importance of the cation-π interaction (the major importance of epitope positive charge and paratope aromatic aminoacids) using the common subgraph analysis. Using this knowledge our study proposed an original methodology able to to generate targeted peptide ligand libraries. The success of our method was observed using the LiD1 protein as target with which we observed binding of 65% of the synthesized EPI-Peptides. The methodology developed in this work could be used to design a new generation of biosensors.

As a perspective of this work, we hope to do a deeper analysis of the LiD1 peptide binder found in this work, including affinity measurements in order to introduce a filter in the program reducing the number of peptides to test. Moreover it would be interesting to further validate this tool using other target antigen with available corresponding antibody, to select reactive EPI-peptides and compare their binding affinity with the antibody.

Another perspective of this work is to add some structural features to this methodology. This would be helped by analyzing antigen-antibody interfaces using IMGT CDRs numbering methodolgy 'collier de perles' (Lefranc et al., 2005) and could also profit from the structural alphabet (Etchebest et al., 2005).

EPI-DB offers a new perspective on the study of Antibody-Antigen interface. The study of correlation of properties shows it exists complexed and relevant relations that can be used to

developed prediction from one side of the interface to the other. The approach of properties prediction of the complementary sequences is of great interest for new interface design and should be tested using different selection methods and a more complete dataset. The properties table could also be improved by adding more features. A property not yet implemented is the flexibility and it would be very interesting to add this classic property to the database.

Research in the field of protein ligands has been proven of great biological importance for their applications as probes as well as potential therapeutics. Nowadays resolution of complexed protein is improving and contribute to the development of new tools based on those data. Antibody-antigen interface complementarity is more and more understood but still lack deep and complex analysis using computational methodologies. We hope that this work will help understand the antibody-antigen interface in terms of molecular interactions, complementary and correlation of properties. This knowledge could help us developed future therapeutics, improving our health and our quality of life.

Bibliography

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28(1):235–242, Jan 2000. xv, 27, 32, 36, 74

A. Stein, A. Ceol, and P. Aloy. 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, 39(Database issue): D718–723, Jan 2011. xv

M. P. Lefranc. IMGT (ImMunoGeneTics) locus on focus. A new section of Experimental and Clinical Immunogenetics. *Exp. Clin. Immunogenet.*, 15(1):1–7, 1998. xv, 2, 27

J. Dunbar, K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi, and C. M. Deane. SAbDab: the structural antibody database. *Nucleic Acids Res.*, 42(Database issue):D1140–1146, Jan 2014. xv, 27

Sharma. Om Prakash, Das. Arindam Atanu, R. Krishna, M. Suresh Kumar, and P. Mathur. Premendu. Structural Epitope Database (SEDB): A Web-based Database for the Epitope, and its Intermolecular Interaction Along with the Tertiary Structure Information. *Journal of Proteomics and Bioinformatics.*, 0(Web Server issue), Feb 2012. xv, 74

J. Adolf-Bryfogle, Q. Xu, B. North, A. Lehmann, and R. L. Dunbrack. PyIgClassify: a database of antibody CDR structural classifications. *Nucleic Acids Res.*, 43(Database issue):D432–438, Jan 2015. xv

U. Kulkarni-Kale, S. Raskar-Renuse, G. Natekar-Kalantre, and S. A. Saxena. Antigen-Antibody Interaction Database (AgAbDb): a compendium of antigen-antibody interactions. *Methods Mol. Biol.*, 1184:149–164, 2014. xvi, 74

G. Neshich, I. Mazoni, S. R. Oliveira, M. E. Yamagishi, P. R. Kuser-Falcao, L. C. Borro, D. U. Morita, K. R. Souza, G. V. Almeida, D. N. Rodrigues, J. G. Jardine, R. C. Togawa, A. L. Mancini, R. H. Higa, S. A. Cruz, F. D. Vieira, E. H. Santos, R. C. Melo, and M. M. Santoro. The Star STING server: a multiplatform environment for protein structure analysis. *Genet. Mol. Res.*, 5(4):717–722, 2006. xvi, 30, 49

V. Kunik, S. Ashkenazi, and Y. Ofran. Paratome: an online tool for systematic identification of antigen-binding regions in antibodies based on sequence or structure. *Nucleic Acids Res.*, 40(Web Server issue):W521–524, Jul 2012a. xvi, 27

P. Marcatili, A. Rosi, and A. Tramontano. PIGS: automatic prediction of antibody structures. *Bioinformatics*, 24(17):1953–1954, Sep 2008. xvi

P. Vanhee, J. Reumers, F. Stricher, L. Baeten, L. Serrano, J. Schymkowitz, and F. Rousseau. PepX: a structural database of non-redundant protein-peptide complexes. *Nucleic Acids Res.*, 38(Database issue):D545–551, Jan 2010. xvi

G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Res.*, 14(6):1188–1190, Jun 2004. xvi

L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. CD-HIT: accelerated for clustering the next-

generation sequencing data. *Bioinformatics*, 28(23):3150–3152, Dec 2012a. xvi, 27

O. Keskin, B. Ma, K. Rogale, K. Gunasekaran, and R. Nussinov. Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach. *Phys Biol*, 2(2):24–35, Jun 2005. 1

J. Maynard and G. Georgiou. Antibody engineering. *Annu Rev Biomed Eng*, 2:339–376, 2000. 1

C. Chothia, A. M. Lesk, A. Tramontano, M. Levitt, S. J. Smith-Gill, G. Air, S. Sheriff, E. A. Padlan, D. Davies, and W. R. Tulip. Conformations of immunoglobulin hypervariable regions. *Nature*, 342(6252):877–883, 1989. 1

I. S. Mian, A. R. Bradwell, and A. J. Olson. Structure, function and properties of antibody binding sites. *J. Mol. Biol.*, 217(1):133–151, Jan 1991. 1, 7, 10

T. T. Wu and E. A. Kabat. An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med.*, 132(2):211–250, Aug 1970. 2

E.A. Kabat, T.T. Wu, H. Bilofsky, M. Reid-Miller, and H. Perry. proteins of immunological interest. Bethesda: National Institute of Health. *J. Exp. Med.*, page 323, 1983. 2

C. Chothia and J. Janin. Principles of protein-protein recognition. *Nature*, 256(5520):705–708, Aug 1975. 2, 6, 30

V. Kunik, B. Peters, and Y. Ofran. Structural consensus among antibodies defines the antigen binding site. *PLoS Comput. Biol.*, 8(2):e1002388, 2012b. 2

J. L. Pellequer, E. Westhof, and M. H. Van Regenmortel. Predicting location of continuous epitopes in proteins from their primary structures. *Meth. Enzymol.*, 203:176–201, 1991. 2, 3

A.M Silverstein. A History of Immunology. . *Science*, 247(4940):347, Jan 1990. 3

A.K. Abbas and A.H. Lichtman, editors. *Cellular and molecular immunology*. Saunders, Philadelfia PA, 2005. 3

J. A. Greenbaum, P. H. Andersen, M. Blythe, H. H. Bui, R. E. Cachau, J. Crowe, M. Davies, A. S. Kolaskar, O. Lund, S. Morrison, B. Mumey, Y. Ofran, J. L. Pellequer, C. Pinilla, J. V. Ponomarenko, G. P. Raghava, M. H. van Regenmortel, E. L. Roggen, A. Sette, A. Schlessinger, J. Sollner, M. Zand, and B. Peters. Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J. Mol. Recognit.*, 20(2):75–82, 2007. 3, 75

T. P. Hopp and K. R. Woods. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. U.S.A.*, 78(6):3824–3828, Jun 1981. 3

E. A. Emini, J. V. Hughes, D. S. Perlow, and J. Boger. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J. Virol.*, 55(3):836–839, Sep 1985. 3

P. M. Moyle and I. Toth. Modern subunit vaccines: development, components, and research opportunities. *ChemMedChem*, 8(3):360–376, Mar 2013. 3

A. Murray, R. G. Smith, K. Brady, S. Williams, R. A. Badley, and M. R. Price. Generation and refinement of peptide mimetic ligands for paratope-specific purification of monoclonal antibodies. *Anal. Biochem.*, 296(1):9–17, Sep 2001. 3

C.V. Rao. *Immunology: A Textbook*. Alpha Science, 2005. ISBN 9781842652558. URL

https://books.google.com.br/books?id=G5QZOIdqde0C. 3
M. Karplus and J. A. McCammon. The dynamics of proteins. *Sci. Am.*, 254(4):42–51, Apr 1986. 3
J. M. Parker, D. Guo, and R. S. Hodges. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry*, 25(19): 5425–5432, Sep 1986. 3
J. L. Pellequer and E. Westhof. PREDITOP: a program for antigenicity prediction. *J Mol Graph*, 11(3):204–210, Sep 1993a. 3, 74
I. a. I. Davydov and A. G. Tonevitski. [Linear B-cell epitope prediction]. *Mol. Biol. (Mosk.)*, 43(1):166–174, 2009. 3
J. L. Pellequer and E. Westhof. PREDITOP: a program for antigenicity prediction. *J Mol Graph*, 11(3):204–210, Sep 1993b. 3
A. J. Alix. Predictive estimation of protein linear epitopes by using the program PEOPLE. *Vaccine*, 18(3-4):311–314, Sep 1999. 3
M. Odorico and J. L. Pellequer. BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. *J. Mol. Recognit.*, 16(1):20–22, 2003. 3
M. J. Blythe and D. R. Flower. Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci.*, 14(1):246–248, Jan 2005. 3
J. Gao and L. Kurgan. Computational prediction of B cell epitopes from antigen sequences. *Methods Mol. Biol.*, 1184:197–215, 2014. 3
S. Saha and G. P. Raghava. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins*, 65(1):40–48, Oct 2006. 3
M. J. Sweredoski and P. Baldi. COBEpro: a novel system for predicting continuous B-cell epitopes. *Protein Eng. Des. Sel.*, 22(3):113–120, Mar 2009. 3
L. Zhao, L. Wong, L. Lu, S. C. Hoi, and J. Li. B-cell epitope prediction through a graph model. *BMC Bioinformatics*, 13 Suppl 17:S20, 2012. 3
H. Denh, G. Runger, and E. Tuv. Bias of importance measures for multi-valued attributes and solutions. *Proceedings of the 21st International Conference on Artificial Neural Networks*, 2011. 3, 75
N. Subramanian and S. Chinnappan. Prediction of promiscuous epitopes in the e6 protein of three high risk human papilloma viruses: a computational approach. *Asian Pac. J. Cancer Prev.*, 14(7):4167–4175, 2013. 3, 75
T. Ramaraj, T. Angel, E. A. Dratz, A. J. Jesaitis, and B. Mumey. Antigen-antibody interface properties: composition, residue interactions, and features of 53 non-redundant structures. *Biochim. Biophys. Acta*, 1824(3):520–532, Mar 2012. 5, 6, 9, 11, 27, 73
L. Lo Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, 285(5):2177–2198, Feb 1999. 6, 7, 8, 10, 30
J. Janin and C. Chothia. The structure of protein-protein recognition sites. *J. Biol. Chem.*, 265(27):16027–16030, Sep 1990. 6
P. Chakrabarti and J. Janin. Dissecting protein-protein recognition sites. *Proteins*, 47(3): 334–343, May 2002. 6, 7
J. Pontius, J. Richelle, and S. J. Wodak. Deviations from standard atomic volumes as a quality

measure for protein crystal structures. *J. Mol. Biol.*, 264(1):121–136, Nov 1996. 6

