

# Exploring on Protein-DNA interactions

Thèse N° 9426

Présentée le 6 mai 2019

à la Faculté des sciences de la vie

Unité du Prof. Dal Peraro

Programme doctoral en biologie computationnelle et quantitative

pour l'obtention du grade de Docteur ès Sciences

par

**Alexandra Styliani KALANTZI**

Acceptée sur proposition du jury

Prof. J. Fellay, président du jury

Prof. M. Dal Peraro, directeur de thèse

Prof. A. Stasiak, rapporteur

Prof. C. Dessimoz, rapporteur

Prof. J. Maddocks, rapporteur

2019



## Acknowledgments

Firstly, I would like to thank sincerely my supervisor, Professor Matteo Dal Peraro for giving me the opportunity to be part of his lab. I have been very lucky to have a supervisor that was patience, encouraging, helpful and who cared that much about my work.

Next, I would like to thank my collaborators, Professor Bart Deplancke, Dr. Alina Isakova, Giulia Fonti and Lucien Fabrice Krapp. It was a pleasure working with them.

I wish to thank all the members of the lab, not only for their input and interesting scientific talks, but also for the nice moments that we had together. Special thanks to Dr. Chan Cao, Deniz Aydin and again Giulia Fonti.

A big thanks to my mentor, Professor Jacques Fellay for being always there, supporting me in difficult times. The advice and assistance given by him was greatly appreciated.

Finally yet importantly, I would like to thank my friends, my mother and my husband Fabrizio. I am so grateful to them, I wouldn't have made it without their precious support, and scientific advises.





## Abstract

A variety of DNA-binding proteins organizes the chromosomal DNA and regulates gene transcription, and DNA replication and recombination. In particular, for gene regulation there is a category of DNA-binding proteins, the transcription factors, which can detect and bind to a specific set of DNA motifs. For the protein-DNA complexes there are some solved structures, but for some cases, it is difficult to define experimentally these structures at the atomistic level. However, there are various *in silico* methods, that alone or in combination with experimental techniques, are capable of predicting proteins structures with high accuracy.

DNA-binding proteins contain DNA-binding domains and unfortunately, there are few protein-DNA complexes that have been characterized at atomistic resolution. Here we have developed artificial neural networks to predict the binding sites and the interface between proteins and the DNA, based on a structure-based approach. Specifically, we used different interface descriptors and we included a variety of protein-DNA complexes, in order to predict the correct localizations of the DNA on a protein.

A large group of proteins that can recognize specific DNA sequences is the KRAB-ZFP family. DNA recognition by zinc finger proteins (ZFPs) plays an important role in gene regulation. However, the molecular determinants defining the recognition of specific DNA nucleotides by ZFP finger repeats has not yet been decoded. Here, we present a method that can predict which ZF repeats specifically bind to a DNA target sequence and what is the most probable DNA target sequence for these repeats. Our method is based on the structural analysis of the binding network of resolved protein DNA complexes, and is validated on a benchmark set of solved ZFP-DNA complexes. We also characterized the binding specificity of two KRAB-ZFPs (ZFP14 and ZNF145) integrating our predictions with SMiLE-seq (Selective Microfluidics-based Ligand Enrichment followed by sequencing) data.

Subsequently, we use our recognition code on ChIP-exo data, in order to give an insight into the poly-ZF domains binding patterns. We determined the most probable binding motifs and we separated the ZFPs in two groups: the ones that have a potential canonical and non-canonical binding.

Altogether, this work gives an insight into the proteins that interact with the DNA, through a better understanding of the various protein-DNA interfaces and the predictions of these complexes at the atomistic level.

## Contents

	Page
Acknowledgments .....	3
Abstract .....	5
Contents .....	6
Chapter 1 Introduction .....	9
1.1 DNA structure and function .....	10
1.2 Protein-DNA interactions .....	12
1.3 Types of Protein-DNA interactions .....	13
1.4 Methods to detect protein-DNA interactions .....	14
1.5 Objectives of the thesis .....	17
Chapter 2 Methods .....	19
2.1 Motif discovery .....	19
2.2. Biomolecular modeling techniques .....	20
2.3. Molecular dynamics simulations .....	22
2.4. <i>pow<sup>er</sup></i> : a parallel optimization workbench to enhance resolution in biological systems .....	26
2.5. Artificial neural networks .....	27
Chapter 3 A machine learning approach to predict protein-DNA interactions .....	35
3.1. Introduction .....	35
3.2. Results .....	36
3.3. Conclusions and perspectives .....	46
Chapter 4 A computational method to predict the DNA specificity of zinc finger proteins .....	48
4.1. Introduction .....	48
4.2. Results .....	52
4.3. Conclusions and perspectives .....	72
Chapter 5 Analysis of the DNA-binding landscape of human KRAB-ZFPs .....	75
5.1. Introduction .....	75
5.2. Results .....	76
5.3. Conclusions and perspectives .....	87
Chapter 6 Conclusions and perspectives .....	89
Chapter 7 Appendix – Related work addressing protein-DNA interactions .....	92
7.1. Molecular modeling of the KAP1 complex .....	92
7.2. Molecular modeling of the CLOCK:BMAL1 complex bound to DNA .....	99
7.3. Molecular modeling and molecular dynamics simulations of the Hop2- Mnd1-DNA complex .....	101
References .....	103
Curriculum Vitae .....	115

## List of figures

- Figure 1 - Complementary DNA chains.
- Figure 2 - DNA types.
- Figure 3 - Bonded interactions.
- Figure 4 - *pow<sup>er</sup>* workflow.
- Figure 5 - Components of the artificial neuron.
- Figure 6 - Feedforward artificial neural network.
- Figure 7 - Roc curves.
- Figure 8 - Illustration of test cases from the protein-DNA benchmark.
- Figure 9 - Histogram of the iRMSDs for each complex, for each of the three methods.
- Figure 10 - ANN results.
- Figure 11 - Histogram of maximum AUC, MCC and accuracy
- Figure 12 -  $\beta\beta\alpha$  C2H2 ZF domains bound to DNA.
- Figure 13 - Representations of the 17 unique ZFP-DNA complexes deposited in PDB.
- Figure 14 - Sliding window for ZNF282 and a SMiLE-seq predicted DNA sequence.
- Figure 15 - Predictions of the ZF domains than bind on the DNA-target.
- Figure 16 - Comparison of the predictions of the four recognition codes.
- Figure 17 - Alignments of the DNAs predicted by SMiLE-seq, by our recognition code, by Najafabadi's, Persikov's and Gupta's recognition codes.
- Figure 18 - Structural models of ZFP14 and ZNF415.
- Figure 19 - Crystal structure of the mouse ZFP568, ZF3-ZF11.
- Figure 20 - ChIP-exo reads of ZNF765.
- Figure 21 - Statistics of the remaining 102 human KRAB-ZFPs.
- Figure 22 - Correlation between the length of the ChIP-exo motif and the cc between our prediction and the MEME motifs.
- Figure 23 - ZFP568 consensus DNA-binding sequence.
- Figure 24 - Consistency of the 3 binding positions.
- Figure 25 - PsychoProt's descriptors.
- Figure 26 - Phylogenetic tree of C2H2-ZF domains.
- Figure 27 - KAP1 domains.
- Figure 28 - The KRAB-ZFP/KAP1 complex and its cofactors.
- Figure 29 - Initial molecular model for KAP1.
- Figure 30 - Possible oligomerization states of the RBCC domain.
- Figure 31 - A model for the CLOCK:BMAL1 heterotetramer.
- Figure 32 - A model for dsDNA binding to the WHD pair of Hop2-Mnd1.

## List of tables

Table 1 - iRMSDs.

Table 2 - Maximum AUC, MCC and accuracy.

Table 3 - ZFP-DNA complexes from PDB.

Table 4 - Amino acid-nucleic acid binding specificity scores based on ZFP-DNA complexes.

Table 5 - Benchmark results of the sliding window method on known ZFP-DNA complexes.

Table 6 - Scores of the alignments between the consensus sequence and the predicted sequences.

Table 7 - ZFPs with their SMiLE-seq predicted DNA sequences and selected truncations.

Table 8 - Results of the sliding window method on 2 ZFPs and their SMiLE-seq predicted binding sequences.

Table 9 - Alignments of truncated ZFPs.

Table 10 - Consistency of the 3 binding positions.

Table 11 - PsychoProt's descriptors and  $k^*$  values.

## Chapter 1      Introduction

Protein polypeptide chains, according to their composition, can fold into a vast number of different three-dimensional structures, which determine, along with the dynamic features, their biological functions (Alberts et al., 2002). Some of the key functions proteins regulate are the catalysis of biochemical reactions (i.e., enzymes), maintenance of mechanical structure in cells (i.e., structural components), protection from viruses and bacteria (i.e., antibodies), signal transmission to coordinate biological processes between different cells, tissues, and organs (i.e., messengers), gene regulation through DNA recognition (i.e., transcription factors), chromatin packing (i.e., nucleosomes). For the last two functions, proteins that can bind to the DNA are involved. In particular, DNA-binding proteins organize the chromosomal DNA and regulate gene transcription, and DNA replication and recombination. Therefore, characterizing the molecular mechanisms of the protein DNA-interactions, through their structure, is crucial for understanding their function.

To this aim one needs to understand the physical and chemical properties of these biomolecules when interacting with their native environment. Recent advances in X-ray crystallography (Smyth and Martin, 2000), Nuclear Magnetic Resonance (NMR) spectroscopy, cryo-Electron Microscopy (cryo-EM) and other techniques are providing a growing body of structural information for protein complexes at atomistic level of resolution. However, direct experimental determination of protein structures is often difficult and each experimental technique has some drawback. For example, protein crystals suitable for X-ray crystallography cannot always be produced when it comes to large assemblies. On the other hand, cryo-EM can overcome this problem, but it is limited by resolution, although this has recently started to change.

Therefore, the combined structural information from different experimental methods can be integrated in computational models in order to build a more accurate description. Integrative modeling is the combination of all kind of biophysics and biochemical experimental data, such as X-ray crystallography, NMR, Cryo-EM structural data and mass spectroscopy (Thalassinos et al., 2013). Molecular modeling is useful in this context to construct missing parts of the structure (loops or domains) using homology modeling. Furthermore, experimental inputs and modeling itself can contribute to define new intermolecular interactions in large assemblies. Moreover, Molecular Dynamics (MD) simulations can be eventually used in order to study the different configurations of a structural complex. Consequently, by combining all these techniques through integrative modeling our understanding of large macromolecular complexes has been largely advanced in recent times. Lastly, as the computational power is steadily increasing, integrative modeling is more widely used, aiding in consistently interpreting a huge amount of data generated from different and heterogeneous sources.

## 1.1. DNA function and structure

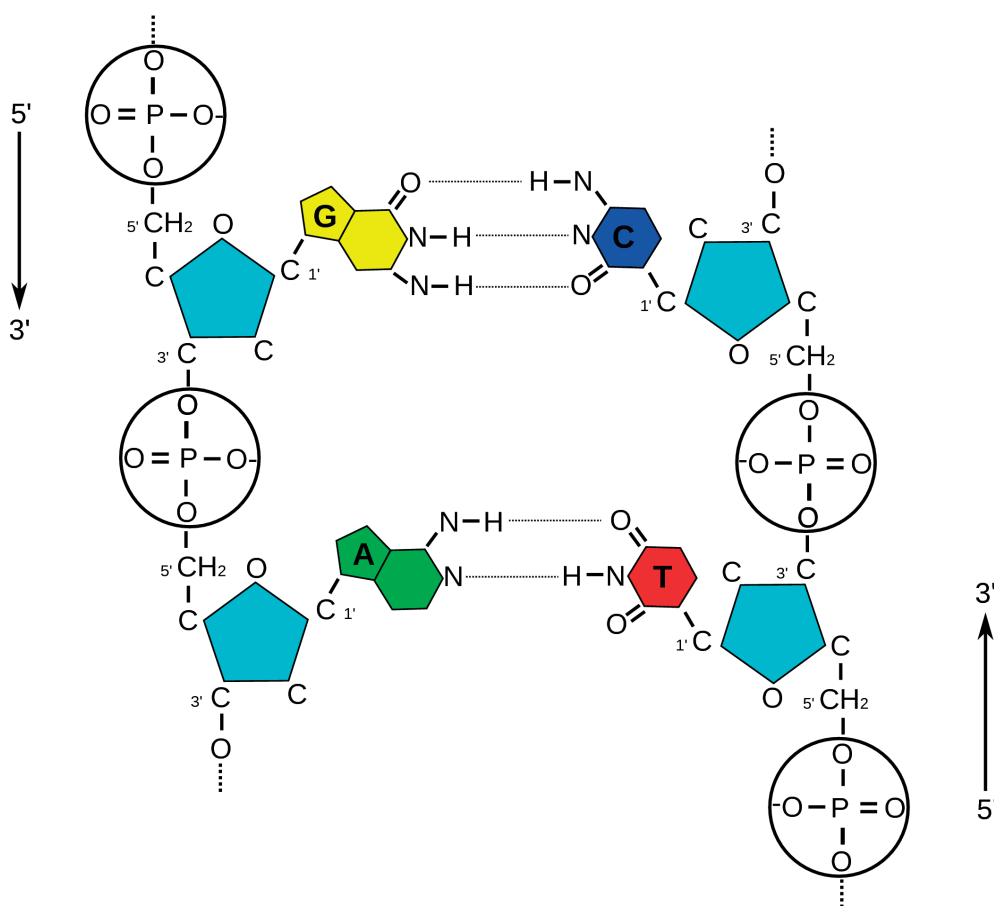
Nucleic acids, proteins, lipids and complex carbohydrates are the four major types of macromolecules that are essential for life. Natural nucleic acids, like DNA (Deoxyribonucleic acid) and RNA (Ribonucleic acid) are long linear macromolecules that contain the genetic information. These macromolecules consist of one or two long chains of monomers, called nucleotides. Each nucleotide contains a sugar, a phosphate group and a nitrogen base. From the natural nucleic acids (DNA-RNA) artificial nucleic acids (PNA, LNA, GNA, TNA) can be constructed with changes in the backbone (Alberts B et al., 2014).

DNA is a chain of nucleotides carrying the genetic instructions used in the growth, development, functioning and reproduction of all known living organisms and many viruses. Most DNA molecules consist of two biopolymer chains coiled around each other to form a double helix. Each chain is made by deoxyribonucleotides, which contain a nitrogenous base, a deoxyribose sugar and a phosphate group (Figure 1). A base and the 1' carbon of a deoxyribose are connected with a covalent bond. Also, the nucleotides are joined to one another in a chain, by covalent bonds between the sugar of one nucleotide and the phosphate of the next nucleotide, forming an alternating sugar-phosphate backbone. The covalent bonds are formed between the phosphate group and the 5' deoxyribose carbon, and they are called 3' to 5' phosphodiester bonds. Always, the first nucleotide of a chain has a free phosphate group attached to its 5' deoxyribose carbon, and the last nucleotide has a free hydroxyl group from its 3' deoxyribose carbon. Thus, the orientation of the nucleotide chain is refereed as 5'-3'. In addition, the two strands of DNA are antiparallel, since they run in opposite directions to each other. Since the strands are not symmetrically located with respect to each other, the DNA grooves have different size. The major groove is 22 Å wide and the minor groove is 12 Å wide. Moreover, the distance between two base pairs is 3.4 Å (around 10 base pairs per turn) and the diameter of the DNA is about 20 Å (Mandelkern et al., 1981) (Figure 2).

The genetic information is stored in the sequence of the bases along the DNA chain. In particular, the four nitrogen bases are adenine (A), cytosine (C), guanine (G) and thymine (T). (Figure 1). Adenine and guanine bases belong in the group of purines, while cytosine and thymine belong to the group of pyrimidines. The nitrogenous bases of the two polynucleotide strands are bound together with hydrogen bonds in order to form the double-stranded DNA. Specifically, an adenine forms two hydrogen bonds with a thymine, and vice versa, while a cytosine forms three hydrogen bonds with a guanine, and vice versa (Nikolova et al., 2013). This base pairing has as a result the formation of a right-handed double helix, which consists of the two threaded polynucleotide chains (Watson and Crick, 1953). The phosphate backbone is polar and is located in the exterior of the molecule, while, in the interior we have the presence of the bases. The base pairs ensure that the two chains of a DNA molecule are complementary, and this implies that the sequence of

one determines the sequence of the other. The complementarity is of great importance for the replication of DNA, a property that makes it the most suitable molecule for the conservation and transfer of the genetic information. Each DNA chain can serve as a matrix for the synthesis of a complementary chain so that two double-stranded DNA molecules can be formed identical to the initial molecule.

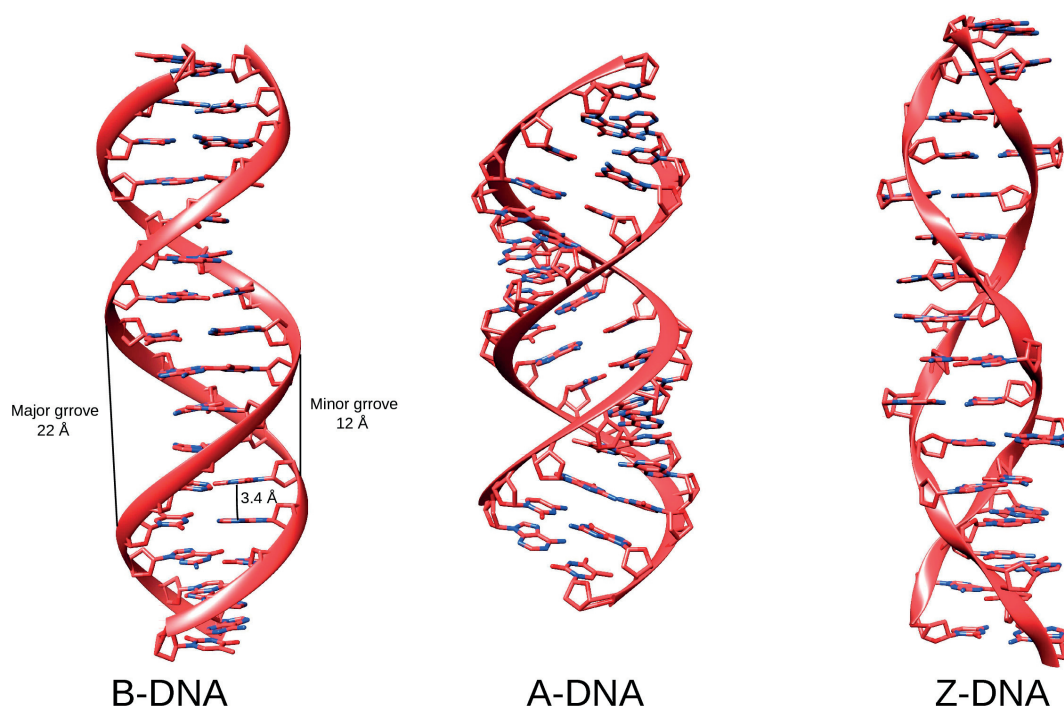
Part of the DNA in the cell has coding properties, which provide instruction for making proteins. In particular, coding DNA transcribes into RNA that translates into protein. However, the majority of the DNA has non-coding features. Initially, non-coding DNA thought it was without purpose. Nevertheless, it was observed that non-coding DNA contains sequences that act as regulatory elements that determine the genes that are turned on and off (transcriptional regulation). Other functions include scaffold attachment regions, origins of DNA replication, centromeres and telomeres. Still, the properties of the non-coding DNA are partially understood.



**Figure 1** - Complementary DNA chains. Guanine and Cytosine form three hydrogen bonds. Adenine and Thymine form two hydrogen bonds.

There are various configurations of the DNA double strand due to its structural flexibility (Figure 2). The most well-known structure is the B-DNA and it was described above. Other structures that are less common are the A-DNA (Wahl and

Sundaralingam 1997), which has a larger diameter than the B-DNA, and the Z-DNA (Rothenburg et al., 2001), which is much smaller in diameter than the B-DNA and it is left-handed.



**Figure 2** - DNA types. B-DNA is the most common DNA type. There are also other types, like the A- and Z- DNA.

## 1.2. Protein-DNA interactions

Protein-DNA interactions play an important role in many biological processes i.e., gene regulation, DNA repair, replication and packing. Therefore, amino acids that interact with the DNA backbone are conserved, while amino acids that interact with DNA bases have different levels of conservation (Luscombe and Thornton, 2002). Moreover, mutations in DNA-binding proteins can lead to neurological disorders (Camero et al., 2013), Axenfeld-Rieger syndrome (Saleem et al., 2001), tumors (Ching Ang et al., 2006) and interference with immunoglobulin transactivation (Nixon et al., 2004). Although it has been reported that DNA-binding proteins are encoded by 2-3% of a prokaryotic genome and 6-7% of a eukaryotic genome (Luscombe et al., 2000; Walter et al., 2009), only 3.3% of the proteins that are deposited in PDB (April 2018) are in complex with the DNA. This makes imperative the need of predicting the 3D atomistic structures of the protein-DNA complexes and the specific interactions between their interfaces, and subsequently combine them with experimental results through integrative modeling.



The 3D structures of the DNA-binding proteins and the DNA play a crucial role for their binding interface. Binding to the DNA is performed by DNA-binding domains that contain specific structural motifs. These motifs usually include a "recognition  $\alpha$ -helix", like in the helix-turn-helix, helix-loop-helix, zinc finger, leucine zipper and other DNA-binding domains. There are exceptions like the  $\beta$ -ribbon proteins that use  $\beta$ -strands to contact the DNA (Stephen 1991; Phillips 1994). In general, DNA-binding proteins in PDB, have been classified in eight different structural/functional groups (Luscombe et al., 2000). The eight groups are: (i) helix-turn-helix (i.e.  $\lambda$  repressor of bacteriophage lambda), (ii) zinc-coordinating (i.e. C2H2 zinc finger proteins), (iii) zipper-type (i.e. leucine zipper family members and Basic-helix-loop-helix proteins), (iv) other  $\alpha$  helix (i.e. Cre protein), (v)  $\beta$  sheet (i.e. TATA box-binding proteins), (vi)  $\beta$  hairpin/ribbon (i.e. MetJ repressor), (vii) other, and (viii) enzymes. In addition, DNA-binding proteins, besides one or more DNA-binding domains, also consist of other functional domains.

The interaction interface between a DNA-binding protein and the DNA shows a high degree of complementarity in terms of shape and electrostatics (Hård et al., 1996). This requires conformational changes in the protein and the DNA. The main interactions of proteins and the DNA include H-bonds (which can induce specificity), water mediated H-bonds and hydrophobic interactions. All these interactions contribute to the stability of the protein-DNA complex.