V. M. Goncalves-Almeida, D. E. Pires, R. C. de Melo-Minardi, C. H. da Silveira, W. Meira, and M. M. Santoro. HydroPaCe: understanding and predicting cross-inhibition in serine proteases through hydrophobic patch centroids. *Bioinformatics*, 28(3):342–349, Feb 2012. 6

J. Novotny, M. Handschumacher, and E. Haber. Location of antigenic epitopes on antibody molecules. *J. Mol. Biol.*, 189(4):715–721, Jun 1986. 6

D. R. Davies, E. A. Padlan, and S. Sheriff. Antibody-antigen complexes. *Annu. Rev. Biochem.*, 59:439–473, 1990. 6

W. G. Laver, G. M. Air, R. G. Webster, and S. J. Smith-Gill. Epitopes on protein antigens: misconceptions and realities. *Cell*, 61(4):553–556, May 1990. 6, 7

M. C. Lawrence and P. M. Colman. Shape complementarity at protein/protein interfaces. *J. Mol. Biol.*, 234(4):946–950, Dec 1993. 7, 10

S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proc. Natl. Acad. Sci. U.S.A.*, 93(1):13–20, Jan 1996. 7

R. M. MacCallum, A. C. Martin, and J. M. Thornton. Antibody-antigen interactions: contact analysis and binding site topography. *J. Mol. Biol.*, 262(5):732–745, Oct 1996. 7, 10

G. H. Cohen, S. Sheriff, and D. R. Davies. Refined structure of the monoclonal antibody HyHEL-5 with its antigen hen egg-white lysozyme. *Acta Crystallogr. D Biol. Crystallogr.*, 52(Pt 2):315–326, Mar 1996. 7

Wiechert DW. Johnson RA. Applied multivariate statistical analy-sis. *New Delhi: Prentice-Hall of India*, 1996. 8

E. J. Sundberg, P. S. Andersen, P. M. Schlievert, K. Karjalainen, and R. A. Mariuzza. Structural, energetic, and functional analysis of a protein-protein interface at distinct stages of affinity maturation. *Structure*, 11(9):1151–1161, Sep 2003. 8

P. Haste Andersen, M. Nielsen, and O. Lund. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.*, 15(11):2558–2567, Nov 2006. 8, 10

N. D. Rubinstein, I. Mayrose, D. Halperin, D. Yekutieli, J. M. Gershoni, and T. Pupko. Computational characterization of B-cell epitopes. *Mol. Immunol.*, 45(12):3477–3489, Jul 2008. 8, 10, 66, 76

S. W. Chen, M. H. Van Regenmortel, and J. L. Pellequer. Structure-activity relationships in peptide-antibody complexes: implications for epitope prediction and development of synthetic peptide vaccines. *Curr. Med. Chem.*, 16(8):953–964, 2009. 8, 10

M. Heinig and D. Frishman. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res.*, 32(Web Server issue):W500–502, Jul 2004. 8, 36, 39, 66

J. V. Kringelum, M. Nielsen, S. B. Padkjaer, and O. Lund. Structural analysis of B-cell epitopes in antibody:protein complexes. *Mol. Immunol.*, 53(1-2):24–34, Dec 2012. 8, 10, 76

I. Sela-Culang, S. Alon, and Y. Ofran. A systematic comparison of free and bound antibodies reveals binding-related conformational changes. *J. Immunol.*, 189(10):4890–4899, Nov 2012. 9, 10

J. W. Stave and K. Lindpaintner. Antibody and antigen contact residues define epitope and

paratope size and structure. *J. Immunol.*, 191(3):1428–1435, Aug 2013. 9

G. Robin, Y. Sato, D. Desplancq, N. Rochel, E. Weiss, and P. Martineau. Restricted diversity of antigen binding residues of antibodies revealed by computational alanine scanning of 227 antibody-antigen complexes. *J. Mol. Biol.*, 426(22):3729–3743, Nov 2014. 10, 49

A. Beck, T. Wurch, C. Bailly, and N. Corvaia. Strategies and challenges for the next generation of therapeutic antibodies. *Nat. Rev. Immunol.*, 10(5):345–352, May 2010. 12

R. R. Porter. The hydrolysis of rabbit y-globulin and antibodies with crystalline papain. *Biochem. J.*, 73:119–126, Sep 1959. 12

R. E. Bird, K. D. Hardman, J. W. Jacobson, S. Johnson, B. M. Kaufman, S. M. Lee, T. Lee, S. H. Pope, G. S. Riordan, and M. Whitlow. Single-chain antigen-binding proteins. *Science*, 242(4877):423–426, Oct 1988. 12

M. Jain, N. Kamal, and S. K. Batra. Engineering antibodies for clinical applications. *Trends Biotechnol.*, 25(7):307–316, Jul 2007. 12

H. M. Azzazy and W. E. Highsmith. Phage display technology: clinical applications and recent innovations. *Clin. Biochem.*, 35(6):425–445, Sep 2002. 13

M. Leenaars and C. F. Hendriksen. Critical steps in the production of polyclonal and monoclonal antibodies: evaluation and recommendations. *ILAR J*, 46(3):269–279, 2005. 13

W. Wang, E. Q. Wang, and J. P. Balthasar. Monoclonal antibody pharmacokinetics and pharmacodynamics. *Clin. Pharmacol. Ther.*, 84(5):548–558, Nov 2008. 14

M. V. Pimm. Drug-monoclonal antibody conjugates for cancer therapy: potentials and limitations. *Crit Rev Ther Drug Carrier Syst*, 5(3):189–227, 1988. 14

A.A. Lugovskoy, K. Hanf, Y. Li, K. Simon, and H. Van Vlijmen. Methods of humanizing immunoglobulin variable regions through rational modification of complementarity determining residues, 2010. URL http://www.patentlens.net/patentlens/patent/US_7678371/. 14

K. J. Hanf, J. W. Arndt, L. L. Chen, M. Jarpe, P. A. Boriack-Sjodin, Y. Li, H. W. van Vlijmen, R. B. Pepinsky, K. J. Simon, and A. Lugovskoy. Antibody humanization by redesign of complementarity-determining region residues proximate to the acceptor framework. *Methods*, 65(1):68–76, Jan 2014. 14

N. London and X. Ambroggio. An accurate binding interaction model in de novo computational protein design of interactions: If you build it, they will bind. *J. Struct. Biol.*, Apr 2013. 14

P. B. Stranges and B. Kuhlman. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. *Protein Sci.*, 22(1): 74–82, Jan 2013. 14

R. J. Pantazes, M. J. Grisewood, and C. D. Maranas. Recent advances in computational protein design. *Curr. Opin. Struct. Biol.*, 21(4):467–472, Aug 2011. 14

D. Pei and A. S. Wavreille. Reverse interactomics: decoding protein-protein interactions with combinatorial peptide libraries. *Mol Biosyst*, 3(8):536–541, Aug 2007. 15

H. Yin, J. S. Slusky, B. W. Berger, R. S. Walters, G. Vilaire, R. I. Litvinov, J. D. Lear, G. A. Caputo, J. S. Bennett, and W. F. DeGrado. Computational design of peptides that target transmembrane helices. *Science*, 315(5820):1817–1822, Mar 2007. 15

P. Vanhee, A. M. van der Sloot, E. Verschueren, L. Serrano, F. Rousseau, and J. Schymkowitz.

Computational design of peptide ligands. *Trends Biotechnol.*, 29(5):231–239, May 2011. 15

S. E. Cwirla, P. Balasubramanian, D. J. Duffin, C. R. Wagstrom, C. M. Gates, S. C. Singer, A. M. Davis, R. L. Tansik, L. C. Mattheakis, C. M. Boytos, P. J. Schatz, D. P. Baccanari, N. C. Wrighton, R. W. Barrett, and W. J. Dower. Peptide agonist of the thrombopoietin receptor as potent as the natural cytokine. *Science*, 276(5319):1696–1699, Jun 1997. 15

J. L. Su, K. P. Lai, C. A. Chen, C. Y. Yang, P. S. Chen, C. C. Chang, C. H. Chou, C. L. Hu, M. L. Kuo, C. Y. Hsieh, and L. H. Wei. A novel peptide specifically binding to interleukin-6 receptor (gp80) inhibits angiogenesis and tumor growth. *Cancer Res.*, 65(11):4827–4835, Jun 2005. 15

R. Hyde-DeRuyscher, L. A. Paige, D. J. Christensen, N. Hyde-DeRuyscher, A. Lim, Z. L. Fredericks, J. Kranz, P. Gallant, J. Zhang, S. M. Rocklage, D. M. Fowlkes, P. A. Wendler, and P. T. Hamilton. Detection of small-molecule enzyme inhibitors with peptides isolated from phage-displayed combinatorial peptide libraries. *Chem. Biol.*, 7(1):17–25, Jan 2000. 15

B. D. Welch, A. P. VanDemark, A. Heroux, C. P. Hill, and M. S. Kay. Potent D-peptide inhibitors of HIV-1 entry. *Proc. Natl. Acad. Sci. U.S.A.*, 104(43):16828–16833, Oct 2007. 15

T. Matsubara, A. Onishi, T. Saito, A. Shimada, H. Inoue, T. Taki, K. Nagata, Y. Okahata, and T. Sato. Sialic acid-mimic peptides as hemagglutinin inhibitors for anti-influenza therapy. *J. Med. Chem.*, 53(11):4441–4449, Jun 2010. 15

S. R. Whaley, D. S. English, E. L. Hu, P. F. Barbara, and A. M. Belcher. Selection of peptides with semiconductor binding specificity for directed nanocrystal assembly. *Nature*, 405 (6787):665–668, Jun 2000. 15

S. Wang, E. S. Humphreys, S. Y. Chung, D. F. Delduco, S. R. Lustig, H. Wang, K. N. Parker, N. W. Rizzo, S. Subramoney, Y. M. Chiang, and A. Jagota. Peptides with selective affinity for carbon nanotubes. *Nat Mater*, 2(3):196–200, Mar 2003. 15

D. J. Rodi, R. W. Janes, H. J. Sanganee, R. A. Holton, B. A. Wallace, and L. Makowski. Screening of a library of phage-displayed peptides identifies human bcl-2 as a taxol-binding protein. *J. Mol. Biol.*, 285(1):197–203, Jan 1999. 15

K. Saar, M. Lindgren, M. Hansen, E. Eiriksdottir, Y. Jiang, K. Rosenthal-Aizman, M. Sassian, and U. Langel. Cell-penetrating peptides: a comparative membrane toxicity study. *Anal. Biochem.*, 345(1):55–65, Oct 2005. 16

G. P. Smith and V. A. Petrenko. Phage Display. *Chem. Rev.*, 97(2):391–410, Apr 1997. 16, 77

R. H. Hoess. Protein design and phage display. *Chem. Rev.*, 101(10):3205–3218, Oct 2001. 16, 77

C. Zahnd, P. Amstutz, and A. Pluckthun. Ribosome display: selecting and evolving proteins in vitro that specifically bind to a target. *Nat. Methods*, 4(3):269–279, Mar 2007. 16

S. W. Cotten, J. Zou, C. A. Valencia, and R. Liu. Selection of proteins with desired properties from natural proteome libraries using mRNA display. *Nat Protoc*, 6(8):1163–1182, Aug 2011. 16

R. Odegrip, D. Coomber, B. Eldridge, R. Hederer, P. A. Kuhlman, C. Ullman, K. FitzGerald,

and D. McGregor. CIS display: In vitro selection of peptides from libraries of protein-DNA complexes. *Proc. Natl. Acad. Sci. U.S.A.*, 101(9):2806–2810, Mar 2004a. 16, 77

L. C. Mattheakis, R. R. Bhatt, and W. J. Dower. An in vitro polysome display system for identifying ligands from very large peptide libraries. *Proc. Natl. Acad. Sci. U.S.A.*, 91(19): 9022–9026, Sep 1994. 16, 77

T. Lamla and V. A. Erdmann. Searching sequence space for high-affinity binding peptides using ribosome display. *J. Mol. Biol.*, 329(2):381–388, May 2003. 16

G. Cho, A. D. Keefe, R. Liu, D. S. Wilson, J. W. Szostak, and J. W. Szostak. Constructing high complexity synthetic libraries of long ORFs using in vitro selection. *J. Mol. Biol.*, 297 (2):309–319, Mar 2000. 16, 77

H. Ingmer, C. Miller, and S. N. Cohen. The RepA protein of plasmid pSC101 controls Escherichia coli cell division through the SOS response. *Mol. Microbiol.*, 42(2):519–526, Oct 2001. 16

R. Odegrip, D. Coomber, B. Eldridge, R. Hederer, P. A. Kuhlman, C. Ullman, K. FitzGerald, and D. McGregor. CIS display: In vitro selection of peptides from libraries of protein-DNA complexes. *Proc. Natl. Acad. Sci. U.S.A.*, 101(9):2806–2810, Mar 2004b. 16

P. Vanhee, F. Stricher, L. Baeten, E. Verschueren, T. Lenaerts, L. Serrano, F. Rousseau, and J. Schymkowitz. Protein-peptide interactions adopt the same structural motifs as monomeric protein folds. *Structure*, 17(8):1128–1136, Aug 2009. 16

N. London, D. Movshovitz-Attias, and O. Schueler-Furman. The structural basis of peptide-protein binding strategies. *Structure*, 18(2):188–199, Feb 2010. 16

T. Clackson and J. A. Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(5196):383–386, Jan 1995. 16

M. C. Honeyman, V. Brusic, N. L. Stone, and L. C. Harrison. Neural network-based prediction of candidate T-cell epitopes. *Nat. Biotechnol.*, 16(10):966–969, Oct 1998. 16

H. H. Lin, G. L. Zhang, S. Tongchusak, E. L. Reinherz, and V. Brusic. Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research. *BMC Bioinformatics*, 9 Suppl 12:S22, 2008. 16

C. T. Wild, D. C. Shugars, T. K. Greenwell, C. B. McDanal, and T. J. Matthews. Peptides corresponding to a predictive alpha-helical domain of human immunodeficiency virus type 1 gp41 are potent inhibitors of virus infection. *Proc. Natl. Acad. Sci. U.S.A.*, 91(21):9770–9774, Oct 1994. 17

F. Naider and J. Anglister. Peptides in the treatment of AIDS. *Curr. Opin. Struct. Biol.*, 19 (4):473–482, Aug 2009. 17

A. Koide, C. W. Bailey, X. Huang, and S. Koide. The fibronectin type III domain as a scaffold for novel binding proteins. *J. Mol. Biol.*, 284(4):1141–1151, Dec 1998. 17, 21

D. Lipovsek. Adnectins: engineered target-binding protein therapeutics. *Protein Eng. Des. Sel.*, 24(1-2):3–9, Jan 2011. 17

T. F. Chen, S. de Picciotto, B. J. Hackel, and K. D. Wittrup. Engineering fibronectin-based binding proteins by yeast surface display. *Meth. Enzymol.*, 523:303–326, 2013. 17

B. J. Hackel, A. Kapila, and K. D. Wittrup. Picomolar affinity fibronectin domains engineered utilizing loop length diversity, recursive mutagenesis, and loop shuffling. *J. Mol. Biol.*, 381 (5):1238–1252, Sep 2008. 17

B. J. Hackel, R. H. Kimura, and S. S. Gambhir. Use of (64)Cu-labeled fibronectin domain with EGFR-overexpressing tumor xenograft: molecular imaging. *Radiology*, 263(1):179–188, Apr 2012. 17

L. Wickstrom, A. Okur, K. Song, V. Hornak, D. P. Raleigh, and C. L. Simmerling. The unfolded state of the villin headpiece helical subdomain: computational studies of the role of locally stabilized structure. *J. Mol. Biol.*, 360(5):1094–1107, Jul 2006. 18

A. Orlova, M. Magnusson, T. L. Eriksson, M. Nilsson, B. Larsson, I. Hoiden-Guthenberg, C. Widstrom, J. Carlsson, V. Tolmachev, S. Stahl, and F. Y. Nilsson. Tumor imaging using a picomolar affinity HER2 binding affibody molecule. *Cancer Res.*, 66(8):4339–4348, Apr 2006. 18

R. P. Baum, V. Prasad, D. Muller, C. Schuchardt, A. Orlova, A. Wennborg, V. Tolmachev, and J. Feldwisch. Molecular imaging of HER2-expressing malignant tumors in breast cancer patients using synthetic 111In- or 68Ga-labeled affibody molecules. *J. Nucl. Med.*, 51(6): 892–897, Jun 2010. 18

V. Tolmachev, D. Rosik, H. Wallberg, A. Sjoberg, M. Sandstrom, M. Hansson, A. Wennborg, and A. Orlova. Imaging of EGFR expression in murine xenografts using site-specifically labelled anti-EGFR 111In-DOTA-Z EGFR:2377 Affibody molecule: aspect of the injected tracer amount. *Eur. J. Nucl. Med. Mol. Imaging*, 37(3):613–622, Mar 2010. 18

V. Tolmachev, M. Friedman, M. Sandstrom, T. L. Eriksson, D. Rosik, M. Hodik, S. Stahl, F. Y. Frejd, and A. Orlova. Affibody molecules for epidermal growth factor receptor targeting in vivo: aspects of dimerization and labeling chemistry. *J. Nucl. Med.*, 50(2):274–283, Feb 2009. 18

E. Nordberg, A. Orlova, M. Friedman, V. Tolmachev, S. Stahl, F. Y. Nilsson, B. Glimelius, and J. Carlsson. In vivo and in vitro uptake of 111In, delivered with the affibody molecule (ZEGFR:955)2, in EGFR expressing tumour cells. *Oncol. Rep.*, 19(4):853–857, Apr 2008. 18

V. Tolmachev, J. Malmberg, C. Hofstrom, L. Abrahmsen, T. Bergman, A. Sjoberg, M. Sandstrom, T. Graslund, and A. Orlova. Imaging of insulinlike growth factor type 1 receptor in prostate cancer xenografts using the affibody molecule 111In-DOTA-ZIGF1R:4551. *J. Nucl. Med.*, 53(1):90–97, Jan 2012. 18

H. Honarvar, N. Jokilaakso, K. Andersson, J. Malmberg, D. Rosik, A. Orlova, A. E. Karlstrom, V. Tolmachev, and P. Jarver. Evaluation of backbone-cyclized HER2-binding 2-helix affibody molecule for in vivo molecular imaging. *Nucl. Med. Biol.*, 40(3):378–386, Apr 2013. 18

J. M. Webster, R. Zhang, S. S. Gambhir, Z. Cheng, and F. A. Syud. Engineered two-helix small proteins for molecular recognition. *Chembiochem*, 10(8):1293–1296, May 2009. 18

D. Rosik, A. Orlova, J. Malmberg, M. Altai, Z. Varasteh, M. Sandstrom, A. E. Karlstrom, and V. Tolmachev. Direct comparison of 111In-labelled two-helix and three-helix Affibody molecules for in vivo molecular imaging. *Eur. J. Nucl. Med. Mol. Imaging*, 39(4):693–702, Apr 2012. 18

R. Tamaskovic, M. Simon, N. Stefan, M. Schwill, and A. Pluckthun. Designed ankyrin repeat proteins (DARPins) from research to therapy. *Meth. Enzymol.*, 503:101–134, 2012. 19

S. J. Moore and J. R. Cochran. Engineering knottins as novel binding agents. *Meth. Enzymol.*,

503:223–251, 2012. 19

R. H. Kimura, Z. Cheng, S. S. Gambhir, and J. R. Cochran. Engineered knottin peptides: a new class of agents for imaging integrin expression in living subjects. *Cancer Res.*, 69(6): 2435–2442, Mar 2009. 19

R. H. Kimura, R. Teed, B. J. Hackel, M. A. Pysz, C. Z. Chuang, A. Sathirachinda, J. K. Willmann, and S. S. Gambhir. Pharmacokinetically stabilized cystine knot peptides that bind alpha-v-beta-6 integrin with single-digit nanomolar affinities for detection of pancreatic cancer. *Clin. Cancer Res.*, 18(3):839–849, Feb 2012. 19

C. Borghouts, C. Kunz, and B. Groner. Peptide aptamer libraries. *Comb. Chem. High Throughput Screen.*, 11(2):135–145, Feb 2008. 19

P. Colas, B. Cohen, T. Jessen, I. Grishina, J. McCoy, and R. Brent. Genetic selection of peptide aptamers that recognize and inhibit cyclin-dependent kinase 2. *Nature*, 380(6574): 548–550, Apr 1996. 19, 21

J. Li, S. Tan, X. Chen, C. Y. Zhang, and Y. Zhang. Peptide aptamers with biological and therapeutic applications. *Curr. Med. Chem.*, 18(27):4215–4222, 2011. 19

M. B. Bickle, E. Dusserre, O. Moncorge, H. Bottin, and P. Colas. Selection and characterization of large collections of peptide aptamers through optimized yeast two-hybrid procedures. *Nat Protoc*, 1(3):1066–1091, 2006. 19

B. Klevenz, K. Butz, and F. Hoppe-Seyler. Peptide aptamers: exchange of the thioredoxin-A scaffold by alternative platform proteins and its influence on target protein binding. *Cell. Mol. Life Sci.*, 59(11):1993–1998, Nov 2002. 19

P. Dumy, I. M. Eggleston, G. Esposito, S. Nicula, and M. Mutter. Solution structure of regioselectively addressable functionalized templates: an NMR and restrained molecular dynamics investigation. *Biopolymers*, 39(3):297–308, Sep 1996. 20

D. Boturyn, J. L. Coll, E. Garanger, M. C. Favrot, and P. Dumy. Template assembled cyclopeptides as multimeric system for integrin targeting and endocytosis. *J. Am. Chem. Soc.*, 126(18):5730–5739, May 2004. 20

K. Nord, E. Gunneriusson, J. Ringdahl, S. Stahl, M. Uhlen, and P. A. Nygren. Binding proteins selected from combinatorial libraries of an alpha-helical bacterial receptor domain. *Nat. Biotechnol.*, 15(8):772–777, Aug 1997. 21

G. Beste, F. S. Schmidt, T. Stibora, and A. Skerra. Small antibody-like proteins with prescribed ligand specificities derived from the lipocalin fold. *Proc. Natl. Acad. Sci. U.S.A.*, 96(5):1898–1903, Mar 1999. 21