### 1.3. Types of protein-DNA interactions

There are two main categories of protein-DNA interactions: those when a protein recognizes specific DNA bases (base readout) and those when a protein recognizes a sequence-independent DNA shape (shape readout) (Rohs et al., 2010), but sometimes it can be a combination of both (i.e. the *Escherichia coli* Trp operator). Usually the proteins that recognize specific DNA bases form H-bonds with the polar parts of these bases (i.e. transcription factors), while the proteins that bind without specificity to the DNA interact with the phosphate backbone (i.e. histones). Therefore, the binding interface of these proteins usually consists of positive amino acids, like Arg or Lys, which can form H-bonds with the negatively charged phosphate backbone. Base readout DNA-binding proteins interact with the bases in both the major and the minor DNA groove. More frequent are the interactions with the major groove, due to the fact that there are present more hydrogen donors/acceptors than in the minor groove (Harteis et al., 2014). Shape readout DNA-binding proteins recognise changes on the DNA structure in global (i.e. narrow minor groove and DNA kinks) or local shape (bending, A-DNA, Z-DNA) (Rohs et al., 2010).

An example of proteins that participate in non-specific interactions with the DNA are the histones. Histones organize the DNA into a compact structure called chromatin (Alberts, 2002). In particular, they form an octamer that consists of two

types of histones H2A, H2B, H3 and H4. Moreover, histones have positively charged amino acids that bind to the negatively charged DNA phosphate backbone. The DNA (~150 bps) wraps around the histone octamer in a left-handed supercoil. This conformation is called nucleosome and it is the basic unit of DNA packing (Reece and Campbell, 2006). Another example is the Hop2-Mnd1 heterodimer, which participates in meiotic DNA recombination (see Appendix). In particular it binds to double-strand DNA and together with Dmc1, it promotes single-strand DNA (Kang et al., 2015).

An example of proteins that participate in specific interactions with the DNA are transcription factors (TFs). TFs are DNA-binding proteins that can bind to a specific set of DNA sequences and activate or repress the transcription of a gene. TFs contain one or more DNA-binding domains, but also protein-interacting domain(s). In order to regulate transcription, a TF uses its protein-interacting domain(s) to interact with other cofactor proteins or the RNA polymerase (Latchman, 1993). TFs, together with their co-regulators, form complex highly interconnected gene regulatory networks. In humans, there are around 1600 TFs (Madan et al., 2004). A class of TFs are the C2H2 zinc finger proteins (see Chapter 4), which are involved in the regulation of many processes in the cell, such as tissue homeostasis (Cassandri et al., 2017). Another TF is the CLOCK:BMAL1 heterotetramer, which regulates the transcription of genes that are participating in the mammalian circadian clock (see Appendix). In particular, it binds to two repeats of E-box sequences (CACGTG), with a distance of 6 or 7 base pairs between them (Nakahata, et al. 2008; Sobel et al 2017).

#### 1.4 Methods to detect Protein-DNA interactions

There are many experimental methods (*in vivo* and *in vitro*) that have been developed to study protein-DNA interactions (Cai and Huang, 2012). In particular, EMSA (Electrophoretic Mobility Shift Assay) is based on the principle that Protein-DNA complexes migrate slower than the unbound linear DNA on a non-denaturing gel and can provide absolute binding affinities (dissociation constants  $K_d$ ) (Garner and Revzin, 1981; Fried and Crothers, 1981). Another technique is the Deoxyribonuclease (DNase) I footprint, which is used for locating the specific binding sites of proteins on DNA (Galas and Schmitz, 1978). This technique is based on the principle that a protein bound to a DNA fragment, will protect that fragment from enzymatic cleavage, which in this case is the DNase. Next, ChIP-seq is an *in vivo* technique for analyzing protein-DNA interactions and histone modifications. Specifically, ChIP-seq combines chromatin immunoprecipitation with parallel DNA sequencing to identify the binding sites of DNA-associated proteins and the chromatin state (Gilmour and Lis, 1983). A modification of the ChIP-seq protocol is the ChIP-exo technique, which uses exonucleases to degrade the strands of the

protein-bound DNA (Rhee and Pugh, 2011). In this case the resolution of the binding sites is improved from hundreds of base pairs to almost one base pair. A relatively new technique that can detect low affinity interactions between protein and the DNA is the MITOMI (Mechanically Induced Trapping of Molecular Interactions), and it has been used for measuring the binding energy landscapes of transcription factors (Maerkl and Quake, 2007; Rockel et al., 2012). A similar technique is SMiLE-seq (Selective microfluidics-based ligand enrichment followed by sequencing), which was developed for the identification of DNA binding specificities and affinities of transcription factors (Isakova et al., 2017). Other techniques for determining the DNA-binding specificities of proteins are the bacterial one-hybrid system (Bulyk, 2005), the yeast one-hybrid system, the SELEX (Systematic Evolution of Ligands by Exponential enrichment) and the protein binding microarrays. Finally, x-ray crystallography can be used to solve the structure of a protein-DNA complex and give an atomistic view of its interactions.

Compared to wet lab experiments, *in silico* methods can be a fast and cheap alternative in the characterization of protein-DNA interactions. These methods have focused on predicting whether a protein can bind to the DNA and/or which are the protein binding residues. The current approaches for these predictions are either sequence-based or structure-based (Si et al., 2015). Sequence-based approaches are using sequence similarity and other sequence features, such as residue type, evolutionary conserved residues and global composition of amino acids. However, they cannot achieve satisfactory predictions because the DNA-binding residues are less conserved, and therefore, structural information is necessary. Structure-based approaches are using the 3D structures of unbound proteins (apo form) and the bound proteins in complex with the DNA (holo form). If no structures are available, but there is a highly homologous complex or protein, then a 3D model can be built via homology modeling. Through the 3D structure, one can have an insight into the structural information of the binding sites. This structural information can be translated into structural descriptors. These descriptors can be the electrostatic potentials, hydrophobicity, accessible surface area, structural motifs etc.

For methods that predict protein-DNA interactions the “state of the art” algorithms are based on machine learning approaches (Lu et al., 2013). Machine learning is a field of computer science that proposes algorithms that can learn from and make predictions on data. In particular, for the protein-DNA interaction predictions ANN (Artificial Neural Networks), SVM (Support Vectors Machines), decision trees/random forests and Bayesian networks have been used. In general, these methods “learn” from a training set of data and attempt to predict the correct output from an unknown test set.

An ANN is a learning process that is inspired by biological neural networks. ANNs are used to model complex relationships and for patterns recognition. For instance, DISPLAR (DNA-Interaction Site Prediction from a List of Adjacent Residues) is a structure-based method that uses an ANN to predict the DNA-binding

sites on protein surfaces (Tjong and Zhou, 2007). In particular, this method takes into account the conservation of the positive residues and the solvent accessibility of the residues at the interface. DISPLAR's predictions on DNA-contacting residues have an accuracy of 76%. Stawiski et al. proposed another method that uses ANN to predict DNA-binding proteins, in 2003. This approach is structure-based and as descriptors, it uses the positively charged electrostatic patches on protein surfaces.

A support vector machine (SVM) is a supervised machine-learning method that is used for a classification of two classes. In the case of protein-DNA interactions, SVMs are used to distinguish DNA-binding residues from non-DNA-binding residues on a protein. For example, BindN+ is a sequence-based tool that uses SVM classifiers to distinguish DNA-binding residues (Wang et al., 2010). The descriptors that are used are PSSMs (Position-Specific Scoring Matrix) and sequence-based evolutionary information, like hydrophobicity, side chain pK<sub>a</sub> value and molecular mass. In this case, distinguishing DNA-binding residues from non-DNA-binding residues has an accuracy of 79.0% (AUC 0.825). Another approach with a SVM is ProteDNA, which is a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors (Chu et al., 2009).

A decision tree is a decision support tool that uses a tree-like graph of decisions and their possible consequences. An extension of that is the random forest, which is a learning method for classification that uses multiple decision trees. For example, DBindR is a sequence-based program that predicts the DNA-binding residues of a protein using a random forest model (Wu et al., 2009). The descriptors that are used are the evolutionary information of the amino acid sequence, secondary structure information and the characteristics the amino acids, in respect of the dipoles and volumes of the side chains. DBindR predicts the DNA-binding residues with accuracy of 91.41% (AUC 0.913).

A Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional independencies via a directed acyclic graph (DAG). A Naïve Bayes classifier has been used from predicting DNA-binding sites of proteins through a sequence-based approach, with 71% accuracy (Yan, 2006).

## 1.5. Objectives of the thesis

In nature, many protein-DNA complexes participate in fundamental biological processes. Currently, there are very few solved atomistic protein-DNA complexes and many determinants of their interacting interface remains unknown. In this thesis, my aim is to study the protein DNA-interactions and propose a recognition code between ZFPs and the DNA. In particular, (i) I give an insight into the protein-DNA interface predictions, next (ii) I investigate the specificity between ZFPs and the DNA (iii) and finally provide an analysis on ChIP-exo binding motifs from poly-ZF proteins and finally.

### A machine learning approach to predict protein-DNA interactions

Since, a large amount of protein sequence data is available, it is useful to develop computational tools that can analyse fast and accurately protein-DNA interactions. In chapter 3, we followed a structure-based approach to predict the DNA and protein binding sites. In particular, we used artificial neural networks to predict among a variety of protein-DNA complexes, which will be the ones that interact. The final goal is to predict the interacting residues in both the DNA and the protein. Therefore, we used the bound and unbound states from a protein-DNA benchmark and investigate their interactions with different number of potential complexes, descriptors and matrices. Eventually, we manage to predict the specific protein-DNA interactions with high accuracy.

### A computational method to predict the DNA specificity of zinc finger proteins

In Chapter 4, my aims are to determine how the C2H2-ZF tandem recognizes 3-4 bps, define the general recognition binding rules between ZFPs and their DNA targets and present a method to analyse and predict ZFP binding specificity. Here I propose (i) a structure based method that can predict which ZF domains of a ZFP can bind to a given DNA sequence and (ii) a recognition code that predicts the most probable DNA target sequence for a ZFP. Our approach is based on the analysis of the H-bond network of solved structures of protein-DNA complexes. We verified our method and our recognition code with known solved ZFP-DNA complexes and compared the results with the results of other recognition codes. For given ZF tandem repeats, our method predicts accurately the ZF domains that bind to the DNA. Our recognition code gives accurate results for most of the cases and competes the current experimental recognition codes. In addition, we further validated our method using ZFPs of the KRAB-ZFP family for which we performed SMiLE-seq experiments of wild type and mutated ZFPs. One additional benefit of our

method is the possibility to eventually built 3D molecular models of these ZFP-DNA complexes providing atomistic insight into the structural binding recognition between ZFPs and DNA.

### Analysis of the DNA-binding landscape of human KRAB-ZFPs

In Chapter 5, raw ChIP-exo data have been analysed to give an insight into the ZFPs binding patterns. We determined the most probable binding motifs and we focused on the poly-ZF proteins that have a potential non-canonical binding. Our results show that it is possible to have multiple ZF domains bound serially on the DNA, but sometimes the binding differs. For example, in the cases that a predicted DNA motif is very long (>18 bps) the canonical binding is disturbed and few zinc finger domains are leading the binding.

## Chapter 2      Methods

Protein-DNA interactions can be studied with experimental methods, but also with computational approaches, or combination of both. For instance, in order to discover the landscape of the DNA sequences that a protein binds, we use experimental methods (i.e. ChIP exo) and a motif discovery algorithm. In particular, motif discovery algorithms can describe the binding specificity of a protein through position weight matrices. Moreover, we can use artificial neural networks (or other machine learning approaches) to predict which proteins can bind to DNA or which residues participate in the interaction interface. This approach works better when we have a benchmark of solved protein-DNA complexes, and therefore it requires high quality data from NMR or X-ray crystallography. However, in the case that there is no solved structure, we can use biomolecular modeling techniques to build a complex. For proteins, we can use as a template a homologous structure or predict a de novo model. Similarly, for the DNA, previous knowledge of its geometry can be used to build accurate 3D models. In addition, to study a molecular system, such as a protein-DNA complex, principles of molecular mechanics can be used. Specifically, we can use molecular dynamics simulations to explore the conformational space of complexes and have a better insight into their binding interface and H-bond network. Furthermore, *pow<sup>er</sup>*, which is an integrative modeling framework, can be used to predict a set of protein-DNA orientations, based on structural information from an initial complex. In this work, we used all the above computational methods to explore protein-DNA interactions.

### 2.1. Motif discovery

Many experimental methods study protein-DNA interactions. Some of these methods (i.e. ChIP-seq) provide us with a large set of DNA-target sequences that a protein can bind to. Among all this information, it is challenging to extract the accurate binding motif (consensus sequence or PWM (position weight matrix)) and detect the "real" specificity of a DNA-binding protein. For this reason, motif discovery algorithms have been developed. These algorithms are classified in three approaches: (i) deterministic optimization, (ii) probabilistic optimization and (iii) enumeration (D'haeseleer, 2006).

Enumerative algorithms cover the entire search space of all possible motifs, for a specific motif model description. These algorithms don't have the risk of getting stuck in a local minimum, but they may overlook some subtle patterns that are present in binding sites. In the probabilistic optimization cases the input data for a given problem is not known accurately, therefore the optimization is considered under the assumption of probability distributions. In particular, the motif model is initialized



with a randomly selected set of sites, and every site in the target sequences is scored against this initial motif model. At each iteration, the algorithm probabilistically decides whether to add a new site and/or remove an old site from the motif model, weighted by the binding probability for those sites. The resulting motif model is then updated, and the binding probabilities recalculated. Given sufficient iterations, the algorithm will efficiently sample the joint probability distribution of motif models and sites assigned to a motif, focusing on the best fitting combinations. On the other hand, in the deterministic optimization cases, the input data for a given problem is known accurately. Specifically, EM (Expectation Maximization) algorithm is used to simultaneously optimize a PWM of a motif and the binding probabilities for its associated sites. EM is an iterative method to find maximum likelihood or maximum a posteriori estimates of parameters in statistical models. The PWM for the motif is initialized with a single n-mer subsequence, plus a small amount of background nucleotide frequencies. Next, for each n-mer in the target sequences, the probability that it was generated by the motif is calculated. Then, EM takes a weighted average across these probabilities to generate a more refined motif model. The algorithm iterates between calculating the probability of each site based on the current motif model and calculating a new motif model based on the probabilities.

An example of motif discovery with a deterministic optimization approach is the MEME algorithm (Multiple EM for Motif Elicitation) (Bailey and Elkan, 1994; Bailey et al., 2009). The algorithm discovers multiple motifs of various lengths in a set of DNA or protein sequences by using EM to fit a mixture model to the set of sequences. MEME requires only a set of unaligned sequences. It returns a model of each motif and a threshold. Both of them can be used to search for occurrences of the motif in other databases. The algorithm can discover several motifs with differing numbers of occurrences in a single dataset. MEME-ChIP is a variation of MEME, which is used for motif discovery in large nucleotide datasets, such as those occurred from ChIP-seq data (Machanick and Bailey, 2011). Here, we used MEME-ChIP to discover binding motifs of ZFPs from SMiLE-seq and ChIP exo data, in Chapters 4 and 5 respectively.

## 2.2. Biomolecular modeling techniques

In living organisms, there is a variety of biomolecules, like proteins and DNA, which interact between them and with the environment, in order to perform important biological tasks. Unfortunately, a fraction of protein structures and therefore functions are known. For determining these unknown structures, besides wet lab techniques, many computational techniques have been developed. In particular, molecular modeling techniques are used to construct missing parts or a complete 3D structural complex, while MD (Molecular Dynamics) simulations are used to study the different configurations of a structure. However, the best approach to build an accurate



biomolecular model is to combine structural information from different experimental methods (X-ray crystallography, NMR and Cryo-EM) with computational methods. By combining all these approaches through integrative modeling, representations of macromolecular assemblies have been more accurate in the last years (Thalassinos et al., 2013).

Homology modeling (or comparative modeling) is the procedure of building a 3D structure of a protein using as a template an experimentally solved structure of a homologous protein. A widely used program for homology modeling of protein structures is Modeller (Šali and Blundell, 1993; Eswar et al., 2006). This program requires an alignment of a sequence that will be modeled, with the sequences of known related structures. Modeller builds a model by satisfying spatial restraints that contain all non-hydrogen atoms. It can also build missing loops in protein structures and optimize various models of protein structures with respect to a flexibly defined objective function. Here, we used Modeller in Chapter 4, in order to model unknown ZFP structures. Specifically, we build the linkers between the ZF domains, ZF domains with length different from 21 amino acids and non-canonical ZF domains.

Another approach in homology modeling of protein structures is Swiss-Model (Biasini et al., 2014). Swiss-Model is an automated protein homology-modeling server, which requires experimental protein structures ("templates") to build models for evolutionary related proteins ("targets"). The Swiss-Model pipeline consist of four steps that are involved in building a homology model of a given protein sequence. Initially, BLAST and HHblits are used to identify structural templates from PDB. After that, the target sequence and the template structures are aligned. Next, models are built based on a rigid fragment assembly approach, followed by energy minimization. Finally, the quality of the models is assessed using a statistical potential of mean force. We used Swiss-Model in order to build models for the domains of KAP1, which had a homologous protein in PDB.

A structure prediction tool is the Rosetta software suite, which includes algorithms for computational modeling and analysis of protein structures (Simons et al., 1997). One of Rosetta's applications is the Loop Modeling, which uses fragments from existing PDB structures in order to guide the building of missing loops (Leaverfay et al., 2011; Wang, Bradley, & Baker, 2007). In addition, a protein structure prediction server, which uses the Rosetta software, is Robetta. In particular, Robetta combines both *ab initio* and homology modeling to predict the 3D structure of protein domains. Specifically, domains that have a PDB homolog are automatically modeled based on that homolog as a template. These various homolog template PDBs are detected and aligned. Next, the alignments are clustered and homologous models are built based on the RosettaCM protocol. On the other hand, domains without PDB homolog are modeled with the Rosetta *de novo* protocol (Simons et al., 1997). *De novo* protein structure prediction refers to structure prediction without using templates, but with knowing only the amino acid sequence and applying first-principles (i.e. physics). Rosetta's protocol uses energy functions in conjunction with

fragments from existing PDB structures, in order to guide the model building. Here, we used both Rosetta and Robetta in order to build a model for the TSS domain of KAP1 and construct the missing loops, respectively.

A software that is used to analyse, build and visualize 3D Nucleic Acid structures, is the 3DNA (Lu and Olson, 2003). At its core, the software uses a simple matrix-based scheme for calculating a complete set of rigid-body parameters that characterize the spatial relationship of the base pairs in DNA and RNA structures. In particular, 3DNA can be used for DNA modeling, given a nucleic acid sequence, while the 3D structure can be also described geometrically with a set of rigid-body parameters. We used 3DNA in Chapters 3 and 4, in order to build DNA models that interact with DNA-binding proteins (i.e. ZFPs).

### 2.3. Molecular dynamics simulations

Molecular dynamics (MD) simulation is a technique for computing the transport properties of a many-body system, in which the nuclear motion of the constituent particles obeys the laws of classical mechanics (Frenkel and Smit, 2001). In particular, simulations can provide the individual particle motions as a function of time. For this reason, simulations are used as a tool for understanding the physical basis of the structure and function of biomolecules.

A widely used parallel molecular dynamics software is NAMD (Nanoscale Molecular Dynamics), which is designed for high-performance simulation of large biomolecular systems (Phillips et al., 2005). The software is implemented in C++ and based on Charm++ parallel objects, while is compatible with AMBER and CHARMM force fields. In this work, we used NAMD to perform energy minimization on the ZFP-DNA complexes in Chapter 4, in order to improve the H-bond network between the ZFP and the DNA.

### Molecular Mechanics

Molecular mechanics uses classical mechanics to model molecular systems. In all-atom MD simulations, each atom is represented as one particle and the electronic motion is ignored. In molecular mechanics, a force field is a set of parameters and functions that are designed to describe the interactions between the atoms in a molecular system. The energy of the system is the numerical solution of the classical equations of motion, which are summed as the interactions between bonded atoms (Figure 3) and non-bonded atoms. The bonded atoms are covalently linked and the non-bonded atoms are the ones that are separated by three bonds at least. The potential energy function is calculated as following:

$$U_{\text{total}} = \{ U_{\text{bond}} + U_{\text{angle}} + U_{\text{propTorsion}} + U_{\text{impropTorsion}} \} + \{ U_{\text{vdW}} + U_{\text{Coulomb}} \}$$

$\uparrow$   
 bonded

$\uparrow$   
 non-bonded

## Bonded interactions

For bond stretching the harmonic approximation is used to describe the potential energy  $U_{\text{bond}}$  between two covalently bonded atoms:

$$U_{\text{bond}} = \sum_{\text{bond}} \frac{1}{2} k_b (r - r_0)^2$$

Where  $K_b$  is a harmonic force constant for a bond and it is an indicator of the stiffness of the bond spring. The energy of a bond changes according to its length  $r$ . The lowest energy of a bond is at a particular length  $r_0$  (equilibrium bond length). If the length is less than  $r_0$  then the bond is compressed and the electron clouds of the two atoms can overlap, which will lead to an increase of the energy. Respectively, if the length is more than  $r_0$  then the bond is stretched and the energy increases.

Angle bending includes the angles between each pair of covalent bonded atoms, sharing a single atom at the vertex. For the angle energy,  $U_{\text{angle}}$  a harmonic potential is also used:

$$U_{\text{angle}} = \sum_{\text{angle}} \frac{1}{2} k_a (\theta - \theta_0)^2$$

Here,  $k_a$  is the angle bending force constant (stiffness of the angle spring),  $\theta$  is the bond angle and  $\theta_0$  is the equilibrium bond angle.

Next, torsional angle rotation include atom pairs separated by exactly three covalent bonded atoms with the central bond subject to the torsion angle. Although very important interactions, some early force fields omitted torsional angle interactions. Proper torsion involves four consecutively bonded atoms. The proper torsion energy  $U_{\text{propTorsion}}$  is modeled by a simple periodic function:

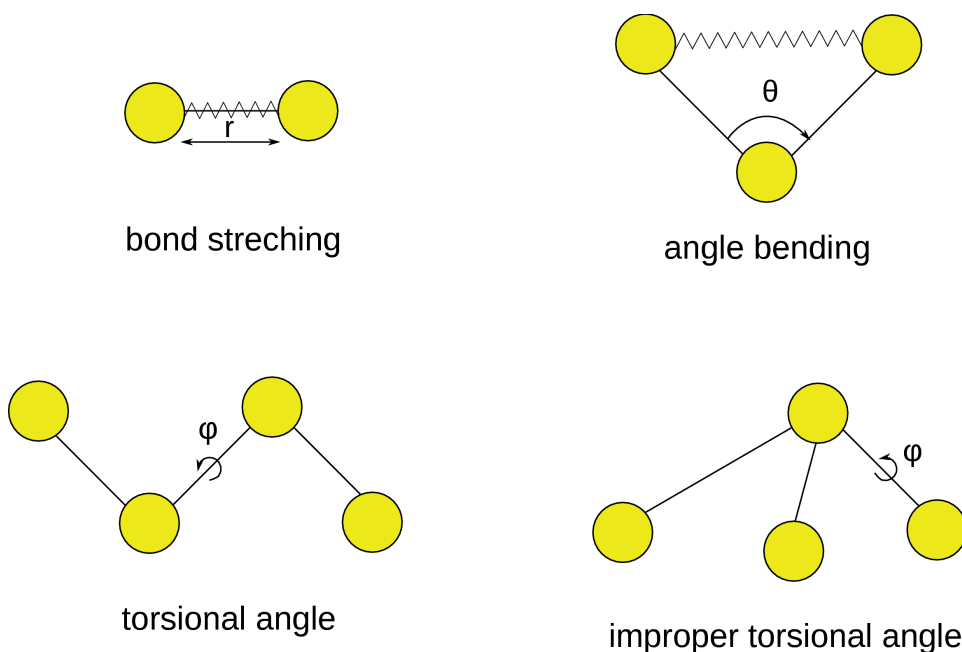
$$U_{\text{propTorsion}} = \sum_{\text{torsion}} k_f [1 + \cos(n\phi - \delta)]$$

Analytically, the rotation dihedral angle is given by the  $\phi$  parameter, the  $K_\phi$  parameter controls the amplitude of the curve, the  $n$  parameter controls its periodicity and the  $\delta$  parameter shifts the entire curve along the rotation dihedral angle.

Improper torsion energy is calculated based on four atoms not successively bonded. The small out-of-plane angle is governed by the so-called "improper" dihedral angle. It is modeled by a harmonic potential:

$$U_{impropTorsion} = \sum_{improp} k_i (\phi - \phi_0)^2$$

The parameter  $\phi$  is the improper torsional angle,  $\phi_0$  is the equilibrium angle and  $k_i$  represents the spring force constant.



**Figure 3** - Bonded interactions. Bond stretching, angle bending, proper and improper torsional angles.

### Non-bonded interactions

Non-bonded interactions are occurring between two non-covalently bonded atoms, in the same or different molecules. Van der Waals interactions are modeled by the Lennard-Jones potential, which is a mathematical approximation of these interactions between a pair of atoms:

$$U_{vdW} = \sum_i \sum_{j>i} 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right]$$

The depth of the potential is determined by the parameter  $\epsilon$ , where  $r_{ij}$  is the distance between the two atoms and  $\sigma_{ij}$  is the distance at which the van der Waals energy is minimized. Sometimes in MD simulations, van der Waals interactions are truncated at a cutoff distance.

The electrostatic interactions are described by Coulomb's law, which is quantifying the amount of force with which stationary electrically charged particles repel or attract each other:

$$U_{Coulomb} = \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

Where  $r_{ij}$  is the distance between two atoms,  $q_i$  and  $q_j$  are the charges of the atoms and  $\epsilon_0$  is the dielectric constant. Usually, electrostatic interactions beyond a specified distance are ignored or assumed zero. It depends if you use a cutoff scheme, in PME they are not ignored, as they are long range and thus important at large distances

## RMSD

After a simulation or an energy minimization, the structural similarity metric that used to compare two molecules is the RMSD (Root-Mean-Square Deviation) of the atomistic positions. In particular, it is the measure of the average distance between the atoms of two superimposed molecules. The atoms that are usually used for proteins are the C $\alpha$  and for the DNA are the Pho. RMSD is calculated in Å or nm (1 nm = 10 Å) and its equation is the following:

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \delta_i^2}$$

Where N is the number of atoms and  $\delta_i$  is the distance between the atoms i of the two superimposed structures.

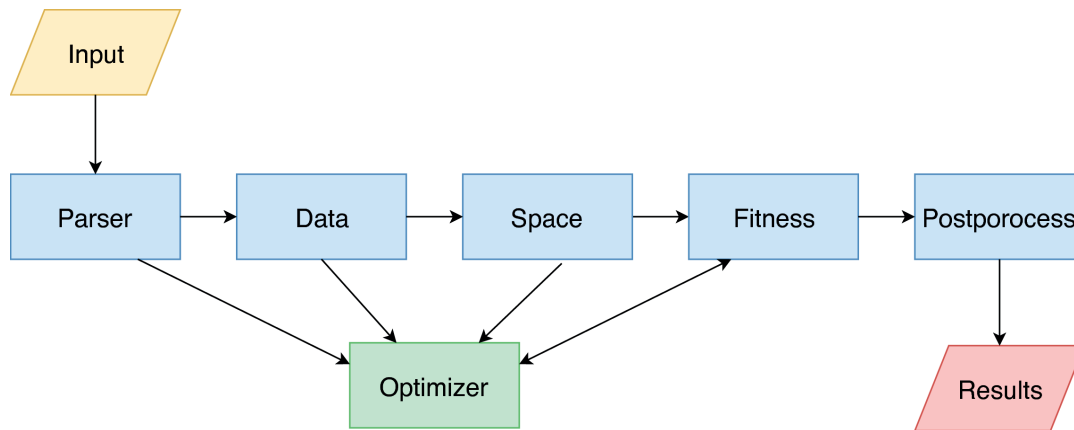
## 2.4 *pow<sup>er</sup>*: a parallel optimization workbench to enhance resolution in biological systems

*pow<sup>er</sup>* (Parallel Optimization Workbench) is an open source optimization framework designed for dynamic integrative modeling of biological systems, proposed by Degiacomi and Dal Peraro in 2013. This method has been used to predict protein assembly of symmetrical complexes, using experimental information as restrains and the protein conformational space from molecular dynamics sampling. Initially, predictions were performed using PSO (Particle Swarm Optimization). PSO is a computational method for the optimization of continuous nonlinear functions and it was initially used to simulate a simplified social model (Kennedy and Eberhart, 1995; Shi and Eberhart, 1998). Specifically, this method solves a problem by having a population of candidate solutions. These solutions are moved around like particles in the search-space according to a mathematical formula. Each particle's movement is affected by its current position, velocity, local best-known position and the best-known positions by other particles. This is expected to move the swarm of particles toward the best solutions. In particular, PSO optimizes a problem by trying to improve a candidate solution with respect to a given measure of quality. Lately, an assembly protocol based on mViE (memetic Viability Evolution) (Maesani et al., 2016) was added in *pow<sup>er</sup>*, by Tamo et al. in 2017. mViE is an optimizer that maintains and recombines multiple sub-populations in order to optimize the balance between local and global search. Here, we used *pow<sup>er</sup>* and mViE in order to optimize the localization of the DNA on a DNA-binding protein and produce multiple poses. We applied this approach on various protein-DNA complexes in Chapter 3.

*pow<sup>er</sup>* is implemented in Python, and MPI for Python is used for high-performance computing. This software consists of six classes: Parser, Data, Space, Fitness, Postprocess and Optimizer (Figure 4). The optimizer, which is a common class for any optimization problem, can be either PSO or MVIE. The user, depending on the optimization problem, can change the rest five classes. A file containing an implementation for these five classes, aiming at solving a specific problem, is called module. In order to use *pow<sup>er</sup>*, a user has to provide the module name and an input file with parameters.

Initially, the class Parser reads, verifies and translates the variables from the input file, which contains specific keywords. Based on the parameters entered, the class Data fetches and loads necessary information (i.e. 3D structure coordinates). Next, the class Space defines the search space of the optimizer. Subsequently, the information defined by the previous classes is transferred to the Fitness and Optimizer classes. These classes are the core of the optimization process, and provide a solution that minimizes the fitness function, once a termination criterion has been reached (i.e. convergence). Finally, the Postprocess class sorts and clusters these solutions, and the final solution or cluster of representative solutions is

returned.



**Figure 4** -  $pow^{ef}$  workflow. Input is provided as a text file containing keywords with associated values. The blue boxes represent the classes, which are specific for the module designed to solve the optimization problem. The optimizer (PSO or MVIE) is represented as a green box.

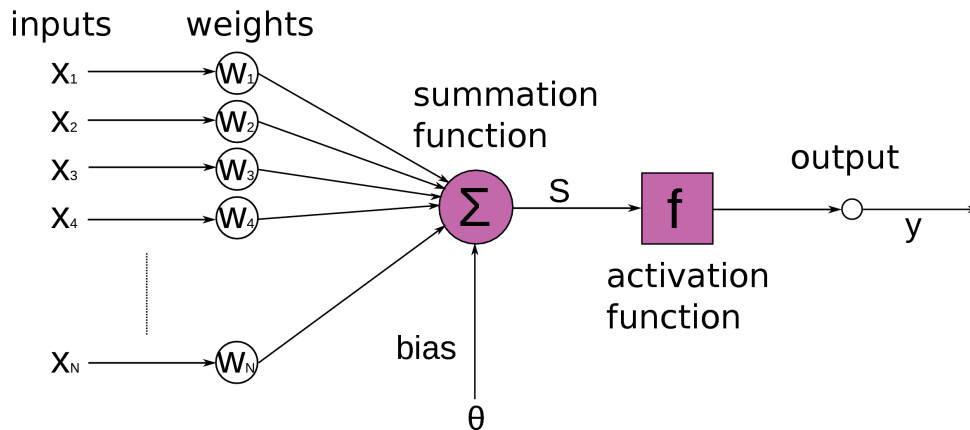
## 2.5. Artificial Neural Networks

### Artificial neuron

The idea of the ANNs (Artificial Neural Networks) is based on the biological neural networks. A typical biological neuron consists of the soma, which is its core, the dendrites through which it receives signals from neighboring neurons and the axon, which is the output of the neuron and its connection with the other neurons. In each dendrite there is a synapse. Synapses through chemical processes accelerate or decelerate the flow of the electrical charges to the neuron's soma. The learning and memory capacity of the brain is due to the ability of the synapses to change their conductivity. The electrical signals that enter the neuron's soma through the dendrites, are combined and if the result exceeds a threshold value the signal is propagated through the axon to other neurons.

Similarly, an artificial neuron is a computational model with components that correspond to the ones of a biological neuron. In particular, an artificial neuron receives some input signs  $x_1, x_2, \dots, x_n$  that are continuous variables, likewise the electrical pulses of the brain (Figure 5). Also, every input sign changes according a (positive or negative) weight  $w_i$ , like in the synapses of the biological neuron (Vlahavas et al., 2011). The soma of the artificial neuron consists of two parts: the "summation function" that sums, the affected by the weights, inputs into an S value, and the "activation function" which forms the final value of the output signal y, in correspondence to the S value and a threshold value. Sometimes, except the input

signals and their weights, the artificial neuron has a weight  $\theta$  (or  $w_0$ ), which is called bias. The only difference of the bias from the rest weights is that it affects constantly an input value of  $x_0 = 1$ . The bias is an external input that is added along with the other input signals/weights. It is used to determine the position of the activation function on the x-y axis. For example, in step functions (see below) the threshold value can be set to zero instead of some  $T$  value and this threshold is assigned to the weight  $w_0$  setting it to  $-T$ .



**Figure 5** - Components of the artificial neuron.

### Common activation functions

The output  $y$  of the step function is 1 if the value  $S$  that is calculated by the "summation function" is greater than a threshold  $T$ , otherwise the output is 0:

$$f(S) = \begin{cases} 0, & S > T \\ 1, & S \leq T \end{cases}$$

The output  $y$  of the sign function is negative (or positive) if the value  $S$  is smaller (or larger) than a threshold value:

$$f(S) = \begin{cases} 1, & S > T \\ 0, & S = T \\ -1, & S < T \end{cases}$$

A sigmoid function has the characteristic "S" shaped graph and a parameter  $\alpha$  to control the transition rate from small to large output values. In particular,  $\alpha$  is a transition rate adjusting factor between the two asymptotic values (slope parameter):

$$F(S) = \frac{1}{1 + e^{-S\alpha}}$$

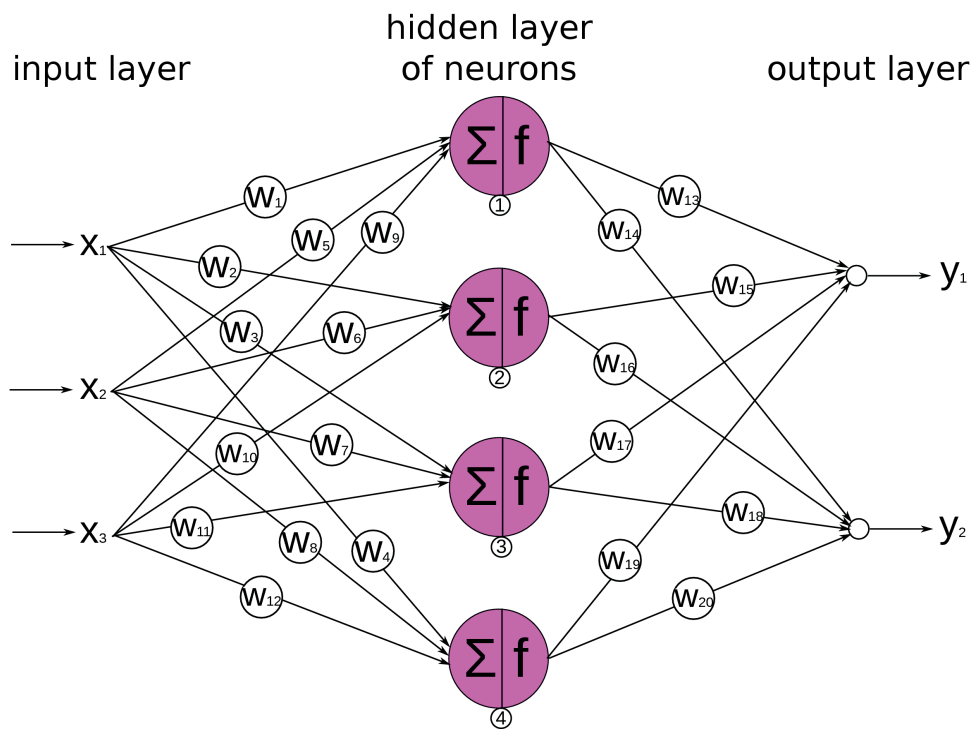


The importance of the sigmoid functions lies on the fact that they are continuous, real valued, differentiable and they have a non-negative first derivative which is "bell" shaped. This "bell" shaped derivative acts as a filter that suppress large input values, while it can provide a satisfactory output for low input values. Moreover, the output values are limited between 0 and 1 (or -1 and 1). In addition, sigmoid functions are non-linear, which is a necessary property for modeling non-linear phenomena. All these properties are very useful for artificial neural networks.

### Artificial neural networks

ANN (Artificial Neural Networks) are data processing systems build up by a number of artificial neurons that are organized into structures similar to those of the human brain. Usually, artificial neurons are organized into a series of layers (Figure 9). The first of these layer is called input layer and it is used to input data. The components of the input layer are not neurons because they do not perform any calculations (no input weights or activation function). This layer is followed by one layer (single-layer neural network) or more hidden layers (multi-layer neural network), while at the end there is the output layer. In Chapter 3, we used ANNs to predict the correct orientation of various DNA-binding proteins and the DNA, in order to interact.

Neurons in ANN may be fully or partially connected with each other. Fully connected neurons, are those that are connected with every neuron of the next layer. In any other case the neurons are partially connected. A feedforward neural network is an ANN where there are no connections between neurons of a layer and the neurons of the previous layer (neurons do not form a graph) (Figure 6). In the case that there are connections between neurons of the same layer or with the previous layer (neurons form a graph), the ANN is called recurrent neural network. In a recurrent neural network, the calculation of the various values occurs in two recurring steps. At the first step the values of the feedforward connections are calculated and at the second step the values for the feedback connections are calculated. Although in some cases recurrent neural networks are very useful, the majority of neural network applications use feedforward networks. The way which a neural network is structured (neuron number, neuron arrangement and neuron connectivity) is directly related to the learning algorithm used to train it.



**Figure 6** - Feedforward artificial neural network. Initial, neuron 1 receives three input signals  $x_1$ ,  $x_2$  and  $x_3$ , from the input layer. These signals are modified by the weights  $w_1$ ,  $w_5$  and  $w_9$  and are added by the "summation function" of neuron 1. Next, the sum is processed by the activation function of neuron 1 and the result is sent to the output layer. Likewise, the function of the rest neurons is similar. Therefore, the ANN receives the signals  $x_1$ ,  $x_2$  and  $x_3$ , as an input vector, and after internal processing from its neurons, it produces the output vector:  $y_1$  and  $y_2$ .

### Learning and recall

ANN perform two basic functions: learning and recall (Russell S.J. and Norvig P., 2009). Learning is the process of modifying the weights of the network so that given a specific input vector (training set) a specific output vector will be produced. This process is also called training. Recall is the process of calculating an output vector for a specific input vector (test set) and specific weight values.

There are three types of learning for an ANN, depending on the way that the weights are modified through training: supervised learning (Russell S.J. and Norvig P., 2009), graded learning and unsupervised learning (Jordan M.I. and Bishop C.M., 2004). In supervised learning, pairs of input and desired output vectors are given to the network. The network uses the initial weights and produces an output that differs from the desired output. This difference is the error, and based on that and a training algorithm the weights are adjusted accordingly. In graded learning, the output is marked as "good" or "bad" based on a numerical scale, and weights are adjusted

according to that scale. In unsupervised learning, the network's response depends on its ability to self-organize based only on the input vectors, since there are no corresponding output vectors. In practice, networks in this category are trying to clusterize the input data. For most ANN supervised learning is used.

There are many algorithms developed for supervised learning. For example, the Delta rule learning algorithm, in which the difference between the actual and the desired output is minimized through a least squares procedure (Russell and Ingrid, 2012). Next, in the back propagation algorithm the weight adjustment is based on the contribution of each weight to the total error (Goodfellow et al., 2016). Another algorithm is the competitive learning algorithm, in which the artificial neurons compete with each other, and only the one with the best response to a given input adjusts its weights (Rumelhart et al., 1986).

In ANN, like in other non-linear prediction methods, underfitting or overfitting may occur. An ANN that is not complex enough can fail to model successfully the training data, leading to incomplete learning (underfitting). On the contrary, a very complex ANN may model the training data and their potential noise too much, like memorizing them. In this case, the network gives a good prediction for the training data but produces completely wrong predictions for other input data (overfitting). Overfitting may also occur in multi-layer ANNs, even if the training data does not contain noise. The best way to limit the above is to use a sufficient number of training data.

The most common way to use training data is in training cycles, which also are called epochs. During each epoch, the network receives as input, one by one, all the training vectors. Next, it sums the changes in the weight values that occurs from each vector and adjusts the weights at the end of each epoch using the accumulated variation. This method is also known as batch learning. Otherwise, weight adjustment can be done after each one of the training vectors, and in this case, it is called incremental learning. Batch learning gives faster results but has higher memory requirements. There are also cases where a combination of the two above methods is used. Independently from which method is used, training ends when the network quality control criterion reaches a desired value. Usually, as a criterion is set the average error or the variation in the mean error of the training set. In both cases, it must be limited to a low value. If this is not possible, training may be terminated after a preselected number of epochs.

Another important issue related to the ANN training is the normalization of the input data, as well as the encoding of the input and output data. The normalization of the input data is mainly related to how these will be treated by the learning algorithm. The normalization can affect the speed and quality of the learning. The two most common normalizations of the input data, is to generate data with mean value 0 and standard deviation 1, or with range 2 and center value 0 (i.e., minimum -1 and maximum 1). Normalization takes place not only in the training data, but also in the test and validation data. However, the normalization parameters are strictly derived

from the training data and then they are used to normalize the test and validation data. Generally, normalization should be done with caution, as it may lead to loss of information that may lead to poor learning.

## Evaluation of ANNs

For the evaluation of a binary classification model, like an artificial neural network with two outputs, there are various measures. These measures use variables, such as the TP (true positive), which is an outcome where the model correctly predicts the positive class (correctly classified as positive). Similarly, a TN (true negative) is an outcome where the model correctly predicts the negative class (correctly classified as negative). A FP (false positive) is an outcome where the model incorrectly predicts the positive class (type I error, wrongly classified as positive). Finally, a FN (false negative) is an outcome where the model incorrectly predicts the negative class (type II error, wrongly classified as negative).

Two common statistical measures of the performance of a classification function are sensitivity and specificity. Sensitivity or TPR measures the proportion of actual positives that are correctly identified as such, while specificity or TNR measures the proportion of actual negatives that are correctly identified as such. TPR and TNR are calculated as the percentage of the correct predictions:

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

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

Other statistical measures are the precision and accuracy. Precision or PPV (positive predictive value) measures the statistical variability (description of random errors), while accuracy measures the statistical bias (systematic errors). In other words, given a set of data points, the set can be said to be precise if the values are close to each other, while the set can be said to be accurate if their average is close to the true value of the quantity being measured. The two concepts are independent of each other, so a particular set of data can be either accurate, or precise, or both, or neither. Mathematically, this can be stated as:

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

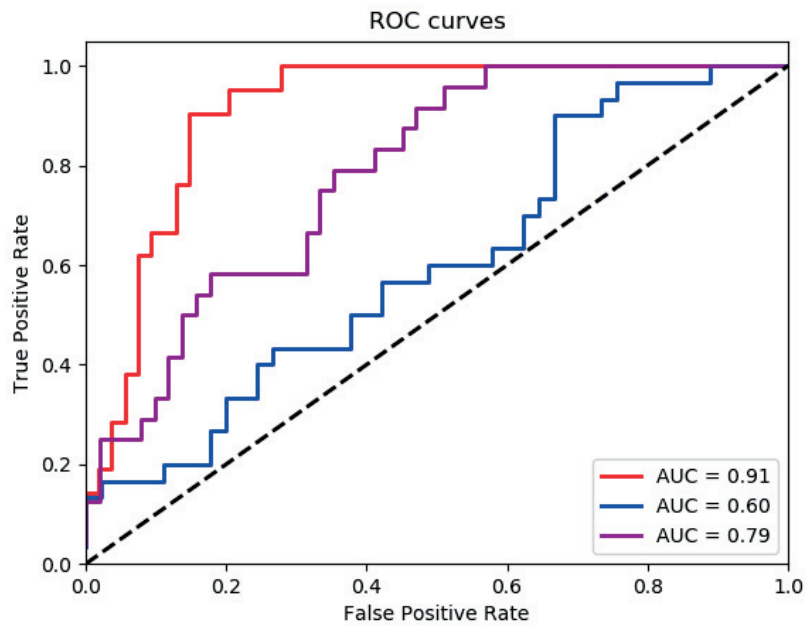
$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Although accuracy takes into account all four parameters (TP, TN, FP and FN) is not so useful when the two classes are of very different sizes. Classification

approaches, like ANNs, assume that the output of the training set is evenly distributed. When the training set for one class is much larger than the training set of the other class problems can arise from what is called class imbalanced classification. This has as a consequence the presence of high TP and very low TN values, or the reverse. A solution to this problem is to use an alternative measure, the Matthews Correlation Coefficient (MCC). The MCC is used in machine learning as a measure of the quality of binary classifiers. It is considered as balanced measure, since it takes into consideration TP, TN, FP and FN, and it can be used if the classes are of very different size. In general, the MCC is a correlation coefficient between the observed and predicted binary classifications. It returns a value between -1 and +1. A coefficient of +1 is an indicator of a perfect prediction, 0 a random prediction and -1 shows total disagreement between prediction and observation. The MCC can be calculated from the following formula:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

Another measure of the performance of a binary classification model is the Area Under a ROC Curve (AUC). In statistics, ROC curve is a graphical plot that illustrates the diagnostic ability of a classification model (Fawcett, 2006). The ROC curve is created by plotting the true positive rate (TPR - sensitivity) against the false positive rate (FPR=FP/(FP+TN)) at various threshold settings (Figure 7). Lowering the classification threshold classifies more items as positive, thus increasing both FP and TP. AUC shows the probability that the model ranks a random positive example higher than a random negative example. In particular, an area of 1 represents a perfect test, while an area of 0.5 represents a bad test.