S. Schneider, M. Buchert, O. Georgiev, B. Catimel, M. Halford, S. A. Stacker, T. Baechi, K. Moelling, and C. M. Hovens. Mutagenesis and selection of PDZ domains that bind new protein targets. *Nat. Biotechnol.*, 17(2):170–175, Feb 1999. 21

M. Nicaise, M. Valerio-Lepiniec, P. Minard, and M. Desmadril. Affinity transfer by CDR grafting on a nonimmunoglobulin scaffold. *Protein Sci.*, 13(7):1882–1891, Jul 2004. 21

L. L. Looger, M. A. Dwyer, J. J. Smith, and H. W. Hellinga. Computational design of receptor and sensor proteins with novel functions. *Nature*, 423(6936):185–190, May 2003. 21

H. K. Binz, P. Amstutz, and A. Pluckthun. Engineering novel binding proteins from nonimmunoglobulin domains. *Nat. Biotechnol.*, 23(10):1257–1268, Oct 2005. 21

K. T. Nam, S. A. Shelby, P. H. Choi, A. B. Marciel, R. Chen, L. Tan, T. K. Chu, R. A.

Mesch, B. C. Lee, M. D. Connolly, C. Kisielowski, and R. N. Zuckermann. Free-floating ultrathin two-dimensional crystals from sequence-specific peptoid polymers. *Nat Mater*, 9 (5):454–460, May 2010. 22

G. K. Olivier, A. Cho, B. Sanii, M. D. Connolly, H. Tran, and R. N. Zuckermann. Antibody-mimetic peptoid nanosheets for molecular recognition. *ACS Nano*, 7(10):9276–9286, Oct 2013. 22, 24

J.L. Kulp, M. Sarikaya, and J. Spencer Evans. Characterization of the Integral Sequence Repeat From the E. coli Gold Binding Protein, GBP-1. *Mater. Chem.*, 14:2325–2332, 2004. 22

P. Timmerman, R. Barderas, J. Desmet, D. Altschuh, S. Shochat, M. J. Hollestelle, J. W. Hoppener, A. Monasterio, J. I. Casal, and R. H. Meloen. A combinatorial approach for the design of complementarity-determining region-derived peptidomimetics with in vitro anti-tumoral activity. *J. Biol. Chem.*, 284(49):34126–34134, Dec 2009. 22, 24

P. J. McEnaney, K. J. Fitzgerald, A. X. Zhang, E. F. Douglass, W. Shan, A. Balog, M. D. Kolesnikova, and D. A. Spiegel. Chemically synthesized molecules with the targeting and effector functions of antibodies. *J. Am. Chem. Soc.*, 136(52):18034–18043, Dec 2014. 23, 24

T. F. Smith, M. S. Waterman, and W. M. Fitch. Comparative biosequence metrics. *J. Mol. Evol.*, 18(1):38–46, 1981. 28, 75

L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, Dec 2012b. 32

Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, S Duvaud, Marc R Wilkins, Ron D. Appel, and Amos Bairoch. Protein identification and analysis tools on the expasy server. *The Proteomics Protocols Handbook*, pages 571–607, 2005. doi: 10.1385/1-59259-890-0:571. 36, 39

R. A. Becker, J. M. CHambers, and A. R. Wilks. The new S language. *Wadsworth and Brooks/Cole*, 1988. 40

R. Suzuki and H. Shimodaira. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22(12):1540–1542, Jun 2006. 40

A.J. Dobson. An introduction to Generalized Linear Models. *London:Chapman and Hall*, 1990. 44

T.J. Hastie and D. Pregibon. Generalized linear models. *Statistical Models in Seds, London: Chapman and Hall*, 6, 1992. 44

P. McCullagh and J. A. Nelder. Generalized Linear Models. *London:Chapman and Hall*, 1989. 44

W. N. Venables and B.D. Ripley. Modern Applied Statistics with S. *New York: Springer*, 2002. 44

S. J. Mason and N. E. Graham. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Q. J. R. Meteorol*, 30:291–303, 1982. 45

B.D. Ripley. Stochastic Simulation. *Wiley*, 1987.

A. Sircar and J. J. Gray. SnugDock: paratope structural optimization during antibody-antigen docking compensates for errors in antibody homology models. *PLoS Comput. Biol.*, 6(1):

e1000644, Jan 2010. 49

K. Krawczyk, T. Baker, J. Shi, and C. M. Deane. Antibody i-Patch prediction of the antibody binding site improves rigid local antibody-antigen docking. *Protein Eng. Des. Sel.*, 26(10): 621–629, Oct 2013. 49

L. Felicori, S. C. Araujo, R. A. de Avila, E. F. Sanchez, C. Granier, E. Kalapothakis, and C. Chavez-Olortegui. Functional characterization and epitope analysis of a recombinant dermonecrotic protein from Loxosceles intermedia spider. *Toxicon*, 48(5):509–519, Oct 2006. 49

S. R. Oliveira, G. V. Almeida, K. R. Souza, D. N. Rodrigues, P. R. Kuser-Falcao, M. E. Yamagishi, E. H. Santos, F. D. Vieira, J. G. Jardine, and G. Neshich. STING_RDB: a relational database of structural parameters for protein analysis with support for data warehousing and data mining. *Genet. Mol. Res.*, 6(4):911–922, 2007. 61

J.H. McDonald. Handbook of Biological Statistics (2nd ed.). pages W239–246, 2009. 68

L. Elden. Matrix Methods in Data Mining and Pattern Recognition (Fundamentals of Algorithms). Society for Industrial and Applied Mathematics. 2007. 68

M. Ruiz and M. P. Lefranc. IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics*, 53(10-11):857–883, Feb 2002. 73

F. Ehrenmann and M. P. Lefranc. IMGT/3Dstructure-DB: querying the IMGT database for 3D structures in immunology and immunoinformatics (IG or antibodies, TR, MH, RPI, and FPIA). *Cold Spring Harb Protoc*, 2011(6):750–761, Jun 2011. 73

R. Vita, J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix, A. Sette, and B. Peters. The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.*, 43(Database issue):D405–412, Jan 2015. 73

M. Kumar, M. Bhasin, N. K. Natt, and G. P. Raghava. BhairPred: prediction of beta-hairpins in a protein from multiple alignment information using ANN and SVM techniques. *Nucleic Acids Res.*, 33(Web Server issue):W154–159, Jul 2005. 74

T. P. Hopp. Protein antigen conformation: folding patterns and predictive algorithms; selection of antigenic and immunogenic peptides. *Ann Sclavo Collana Monogr*, 1(2):47–60, 1984. 74

P. Manavalan and P. K. Ponnuswamy. Hydrophobic character of amino acid residues in globular proteins. *Nature*, 275(5681):673–674, Oct 1978. 74

B. J. DeKosky, G. C. Ippolito, R. P. Deschner, J. J. Lavinder, Y. Wine, B. M. Rawlings, N. Varadarajan, C. Giesecke, T. Dorner, S. F. Andrews, P. C. Wilson, S. P. Hunicke-Smith, C. G. Willson, A. D. Ellington, and G. Georgiou. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat. Biotechnol.*, 31(2):166–169, Feb 2013. 76

I. Sela-Culang, Y. Ofran, and B. Peters. Antibody specific epitope prediction-emergence of a new paradigm. *Curr Opin Virol*, 11:98–102, Apr 2015. 76

V. Sobolev, A. Sorokine, J. Prilusky, E. E. Abola, and M. Edelman. Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15(4):327–332, Apr 1999. 76

Z. Xia, T. Huynh, S. G. Kang, and R. Zhou. Free-energy simulations reveal that both hydrophobic and polar interactions are important for influenza hemagglutinin antibody bind-

ing. *Biophys. J.*, 102(6):1453–1461, Mar 2012. 77

K. J. Hanf, J. W. Arndt, L. L. Chen, M. Jarpe, P. A. Boriack-Sjodin, Y. Li, H. W. van Vlijmen, R. B. Pepinsky, K. J. Simon, and A. Lugovskoy. Antibody humanization by redesign of complementarity-determining region residues proximate to the acceptor framework. *Methods*, Jun 2013. 77

Siegfried Nijssen and Joost Kok. A quickstart in frequent structure mining can make a difference. proceedings of the sigkdd. 2004. 77

R. W. Roberts and J. W. Szostak. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci. U.S.A.*, 94(23):12297–12302, Nov 1997. 77

M. Zemlin, M. Klinger, J. Link, C. Zemlin, K. Bauer, J. A. Engler, H. W. Schroeder, and P. M. Kirkham. Expressed murine and human CDR-H3 intervals of equal length exhibit distinct repertoires that differ in their amino acid composition and predicted range of structures. *J. Mol. Biol.*, 334(4):733–749, Dec 2003. 78

M. P. Lefranc, V. Giudicelli, Q. Kaas, E. Duprat, J. Jabado-Michaloud, D. Scaviner, C. Ginestoux, O. Clement, D. Chaume, and G. Lefranc. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.*, 33(Database issue):D593–597, Jan 2005. 79

C. Etchebest, C. Benros, S. Hazout, and A. G. de Brevern. A structural alphabet for local protein structures: improved prediction methods. *Proteins*, 59(4):810–827, Jun 2005. 79

Annexes

Annex 1: Classification epitopes in groups based on their protein family

Collaborative work on epitope prediction have led to the following article submitted and accepted to BMC Bioinformatics.

Classification epitopes in groups based on their protein family

Edgar Ernesto Gonzalez Kozlova[1], Benjamin Thomas Viart[1], Ricardo Andrez Machado de Avila[2], Liza Figueredo Felicori[1], Carlos Chavez-Olortegui[1]

Carlos Chavez-Olortegui[1]

olortegi@icb.ufmg.br

Liza Figueredo Felicori[1]

liza@icb.ufmg.br

Ricardo Andrez Machado de Avila[2]

r_andrez@yahoo.com.br

Benjamin Thomas Viart[1]

benjamin_viart@hotmail.fr

Edgar Ernesto Gonzalez Kozlova[1]

eegonzalezk@gmail.com

Corresponding Author: olortegi@icb.ufmg.br

[1]Laboratório de Imunoquímica de Proteínas, Departamento de Bioquímica-Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, CP: 486; CEP: 31270-901, Belo Horizonte-MG, Brazil.

[2]Laboratório de Biologia Celular e Molecular, Programa de Pós-Graduação em Ciências da Saúde, Unidade Acadêmica de Ciências da Saúde, Universidade do Extremo Sul Catarinense, CEP: 88806-000. Criciúma-SC, Brazil.

Abstract

Background : The humoral immune system response is based on the interaction between antibodies and antigens for the clearance of pathogens and foreign molecules. The interaction between these proteins occurs at specific positions known as antigenic determinants or B-cell epitopes. The experimental identification of epitopes is costly and time consuming. Therefore the use of *in silico* methods, to help discover new epitopes, is an appealing alternative due the importance of biomedical applications such as vaccine design, disease diagnostic, anti-venoms and immune-therapeutics. However, the performance of predictions is not optimal been around 70% of accuracy. Further research could increase our understanding of the biochemical and structural properties that characterize a B-cell epitope.

Results: We investigated the possibility of linear epitopes from the same protein family to share common properties. This hypothesis led us to analyze physico-chemical (PCP) and predicted secondary structure (PSS) features of a curated dataset of epitope sequences available in the literature belonging to two different groups of antigens (metalloproteinases and neurotoxins). We discovered statistically significant parameters with data mining techniques which allow us to distinguish neurotoxin from metalloproteinase and these two from random sequences. After a five cross fold validation we found that PCP based models obtained area under the curve values (AUC) and accuracy above 0.9 for regression, decision tree and support vector machine.