**Figure 7** - Roc curves: plots of TPR VS FPR show three different ROC curves. The red curve with 0.91 AUC represents an excellent test, the purple curve with 0.79 AUC represents a relatively good test, and the blue curve with 0.60 AUC represents a fair test.

## Chapter 3      A machine learning approach to predict protein-DNA interactions

### 3.1. Introduction

Macromolecular assemblies are important for the biological functions of cells. However, because of their complexity, it is challenging to solve their structures by only using experimental methods. Therefore, by combining experimental with *in silico* methods, we can have a better insight into the structures and dynamics of macromolecular assemblies.

In particular, to study protein-protein interactions, the best approach is to combine algorithms with various experimental data to generate high-resolution models of protein complexes (Rodrigues and Bonvin, 2014). A widely used approach is the Integrative Modeling (IM). With IM we can predict the structure of a molecular complex by using the individual structures of its participants. This approach integrates experimental information with the conformational and interaction landscape of all possible orientations that the monomers of a complex can adopt. It also uses a scoring function to discriminate the highest affinity complexes from the rest (Gromiha et al., 2016). In addition, in order to limit the search space, restraints from experimental information can be used. Although IM has successfully predicted several large molecular assemblies from their monomers (i.e. CAPRI competition (Lensink and Wodak, 2013)), it is very difficult to predict the “near-native” model assemblies.

Protein-DNA interactions are fundamental for all living organisms. In nature, there are different structural motifs of proteins that can bind to the DNA. Some proteins can recognise specific DNA base pairs and others can recognise specific DNA shapes (see Chapter 1). Both, proteins and DNA change their conformation when they interact, in order to form a complementary interface and therefore a more stable complex. It is possible that predicting protein-DNA interactions is more difficult than predicting protein-protein interactions, since we have very few solved structures. However, the possible molecular assemblies and binding modes are more limited, since the DNA's structure has less degrees of freedom than a protein.

To study the protein-DNA interfaces, there are many experimental and computational techniques. Some experimental techniques include EMSA, which can provide absolute binding affinities between a TF and its binding DNA (Garner and Revzin, 1981; Fried and Crothers, 1981), and ChIP-seq, which can identify the binding sites of DNA-binding proteins (Gilmour and Lis, 1983). Nevertheless, *in silico* methods are an alternative to experimental techniques, in order to characterize protein-DNA interactions. These methods are either sequence-based or structure-based (Si et al., 2015), and some can predict whether a protein can bind to the DNA and others which are the protein binding residues. Structure-based approaches

appear to be more efficient than sequence-based approaches, since the sequence is less conserved than the structure. In particular, these approaches use the 3D structures of unbound proteins and the bound proteins in complex with the DNA, or high-quality homologous models. The structural information of the binding sites is converted to structural descriptors (i.e. electrostatic potentials, accessible surface area, etc.). Such methods include ANN, SVM, decision trees/random forests and Bayesian networks. These methods "learn" from a training set of the protein-DNA interface descriptors and attempt to predict the correct protein-DNA interactions from an unknown test set.

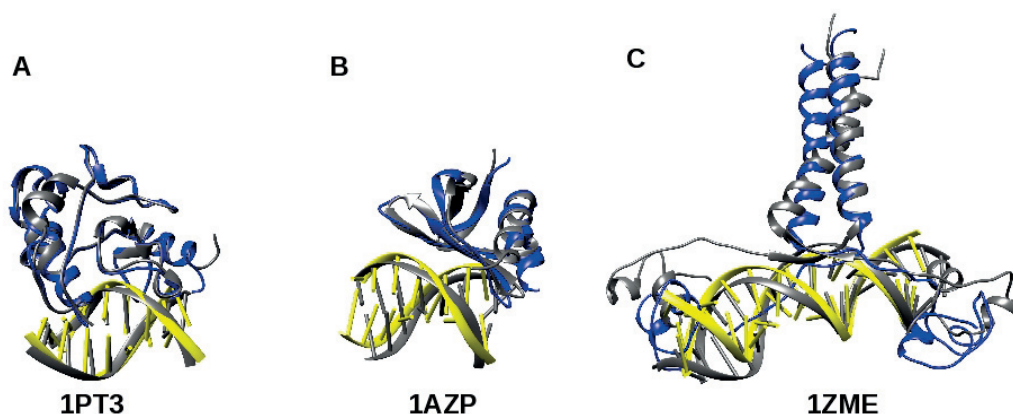
Here, in order to predict protein-DNA interfaces we used a structure-based computational approach, based on ANNs. In particular, to describe the protein-DNA interface of a set of protein-DNA complexes we used the distances, the electrostatic interactions energies and the van der Waals interactions energies between the residues on the interface. We tested our method on different protein-DNA complexes from a benchmark of bound and unbound proteins. Our goal is to predict the correct orientations of the DNA on a protein, and therefore, predict the interacting residues in both protein and the DNA.

## 3.2. Results

### A protein-DNA benchmark

In our approach, we used artificial neural networks to predict which protein/DNA residues participate in the interaction interface. Artificial neural networks seem to give better predictions when the interface descriptors are based on a benchmark of solved protein-DNA complexes (Si et al., 2015). Here, we will use a protein-DNA benchmark with 47 bound-unbound proteins proposed by van Dijk and Bovin in 2008. In this benchmark 13 complex are classified as easy cases, 22 as intermediate cases and 12 as difficult (Figure 8). This classification was done by calculating the iRMSD (RMSD at the protein-DNA interface). Specifically, residues belonging to the interface are identified as those having atoms within 5 Å intermolecular distance from one another. All the classes of protein-DNA complexes, except the zipper-type class, are included in this benchmark (Luscombe et al., 2000).





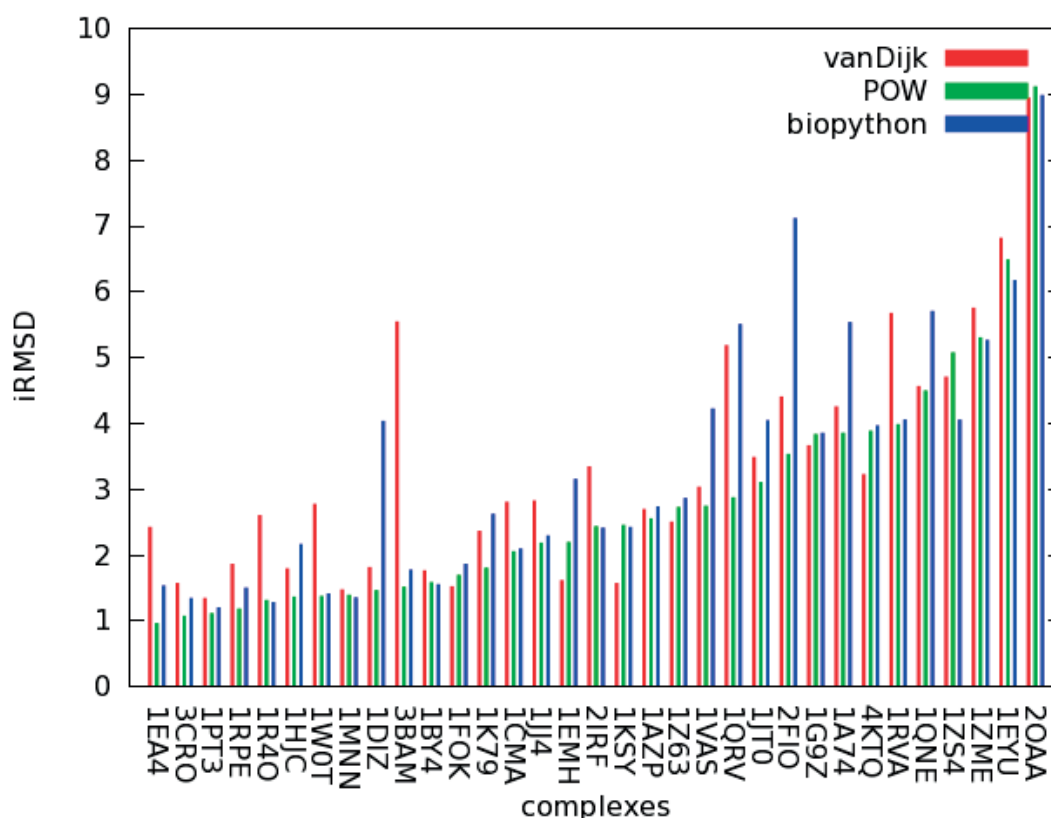
**Figure 8** - Illustration of test cases from the protein-DNA benchmark A. of “easy” ( $i\text{RMSD} < 2.0 \text{ \AA}$ ), B. “intermediate” ( $2.0 \text{ \AA} \leq i\text{RMSD} < 5.0 \text{ \AA}$ ) and C. “difficult” ( $i\text{RMSD} \geq 5.0 \text{ \AA}$ ). The bound form of the complex is shown in gray, the unbound protein in blue and the B-DNA model in yellow.

Initially, we processed and analysed the PDB files in van Dijk's benchmark. In particular, we kept one occupancy for multiple amino acids. In addition, we removed flanking N- or C- terminal parts that were present in the bound protein but were missing from the unbound or the reverse. Next, in case of missing loops, we added them with the Rosetta loop modeling application. For the DNA molecules, we removed unpaired terminal nucleic acids and we built the unbound DNA with the 3DNA software. In general, we ensure that we have the same type and number of residues (C $\alpha$  and Pho) for the bound and unbound states, as well as the same stoichiometry. Complexes that were classified as difficult cases and some others with very high  $i\text{RMSD}$  were excluded. Finally, we calculated through different approaches the  $i\text{RMSD}$  of the 33 remaining complexes. In the first approach, we used the Biopython library (Bio.PDB module), in order to superimpose our molecules and calculate the  $i\text{RMSD}$ . In the second approach, we used  $pow^{er}$ , to optimize the position of the unbound DNA on the unbound protein, using as a reference the bound protein-DNA complex. The best position of the DNA on the protein is considered the one that minimizes the  $i\text{RMSD}$ . In both cases, we identified as interface residues the C $\alpha$  and Pho that are within a distance of 8  $\text{\AA}$  from each other. The calculated  $i\text{RMSDs}$  are presented in the following table (Table 1) and in Figure 9:

Complex	$i\text{RMSD}$ - van Dijk and Bovin	$i\text{RMSD}$ - $pow^{er}$	$i\text{RMSD}$ - Biopython
1EA4	2.43	0.97	1.54
3CRO	1.58	1.08	1.35
1PT3	1.35	1.12	1.21
1RPE	1.87	1.19	1.51

<b>1R4O</b>	2.61	1.32	1.29
<b>1HJC</b>	1.80	1.37	2.17
<b>1W0T</b>	2.78	1.38	1.42
<b>1MNN</b>	1.48	1.40	1.36
<b>1DIZ</b>	1.82	1.47	4.04
<b>3BAM</b>	5.55	1.52	1.78
<b>1BY4</b>	1.77	1.59	1.56
<b>1FOK</b>	1.53	1.70	1.87
<b>1K79</b>	2.37	1.81	2.63
<b>1CMA</b>	2.81	2.06	2.10
<b>1JJ4</b>	2.83	2.19	2.30
<b>1EMH</b>	1.62	2.20	3.16
<b>1KSY</b>	1.58	2.46	2.43
<b>1AZP</b>	2.70	2.56	2.74
<b>2IRF</b>	3.35	2.44	2.42
<b>1Z63</b>	2.51	2.73	2.87
<b>1VAS</b>	3.04	2.75	4.23
<b>1QRV</b>	5.19	2.88	5.51
<b>1JT0</b>	3.49	3.11	4.05
<b>2FIO</b>	4.41	3.54	7.12
<b>1G9Z</b>	3.67	3.84	3.86
<b>1A74</b>	4.26	3.86	5.54
<b>4KTQ</b>	3.23	3.89	3.97
<b>1RVA</b>	5.68	3.99	4.06
<b>1QNE</b>	4.57	4.50	5.71
<b>1ZS4</b>	4.71	5.08	4.06
<b>1ZME</b>	5.76	5.31	5.27
<b>1EYU</b>	6.82	6.49	6.18
<b>2OAA</b>	8.95	9.12	8.99

**Table 1** - iRMSDs. The pdb ids of the protein-DNA complexes in the Benchmark and their iRMSDs through different calculations. The gray ones were excluded from the analysis, since their calculated iRMSDs are higher than 2.8 Å. We assume that the protein and the DNA don't interact if their iRMSD is higher than 3 Å.



**Figure 9** - Histogram of the iRMSDs for each complex, for each of the three methods.

It seems that the three methods that calculate the iRMSD tend to agree, since the average iRMSD is 3.34 Å for van Dijk, 2.82 Å for POW and 3.34 Å for biopython. However, the complexes of 3BAM and 2FIO appear to have differences in their iRMSD calculations. In particular, for 3BAM (restriction endonuclease BAMHI), van Dijk's method calculate a higher iRMSD from the other two methods (~5.5 Å vs. ~1.5 Å). This could be due the fact that van Dijk's method takes into account all atoms, while the other two methods consider only the backbone atoms (Ca and Pho). Since restriction endonuclease BAMHI has many flexible loops, it has higher iRMSD when all atoms are considered. Also, for 2FIO (phage PHI29 transcription regulator P4) there is a heterogeneity between the iRMSDs of the three methods. This could be due the fact that phage PHI29 transcription regulator P4 interacts with a flexible and very long DNA of 39 base pairs (the longest in this benchmark). For such a long DNA it is difficult to superimpose a B-DNA model and this could have an effect on the iRMSD calculation.

## Making the ANN input matrices

The goal is to predict which orientations of the DNA on its binding protein are closer to the native-like complex. For this reason, we used ANNs that were trained to distinguish between the protein-DNA complexes that interact ( $i\text{RMSD} \leq 3 \text{ \AA}$ ) and those that do not ( $i\text{RMSD} > 3 \text{ \AA}$ ), for different data sets in our benchmark. The benchmark includes two versions of DNA-binding proteins: bound and unbound state, as well as the bound DNA. Initially, we used POW to place the unbound DNA on different positions on the unbound protein, for all the complexes in the benchmark. We used as a reference structure the protein-DNA complex and we calculated the various  $i\text{RMSDs}$ . For  $i\text{RMSDs}$  that are less or equal to  $3 \text{ \AA}$ , we considered that the protein and the DNA interact in the proper localization. For  $i\text{RMSDs}$  that are higher than  $3 \text{ \AA}$ , we considered that the protein and the DNA don't interact. POW produced multiple positions for every complex in the benchmark. Therefore, complexes for which their calculated  $i\text{RMSD}$  was higher than  $2.8 \text{ \AA}$  were excluded from the data set (11 complexes). For the remaining 17 complexes, the positions of the DNA on the protein, that included clashes or knots, were removed with Chimera software. The remaining complexes were parametrized with AMBER force field and minimized with NAMD software for 2000 steps. The minimized pdb files and their force field parameter files (prmtop files) were used to extract information and make the input matrices for the ANN. In particular, from the parameter files we extracted the atom charges and the Lennard Jones terms for all possible atom type interactions. From the minimized pdb files, we calculated the distances between the atoms of the DNA and the atoms of the protein. Based on the atom distances, we calculated the distance between all the amino acid-nucleic acid pairs as the average distance between their atoms. Due to these distances, we defined the 15 nucleic acids and the 20 amino acids that consist the interacting interface, and therefore to be used to build the neural network matrices. Also, based on the atom charges and the atom distances, we calculated the energy of the electrostatic interactions. In addition, based on the Lennard Jones terms for all possible atom type interactions and the atom distances, we calculated the energy of the van der Waals interactions. Finally, for 18 protein-DNA complexes, we made matrices of  $15 \times 20 \times 3$  elements, with 15 being the interface nucleic acids, 20 being the interface amino acids and 3 the descriptors of the interactions: distances, electrostatic interactions energies and van der Waals interactions energies.

## Artificial neural network

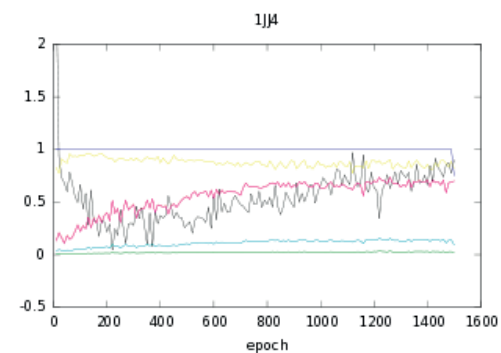
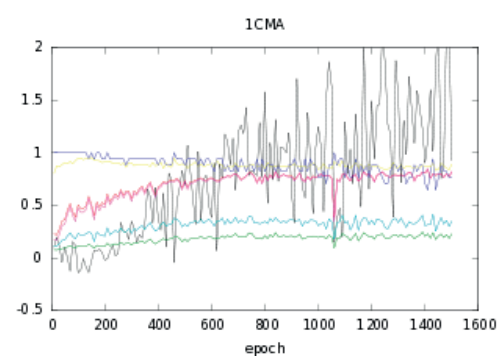
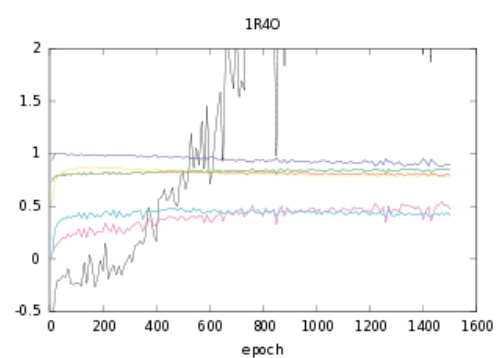
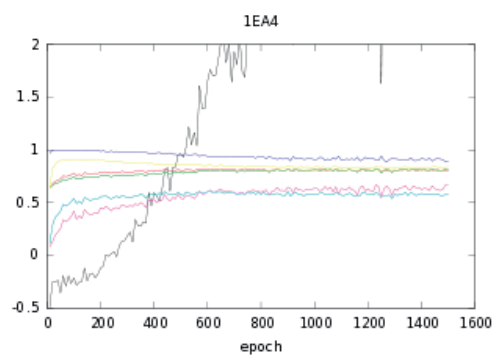
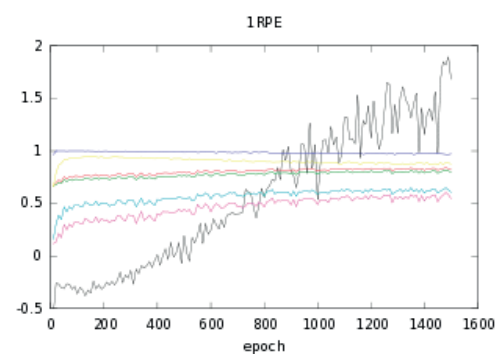
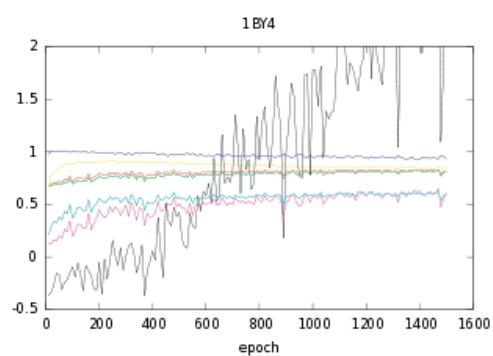
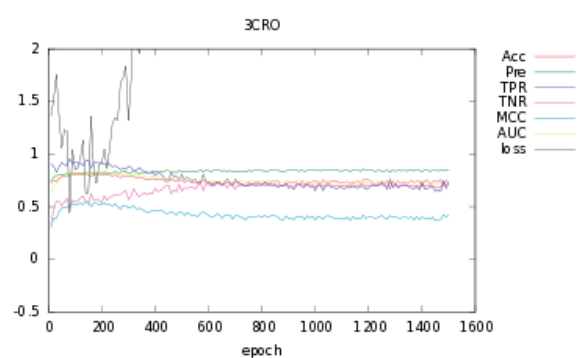
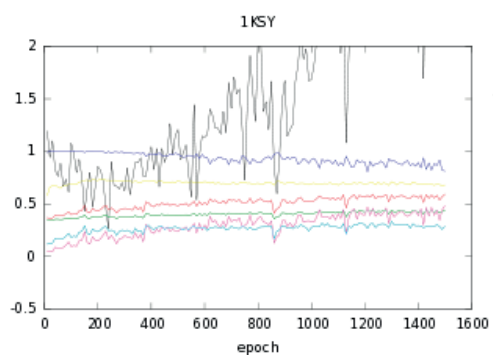
The artificial neural network was implemented with Python's tensorflow. The ANN has 900 ( $15 \times 20 \times 3$ ) input neurons, one hidden layer of neurons and two output neurons. Its architecture is feedforward and all neurons are fully connected. The

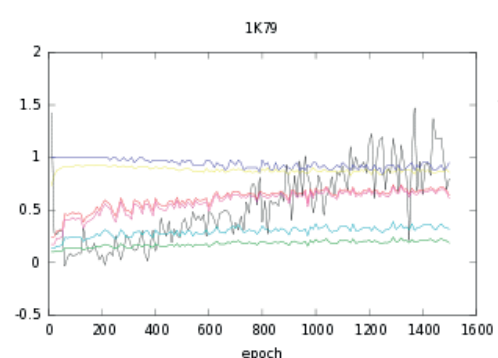
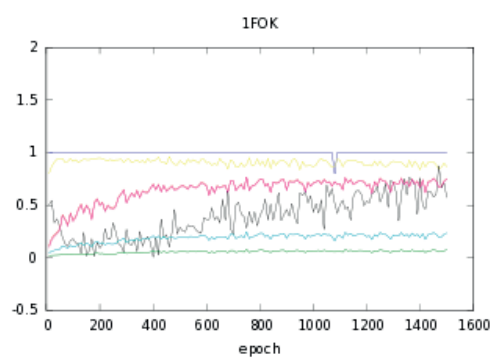
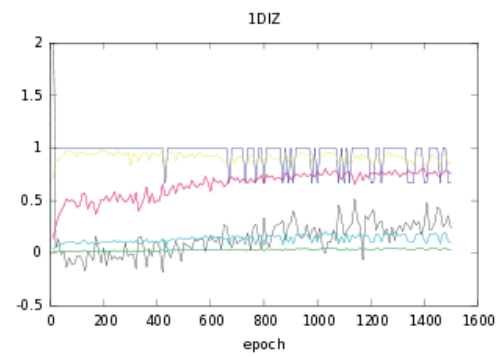
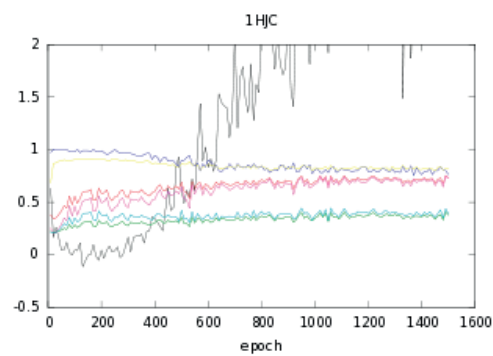
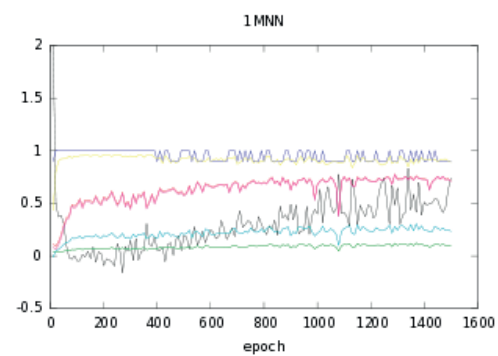
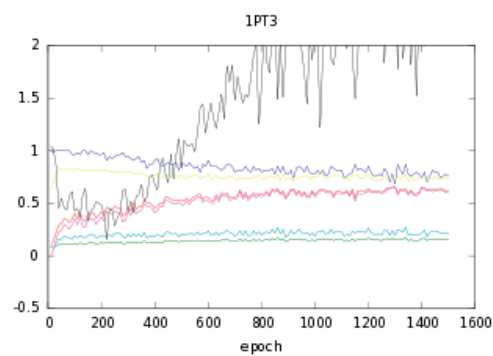
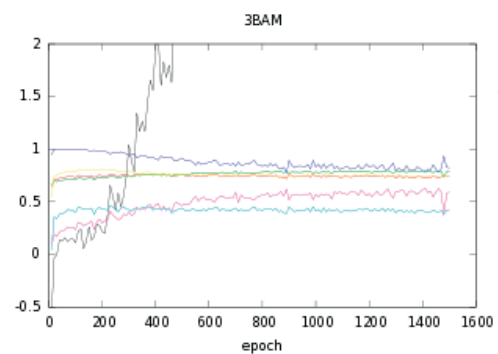
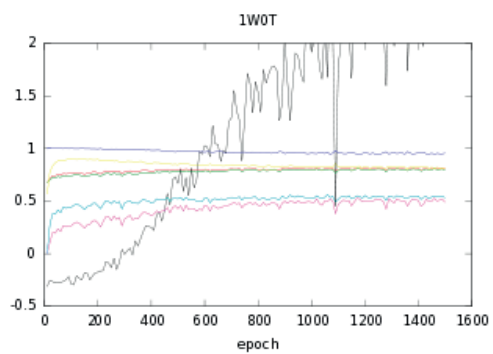
activation function that is used is the sigmoid function, while the calculated loss is the mean sigmoid cross entropy. In addition, the learning rate is 0.01, the batch size is 2048 and the number of epochs (training cycles) is 1500.

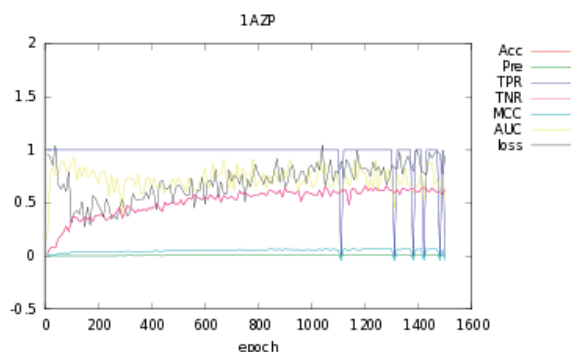
The matrices, that were described before, were used as inputs for the neural network and the output was defined as '0' in the case that there is no interaction ( $i\text{RMSD} > 3 \text{ \AA}$ ) and as '1' in the case that the protein and the DNA interact ( $i\text{RMSD} \leq 3 \text{ \AA}$ ). In average, we have 12545 cases of '0' and 17548 cases of '1'. This type of training is supervised, since we provide to the neural network the desired output in order to adapt its weights accordingly. Because we have only 18 protein-DNA complexes, we decided each time to train our neural network with 17 complexes and keep 1 complex as a test set. We did that for all the 18 training-test sets and average our results.

### Training and testing the ANNs

Subsequently, we trained and tested our neural network for all our sets. Only one set didn't manage to give an output (pdb 1EHM). From the 17 remaining sets, we calculated various metrics to evaluate the performance of our predictions. In particular, we calculated the sensitivity or TPR (true positive rate), the specificity or TNR (true negative rate), the accuracy, precision, the MCC (Matthews correlation coefficient), the AUC (the Area Under a ROC Curve) and the loss. Here, as TP cases are considered the complexes that their DNA and their protein interact and that are predicted correctly to interact (output '1'). As TN cases are considered the complexes that their DNA and their protein don't interact, and that are predicted correctly to not interact (output '0'). As FP cases are considered the complexes that their DNA and their protein don't interact, but are predicted incorrectly to interact. And as FN cases are considered the complexes that their DNA and their protein interact, but are predicted incorrectly to not interact. In the following figures, we plotted the values of the various metrics for each of the 17 sets (Figure 10):







**Figure 10** - ANN results. Plots of TPR, TNR, accuracy, precision, MCC, AUC and loss VS. epochs.

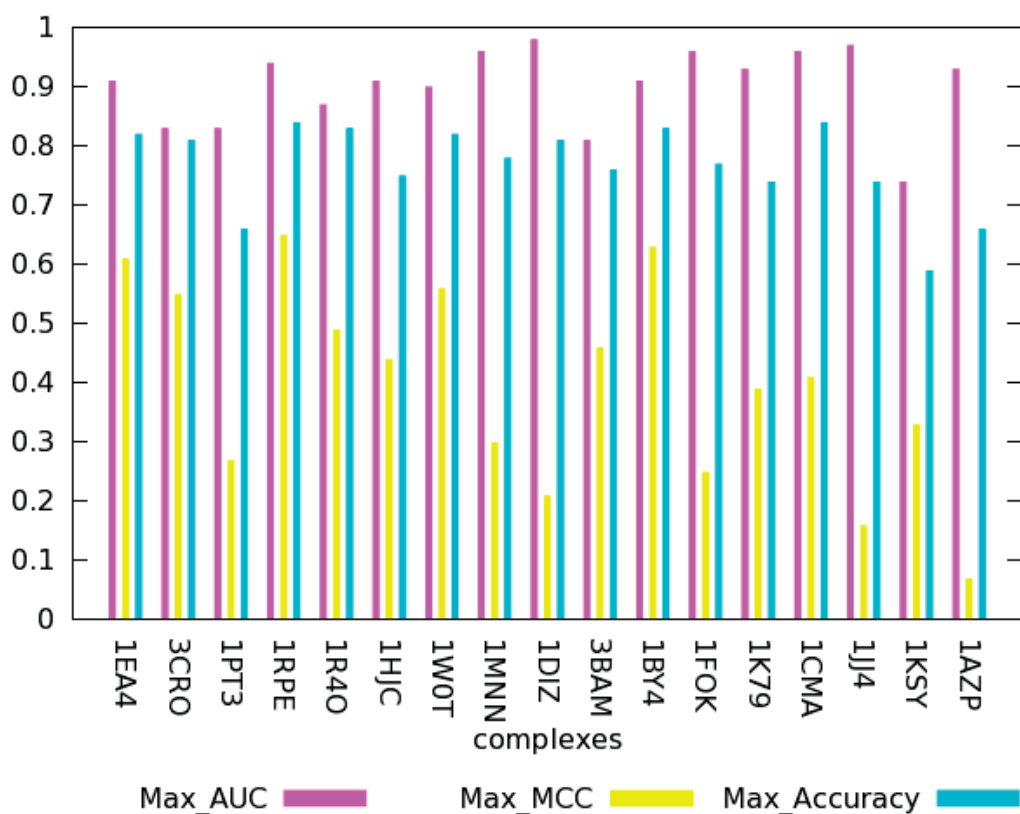
The above plots indicate that, for all sets, the learning processes finished at the first epochs (less than 1000), before the neural networks start to over-fit from the learning data sets. The AUC is high in most of the cases, shows that the predictions are quite accurate. The TPR is also very high and it is usually slightly higher than the TNR. This shows that in our data sets there are more positive than negative examples. The precision is low, while the recall is high in almost all of the cases. This could be an indication that the neural networks give many false positive predictions, but still many true positives are recovered. Specificity and accuracy are average, likely because there are many negative examples compared to positive examples in the training set, and therefore TN dominate. More analytically, we present in the following table the maximum AUC, the maximum MCC and the maximum accuracy (Table 2 and Figure 11):

	Max AUC	Max MCC	Max Accuracy
<b>1EA4</b>	0.91	0.61	0.82
<b>3CRO</b>	0.83	0.55	0.81
<b>1PT3</b>	0.83	0.27	0.66
<b>1RPE</b>	0.94	0.65	0.84
<b>1R4O</b>	0.87	0.49	0.83
<b>1HJC</b>	0.91	0.44	0.75
<b>1W0T</b>	0.90	0.56	0.82
<b>1MNN</b>	0.96	0.30	0.78
<b>1DIZ</b>	0.98	0.21	0.81
<b>3BAM</b>	0.81	0.46	0.76
<b>1BY4</b>	0.91	0.63	0.83
<b>1FOK</b>	0.96	0.25	0.77



<b>1K79</b>	0.93	0.39	0.74
<b>1CMA</b>	0.96	0.41	0.84
<b>1JJ4</b>	0.97	0.16	0.74
<b>1KSY</b>	0.74	0.33	0.59
<b>1AZP</b>	0.93	0.07	0.66
<b>Average</b>	<b>0.90</b>	<b>0.40</b>	<b>0.77</b>

**Table 2** - Maximum AUC, MCC and accuracy.



**Figure 11** - Histogram of maximum AUC, MCC and accuracy

Since we had 18 complexes and each time we trained our neural network with 17 complexes and test it with the remaining complex, we averaged their results. For one of our complex we didn't get results, therefore we averaged the results for the 17 remaining cases. The average maximum average AUC is very high 0.90, which shows that the predictions are accurate and our ANNs can distinguish between the protein-DNA complexes that interact and those that do not. Also, the maximum average MCC is 0.4, which indicates a moderate positive correlation of the predictions with the actual data. In addition, the maximum average accuracy is 0.77, which is quite high. Moreover, there seems to be a negative correlation between the iRMSD calculated by POW and the maximum MCC and maximum accuracy (ccs are

-0.62 and -0.57, respectively). This correlation is relatively strong and it shows that for lower iRMSD we can have higher MCC and accuracy for our predictions. For example, for the 1RPE complex that we have a low iRMSD of 1.19 Å, the maximum MCC and the maximum accuracy are 0.65 and 0.84, respectively. Both of them are higher than the average calculations for all the complexes. On the contrary, for the 1AZP complex that we have a high iRMSD of 2.56 Å, the maximum MCC and the maximum accuracy are 0.07 and 0.66, respectively. Both of them are lower than the average calculations for all the complexes. Finally, it seems that there is no linear relationship between the iRMSD and the maximum AUC (cc is 0.02).

### 3.3. Conclusions and perspectives

Protein-DNA complexes are important for the biological functions of a cell, but there are not many complexes solved experimentally. In addition, it is difficult to predict computationally these complexes, since proteins and DNA change their conformation when they form a complex. The best approach is to combine computational and experimental approaches, through IM, in order to make atomistic models of protein-DNA complexes.

Here, we used an IM approach to predict the structure of a set of protein-DNA complex and their interacting interface. Specifically, we used experimentally solved DNA-binding proteins (unbound state) and we built the 3D models of the corresponding DNA binding sequences. Next, we produced all possible orientations that these monomers can adopt. With the experimentally solved structures of these protein-DNA complexes (bound state), we trained ANNs to distinguish the interacting from the non-interacting orientations. In particular, we used these ANNs to predict the protein-DNA interfaces of different complexes, with high accuracy (0.77) and AUC (0.9). These ANNs, can be used as a scoring function to discriminate the “best” complexes from the rest and therefore, given a DNA-binding protein and a DNA model, they can predict a new protein-DNA complex and its interacting interface.

In our approach, there are some limitations, but also some improvements can be made. In particular, although the DNA has limited degrees of freedom, it is not the same for the DNA-binding domains of the proteins. Moreover, since we have many different types of DNA-binding domains, it is difficult to predict how they will interact with the DNA. Furthermore, some proteins can recognise their DNA target by its shape and other bind to a specific set of DNA sequences. A solution to that would be to divide proteins into groups with similar properties and train separately an ANN. This is not doable for the moment, since we have a limited data set of protein-DNA complexes. Further extension would be useful to include more interface descriptors, such as the H-bonds between the binding protein and the DNA, and the solvent accessibility surface area (SASA). H-bonds in the binding interface can be a useful descriptor in the cases that the binding is specific. In addition, the SASA can be used

to model the presence of water molecules and therefore the formation of water-mediated H-bonds in the interface. These new descriptors could improve the robustness and accuracy of the interface predictions.

## 4. A computational method to predict the DNA specificity of zinc finger proteins

(Kalantzi A.S., Isakova A., Deplancke B., and Dal Peraro M. A computational method to predict the DNA specificity of zinc finger proteins – To be submitted)

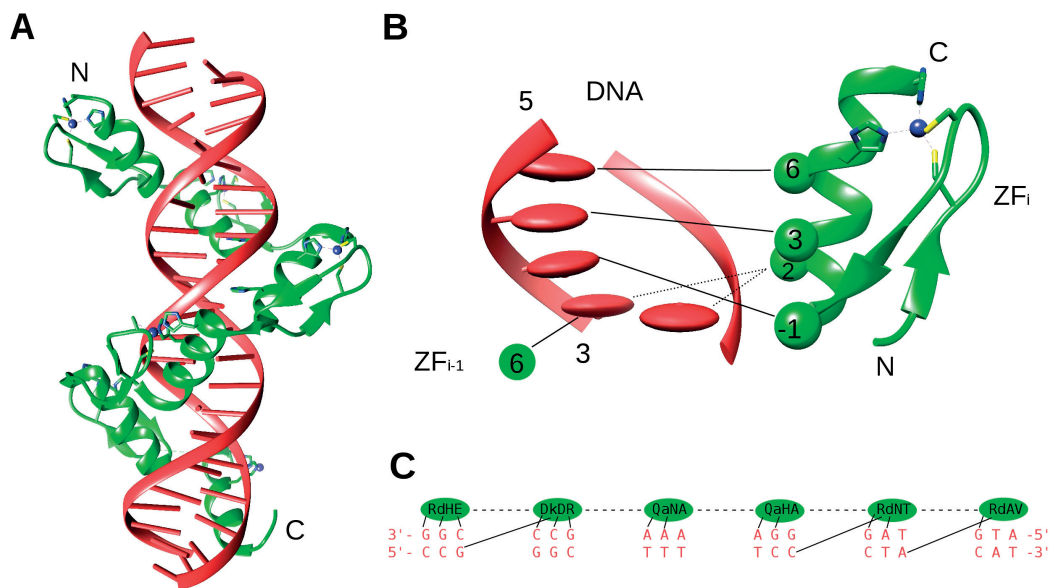
### 4.1. Introduction

#### C2H2-ZF proteins

Gene expression is regulated by a large number of TFs, which contain at least one DNA-binding domain, which can bind with high affinity to a specific set of DNA sequences close to the genes that they regulate. A DNA-binding domain that is mostly common in higher vertebrates is the classical C2H2-Zinc Finger domain: a protein structural motif that consists of an  $\alpha$ -helix and an antiparallel  $\beta$ -sheet, while two Cys and two His coordinate a zinc ion. C2H2 Zinc Finger Proteins (ZFPs) have a variety of functions such as binding RNA and mediating protein-protein interactions, but more often they participate in transcription regulation via sequence-specific DNA-binding.

The first C2H2-ZF domains were discovered in frog eggs, which contain the transcription factor TFIIIA (Miller et al., 1985). A  $\beta\beta\alpha$  fold, featuring two cysteine and two histidine residues forming a Cys-X 2-5-Cys-X 12-15-His's-X 3-5-His motif that coordinates tetrahedrally a zinc metal ion (Figure 12), characterizes the canonical C2H2-ZF domain. ZF domains connect to each other by a small linker, usually 7 amino acids in length from the last His of a ZF domain to the first Cys of the next ZF domain. This tandem of repeats can have up to 40 ZF domains (Stubbs et al., 2011). ZF domains tandem binds to the major groove of the DNA using the  $\alpha$ -helix of each domain. Each  $\alpha$ -helix has a DNA-binding structural interface that can recognize a target sequence of ~3-4 bps, with high affinity and specificity. The amino acids occupying key positions -1, 2, 3 and 6 (Figure 12) (numbered with respect to the start of the  $\alpha$ -helix) form hydrogen bonds, water-mediated H-bonds and van der Waals interactions with the nucleotides of the DNA target sequence. There is a variety of ways in which a ZF domain can bind to its DNA sequence (Wolfe et al., 2000), but the "canonically accepted" binding mode involves the amino acids in positions -1, 3 and 6, bound to a 3 bps nucleotide in the 3' to 5' strand direction. Sometimes amino acids in position 2 interact with a nucleotide in the opposite strand (Figure 12). However, this canonical mode of binding is not always respected. For example, ZFP568 deviates from the usual 3bps recognition by a ZF domain (Patel et al., 2018). In

particular, it includes ZFs that contact 2, 3, or 4 bases and recognise nucleotides on the opposite DNA strand. In general, in TFs characterized by long ZF tandem repeats not all ZF domains participate in DNA binding. In most cases, although there are many ZF domains in a tandem, a short DNA-target is recognized (Patel et al., 2016), while in some other cases two different DNA-targets can be recognized by separate zinc finger clusters (Han et al., 2016).



**Figure 12** - A.  $\beta\beta\alpha$  C2H2 ZF domains bound to DNA. B. Amino acids in positions -1, 3 and 6 (see inset) can bind serially 3 bps in the 3'-to-5' direction. The amino acid in position 2 can bind a nucleotide in the opposite strand. C. H-bonds between the binding amino acids and the DNA target. Red: DNA, Green: ZN domain, Blue: Zinc ion.

### KRAB C2H2-ZF proteins

From the ~800 different C2H2-ZFPs present in the human genome one third contains at least one KRAB (Krüppel associated box) domain (Urrutia, 2003). Bellefroid et al. first discovered KRAB-ZFPs in 1991 and they constitute the largest family of transcriptional regulators encoded by higher vertebrates (tetrapods) (Bellefroid et al., 1991). KRAB-ZFPs genes are one of the most recent and fastest growing gene families in primates and this expansion is hypothesized to enable primates to respond to newly emerged retrotransposons (Jacobs et al., 2014). KRAB-ZFPs typically contain one or two KRAB domains, which can interact with their

universal cofactor KAP1 (KRAB associated protein-1; also called TRIM28), and a tandem repeat of C2H2 zinc fingers motifs that enable sequence-specific DNA-binding. The functions of the currently known KRAB-ZFPs family members include the regulation of crucial physiological and pathological processes as development, differentiation, metabolism, apoptosis and cancer (Lupo et al., 2013). In addition, a big majority of human KRAB-ZFPs regulate transposable elements (Imbeault et al., 2017).

The KRAB domain is around 75 amino acids long and is found in the N-terminal part of about one third of eukaryotic ZNF proteins. Two types of KRAB domains have been observed: A and B, while other KRAB domain types might be found in the future. There has not been any solution of a KRAB domain structure, but it has been predicted to fold into two amphipathic  $\alpha$ -helices (Bellefroid et al., 1991). The KRAB domain functions as a transcriptional repressor by interacting with the KAP1 protein, which acts as a scaffold for chromatin-modifying complexes. A sequence of 45 amino acids in the KRAB A subdomain has been shown to be necessary and sufficient for transcriptional repression. The B box does not repress by itself but does potentiate the repression exerted by the KRAB A subdomain (Margolin et al., 1994; Witzgall et al., 1994). In addition, some KRAB-ZFPs contain another conserved motif called SCAN, with at least 87 amino acids in length. The SCAN domain is not associated with transcriptional regulation but instead allows homo- and hetero-dimerization with other SCAN- containing zinc-finger proteins (Collins & Sander, 2005). Depending on the type of the KRAB domain and the presence or not of SCAN domain, the KRAB-ZNF proteins have been divided into six subfamilies, the A + B, the A + b, the A only, the A + A + B, the SCAN + A and the A + B + SCAN + A (Looman et al., 2002).

There are 343 human KRAB-ZFP sequences in Uniprot, of which the 3/4 has a linker of length of 4 to 12 amino acids between its ZN domains (a canonical linker is 7 amino acids long). In the rest 1/4 proteins, the linker's lengths can vary from 13 to more than 90 amino acids. Also the number of ZF domains in each protein can vary from 2 to 37, with 96% being canonical (C-X<sub>2-5</sub>-C-X<sub>12-15</sub>-H-X<sub>3-5</sub>-H), 2.9% being degenerate (one or two Cys or His are replaced by other amino acids), 0.8% being atypical (the spacing between Cys or His residues is slightly altered), and 0.2% being half (truncated ZF).