Conclusions: We demonstrated that antigen's family can be inferred from properties within a single group of linear epitopes (metalloproteinases or neurotoxins). Also we discovered the characteristics that represent these two epitope groups including their similarities and differences with random peptides and their respective amino acid sequence. These findings open new perspectives to improve epitope prediction by considering the specific antigen's protein family. We expect that these findings will help to improve current computational mapping methods based on physico-chemical due it's potential application during epitope discovery.

Keywords: Data mining – B cell epitopes – metalloproteinases – neurotoxins – protein family – epitope prediction.

1 Background

Living organisms often encounter a pathogenic virus, microbe or any foreign molecule during it's lifetime[1]. The B cells of the immune system recognize the foreign body or pathogen's antigen by their membrane bound immunoglobulin receptors, which later produce antibodies against this antigen[2][3]. The recognized sites on the antigen's surface, known as epitopes, represent the minimum wedge recognized by the immune system[4]. Therefore, epitopes lie at the heart of the humoral immune response[5]. The rapid reaction to a previously encountered antigen depends on the binding ability of the antibodies found in the immune system of the organism[6], the physico-chemical properties of the epitope and it's structural conformation[7]. Thus, understanding epitope characteristics and how they are recognized, in sufficient detail, would allow us to identify and predict their position in the antigen[8].

The main objective of epitope prediction is to design a molecule that can replace an antigen in the process of either antibody production or antibody detection[4][9][10][11]. Such a protein can be synthesized in case of peptides or in case of a larger protein, produced by yeast after the gene is cloned into an expression vector[12]. After 30 years of research, it is known that the optimum size of peptides possessing cross-reactive immunogenicity is between 10-15 amino acids[13]. The earliest efforts made to understand and predict B-cell epitopes were based on the amino acid properties, such as flexibility[14], hydrophaty[15], antigenicity[7], beta turns[16] and accessibility[17]. Epitope prediction is important to design epitope-based vaccines and precise diagnostic tools such as diagnostic immunoassay for detection, isolation and characterization of associated molecules for various disease states. These benefits are of undoubted medical importance[18][19].

Recently developed prediction methods face several challenges like data quality[20][7], a limited amount of positive learning examples[21] or difficulty in choosing an appropriate negative learning examples[22]. These negative training samples may harbor genuine B cell epitopes and affect the training procedure, resulting in a poor classification performance[23][24]. Moreover, none of the published work took into account the protein family or function to predict epitopes[25].

The present study explores the possibility of epitopes belonging to same protein family share common properties. For these purpose, the amino acid statistics, physico-chemical and structural properties were compared within each other[26] for two protein's group. This assumption is based on previous studies showing that it exists amino acid trends in composition and shared properties for intravenous immunoglobulins[27]. Despite the difficulty of distinguishing epitopes from non epitopes[28] the addition of information, such as evolutionary and propensity scales, proved to be helpful for epitope prediction[21]. Therefore, it is interesting to assume including information about the protein antigen's family may be resourceful to improve prediction.

2 Methods

Dataset composition

We have obtained experimentally validated 106 linear B-cell epitopes for two groups of antigens (metalloproteinases and neurotoxins) extracted from Pubmed (http://www.ncbi.nlm.nih.gov/pubmed/).

They were manually curated until September 2012 following several search criteria based on the keywords: epitope, metalloproteinase, proteinase, peptidase, toxin and neurotoxin in a joint and disjoint manner. The redundancy was removed for repeated sequences using 100% identity as threshold and the maximum size of the epitopes was fixed to be equal or less than 32. As non epitope data, we created 49 linear random peptides proportional number to the mean of the amount of epitopes in the groups metallorproteinase and neurotoxin. These random peptides are based on the statistics from the dataset UniProtKB/Swiss-Prot, meaning that the sum of the random peptides amino acids are equal to the percentages found in uniprot database. The final set contained 99 non redundant epitopes, containing 29 metalloproteinases, 70 neurotoxins and 49 random peptides as showed in Additional file 1.

Feature selection for data mining analysis

In this study, we generated and used 33 physico-chemical parameters composed by aliphatic index, GRAVY, isoelectric point, amino acid content in percentages, amino acid groups such as hydrophobic (AVILMFYW), positive charged (RHK), negative charged (DE), not charged (STNQ) and specials (SGP) as described by Gasteiger with the difference that each feature was transformed to percentage removing the length difference for the epitope sequences[29]. Also 6 predicted secondary structure properties such as strand, helix, coil, relative surface accessibility, absolute surface accessibility and z-fit which were calculated with Netsurf algorithm[29]. These parameters were calculated for the three groups in study (Metalloproteinase, Neurotoxin and Random) and the results where compared using Welch two sample t-test available in the statistical software R. In total, we evaluated 3 different matrices for the classification purpose of discover how much sequence-derived information was needed to obtained a good classification. The first matrix based of purely PCP information, a second with only PSS data and a third one which was merely the addition of the PSS features to the PCP matrix.

Selection of data mining methods and statistical analysis

The Konstanz Information Miner (KNIME)[30] was used to evaluate Kmeans (KM), decision tree[31] (DT) , naive bayes classifier (NB), support vector machine[32] (SVM) for the matrices generated with our dataset. The free software environment R for statistical computing and graphics was used to create the multiple regression models (LMR). For LMR the nominal class variable was transformed into a numerical variable for the two groups, a positive with value log(0.99/(1-0.99)) for metalloproteinases and a negative been log(0.01/(1-0.01)) for neurotoxins. The linear model function available in R was used to solve a series of equations where the class variable was equal to the feature variables. After solving the equations, a linear multiple regression model was generated, a p-value was calculated and the model was rejected for any p-value superior to 0.005. The predicted resulting score of the model was scaled (0 to 1) by using exp(predicted value./(1+predicted value)) formula. The performance of all the generated models was evaluated for every possible decision threshold with ROCR package by using the parameters AUC (area under the curve formed by true and false positive rates) and accuracy, which gives an overall view of the performance of the classification method used [33].

3 Results

Statistical differences of amino acid composition between metalloproteinase and neurotoxin linear epitopes compared with random sequences

The dataset contain 11 metalloproteinases and 16 neurotoxins. The two protein families (or group) respectively contains 29 and 70 epitopes with an average sequence length of 13.8 amino acids (aa). The minimum length was 4 aa and maximum 32 aa. The negative or non epitope set contained 49 sequences of 14 aa length (Table 1).

Table 1. Dataset composition

Groups	Proteins	Epitopes	Non epitopes
Uniprot	544996	--	--
Neurotoxin	16	29	0
Metalloproteinase	11	70	0
Negative examples	13	0	49

The metalloproteinase and neurotoxin epitopes showed to be different from each other showing a statistical dissemblance for a confidence interval of 95% for the amino acids R, K, M and Y (Table 2, column 1). Also when compared these epitopes to their respective proteins they showed differences for the amino acids R, Q, V and M for metalloproteinases (Table 2, column 4) and D and C for neurotoxins (Table 2, column 5).

These epitope groups also indicated variation when compared to our non epitope control for the amino acids K, C, A, V and I for metalloproteinases and R, K, D, N, Q, C, A, I, K, M and W for neurotoxins (Table 2, columns 2 and 3). As expected, we also detected differences in other parameters such as aliphatic index, grand average of hydropaty and isoelectric point (Table 2, last three rows). Therefore, we were able to identify common characteristics in epitope's composition within unique antigen groups and differences between neurotoxin and metalloproteinase epitope groups.

Table 2: Analysis of means for all datasets with Welch two sample T-test

Parameter	p – values for a confidence interval of 95%				
	(1)ME vs NE	(2)Random vs ME	(3)Random vs NE	(4) MP vs ME	(5) NP vs NE
R (Arg)	**0.0029**	0.0762	**0.0001**	**0.0241**	0.4226
H (His)	**0.0362**	0.1046	0.1074	0.5636	0.7906
K (Lys)	**0.0000**	**0.0113**	**0.0000**	0.4098	0.4818
D (Asp)	0.0890	0.6994	**0.0079**	0.7091	**0.0030**
E (Glu)	0.9289	0.2681	0.0838	0.6696	0.4072
S (Ser)	0.2953	0.5024	0.3546	0.9630	0.8954
T (Thr)	0.4077	0.1867	0.3509	0.2199	0.4523
N (Ans)	0.1878	0.7647	**0.0101**	0.5880	0.4944
Q (Gln)	0.1509	0.9483	**0.0039**	0.8471	0.8185
C (Cys)	0.1821	**0.0003**	**0.0000**	**0.0316**	**0.0075**
G (Gly)	0.6979	0.2576	0.4620	0.3509	0.8450
P (Pro)	0.3156	0.5165	0.3781	0.2103	0.4271
A (Ala)	0.2121	**0.0066**	**0.0000**	0.1092	0.0756
V (Val)	0.0993	**0.0019**	0.2903	0.0550	0.1854
I (Ile)	0.2657	**0.0068**	0.0352	0.1286	0.3275
L (Leu)	0.1374	0.1182	**0.0000**	0.5549	0.2322
M (Met)	**0.0017**	0.0725	**0.0000**	**0.0282**	0.2477
F (Phe)	0.6997	0.4713	0.0765	0.7890	0.5818
Y (Tyr)	**0.0023**	0.5245	**0.0000**	0.8318	0.0938
W (Trp)	0.0889	0.9443	**0.0244**	0.5782	0.1221
Isoe.Point	0.0425	0.5190	0.5190	**0.0425**	0.3221
gravy	0.0672	**0.0010**	**0.0000**	0.0672	**0.0514**
Aliph. Index	**0.0086**	**0.0000**	**0.0000**	**0.0086**	0.8550

Values under p-value under 0.05 are writen in bold. IC=95%, H0 = Difference in means is cero. Hi = Difference in means is not equal to zero. Metalloproteinases epitopes = ME, Neurotoxin epitopes = NE, Metalloproteinase proteins = MP, Neurotoxin proteins = NP, Random = Random sequences.

Decision tree and multiple regression models can distinguish linear B-cell epitopes from two different antigen groups

We investigated our capacity to discriminate if an epitope belonged to neurotoxin or metalloprotease based on the statistical significant differences observed in epitopes amino acids composition, isoeletric point, gravy and aliphatic index (Table 2). For this purpose, we used five different methods: SVM, NB, DT, KM and LMR.

Our analysis used three different input matrices as described before: Only physico-chemical properties (PCP), only secondary structure (PSS) and the combination of both (PCP+PSS) for each algorithm. The performances displayed as AUC values for all data mining methods are showed in table 3. All the methods with the exception of KM were able to group and distinguish correctly both groups of epitopes. As expected, the best results were for SVM followed by similar performance by much simpler techniques, LMR and DT.

Table 3: Performance of all data mining methods showed in AUC and accuracy.

Matrix	**PCP**		**PSS**		**PCP+PSS**	
Models	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**
SVM	1	1	1	1	1	1
MLR	0.986	0.952	0.655	0.714	1	1
DT	0.957	0.962	0.921	0.943	0.943	0.952
NB	0.8	0.838	0.521	0.667	0.793	0.838
KM	0.493	0.667	0.509	0.681	0.507	0.667

During the use of PSS features as input, a reduction in the performance of 0.1-0.3 AUC value was noticed for MLR and NB techniques (Table 3). Only SVM and DT obtained an AUC superior to 0.9 while all the other methods performed poorly with AUC of 0.65 for LMR and close to 0.5 for the others. The SVM technique performed with an AUC of 1.0 for combined properties while LMR showed a slight increase from 0.9 to 1.0. By the other hand DT, NB and Kmeans stayed the same (Table 3). These results indicate that the type of input used (PSS or PCP) were not significant, where the models based on the PCP were the simplest to analyze and understand. The most stable AUC results were obtained with DT method where all the matrices analyzed resulted in an AUC value around 0.95.

The techniques DT and LMR are statistical approaches that showed results similar to SVM which is a non statistical classifier. These methods allowed us to discriminate the epitopes belonging to metalloproteinases or neurotoxins and to identify the important properties inside these groups. The relevant features to classify the epitope groups for the LMR and DT models can be found in table 4.

We observed which amino acids were critical to differentiate epitopes from neurotoxins and metalloproteinases. In the case of LMR model, the amino acids asparagine (N), glutamine (Q) and serine (S), and in the case of DT model the amino acids lysine (K), aspartate (D) and methionine (M) were the key to achieve good classification (above 0.9 AUC) (Table 4).

Table 4: Properties used by the classification models until 8º order out of 39.

Classification Model: Linear Multiple Regression			
Order	**PCP**	**PSS**	**PCP+PSS**
1º	Statistic of N	Z-fit	Statistic of E
2º	Statistic of Q	ASA	Statistic C Atoms
3º	Statistic of S	RSA	Statistic of N
4º	Statistic of T	Strand index	Statistic of Q
5º	Uncharged STNQ	Helix index	Statistic of S
6º	Special CGP	Coil index	Statistic of T
7º	Statistic H Atoms	--	Uncharged STNQ
8º	Statistic C Atoms	--	Statistic H Atoms
Classification Model: Decision Tree			
Order	**PCP**	**PSS**	**PCP+PSS**
1º	Statistic of K	Z-fit	Statistic of K
2º	Statistic of D	RSA	Statistic of D
3º	Statistic of M	ASA	Statistic of M
4º	Statistic S Atoms	Strand index	Statistic S Atoms
5º	Statistic of I	Coil index	Statistic of I
6º	Statistic of W	--	Statistic of W
7º	Statistic of Y	--	Coil index
8º	Isoelectric point	--	--

4 Discussion

The amino acid composition has been investigated for proteins related to the B-cell response [34] and as key for understanding protein-protein interactions[35][36] alongside their role during prediction of epitopes for both T and B-cells[37]. Epitopes are rich in charged and polar amino acids and low in aliphatic hydrophobic amino acids, when comparing the epitope amino acid distribution to either the entire PDB database [38] or to the antigen [39][40]. Also Rubinstein [39] suggested that the amino acid Tyr is significantly over-represented in epitopes and that Val is significantly depleted. Interestingly, the residues Arg and Lys are more frequent in the epitopes of our dataset along other differences as aliphatic index and gravy. This particularities are probably a result of focusing common features in a diverse epitope group, phenomena which was evidenced in the amino acids composition found in epitopes for papilloma viruses [22]. The PCP based methods have been explored in detail for epitope prediction [40] with some limitations in terms of specificity and precision as seen in models for SVM with AUC values of 0.85 for amino acid composition and 0.58, where the accuracy never surpass 0.8 [26].

Our study suggests an improvement in performance when a single epitope group is targeted, resulting in AUC and accuracy superior to 0.9. We included groups of amino acids based on type of charge and lateral chain due to the the concept of amino acids working cooperatively in protein:protein interfaces[41]. Our results indicate that these amino acid groups such as hydrophobic, polar, or special amino acids (CGP), do not posses significance for the prediction models by themselves but may add value when combined with single amino acid statistics.

The secondary structure of epitopes was also investigated by several authors[42][43][44], and epitopes are in general reported to have significantly less strands and helices and significantly more loops compared to the rest of the antigen[8] [38]. The over-representation of loops is small but significant and in agreement with the perception that protein–protein binding sites are flexible regions[41]. The overall secondary structure of epitopes has been reported to been different from regular protein–protein interfaces[23] based on crystals available on the PDB indicating some structural particularities of the Ab-Ag interaction[45]. These particularities could be also family restrictred which could be interesting to explore with computational methods despite of having an accuracy of 79% when predicted from sequence [46] but the DT outcome showed no real relevance in PSS features when applied to epitope classification. The inclusion of predicted secondary structure as commonly done[40] could be a source of misleading results for the prediction, issue which has been reviewed briefly in the literature[47].

The features that characterize each epitope's group could represent the complementary data needed to improve epitope prediction. For example, when adding evolutionary information to the prediction the performance was improved[48] despite recent studies that explain no relation exits between epitope and antigens sequences[28]. Therefore, we showed that a wide range of data mining methods including support vector machine[21], decision tree[48], regression[26] and Naive Bayer classifier had similar successful results bringing some light to the question of which characteristics are important for these epitope groups. It's important to note that we used amino acid percentage[4] in comparison with some recent epitope prediction methods that prefer propensities[12]. The data normalization made in the present study are based on the assumption that each feature is equally relevant for any protein sequence based analysis[9]. We also demonstrate that despite the method, it was possible to classify the studied groups, pointing out the importance of the quality of the used data[49].

5 Conclusions

Our study indicates that linear epitopes that belong a single protein family share common properties but different when compared to epitopes from different families, as demonstrated for neurotoxins and metalloproteinases. We confirmed our hypothesis with five different data mining algorithms, probabilistic and non probabilistic, showing similar results except for Kmeans. The proposed models allowed to separate the studied groups from random sequences based on Uniprot statistics. The models based only in PCP features were enough to show and identify the differences between epitope groups. Therefore, we demonstrate that considering the epitope's protein family can reveal unseen patterns within epitope groups that could be used to improve epitope discovery.

List of abbreviations

SVM: Support Vector Machine

NB: Naive Bayes

DT: Decision Tree

KM: K-Means

LMR: Linear Multiple Regression

PDB: Protein Data Bank

PSS: Position Specific Matrix

PCP: Physico-Chemical-Properties

ASA: Absolute Surface Area

RSA: Relative Surface Area

AUC: Area Under the Curve

ROC: Receiver Operating Characteristic

ME: Metalloproteinase epitopes

MP: Metalloproteinase proteins

NE: Neurotoxin epitopes

NP: Neurotoxin proteins

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

The five authors conceived th study and wrote the manuscript. EEGK and BTV implemented the methods and analyzed the results. CCO, RAMA and LFF revised and oriented the experiments. This work is a revised and extended version of a manuscript presented at the Brazilian Symposion on Bioinformatics in Belo Horizonte, Oct 2014.

Acknowledgements

This research was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, (CAPES-Brazil), (Toxinologia No 23038000825/2011-63). Fundação de Amparo a Pesquisa do Estado de Minas Gerais, Brazil (FAPEMIG-Brazil) and *Conselho Nacional* de *Desenvolvimento* Científico e Tecnológico (CNPq–Brazil).

References

1. Charles Norris Cochrane. Thucydides and the Science of History. Oxford University Press, 35(3):584–585, Apr 1929.
2. FM. Burnet. A modification of Jerne's theory of antibody. Australian Journal of Science, 20:67–69, 1957.
3. NK Jerne. The natural-selection theory of antibody formation. Proceedings of the National Academy of Sciences, 41:849–857, 1955.
4. Perlow DS Boger J. Emini EA, Hughes JV. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. J Virol., 55(3):836–839, 1985.
5. AM. Silverstein. A History of Immunology. Academic Press, San Diego, 1989.
6. Andrew H. Abbas, Abul K. Lichtman. Cellular and Molecular Immunology. 5Th(1):3–14, 2005.
7. J. A. Greenbaum, P. H. Andersen, M. Blythe, H. H. Bui, R. E. Cachau, J. Crowe, M. Davies, A. S. Kolaskar, O. Lund, S. Morrison, B. Mumey, Y. Ofran, J. L. Pellequer, C. Pinilla, J. V. Ponomarenko, G. P. Raghava, M. H. van Regenmortel, E. L. Roggen, A. Sette, A. Sch-lessinger, J. Sollner, M. Zand, and B. Peters. Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. J. Mol. Recognit., 20(2):75–82, 2007.
8. Yang J Chou KC. Chen J, Liu H. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids., 33(3):423–428, Jan 2007.
9. T. P. Hopp and K. R. Woods. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. U.S.A., 78(6):3824–3828, Jun 1981.
10. Toth I. Moyle, PM. Modern subunit vaccines: development, components, and research oppor- tunities. ChemMedChem., 8(3):360–376, Mar 2013.
11. Ditzel HJ Williamson RA Burton DR. Parren PW, Poignard P. Antibodies in human infectious disease. Immunol Res, 21(2-3):265–278, 2000.
12. V. L. Patel, E. H. Shortliffe, M. Stefanelli, P. Szolovits, M. R. Berthold, R. Bellazzi, and A. Abu-Hanna. The coming of age of artificial intelligence in medicine. Artif Intell Med, 46(1):5–17, May 2009.
13. G. N. Sivalingam and A. J. Shepherd. An analysis of B-cell epitope discontinuity. Mol. Immunol., 51(3-4):304–309,

Jul 2012.

14. M. Karplus and J. A. McCammon. The dynamics of proteins. Sci. Am., 254(4):42–51, Apr 1986.
15. J. M. Parker, D. Guo, and R. S. Hodges. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. Biochemistry, 25(19):5425–5432, Sep 1986.
16. J. L. Pellequer and E. Westhof. PREDITOP: a program for antigenicity prediction. J Mol Graph, 11(3):204–210, Sep 1993.
17. I. a. I. Davydov and A. G. Tonevitski. Linear B-cell epitope prediction. Mol. Biol. (Mosk.), 43(1):166–174, 2009.
18. M. Z. Atassi. H. M. Azzazy and W. E. Highsmith. Phage display technology: clinical applications and recent innovations. Clin. Biochem., 35(6):425–445, Sep 2002.
19. M. J. Blythe and D. R. Flower. Benchmarking B cell epitope prediction: underperformance of existing methods. Protein Sci., 14(1):246–248, Jan 2005.
20. Houtao Deng, George Runger, and Eugene Tuv. Bias of importance measures for multi-valued attributes and solutions. Lecture Notes in Computer Science, 6792:293–300, 2011.
21. Wang HW1, Lin YC, Pai TW, Chang HT. Prediction of B-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification. J Biomed Biotechnol. 2011;2011:432830. doi: 10.1155/2011/432830. Epub 2011 Aug 23.
22. N. Subramanian and S. Chinnappan. Prediction of promiscuous epitopes in the e6 protein of three high risk human papilloma viruses: a computational approach. Asian Pac. J. Cancer Prev., 14(7):4167–4175, 2013.
23. E. Zhou Y. Ruan J. Kurgan L. Gao, J. Faraggi. BEST: improved prediction of B-cell epitopes from antigen sequences. PloS One., 7(6):e40104, Jun 2012.
24. Y. El-Manzalawy, D. Dobbs, and V. Honavar. Predicting linear B-cell epitopes using string kernels. J. Mol. Recognit., 21(4):243–255, 2008.
25. PC. Kolaskar, AS. Tongaonkar. A semi-empirical method for prediction of antigenic determi-nants on protein antigens. FEBS Lett., 276:172–174, 1990.
26. H. Singh, H. R. Ansari, and G. P. Raghava. Improved method for linear B-cell epitope prediction using antigen's primary sequence. PloS ONE, 8(5):e62216, 2013.
27. Luštrek M, Lorenz P, Kreutzer M, Qian Z, Steinbeck F, Wu D, Born N, Ziems B, Hecker M, Blank M, Shoenfeld Y, Cao Z, Glocker MO, Li Y, Fuellen G, Thiesen HJ. Epitope predictions indicate the presence of two distinct types of epitope-antibody-reactivities determined by epitope profiling of intravenous immunoglobulins. PloS One. 2013 Nov 11;8(11):e78605. Doi: 10.1371/journal.pone.0078605. ECollection 2013.
28. Ofran Y. Kunik V. The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. Protein Eng Des Sel., 26(10):599–609, Oct 2013.
29. Bent Petersen, Thomas Nordahl Petersen, Pernille Andersen, Morten Nielsen and Claus Lundegaard1. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Structural Biology 2009, 9:51 doi:10.1186/1472-6807-9.
30. Michael R. Berthold and Nicolas Cebron and Fabian Dill and Thomas R. Gabriel and Tobias Otter and Thorsten Meinl and Peter Ohl and Christoph Sieb and Kilian Thiel and Bernd Wiswedel. KNIME: The Konstanz Information Miner. Studies in Classification, Data Analysis, and Knowledge Organization. Springer. ISSN:1431-8814. 2007.
31. EJ. Bremel, RD. Homan. An integrated approach to epitope analysis I: Dimensional reduction, visualization and prediction of MHC binding using amino acid principal components and regression approaches. Immunome Res., 6(7):1745–7580, Nov 2010.
32. D. Kam YW. Tong JC. Wee, LJ. Simarmata. SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction. BMC Genomics, 2(11):1471–2164, 2010.
33. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2014.