### Protein-DNA recognition codes

The recognition of specific DNA sequences by ZFPs plays a fundamental role in gene regulation (Tan et al., 2003), determining which ZF domains can bind to the DNA remains a challenge. Suzuki did a first description of the stereo-chemical rules that apply to the protein-DNA recognition in 1994 (Suzuki et al., 1994). More recently, *in silico* recognition codes between DNA and ZFPs had been proposed. Jiajian et al. (Jiajian et al., 2008) presented a method that estimates the DNA binding specificities

of C2H2-ZFPs based on an artificial neural network, but this was limited by the lack of sufficient training data. Similarly, Molparia et al. (Molparia et al., 2010) proposed an artificial neural network approach to predict DNA-binding specificity of ZFPs. In this approach, the binding sites of the classical C2H2 fingers were identified for a known DNA sequence. In particular, this recognition code predicts the protein recognition helices that bind to a known DNA sequence. Therefore, it does not give an overall view of all the possible ZF/DNA combinations, since diverse sets of ZF can bind common targets (Persikov et al., 2014).

More recently, three experimental-based recognition codes were proposed by Gupta et al. (Gupta et al., 2014), Persikov et al. (Persikov et al., 2014) and Najafabadi et al. (Najafabadi et al., 2015). In the first approach, Gupta et al. constructed a random forest predictive code based on the DNA-binding specificities of natural and artificial two-finger modules, using data from a bacterial one-hybrid analysis. Their code is based on limited number of (678) ZF domains, leading to only partially correct predictions. Persikov et al. proposed a recognition code that was based on large synthetic protein libraries that were screened in order to select binding C2H2-ZF domains for each possible 3 bps nucleotide target. This resulted to a selection of unique domain–DNA interactions that are used to predict the DNA-binding specificity of other ZF domains. Finally, Najafabadi et al. presented a recognition code that was derived by analyzing DNA-binding data for 8138 natural C2H2-ZF domains through the combination of data from a modified bacterial one-hybrid system with protein-binding micro-array and chromatin immunoprecipitation analysis. In all cases, their results indicate that the majority of ZFPs show widespread binding to regulatory regions by targeting a diverse range of genes and pathways. In addition, these data illustrate the complex binding landscape of the ZF domains and provide insights into their DNA-binding specificities. All these recognition codes tend to agree in the usual amino acid/nucleotide pairing prediction, for example, that Arg residues bind to G, but there is no absolute agreement in the non-canonical pairing. These inaccuracies could be due to the fact that these methods do not consider the role of flanking ZF domain for determining the DNA specificity, neither the role of long linkers between ZF domains. A structural approach by Garton et al. (Garton et al., 2015), that eventually is integrated in Najafabadi's recognition code, gave later insights into the influence of neighboring ZF domains, but this is limited due the small number of ZFP-DNA complexes that have solved structures. In a recent study (Barazandeh et al., 2018), compared KRAB-ZFPs binding motifs from ChIP-seq (Schmitges et al. 2016) and from ChIP-exo (Imbeault et al., 2017). These motifs were enriched by motifs that were predicted by Najafabadi's et al. recognition code (Najafabadi et al., 2015b). In general, there is an agreement between these motifs, but there are still discrepancies. Finally, the recognition codes take into account only amino acids that interact with the DNA, while also non DNA-base contacting amino acids might play a role in the binding mode (Najafabadi et al., 2017). As it seems from the most recent literature (Imbeault et al., 2017; Patel et al., 2018) the binding landscape appears

much more complicated than these codes can capture.

The binding recognition between ZFPs and DNA is still not fully decoded. Our aims are to determine how the C2H2-ZF tandem recognizes 3-4 bps, define the general recognition binding rules between ZFPs and their DNA targets and present a method to analyse and predict ZFP binding specificity. To this goal, we propose a structure-based method that can predict which ZF domains of a ZFP can bind to a given DNA sequence, and a recognition code that predicts the most probable DNA target sequence for a ZFP.

## 4.2. Results

### Data sets

We used 3614 protein-DNA complexes (9014 protein chains) deposited in PDB in September 2016. From these protein-DNA complexes, we extracted a non-redundant set (homology <95%) using the CD-HIT server (Huang et al., 2010), in order to remove duplicate chains and high homologous proteins. The final dataset consists of 969 complexes (1124 unique protein chains, PDB ids are available in Supplementary Material). For structures solved by NMR only the first model of the ensemble was considered. The analysis of the interactions of the structures of this dataset was used to develop a scoring method for ZFP-DNA specificity.

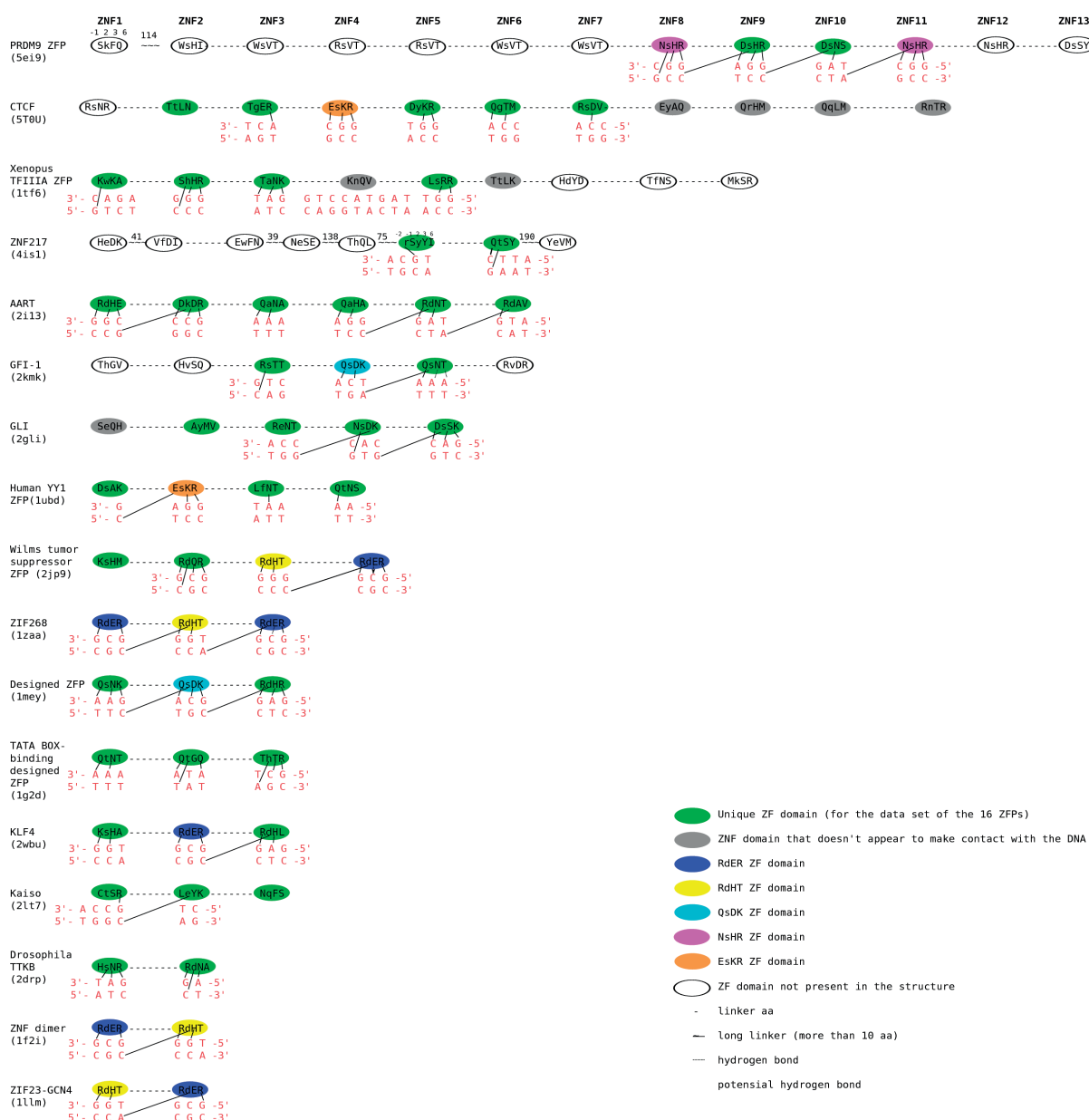
Within the protein-DNA complexes dataset there were C2H2-ZFPs structures. From this dataset, 17 complexes are unique and their ZF domains interact with the major groove of the DNA (Table 3). Of these 17 ZFPs there are 86 C2H2-ZF domains, of which 76 are unique (Figure 13). The dataset of these 17 complexes was used to analyze the binding properties of the ZF domains to DNA, and it was further used as a benchmark dataset to validate our predictions of ZF binding specificity.

ZF protein (pdb id)	Num of ZF domains	ZF domains that bind the DNA target	Consensus DNA sequence	DNA sequence from pdb structure
CTCF (5t0u)	11	ZF3, ZF4, ZF5, ZF6, ZF7	NCANNAGRNGCG RSY	CCAGCAGGGGGC GCT
Xenopus TFIIIA ZFP (1tf6)	9	ZF1, ZF2, ZF3	GGANGGNNGN	GGATGGGAGA
ZNF217 (4is1)	8	ZF6, ZF7	VTTCTGYW	ATTCTGCA



AART ZFP (2i13)	6	ZF1, ZF2, ZF3, ZF4, ZF5, ZF6	ATGKAGRGAAAA GCCCNN	ATGTAGGGAAAAAG CCCGG
GFI-1 ZFP (2kmk)	6	ZF3, ZF4, ZF5	TAAAKCACNGCA	AAATCACTG
GLI ZFP (2gli)	5	ZF3, ZF4, ZF5	GACCNCCCA	GACCACCCA
PRDM ZFP (5ei9)	13 (+1 degene rate)	ZF8, ZF9, ZF10, ZF11	CGTGGCTAGGGA GGC	NGNGGNNANGGN GGN
Human YY1 ZFP (1ubd)	4	ZF1, ZF2, ZF3, ZF4	AANATGGMG	AAAATGGAG
Wilms tumor suppressor ZFP (2jp9)	4	ZF2, ZF3, ZF4	GCGKGGGMG	GCGGGGGCG
ZIF268 (1zaa)	3	ZF1, ZF2, ZF3	GCGKGGGCG	GCGTGGGCG
Designed ZFP (1mey)	3	ZF1, ZF2, ZF3	GRGKCRGAA	GAGGCAGAA
TATA BOX-binding designed ZFP (1g2d)	3	ZF1, ZF2, ZF3	-	GCTATAAAA
KLF4 ZFP (2wbu)	3	ZF1, ZF2, ZF3	RRGGYGY	GAGGCGTGG
Kaiso ZFP (2lt7)	3	ZF1, ZF2	CTGCNA	CTGCCA
Drosophila TTKB (2drp)	2	ZF1, ZF2	AGGRY	AGGAT
Designed ZFP dimer (1f2i)	2	ZF1, ZF2	GGGCN	TGGGCG
ZIF23-GCN4 (1llm)	2	ZF1, ZF2	-	GCGTGG

**Table 3** - ZFP-DNA complexes from PDB. The ZF domains that appear to bind DNA are presented in the 3<sup>rd</sup> column. The DNA binding sequences are also presented in last two columns.



**Figure 13** - Representations of the 17 unique ZFP-DNA complexes deposited in PDB. In particular, for each ZFP the protein name, pdb id, the ZF domains with the amino acids in the binding positions (-1, 2, 3 and 6), the linkers between the ZF domains and the contacts between ZF domains and their DNA-target are reported. Solid lines represent H-bonds (the donor-acceptor distance is less than 3.5 Å).

## Scoring methods for ZFP-DNA interactions

We developed a scoring scheme that is based on a structural analysis of a dataset of 969 protein-DNA complexes. In this dataset, the H-bonds network between nucleic acids and amino acids was defined using a distance cut-off of 3.5 Å between

donor-acceptor heavy atoms and an angle range between 90 and 270 degrees between donor-hydrogen-acceptor. Only the atoms of the nucleic acids that are present in the major groove were taken into consideration. For each amino acid-nucleic acid pair a binding score was assigned based on the relative occurrence in the dataset. In particular, the binding specificity score of each amino acid was based on a nucleic acid weighted sum of the occurring number of H-bonds. In the case interactions are repulsive (i.e., Arg-C) the score is defined to be close to 0.

In order to investigate the possible preferences between non-polar amino acids and nucleic acids we analyzed the existing, corresponding pairs in the dataset, as defined by a cut-off of 3.5 Å (Table S2). From the total amount of 238 encountered hydrophobic pairs, 118 were with T, and in all of the 4 possible pairs for each amino acid T was the highest preference. Since hydrophobic interactions (~0.7 kcal/mol) that can be formed between proteins and DNA are ~6 times weaker than the hydrogen bonds (~4.1 kcal/mol) (Southall et al., 2001), the score for the T-non-polar amino acids pairs were set as the average occurrence of hydrophobic contacts for each of the four nucleic acids. The scores for all the pairs are presented in Table 4:

<b>Amino acid</b>	<b>Adenine score (num of pairs)</b>	<b>Cytosine score (num of pairs)</b>	<b>Guanine score (num of pairs)</b>	<b>Thymine score (num of pairs)</b>
A	0.0345 (6)	0.018 (4)	0.0335 (9)	0.084 (22)
C	0.0345 (2)	0.018 (2)	0.0335 (1)	0.084 (3)
D	0.17 (15)	0.83 (72)	0.01 (0)	0.01 (0)
E	0.13 (12)	0.87 (83)	0.01 (0)	0.01 (0)
F	0.0345 (7)	0.018 (3)	0.0335 (8)	0.084 (19)
G	0.0345 (4)	0.018 (4 )	0.0335 (19)	0.084 (20)
H	0.23 (14)	0.02 (1)	0.57 (34)	0.18 (11)
I	0.0345 (13 )	0.018 (3)	0.0335 (3)	0.084 (15)
K	0.08 (19)	0.01 (0)	0.80 (198)	0.12 (30)
L	0.0345 (8 )	0.018 (3)	0.0335 (2)	0.084 (17)
M	0.0345 (4)	0.018 (6)	0.0335 (1)	0.084 (6)
N	0.51 (87)	0.13 (22)	0.19 (32)	0.18 (30)
P	0.0345 (3)	0.018 (0)	0.0335 (4)	0.084 (12)
Q	0.51 (75)	0.15 (22)	0.18 (26)	0.16 (23)
R	0.07 (43)	0.01 (0)	0.83 (494)	0.084 (59)
S	0.30 (33)	0.16 (17)	0.34 (37)	0.20 (22)
T	0.19 (13)	0.28 (19)	0.25 (17)	0.29 (20)
V	0.0345 (1)	0.018 (0)	0.0335 (0)	0.084 (4)
W	0.20 (2)	0.01 (0)	0.30 (3)	0.50 (5)

Y	0.53 (27)	0.20 (10)	0.16 (8)	0.12 (6)
---	-----------	-----------	----------	----------

**Table 4** - Amino acid-nucleic acid binding specificity scores based on ZFP-DNA complexes. Scores that occurred after the analysis of the amino acid-nucleic acid pairs of the protein-DNA complexes that are deposit in PDB. As expected some amino acids tend to have strong binding preferences for a specific nucleic acid, like Arg and Lys to bind to G, and it is possible to have 2 hydrogen bonds contributing to this binding. On the contrary some amino acid-nucleic acid pairs that have the same charge tend to have a negative contribution to the binding and have a low binding specificity score, like Asp and Glu with G and T. While other amino acids can bind all four nucleic acids with no strong preference, like Ser and Thr. In addition, the binding specificity scores of Tryptophan-nucleic acid pairs could be not representative since there were only 7 pairs encountered.

As the previous score is based on all the available solved structures of protein-DNA complexes in PDB, a ZFP-specific score was developed based on the dataset containing 17 ZFP-DNA complexes. The available ZFP-DNA structures in the PDB are very few, so their H-bond network could not be used to generate robust scores, as was previously done for all the protein-DNA complexes. For this reason, we extracted from the 17 ZFP-DNA complexes every possible combination of the amino acids in the 3 critical binding positions, with their pairing nucleic acids. These amino acid-nucleic acid pairs were used as a parameters (scores) in the following function:

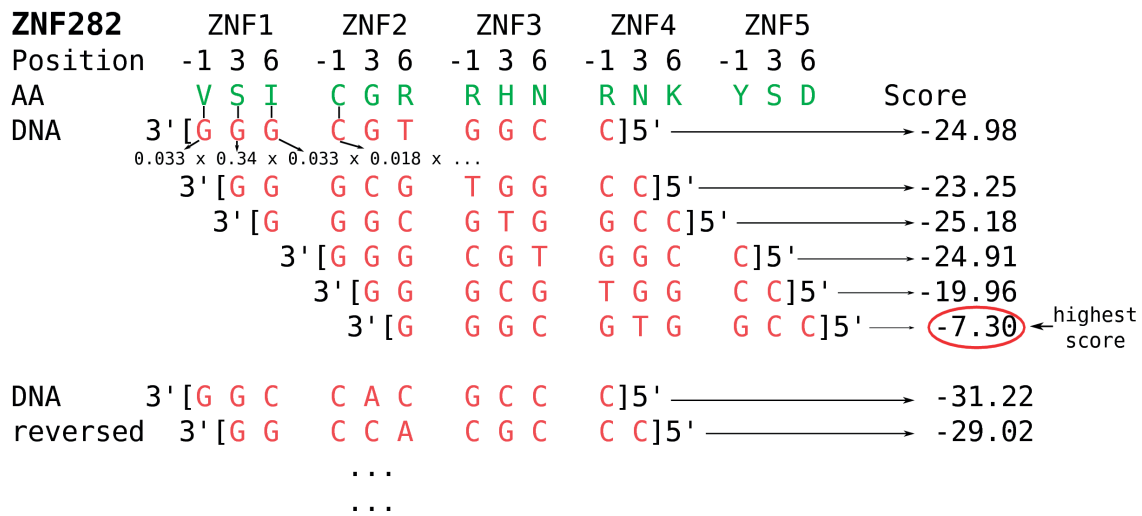
$$\sum_{complex=1}^{13} \left( \prod_{slidingWindow}^{\#slidingWindows} \frac{\sum_{nucAcid}^{\#nucAcids} (-\log(pairsAppearInComplex))}{\sum_{nucAcid}^{\#nucAcids} (-\log(potentialPairsAppearInComplex))} \right)$$

The weighed function accounts for all the possible complexes that can be formed by placing a ZFP on its DNA target. We used a sliding window method (see next section) to generate all the possible amino acid-nucleic acid combinations. Every amino acid-nucleic acid pair that is created is a parameter in our function (i.e. we have 73 pairs). In order to estimate the parameters, we used a Jackknife resampling approach. In particular, we adapted the function to represent the pairing parameters of 13 ZFP-DNA complexes and we kept as a test the artificial ZFP dimer, the ZIF23-GCN4, ZIF268 and CTCF. Each time, the function was minimized by optimizing its parameters that represent all the possible nucleic acids-amino acid pairs for 13 ZFP-DNA complexes. We kept the parameters that predicted successfully the binding of the 4 test complexes. After this procedure was followed for all the 13 complexes, the average of the estimated parameters gave the final bindings scores. Some pairs for Pro, Cys (Cys-A and Cys-T) and Gly (Gly-C) were missing from the 17 ZFP-DNA complexes, and their score was arbitrarily set to 0.1 (future availability of structures including such pairs will be useful to improve the parameters set). The minimization of the function was performed by the differential

evolution function of scipy.optimization library (Storn et al., 1997), which is able to search large spaces and therefore is more unlikely to get trapped into a local minimum, compared to other methods.

### Sliding window method to predict the ZF domains binding specificity

The designation of how a ZFP bind to its target DNA sequence is of high complexity, due to the multiple determinants of the binding specificity between the recognition amino acids and the DNA sequence. In order to find an accurate prediction of the ZF specificity a protocol was developed that was tested on a set of known structures. This is based on overlapping sliding windows of the DNA sequence on the main binding positions of the ZF sequence (Figure 14). For each position of the DNA sliding window on the binding positions of the ZF sequence a final score is calculated. This final score is the sum of the logs of all the amino acid-nucleic acid pair scores at that position. Due to the fact that a ZF domain can bind to the DNA in different ways, we take into consideration the most common binding motifs: (i) positions -1, 3, 6 bind 3 nucleotides, (ii) positions -1, 2, 3, 6 bind 3 or 4 nucleotides (for the position 2 the nucleic acid of 3' or 4' position of the opposite strand is taken into account, depended on which pair has the highest binding score, Figure 1). In addition, position -2 is taken into consideration if it has an Arg residue, and therefore a prediction for G is favored. If there are long linkers (> 15 aa) then the ZFP protein is broken at the linkers into different parts. In particular, if there are N long linkers, then the protein is broken in N+1 parts, and each part is treated as a different ZFP and tested for the DNA-target separately. As a result, a prediction for each part is performed, and the prediction with the lowest score will be the one for the overall ZFP. Finally, to have a better insight into the binding specificity we use as input not only the most frequent DNA binding sequence, but also the reversed strand, as well as the consensus (if there is one). The maximum final score, indicates the DNA sliding window's position and reveals the nucleic acid-amino acid pairs. The nucleic acid-amino acid pairs provide a prediction for the binding relationship between a ZFP and its binding DNA sequence.