34. T. Kurosaki. Regulation of B-cell signal transduction by adaptor proteins. Nat. Rev. Immunol., 2(5):354–363, May 2002.

35. S. Jones and J. M. Thornton. Principles of protein-protein interactions. Proc. Natl. Acad. Sci. U.S.A., 93(1):13–20, Jan 1996.

36. CW. Su EC. Lin, SY. Cheng. R. Liu and J. Hu. Computational prediction of heme-binding residues by exploiting residue interaction network. PloS ONE, 6(10):e25560, 2011.

37. Greenbaum JA Emami H Hoof I Salimi N Damle R Sette A Peters B. Vita R, Zarebski L. The immune epitope database 2.0. Nucleic Acids Res., D:854–862, Nov 2010.

38. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. Nucleic Acids Res., 28(1):235–242, Jan 2000.

39. N. D. Rubinstein, I. Mayrose, D. Halperin, D. Yekutieli, J. M. Gershoni, and T. Pupko. Computational characterization of B-cell epitopes. Mol. Immunol., 45(12):3477–3489, Jul 2008.

40. Zhao M Li Q. Zhang W, Liu J. Predicting linear B-cell epitopes by using sequence-derived structural and physicochemical features. Int J Data Min Bioinform., 6(5):557–569, 2012.

41. J. Janin and C. Chothia. The structure of protein-protein recognition sites. J. Biol. Chem., 265(27):16027–16030, Sep 1990.

42. U. Reimer. Prediction of linear B-cell epitopes. Methods Mol Biol., 524:335–344, 2009. N. D. Rubinstein, I. Mayrose, D. Halperin, D. Yekutieli, J. M. Gershoni, and T. Pupko. Computational characterization of B-cell epitopes. Mol. Immunol., 45(12):3477–3489, Jul 2008.

43. C. P. Toseland, D. J. Clayton, H. McSparron, S. L. Hemsley, M. J. Blythe, K. Paine, I. A. Doytchinova, P. Guan, C. K. Hattotuwagama, and D. R. Flower. AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. Immunome Res, 1(1):4, Oct 2005.

44. L. Zhao, L. Wong, L. Lu, S. C. Hoi, and J. Li. B-cell epitope prediction through a graph model. BMC Bioinformatics, 13 Suppl 17:S20, 2012.

45. O. Keskin, B. Ma, K. Rogale, K. Gunasekaran, and R. Nussinov. Protein-protein interactions: organization, cooperativity and mapping in a bottom-up Systems Biology approach. Phys Biol, 2(2):24–35, Jun 2005.

46. J. L. Pellequer, E. Westhof, and M. H. Van Regenmortel. Correlation between the location of antigenic sites and the prediction of turns in proteins. Immunol. Lett., 36(1):83–99, Apr 1993.

47. Bourne PE. Ponomarenko JV. Antibody-protein interactions: benchmark datasets and prediction tools evaluation. BMC Struct Biol., 2:7–64, Oct 2007.

48. S. Saha and G. P. Raghava. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. Proteins, 65(1):40–48, Oct 2006.

49. S. Saha, M. Bhasin, and G. P. Raghava. Bcipep: a database of B-cell epitopes. BMC Genomics, 2005.

Aditional file 1: The datasets composed of the sequences used in this work is available in this .csv file, containing four columns. First column shows the pubmedID of the paper from which the sequence was extracted. The second column contains the sequence. The third collumn contain the sequence IDs from genebank, uniprot or pdb, databases. The fourth column contains the class of the sequences which can be neurotoxin, metalloproteinase or random. The column separator in this .csv file is a standart semicolon ";".

Journal_ID	Epitope_sequence	Uniprot, PDB, genebank	Class
21763377	SCMLDQGRSRCR	P22796	metalloproteinase
21763377	HCTMDQGRLRCR	P22796	metalloproteinase
21763377	HCFHDQGRVRCA	P22796	metalloproteinase
21763377	TCATDQGRLRCT	P22796	metalloproteinase
21763377	QCTMDQGRLRCR	P22796	metalloproteinase
16212890	MEASHTHARPAP	Q5C1N0	metalloproteinase
16212890	TLAHTSQIGLTA	Q5C1N0	metalloproteinase
16212890	TSFGSMLSKWQK	Q5C1N0	metalloproteinase
16212890	ITSHTGYLQLRL	Q5C1N0	metalloproteinase
16212890	SNPPGMALSAPP	Q5C1N0	metalloproteinase
20093370	GFEESLEVDTNPL	P10845	metalloproteinase
19509157	YTFRYPLSL	B3KQS8	metalloproteinase
15607634	IRIKRDMS	AAG32166	metalloproteinase
15607634	GTSMATPHVAG	AAG32166	metalloproteinase
16428330	IADCTYRWHVGTWMECSVSCGD	Q76LX8	metalloproteinase
16737347	DVKCGRLYC	(EOC0028-06-63-24),(EOC0063-24)	metalloproteinase
16737347	GTICKMARGDNMHDYCN	(EOC0028-06-63-24),(EOC0006)	metalloproteinase
16737347	GTKCEDGKVC	(EOC0028-06-63-24),(EOC0063-24)	metalloproteinase
16737347	TECRGIRSECDLPEYCTGQ	(EOC0028-06-63-24),(EOC0063-24)	metalloproteinase
16737347	NCRDPCCDAASCKLHSW	(EOC0028-06-63-24),(EOC0063-24)	metalloproteinase
16737347	GEECDCGSPENCQ	(EOC0028-06-63-24),(EOC0063-24)	metalloproteinase
16737347	HNLGMNHDGNQCNCGAAGCIMSAIISQYRS	(EOC0028-06-63-24),(EOC0028-06-63)	metalloproteinase
19084031	HNLGMEHDGKDCL	Q9I9R4	metalloproteinase
19084031	NTVNGFFRSMN	Q9I9R4	metalloproteinase
17014879	SEGPSYEFSDCS	P22796	metalloproteinase
17014879	LKTFGEWRERVL	P22796	metalloproteinase
17014879	VVADHGMFTKYN	P22796	metalloproteinase
18061641	IVNTLNEIYRYLYVR	2ERO(B);Q8JIR2	metalloproteinase
18061641	EQQRYLNNFRFIELV	2ERO(B);Q8JIR2	metalloproteinase
7690110	VKDGYIVD	P01484;1AHO	neurotoxin
9784249	KKYRYYLKPLCKK	1CLP	neurotoxin
9276446	IVDDVNCTYFCGRNAYC	1AHO;P01484	neurotoxin
9276446	NEECTKLKGESGYCQ	1AHO;P01484	neurotoxin
9276446	ACYCYKLPDHVRTKG	1AHO;P01484	neurotoxin
9276446	YKLPDHVRTKGPGRCH	1AHO;P01484	neurotoxin
9276446	ACYCYKLPDHVRT	1AHO;P01484	neurotoxin
22922018	FTNPEEGDLNPPPEAKQVPVSYYDSTYLST	2ILP;Q7B8V4	neurotoxin
22922018	VPVSYYDSTYLSTDNEKDNYLKG	2ILP;Q7B8V4	neurotoxin
22922018	SPDFTFGFEESLEVDTNPLLGAGKFATDP	2ILP;Q7B8V4	neurotoxin
22922018	DFTFGFEESLEVDTNPLLG	2ILP;Q7B8V4	neurotoxin
22922018	KMLTEIYTEDNFVFFKVLNRKTYLNFDKAVFK	2ILP;Q7B8V4	neurotoxin
22922018	PKVNYTIYDGFNLRNTNLAANFNGQNTEINNMNFTK	2ILP;Q7B8V4	neurotoxin
22922018	FNGQNTEINNMNFTKLKNFTGLFEF	2ILP;Q7B8V4	neurotoxin
22922018	FNGQNTEINNMNFTKLKNFTGLFEFYK	2ILP;Q7B8V4	neurotoxin
11275260	KDLYG	Q9TXD1, P08815	neurotoxin
11275260	NVKTSPKQSKP	Q9TXD1, P08815	neurotoxin
9517541	KVWRDHRGTIIE	3NDS,1IQ9;P01426	neurotoxin
9517541	KPGI	3NDS,1IQ9;P01426	neurotoxin
21149386	YNQYTEEEK	2ILP;Q7B8V4	neurotoxin
21149386	YKKYSGSDK	2ILP;Q7B8V4	neurotoxin
22149274	WTLQDTQEIKQRVVF	2ILP;Q7B8V4	neurotoxin
22952786	SKWY	2NM1;P10844	neurotoxin
22952786	SDEFY	2NM1;P10844	neurotoxin
22952786	KSDP	2NM1;P10844	neurotoxin
19162253	NPVEWFMSTVNT	1CTX;P01391	neurotoxin
16647121	EENISLDLIQQYYLTFNFI	2ILP;Q7B8V4	neurotoxin
16647121	SGAVILLEFIPEIAIPVLG	2ILP;Q7B8V4	neurotoxin
16647121	TKAIINYQYNQYTEEENN	2ILP;Q7B8V4	neurotoxin
16647121	NKFLNQCSVSYLMNSMIPY	2ILP;Q7B8V4	neurotoxin
16647121	CMENNSGWKVSLNYGEIIW	2ILP;Q7B8V4	neurotoxin