**Figure 14** - Sliding window for ZNF282 and its predicted DNA target sequence (CCGGTGCGGG)

To validate our method, we used 17 ZFP-DNA complexes that are deposited in PDB. We used these complexes to assess the prediction of which ZF domains bind to the DNA target sequence. For the predictions, the full sequences of the ZFPs and the DNA core motifs were used and the results are presented briefly in Figure 15 and analytically in Table 5:

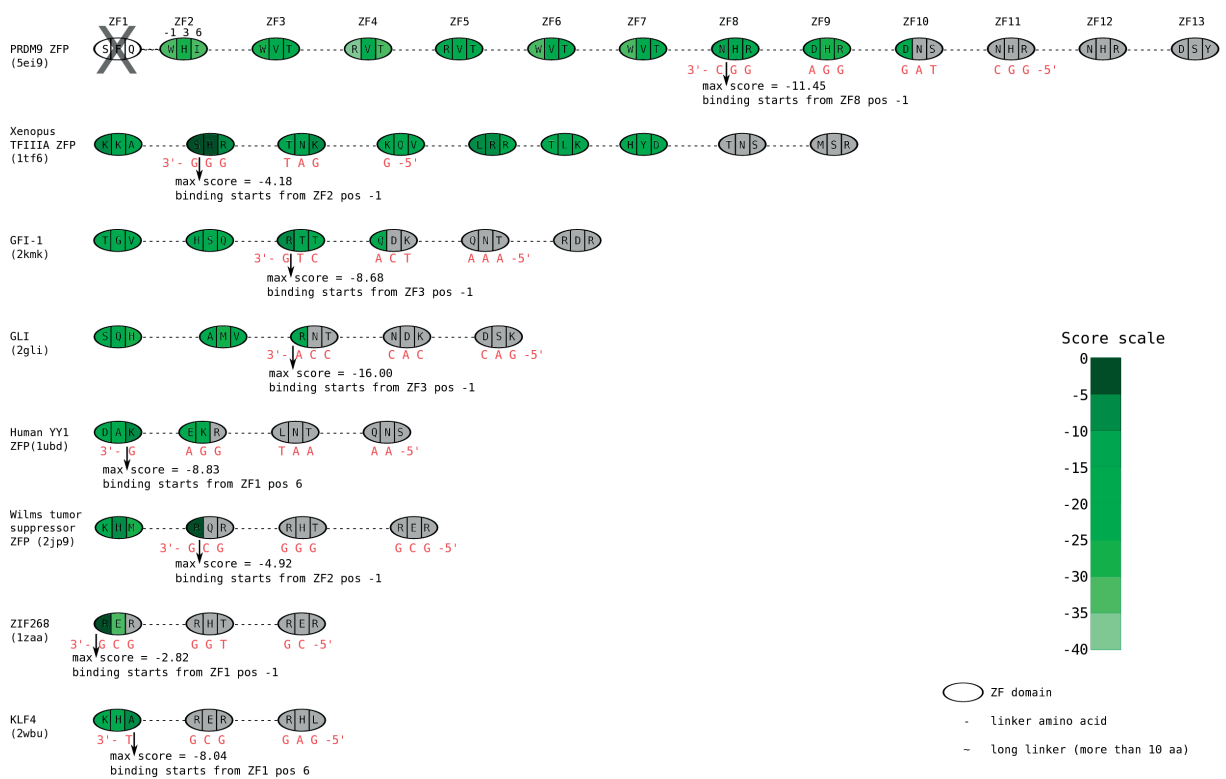
ZFP (pdb id)	DNA sequence from the pdb structure	ZF domain (position) that the binding starts	Most frequent sequence, scores from prot/DNA complexes, positions -1,3,6	Most frequent sequence, scores from prot/DNA complexes, positions -1,2,3,6	Consensus sequence, scores from prot/DNA complexes, positions -1,3,6	Most frequent sequence, optimized scores from jackknife, positions -1,3,6
CTCF (5t0u)	CCAGCAGGG G GCGCT	ZF3(-1) +	ZF3(-1)+ (S= -14.91) ZF5(-1)- (S= -28.98)	ZF3(-1)+ (S= -24.63) ZF5(3)- (S= -37.73)	ZF3(-1)+ (S= -17.02) ZF2(3)+ (S= -25.96)	ZF3(-1)+ (S= -1.94) ZF6(6)- (S= -10.13)
Xenopus TFIIIA ZFP (1tf6)	GGATGGGAG A	ZF2(-1) +	ZF2(-1)+ (S= -4.18) ZF2(3)+ (S= -8.31)	ZF7(3)- (S= -13.16) ZF2(-1)+ (S= -14.28)	ZF2(-1)+ (S= -4.63) ZF5(-1)+ (S= -8.68)	ZF2(-1)+ (S= -0.11) ZF5(-1)+ (S= -0.38)
ZNF217 (4is1)	ATTCTGC A	ZF6(-1) +	ZF6(-1)+ (S= -10.4) ZF6(-1)- (S= -10.83)	ZF6(-1)- (S= -12.71) ZF6(-1)+ (S= -13.40)	ZF6(-1)- (S= -10.26) ZF6(-1)+ (S= -12.84)	ZF6(-1)+ (S= -2.31) ZF2(-1)+ (S= -3.90)
AART ZFP (2i13)	ATGTAGGGA A AAGCCCGG	ZF1(-1) +	ZF1(-1)+ (S= -13.08) ZF1(3)+ (S= -30.65)	ZF1(-1)+ (S= -18.81) ZF1(3)+ (S= -39.32)	ZF1(-1)+ (S= -15.15) ZF1(3)+ (S= -27.31)	ZF1(-1)+ (S= -0.50) ZF1(6)+ (S= -25.89)

GFI-1 ZFP (2kmk)	<b>AAATCACT G</b>	ZF3(-1) +	<b>ZF3(-1)+</b> (S= - 8.68) ZF1(6)- (S= - 10.14)	ZF1(6)- (S= - 15.70) <b>ZF3(-1)+</b> (S= - 16.28)	<b>ZF3(-1)+</b> (S= - 7.48) ZF1(6)- (S= - 11.29)	<b>ZF3(-1)+</b> (S= - 0.18) ZF1(-1)- (S= - 0.55)
GLI ZFP (2gli)	<b>GACCACCC A</b>	ZF3(-1) +	<b>ZF3(-1)+</b> (S= - 16.00) ZF1(-1)- (S= - 16.58)	ZF1(3)- (S= 18.74) ZF2(3)- (S= 21.59)	<b>ZF3(-1)+</b> (S= - 15.60) ZF1(-1)- (S= - 17.26)	<b>ZF3(-1)+</b> (S= - 1.35) ZF1(6)+ (S= - 5.82)
PRDM ZF (5ei9 )	<b>GTGGCTAGG G AGGC</b>	ZF8(-1) +	<b>ZF8(-1)+</b> (S= - 11.45) ZF8(3)+ (S= - 14.99)	<b>ZF8(-1)+</b> (S= - 17.05) ZF7(6)+ (S= - 20.90)	ZNF8(3)+ (S= - 9.23) ZNF8(-1)+ (S= - 9.81)	<b>ZF8(-1)+</b> (S= - 0.55) ZF3(6)+ (S= - 2.34)
Human YY1 ZF P (1ubd)	<b>AAAATGGA G</b>	ZF1(6) +	<b>ZF1(6)+</b> (S= - 8.83) ZF2(-1)+ (S= - 16.64)	<b>ZF1(6)+</b> (S= - 14.12) ZF2(-1)+ (S= - 22.78)	<b>ZF1(6)+</b> (S= - 10.05) ZF2(-1)- (S= - 12.22)	<b>ZF1(6)+</b> (S= - 0.26) ZF1(-1)+ (S= - 2.84)
Wilms tumor suppressor ZFP (2jp9)	<b>GCGGGGGC G</b>	ZF2(-1) +	<b>ZF2(-1)+</b> (S= - 4.92) ZF1(3)+ (S= - 8.87)	<b>ZF2(-1)+</b> (S= - 7.12) ZF1(3)+ (S= - 10.33)	<b>ZF2(-1)+</b> (S= - 4.05) ZF1(3)+ (S= - 9.08)	<b>ZF2(-1)+</b> (S= - 0.48) ZF1(3)+ (S= - 7.22)
ZIF268 (1zaa)	<b>GCGTGGGC G</b>	ZF1(-1) +	<b>ZF1(-1)+</b> (S= - 2.82) ZF1(-1)- (S= - 13.10)	<b>ZF1(-1)+</b> (S= - 3.38) ZF1(-1)- (S= - 10.08)	<b>ZF1(-1)+</b> (S= - 2.90) ZF1(-1)- (S= - 13.66)	<b>ZF1(-1)+</b> (S= - 0.10) ZF1(-1)- (S= - 10.06)
Designed ZFP (1mey)	<b>GAGGCAGA A</b>	ZF1(-1) +	<b>ZF1(-1)+</b> (S= - 4.31) ZF1(3)+ (S= - 14.7)	<b>ZF1(-1)+</b> (S= - 7.71) ZF1(3)+ (S= - 17.39)	<b>ZF1(-1)+</b> (S= - 4.70) ZF1(3)+ (S= - 14.05)	<b>ZF1(-1)+</b> (S= - 0.12) ZF1(3)+ (S= - 8.41)
TATA BOX- binding designe d ZFP (1g2d)	<b>GCTATAAA A</b>	ZF1(-1) +	<b>ZF1(-1)+</b> (S= - 9.34) ZF1(-1)- (S= - 11.08)	<b>ZF1(-1)+</b> (S= - 13.29) ZF1(-1)- (S= - 15.07)	-	<b>ZF1(-1)+</b> (S= - 0.16) ZNF1(-1)- (S= - 6.16)
KLF4 ZFP (2wbu)	<b>GAGGCGTG G</b>	ZF1(6) +	<b>ZF1(6)+</b> (S= - 8.04) ZF1(-1)+ (S= - 14.34)	<b>ZF1(6)+</b> (S= - 9.43) ZF1(-1)+ (S= - 15.91)	<b>ZF1(6)+</b> (S= - 8.65) ZF1(-1)+ (S= - 12.66)	<b>ZF1(6)+</b> (S= - 0.30) ZF1(-1)+ (S= - 3.98)
Kaiso ZFP (2lt7)	<b>CTGCC A</b>	ZF1(-1) +	ZF1(6)- (S= - 7.05) ZF1(3)- (S= - 9.81)	ZF1(6)- (S= - 8.42) ZF1(3)- (S= - 11.05)	ZF1(6)- (S= - 6.76) ZF1(-1)+ (S= - 9.26)	<b>ZF1(-1)+</b> (S= - 1.68) ZF1(3)- (S= - 3.52)
Drosophila TTK B (2drp)	<b>AGGA T</b>	ZF1(-1) +	<b>ZF1(-1)+</b> (S= - 3.43) ZF1(3)+ (S= - 9.59)	<b>ZF1(-1)+</b> (S= - 4.82) ZF1(3)+ (S= - 11.2)	<b>ZF1(-1)+</b> (S= - 4.40) ZF1(3)+ (S= - 7.88)	<b>ZF1(-1)+</b> (S= - 0.12) ZNF1(3)+ (S= - 8.61)
Designed ZFP dimer (1f2i)	<b>TGGGC G</b>	ZF1(-1) +	<b>ZF1(-1)+</b> (S= - 1.26) ZF1(3)- (S= - )	<b>ZF1(-1)+</b> (S= - 1.63) ZF1(3)+ (S= - )	<b>ZF1(-1)+</b> (S= - 2.45) ZF1(3)+ (S= - )	<b>ZF1(-1)+</b> (S= - 0.06) ZF1(3)+ (S= - )

			11.18)	11.53)	8.11)	11.08)
ZIF23-GCN4 (1llm)	<b>GCGTGG</b>	ZF1(-1)+	<b>ZF1(-1)+</b> (S= - 2.31) ZF1(3)- (S= - 9.24)	<b>ZF1(-1)+</b> (S= - 2.68) ZF1(3)+ (S= - 13.65)	-	<b>ZF1(-1)+</b> (S= - 0.06) ZNF1(3)- (S= - 8.73)

**Table 5** - Benchmark results of the sliding window method on known ZFP-DNA complexes. The DNA sequences in bold are the binding core motifs. The starting binding ZF domains that are in bold show that the method's prediction was successful.

Scores of the sliding window for one DNA strand



**Figure 15** - Predictions of the ZF domains than bind on the DNA-target. For each sliding window of the DNA on the ZF's binding amino-acid sequence there is a green indicator of the score. The darker the green is the higher is the score of the sliding window. The highest score shows a prediction of the localization of the protein on the DNA. For each ZFP it is shown: its DNA-target, the amino acids in positions -1, 3 and 6, the maximum sliding window score and the start of the binding (i.e. For the PRDM ZFP the maximum sliding window score indicates that the binding starts from the position -1 of ZF8 and that ZF8-ZF11 bind on the DNA-target).

According to the results reported in Table 5, the specificity scores from protein-DNA complexes using binding positions -1, 3, 6 and the most frequent DNA



sequence (derived from the pdb file; i.e. AGGAT for Drosophila TTKB) was successfully predicted for 16 out of 17 ZFPs. When the amino acids for position 2 were introduced, 12 out of 17 proteins had their binding ZF domain successfully predicted (Table S6). This could be the indication that the amino acids in the position 2 do not always interact with the DNA or that they provide more stability than specificity for the complex. Moreover, usually in position 2 we have the presence of Ser that can act as H-bond donor and acceptor and therefore can interact with all four nucleotides. Next, by applying the specificity scores using binding positions -1, 3, 6 on the consensus DNA sequences (i.e. AGGRY for Drosophila TTKB), we successfully predicted the binding ZF domains of 12 out of 15 proteins. This might be due to the fact that the consensus sequences cannot describe in what frequency a nucleotide is represented in the motif, on the contrary a PWM might have been more precise. As it was expected, the optimized binding specificity scores from the complexes successfully predicted the binding of 17 ZFPs. But these scores might be over-fitted in the 17 complexes and the efficiency of this approach on other ZFP-DNA complexes might not be robust.

In most of the cases, we failed to predict the binding ZF domains of ZNF217. This might be because there are multiple ZF domains (8) that could bind to the small DNA target sequence of 6 nucleotides. Since there is an increased number of possible ZF-domains/DNA combinations, it makes the prediction more difficult. Also, ZNF217 has 5 long linkers (more than 10 amino acids length) that can affect the flanking of neighbouring ZF domains. In general, our predictions agree in the way of binding for ZNF217, which might be an indication that ZFPs can have more than one way to bind to their target sequence or that some might not have a strong specificity for a DNA target. For Kaiso protein only the method with the optimized scores had a successful prediction. The failing of the other method might be because of Kaiso's Arg595 in the C-terminal tail that bind to a nucleotide base at the minor groove of the DNA and leads to a non-canonical binding. Therefore, Kaiso's binding properties could differ from the ZFP that don't use their C-terminal amino acids to bind nucleotides.

### [A method to predict the ZF domains binding specificity](#)

We derived a recognition code from the analysis of the binding properties of the protein-DNA complexes and the 17 ZFP-DNA complexes in PDB. The recognition code can provide a PWM and predict a set of DNA binding sequences on which a ZFP specifically binds. There are strong preferences between Arg and G (score: 0.83), Lys and G (score: 0.8), His and G (score: 0.57), Asn and A (score: 0.51), Gln and A (score: 0.51), Asp and C (score: 0.83), and Glu and C (score: 0.87), while weaker binding has been observed for some amino acids, like Thr that binds mainly to T but also to C, and Ser that binds mainly to G but also to A. For both Thr and Ser,

that occur in position 6, their binding seems to be influenced by the amino acid in the position 2 of the next ZF domain (if any). In general, the assignments for neighboring ZF domains and the contribution of an Arg in position -2, are based on the existing ZFP structures.

The nucleic acids preferences of non-polar amino acids are not easy to be determined. Although, non-polar amino acids tend to contact mainly T, if they occur in position 6 this tendency could be influenced but the amino acid in the position 2 of the next ZF domain (if any). We assume that if in the position 2 of the previous ZF domain is an Arg, Lys or His then they make contact with C, if there is an Asn or a Gln, they make contact with T, if there is an Asp or Glu they make contact with G, and if there is a non-polar amino acid they make contact with A. In addition to that, Gly appears to be "replaced" by other amino acids. In particular, if a Gly is in position -1 or 3 it will be the amino acid in position -2 or 2, respectively, that will interact with a nucleotide (like the Gly in the ZF2 position 3 of the TATA-binding ZFP, that the T in position 2 "replaces" it, since it makes a hydrophobic contact with the T).

To validate our method, we compared it with the predicted sequences by Persikov et al (Persikov et al., 2014) Najafabadi et al (Najafabadi et al., 2015a) and Gupta et al. (Gupta et al., 2014), using the respective online prediction tools. We made the alignments of the consensus DNA binding sequence of 17 known ZFP-DNA complexes with the most frequent predicted DNA sequence of our prediction and the three recognition codes and calculated how successful the prediction was by counting the cases that the predicted nucleic acid in a specific position was included in the corresponding position of the consensus sequence, divided by the overall length of the sequence. Therefore, the prediction validating score is the sum of the nucleic acids that are present in specific positions of the consensus sequences, divided by the length of the consensus sequence. Likewise, we applied our method on the "Gold Standard" motifs, that are described by Najafabadi et al (Najafabadi et al., 2015a). We aligned the PWM of our prediction with the PWM of the experimental motifs and we chose the best alignment in order to maximize the Pearson correlation. Finally, our predictions were compared with the predictions of the other recognition codes.

Our recognition code predicts a possible PWM for all the ZF domains of a ZFP. It is based on the preferences of the binding amino acids, that were derived from a structural analysis of protein-DNA complexes, and also on observations on how neighbouring ZF domain can affect the binding of each other. We compared our method's predictions with the predictions of Najafadi's (Najafabadi et al., 2015a), Persikov's (Persikov et al., 2015) and Gupta's (Gupta et al., 2014) recognition codes by aligning the predicted sequences. We also involved in the alignment the experimentally defined consensus sequence for each ZFP. The alignment could also be an indication of which ZF domains bind to DNA, by showing the ZF localization on the predicted sequences. The accuracy of the predictions was assessed based on the percentage of the nucleotides that were predicted successfully by included in the

consensus DNA motif (Table 6 - Figure 16). ZIF23-GCN4 was excluded since there is not known consensus sequence and from the TATA BOX-binding designed ZFP, only the TATA core binding motif was used.

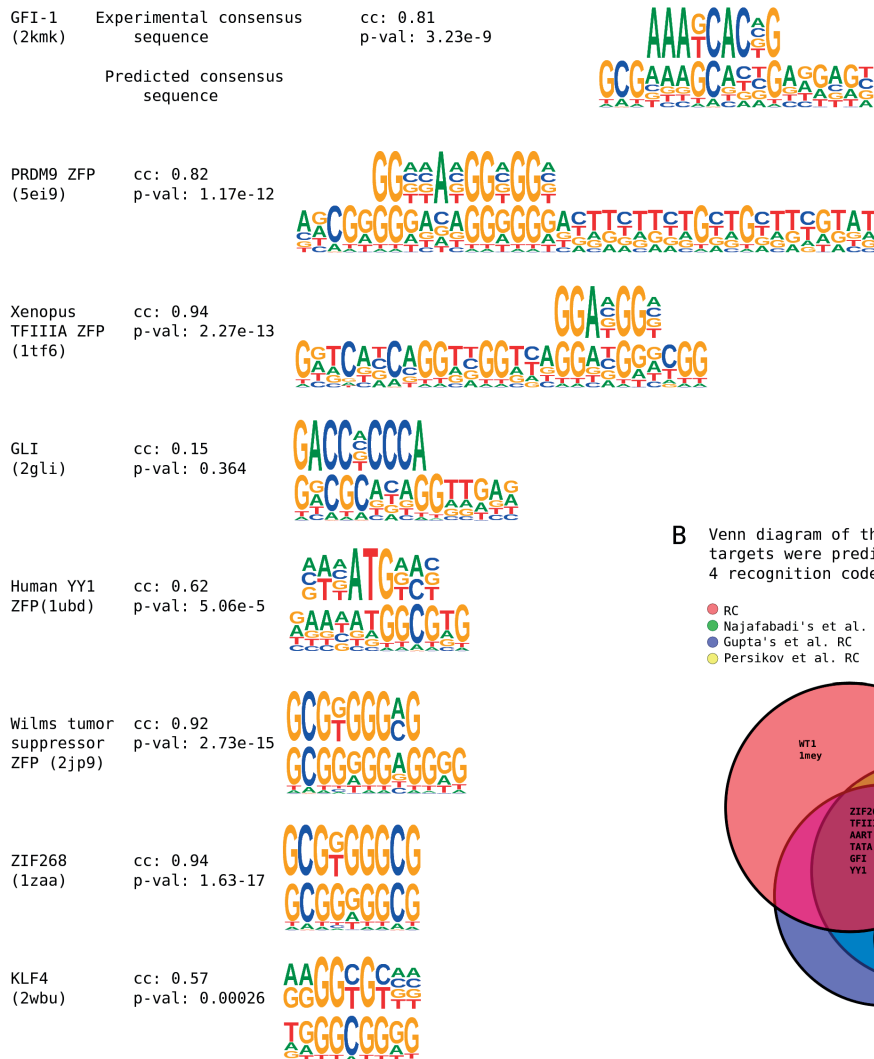
<b>ZFP (pdb id)</b>	<b>Alignments</b>	<b>Method – success percentage</b>	<b>Najafabadi – success percentage</b>	<b>Persikov – success percentage</b>	<b>Gupta – success percentage</b>
Xenopus TFIIIA ZFP (1tf6)	Consens.: <b>GGANGGNNGN</b> Predict.: <b>GGATGGGCGG</b> Najafab.: <b>GGATGGTTGG</b> Persikov: <b>NTACGGACAG</b> Gupta: <b>GGATGGTCGG</b>	<b>100%</b>	<b>100%</b>	70%	<b>100%</b>
CTCF (5t0u)	<b>NCANNAGRNGGCRSY</b> <b>TCGCTAGGGGGCGCT</b> <b>TCGATAGGGGGCGTT</b> <b>CCGACAGAGGGCGCA</b> <b>GCGCCAGGCGGCGTT</b>	<b>93.3%</b>	86.7%	86.7%	86.7%
ZNF217 (4is1)	<b>VTTCTGYW</b> -AGATAG- <b>GATATTGG</b> <b>NACCCTCA</b> -----	16.7%	35.5%	35.5%	-
AART ZFP (2i13)	<b>ATGKAGRGAAAAGCCNN</b> <b>TTGGAGGGAAAAGCCCGG</b> <b>ATGTAGAGAAAAGCCTGG</b> <b>AAGCAGAGACAAGGCCGG</b> <b>GTGTAGTGAAAAGCCTGG</b>	<b>94.4%</b>	<b>94.4%</b>	78.8%	83.3%
GFI-1 ZFP (2kmk)	<b>AAAKCACNG</b> <b>AAAGCACTG</b> <b>AAATCACTG</b> <b>CAANNNATG</b> <b>AAATCACAG</b>	<b>100%</b>	<b>100%</b>	44.4%	<b>100%</b>
GLI ZFP (2gli)	<b>GACCNCCCA</b> <b>GGCGCACAG</b> <b>TCCTCCAAG</b> <b>NCCTACCCG</b> <b>GTCTCTTAG</b>	44.4%	33.3%	55.6%	33.3%
PRDM ZFP (5ei9)	<b>NGNGGNNANGGNGGN</b> <b>GGGGGACAGGGGGGA</b> <b>GGAGGACACGGCGGA</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>

	<b>GGCGGCCACGGCGGC GGTGGTGACGGCGGT</b>				
Human YY1 ZFP (1ubd)	<b>AANATGGMG AAAATGGCG AAAATGGCG AACAGGGCT AAAATGGCG</b>	<b>100%</b>	<b>100%</b>	77.8%	<b>100%</b>
Wilms tumor suppressor ZFP (2jp9)	<b>GCGKGGGMG GCGGGGGAG GCGTGGGTG GCGAGGGGG GCGTGGGTG</b>	<b>100%</b>	88.9%	77.8%	88.9%
ZIF268 (1zaa)	<b>GCGKGGGCG GCGGGGGCG GCGTGGGCG GCGAGGGCG GCGTGGGCG</b>	<b>100%</b>	<b>100%</b>	88.9%	<b>100%</b>
Designed ZFP (1mey)	<b>GRGKCRGAA GGGGCAGAA GGGTCAAAA GGGNNTAC GGGTCATAA</b>	<b>100%</b>	88.9%	44.4%	88.9%
TATA BOX- binding designed ZFP (1g2d)	<b>TATA TATA TATA AATA TATA</b>	<b>100%</b>	<b>100%</b>	75%	<b>100%</b>
KLF4 ZFP (2wbu)	<b>RRGGYGYNN TGGGCGGGG TGGGCGTGG TGGGCGAGG TGGGCGTGG</b>	77.8%	<b>88.9%</b>	77.8%	<b>88.9%</b>
Kaiso ZFP (2lt7)	<b>CTGCNA GATGGT TGCGTT TGCGTT TATGTT</b>	16.7%	16.7%	16.7%	16.7%
Drosophila TTKB	<b>AGGRY AGGAG</b>	80%	<b>100%</b>	<b>100%</b>	<b>100%</b>

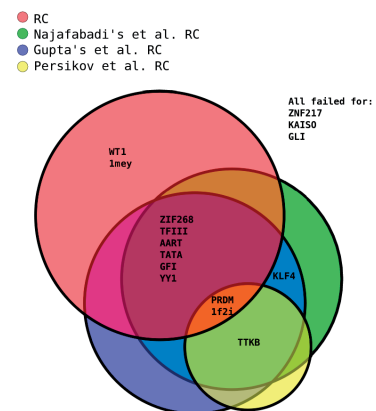
(2drp)	<b>AGGAT</b> <b>AGGAC</b> <b>AGGAT</b>				
Designed ZFP dimer (1f2i)	<b>NGGGCN</b> <b>TGGGCG</b> <b>TGGGCG</b> <b>AGGGCG</b> <b>TGGGCG</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>

**Table 6** - Scores of the alignments between the consensus sequence and the predicted sequences. In the column of the alignments the first line is the consensus binding DNA sequence, the second line is the sequence predicted by our recognition code, the third line is the sequence predicted by Najafabadi's recognition code, the fourth line is the sequence predicted by Persikov's recognition code and the fifth line is the sequence predicted by Gupta's recognition code.