16647121	GEIIWTLQDTQEIKQRVVF	2ILP;Q7B8V4	neurotoxin
16647121	NNIMPKLGCRDTHRYIWI	2ILP;Q7B8V4	neurotoxin
16647121	KYVDVNNVGIRGYMYLKGP	2ILP;Q7B8V4	neurotoxin
16647121	SRTLGCSWEFIPVDDGWGERPL	2ILP;Q7B8V4	neurotoxin
11425742	KGTFDPLQEPRT	2ILP;Q7B8V4	neurotoxin
8576079	TNCYKKRWRDHRGYRTE	P60770	neurotoxin
7945236	CAPGQNLCY	1NTN;P01382	neurotoxin
7945236	PGQNLCYTK	1NTN;P01382	neurotoxin
7945236	KTWCDAWCG	1NTN;P01382	neurotoxin
7945236	DAWCGSRGK	1NTN;P01382	neurotoxin
11602284	LPDSEPTKTNGKCKS	2sn3;P15226	neurotoxin
11602284	GREGYPADSKGCKIT	2sn3;P15226	neurotoxin
11602284	TLKKGSSGYCAWPAC	2sn3;P15226	neurotoxin
11602284	PDSVKIWTSETNKCG	2sn3;P15226	neurotoxin
15302529	VPDHIKVWDYATNK	2sn3;P15226	neurotoxin
15302529	GLPDSEPTKTNGKCK	2sn3;P15226	neurotoxin
15302529	LPNWVKVWDRATNKC	2sn3;P15226	neurotoxin
15970301	KEGYAMDHEGCKFSC	2sn3;P15226	neurotoxin
15970301	CDGYCKTHLKASSGY	2sn3;P15226	neurotoxin
15970301	PDHIKVMDYATNKKC	2sn3;P15226	neurotoxin
15970301	KEGYLMDHEGCKLSC	2sn3;P15226	neurotoxin
15970301	IRPSGYCGRECGIKK	2sn3;P15226	neurotoxin
15970301	LPNWVKVWDRATNKC	2sn3;P15226	neurotoxin
15970301	KKDGYPVEYDMCAYI	2sn3;P15226	neurotoxin
15970301	WNYDNAYCDKLCKDK	2sn3;P15226	neurotoxin
9022703	GYIVDDV	P01484	neurotoxin
9022703	IVDDVNC	P01484	neurotoxin
9022703	LKGESGY	P01484	neurotoxin
9022703	VKDGYIVD	P01484	neurotoxin
9022703	YIVDDVN	P01484	neurotoxin
9276446	IVDDVNCTYFCGRNAYC	P01484	neurotoxin
9276446	NEECTKLKGESGYCQ	P01484	neurotoxin
9276446	PDHVRTKGPGRCH	P01484	neurotoxin
9276446	YKLPDHVRT	P01484	neurotoxin
11750040	KELYGSSA	P01484	neurotoxin
11750040	TSPKQCSKPC	P01484	neurotoxin
19962461	GRNAYCN	Q7YXD3	neurotoxin
19962461	YIVDDVNCT	Q7YXD3	neurotoxin
UniprotKB	HRMSMRIFLRFQPRP	random peptide 1	random
UniprotKB	FNYGKDATGASAPYS	random peptide 2	random
UniprotKB	GMELYTMVAMIWGAG	random peptide 3	random
UniprotKB	EAQGQLKREWKNAPF	random peptide 4	random
UniprotKB	SDNGSSEALYQSQLS	random peptide 5	random
UniprotKB	VDGPLMFFNFKTFPS	random peptide 6	random
UniprotKB	QSGWEEEEKTKERQV	random peptide 7	random
UniprotKB	EDPNSYVERRLGGVR	random peptide 8	random
UniprotKB	SAAAFIYLLASKSRQ	random peptide 9	random
UniprotKB	PSMAQPAAKAGSEEL	random peptide 10	random
UniprotKB	EISGAEVNDFTRKSI	random peptide 11	random
UniprotKB	EHAPSYAADVAQDVD	random peptide 12	random
UniprotKB	TYRSRNRTLPKDAAE	random peptide 13	random
UniprotKB	PAVGVKKATEQKTVD	random peptide 14	random
UniprotKB	IPSEWKFFIALIGVP	random peptide 15	random
UniprotKB	GVPMQEDTQAGYSVQ	random peptide 16	random
UniprotKB	IRRTDTINDIPMQCL	random peptide 17	random
UniprotKB	HGASGYDNEQVSGSK	random peptide 18	random
UniprotKB	SGDNAKAGKENTDGR	random peptide 19	random
UniprotKB	IKNRGIEFPTDAGGR	random peptide 20	random
UniprotKB	LMTWKERSVFDGTMD	random peptide 21	random
UniprotKB	GECQYDKQKPILTGC	random peptide 22	random
UniprotKB	VPSACDFVEGTSLGD	random peptide 23	random

UniprotKB	QTVADEASLGHRTRA	random peptide 24	random
UniprotKB	SVVDNAAAKFKKGPA	random peptide 25	random
UniprotKB	RAARLPRKGVVYAFK	random peptide 26	random
UniprotKB	FRVNETYRIYPWYIG	random peptide 27	random
UniprotKB	IRSLLQGDIMRQLEQ	random peptide 28	random
UniprotKB	QVAIVRGLSGGERGV	random peptide 29	random
UniprotKB	VGPSLELSGSITVVI	random peptide 30	random
UniprotKB	IVYRQDGDQFPIYSS	random peptide 31	random
UniprotKB	IFKIVDKSLIRVMGN	random peptide 32	random
UniprotKB	LSAWGGAHYLGSGRS	random peptide 33	random
UniprotKB	ARTVLLTPRAGDLVI	random peptide 34	random
UniprotKB	RSSNYEFGDGMLKRL	random peptide 35	random
UniprotKB	LRRADGQKVVDAEAL	random peptide 36	random
UniprotKB	KMWIGSPQSDQLGQM	random peptide 37	random
UniprotKB	ANVPVLENSLKTTGN	random peptide 38	random
UniprotKB	FYKTVKLAEFDMETT	random peptide 39	random
UniprotKB	KFGFTNRLGEKSAGA	random peptide 40	random
UniprotKB	RVFDPSEISESWASQ	random peptide 41	random
UniprotKB	VAIVTAIERMSPSLF	random peptide 42	random
UniprotKB	YSEEAIAARKMMNRF	random peptide 43	random
UniprotKB	HPILELSYVPVVSLS	random peptide 44	random
UniprotKB	PAIGKSAVRRYFEVK	random peptide 45	random
UniprotKB	IASTMNPAVVFFKQY	random peptide 46	random
UniprotKB	GAASFTRLGSYANVG	random peptide 47	random
UniprotKB	ERKFLESKLIMDWKE	random peptide 48	random
UniprotKB	LGATALATNEATGTR	random peptide 49	random

Annex 2:EPI-Peptide Designer README file

README file from the EPI-Peptide Designer program.

README,
Viart Benjamin, benjamin.viart@gmail.com July 2015

WELCOME TO EPI-PEPTIDE DESIGNER

1) Execution

In the terminal, in the folder of the EPIDESIGNER.jar file execute :

java -jar EPIPEPTIDE-DESIGNER.jar

2) Usage

EPI-PEPTIDE DESIGNER can be used for two main purposes,

a) Compute graph representation of Antibody – Antigen interfaces from the PDB and BLUE STAR STING server.

- **Input** : the input format is organized as follow :
  PDB;CHAIN-A;CHAIN-B
  where CHAIN-A correspond to the one letter code of the protein of one side of the interface form the PDB and CHAIN-B the corresponding one letter code of the other side of the interface.
  In the folder you will find a file (allABAGinterface.txt) containing all the interfaces present in our databases.
- **Options** : Pajek or Gaston
  Those option will modify the format of output of the graph

b) Generate EPI-peptides. EPI PEPTIDE DESIGNER can generate EPI-PEPTIDES from a set of AB-AG interfaces, a putative or real epitope sequence and a score of similarity.

- **Input** : the input format is organized as follow :
  PDB;CHAIN-A;CHAIN-B
  where CHAIN-A correspond to the one letter code of the protein of one side of the interface form the PDB and CHAIN-B the corresponding one

letter code of the other side of the interface.

In the folder you will find a file (peptideInput.txt) peptide Antigen interfaces code to be used as input..

!!! => THE DESIGN IS BASED ON FIRST SIDE DEFINED!!!

ex : PDB;CHAIN-A;CHAIN-B design will be based on CHAIN-A

for AB-AG interfaces the Antibody chain(s) comes first !

1A3R;LH;P

- **Options** : Epitope Sequence (Real or Putative)

  A sequence of epitope has to be define for EPI-PEPTIDE DESIGNER to base the design on.

  The sequence can contain any of the 20 Amino – Acids in the one letter format and '-' use in this case to create non-linear sequence.

  Ex: AFTG-GIMNCPLTR-RG

  Size

  The Size indicate the size of the EPI-PEPTIDE to be generated.

  Number

  The number of EPI-PEPTIDE to be generated

  Score

  The score (expressed between 1 to 100) represent the importance of the inputed epitope sequence in the design of the EPI-PEPTIDES.

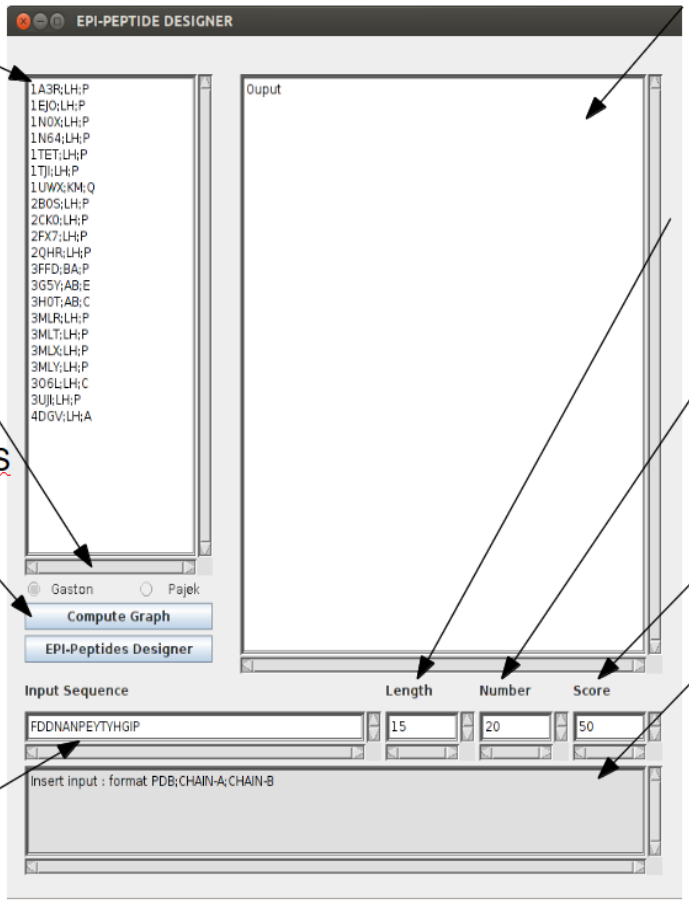
Input Field:
PDB;CHAIN-A;CHAIN-B
file: peptideInput.txt
Ex: 1A3R;LH;P

! The design will be based on the first side!

Option to use with the compute graph

Program mode:
-Compute graph
-Desing of EPI-PEPTIDES

Epitope (real or putative) Sequence



Output Field

Size of peptide (At least 1)

Number of peptides (At least 1)

Score From 1 to 100

Logs field Display error, details of processes...