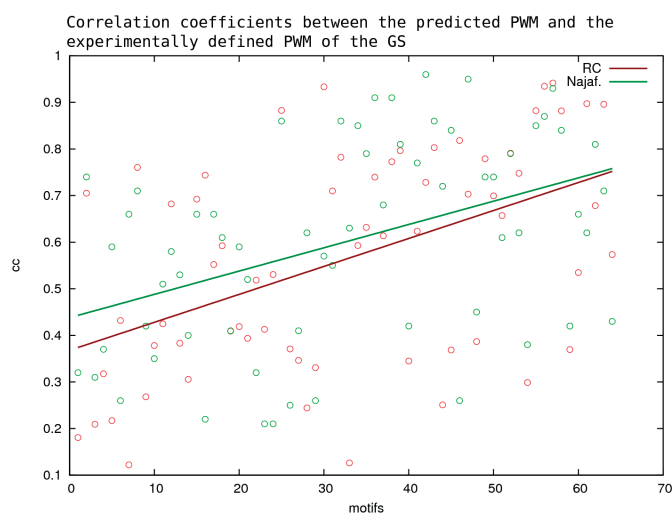
# **A** Alignments of the experimental and predicted consensus sequences



# **B** Venn diagram of the ZFPs that their DNA targets were predicted successfully by the 4 recognition codes:



# **C**



**Figure 16** - Comparison of the predictions of the four recognition codes. A. Alignment of the PWM (position weight matrix) logos of the experimentally-derived consensus

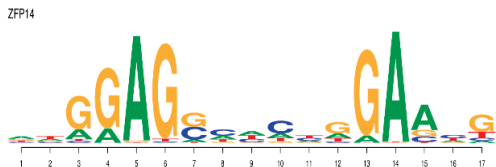
sequence and the consensus sequence predicted by our recognition code. B. Venn diagram of the ZFPs that their DNA targets were successfully predicted by the recognition codes. C. The correlation coefficients for each of the PWM alignments.

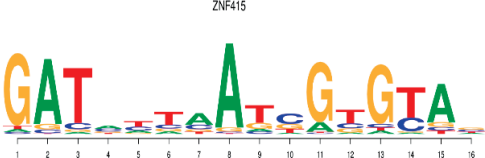
From the assessment of the four recognition codes based on the present benchmark set we obtain that Najafabadi's, Gupta's and our methods have the same robustness, while they outperform Persikov's recognition code (Figure 16A-B). In 9 and 8 out of 17 proteins the DNA binding sequence is predicted with 100% success from our and Najafabadi's-Gupta's recognition codes respectively, while Persikov's recognition codes predicts 3 out of 17 correctly. All recognition codes failed to predict the DNA target sequence in the cases of Kaiso, GLI-1 and ZNF217. It seems like these three ZFPs have their own unique way to bind to their DNA sequence that cannot be modeled by any of the known recognition codes. In two cases (WTS ZFP and designed ZFP) our recognition code gives better results than the other codes.

Finally, we used our recognition code to predict the DNA motifs of ZFPs as they were defined by Najafabadi et al., 2016 (64 Gold-standard motifs). It appears that our recognition code performs almost as robust as Najafabadi's recognition code (Figure 16C).

#### Validation of our methods with ZFPs of unknown structure

SMiLE-seq analysis was performed by our collaborators (i.e., Dr. Alina Isakova from the Deplancke's lab at the EPFL), on two ZFPs: ZFP14 and ZNF415. In particular, the full-length open reading frames (ORFs) of human ZFP14 and ZNF415 containing canonical C2H2-ZF domains and canonical linkers (Table 7) were subcloned into pF3A-eGFP in vitro expression vector (Isakova et al., 2016) by Gateway cloning. Next, the constructs were expressed in vitro and subjected to SMiLE-seq analysis as described in Isakova et al., 2017. Furthermore, the ZFPs were truncated, by keeping those ZF domains that are predicted to bind to the DNA target.

ZFP	Number of ZF domains	SMiLE-seq highest frequent DNA	Logos	Truncations
ZFP14	13	ATGGAGGCACTGGA ACG		ZF1-ZF6

ZNF415	11	GATATTAATCGTGTA G	
--------	----	----------------------	--

**Table 7** - ZFPs with their SMiLE-seq predicted DNA sequences and selected truncations.

To test further our method, we looked at two ZFP proteins, ZFP14 and ZNF415, and their predicted DNA-target sequence as obtained from SMiLE-seq (see Table 7). For each ZFP a prediction of which and how its ZF domains bind to their DNA target sequence was made (Table 8). In particular, we analysed the binding of the two ZFPs with the binding specificity scoring sliding window method, taking into consideration the amino acids in positions -1, 3, 6 and -1, 2, 3, 6 with the specificity scores that occurred from the protein-DNA complexes analysis, and the positions -1, 3, 6 with the optimized binding specificity jackknife scores from the 17 ZFP-DNA complexes. All three scoring predictions came in agreement with each other on which ZF domains bind for ZNF415, but not for ZFP14. ZFP14 has in fact a Cys residue in a critical binding position and this might affect the prediction, since there is no estimated score for the Cys-A and Cys-T pairs. For ZFP14 the binding is predicted to start from the position -1 of ZF1, and the ZF domains from 1 to 6 are predicted to bind to the DNA. For ZNF415 the binding is predicted to start from the position 6 of ZF5, and the ZF domains from 6 to 10 are predicted to bind to the DNA.

ZFP (pdb id)	SMiLE-seq sequence	Most frequent sequence, scores from prot/DNA complexes, positions -1,3,6	Most frequent sequence, scores from prot/DNA complexes, positions -1,2,3,6	Most frequent sequence, optimized scores from jackknife, positions -1,3,6
ZFP14	ATGGAGGCACTGGAACG	<b>ZF1(-1)+</b> (S= -19.02)	<b>ZF1(-1)+</b> (S= -30.32)	ZF1(3)+ (S= -9.75)
ZNF415	GATATTAATCGTGTA	<b>ZF5(6)+</b> (S= -22.27)	<b>ZF5(6)+</b> (S= -28.57)	<b>ZF5(6)+</b> (S= -1.10)

**Table 8** - Results of the sliding window method on 2 ZFPs and their SMiLE-seq predicted binding sequences. The scoring matrix of the optimized scores don't agree with the predictions from the scoring matrix of the analysis of the protein-DNA complexes.

We set out to further investigate whether the predicted zinc fingers of the C2H2 ZFPs are indeed involved in DNA binding. This analysis suggested that six fingers out of thirteen contribute to the ZFP14 binding specificity and indicated ZF1, ZF2, ZF3, ZF4, ZF5 and ZF6 as potentially bound fingers. Similarly, five fingers out of eleven of ZFP415 are predicted bind the DNA motif (ZF6 ZF7 ZF8 ZF9 and ZF10). To validate our predictions, we truncated the ZF domains that have predicted to not bind to their DNA target sequence for each of the ZFP. For ZPF14 we kept the domains



ZF1 to ZF6 and for ZNF415 we kept the domains ZF6 to ZF9. Although ZF10 is predicted to bind to the DNA, we decided to truncate it from ZNF415 because its 3bp-binding target was not always appeared in the SMiLE-seq motif. Our collaborators then generated synthetic constructs of ZFP14 and ZFP415 in which we only kept the fingers that are predicted to bind and removed the remaining ones. SMiLE-seq analysis on this synthetic construct yielded a motif that exactly matched the previously identified motif for the full length ZFP14 (Table 8). Thus, we demonstrated that the specificity of C2H2-ZFPs is achieved through DNA binding of specific zinc fingers. In addition, for ZFP14 and ZNF415, we predicted a PWM that indicates further potential DNA targets. It is also remarkable, that our recognition code comes in agreement with the other recognition codes in the positions that the SMiLE-seq signal in higher (Figure 17). This could be an indication that for some positions/amino acids there is higher specificity than others.

SMiLEseq

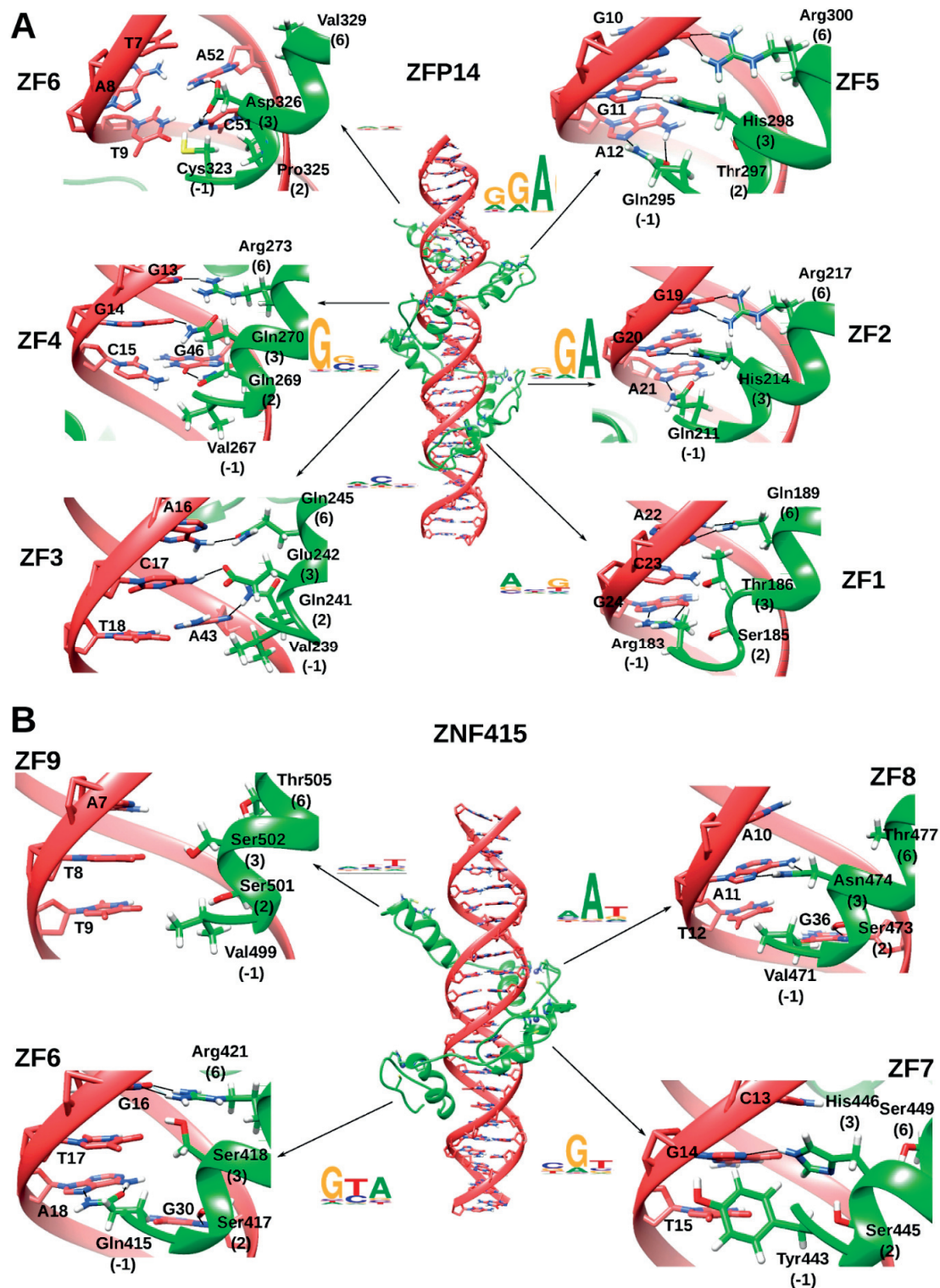
Our pred	AGAATGAAGAAGCAGGAGTAGTCTGGAGAGACTGGAATG
Najafaba	AGAAATATTATTATGGGTATTACCGGAGATACTGGAATG
Persikov	TGAANTAGTAGTCGTGATNNNNNNGGAGATCCTGGAATG
Gupta	AAAAATAATAATTATGATGATTCCGGAGATACTGGAACG

SMiLEseq

Our pred GATGATAGTCATCGAGGACGAATTCTGCTGGAA  
Najafaba GATGATATTAATTGTGTAAATACTGTGGTGGAA  
Persikov GATGATATCCATAGTGTAAACAACGTTGTTGGAA  
Gupta GATGATTTTTATTGTGTAAATACTGTGGTGGAA

structural models were built with the 3DNA software (v2.0) (Lu & Olson, 2003). Each of the 2 models was placed in an explicit solvent box with 0.15M NaCl concentration under periodic boundary conditions. The complexes were parametrized using the AMBER force field (ff99bsc0) and ZAFF force field (Peters et al., 2010) (for the zinc ion and the 4 residues that coordinate it), and the TIP3P model was used for water molecules (LeaP program of Amber tools V11 - Case et al., 2005). Geometry optimization via energy minimization was performed using the NAMD program (v2.8) (Phillips et al., 2005). Initially, an energy minimization was performed for 5000 steps with constraints on the atoms of the side chains that are involved in the H-bonds formation between the protein and the DNA (positions -1, 3 and 6), followed by an energy minimization of 1000 steps with no constraints. In all phases a time step of 1 fs was used, the covalent bonds involving hydrogen atoms were constrained by the RATTLE algorithm and the van der Waals interaction cut-off was set at 12 Å.

Based on these results, we built the models for ZFP14 and ZNF415 based on the first ZF domain of ZIF268. Next, we placed these protein models in the respective DNA models that represent the SMiLE-seq predicted sequences. After geometry optimization through energy optimization we obtained the final complexes (Figure 18):



**Figure 18** - Structural models of ZFP14 (A) and ZNF415 (B). The model of the complex appears in the middle of the image, while at each side the ZF domains appears. For each ZF domain it appears: SMiLE-seq logo, binding positions, binding amino acids, nucleic acids that participate in the binding and hydrogen bonds between proteins and DNA.

The modeled ZFs domains appear to form H-bonds with the DNA, with the amino acids in the binding positions -1, 2, 3 and 6. When SMiLE-seq experiments predict G in the DNA consensus it appears that a His or an Arg are the interacting residues. This confirms the prediction for both ZFPs, not only because His and Arg have a high sliding window score in binding Gs, but can also form two hydrogen bonds with a G and contribute to the stability of the complex.

### 3.3. Conclusions and perspectives

DNA recognition by ZFPs plays an important role in gene silencing. In the last years, many recognition codes between ZF domains and the DNA, that follow different approaches, have been proposed. However, these codes cannot capture the complex DNA binding landscape. Therefore, determining of which ZF domains and how do they bind on the DNA remains still a challenge.

Based on a structural analysis of protein-DNA and ZFP-DNA complexes deposited in the PDB, we extracted some general recognition binding rules between ZFPs and their DNA targets, encoded them in scoring functions used to predict (i) which ZF domains of multi-domain ZFPs bind to a given DNA sequence and (ii) the most probable DNA target sequence for a given ZFP.

Our approach showed that for a ZF domain that recognize 3-4 nucleotides in a canonical mode, it is possible to predict its DNA binding specificity. For example, the sliding window method detected which ZF domains bind to a given DNA target, when the binding follows the canonical rules. Also, our recognition code, predicted correctly the DNA target sequences of half of the tested solved ZFP-DNA complexes. In addition, we further tested our recognition code with two ZFPs with unsolved structures, and we managed to predict accurately their binding DNA targets, in respect with the results from ChIP-exo experiments. But, it is important to further test our recognition code with a much larger data set of unsolved complexes.

Our recognition code slightly outperformed the current approaches for solved complexes: 9 out 17 DNA targets of canonical-binding ZFPs were predicted correctly, in contrast to 8, 8 and 3 out of 17 DNA targets for the other three recognition codes. In addition, besides Najafabadi's recognition code, our recognition code is one of the few recognition codes that takes into consideration the neighboring ZF domains effect on binding specificity. Also, our recognition code is the only one that takes into considerations the Arg of the position -2 and its effect on the binding.

On the other hand, there are some limitations to our methods. For example, since our method is based on the assumption that one ZF domain binds to three nucleotides of the same strand, is limited to predict the canonical binding only. However, the predictions of the binding specificity sliding window seems to be quite robust for the canonical binding, but its efficiency depends also on the quality of the ChIP-seq/SMiLE-seq data. Furthermore, the scoring matrix, that is based on known solved protein-DNA complexes, gives more importance to the H-bonds and possibly underestimates the contribution of hydrophobic residues. In addition, Trp was underrepresented in the H-bonds network. Those two reasons could sometimes lead to slightly incorrect predictions. Also, the scoring matrix that is based on the

optimization of the parameters doesn't include all the possible amino acid-nucleic acid pairs and can lead to false predictions. Finally, part of the rules of our recognition code is based on known solved ZFP-DNA complexes and thus might not be so efficient when applied to unknown data. In the future, this limited statistics could overcome with more solved ZFP/DNA complexes, that will give more insights into the non-canonical binding.

From the dataset of the 17 ZFP-DNA complexes, ZNF217 and Kaiso are not binding in a canonical way, and as a consequence every method and recognition code failed to characterize their specificity. In particular, in the structure of Kaiso, a tail of amino acids in the C-terminal binds to the minor groove of the DNA. This tail might also occur in other ZFPs and could participate in the DNA-binding and affect their specificity. Moreover, ZNF217 has multiple long linkers (over 10 aa) between its ZF domains, a non-canonical DNA-binding and low specificity (Vandevenne et al., 2013). It is possible that a long linker creates subtle displacements in the flanking ZF domain. Hence, the binding specificity might be affected in ZFPs that contain long linkers or a tail of polar amino acids in their C-terminal, or other non-canonical arrangement that have not been observed yet in high-resolution ZFP structures.

Moreover, ZFPs can bind to a variety of DNA sequences with different degrees of specificity. It is also possible that all the ZF domains that have polar amino acids in their binding positions can bind to DNA, but it is not always visible with experiments. As a consequence, this can have affected the DNA targets that have experimentally defined for a given ZFP. Also, some ZFPs could use different ZF domains of their extended arrays to bind to different DNA targets: for example, Zfp335 recognizes two consensus motifs using separate ZF arrays (Han et al., 2016). In most of the results produced by existing recognition codes, ZFPs with fewer ZF domains are the ones for which is easier to predict their DNA target sequence. This could be due to the fact that is easier to find their "ideal" DNA sequence in the genome than in the poly-ZF proteins. For example, for a tandem of 3 ZF domains the potential DNA binding sequence is 9 nucleotides length and therefore  $4^9$  possible combinations, while for a tandem of 7 ZF domains the DNA binding sequence is 21 nucleotides length and therefore  $4^{21}$  possible combinations. Hence, it will be easier for the shorter tandem to detect and bind on its set of DNA sequences.

Furthermore, some amino acids in the binding positions seem to generate higher specificity for DNA targets than others. Specifically, Arg, Lys, His, Asp, Asn, Glu and Gln have strong preferences in specific nucleotides, on the contrary Ser and Thr appear to be more promiscuous. In addition, the binding position of an amino acid could play a role to its specificity. In particular, based on our analysis, using the amino acids in position 2 to predict which ZF domains bind on the DNA, didn't improve our prediction, although they bind to the DNA. This could be an indication that amino acids in position 2 provide more stability than specificity to the ZFP-DNA complex. Moreover, position 2 is usually occupied by a Ser with no strong preference to a specific nucleotide (can form H-bonds with all four nucleotides). Also sometimes, the amino acid in position 2 binds to a nucleic pair, in collaboration with the amino acid in position 6 of the previous ZF domain, and this could have an effect on the binding recognition.

As it is, our recognition code and others fail to predict the DNA target of ZFP that have shown to bind to a long DNA target. For example, in the case of ZNF648, ZNF765, ZNF93 and ZNF649 that bind to primate-specific L1 sequences (length ~45 bps), recognition codes fail to predict these sequences (Imbeault et al., 2017) (data not shown). Also, the recognition codes have inaccurate predictions (data not shown)

in the case of ZFP568 that binds on a 24 bps sequence (Patel et al., 2018). The crystal structure of the complex shed light on the binding of ZFP568 to its DNA target. Only three ZF domains bind canonically to the DNA and can be accurately predicted by the recognition codes. The remaining seven ZF domains bind non-specifically (phosphate backbone or contacts with the opposite strand in a shape-readout context) or they recognize 2 or 4 bps. The ZF domains of ZFP568 that follow the canonical binding occur at the N- and C- terminal of the ZF tandem and contain Arg that bind to guanine. This could be an indication that the binding starts from the ZF domains that are at the edges of the tandem and contain amino acids that have high affinity for a specific nucleic acid (like Arg have high specificity for Gs). This is an evidence that besides the canonical binding, it is possible to have different binding modes. But in order to investigate these alternative modes, more X-ray structures of ZFP/DNA complexes are needed.

Most of the ZFPs in this analysis are KRAB-ZFPs, which indicates that they contain at least one KRAB domain in the N-terminal. KRAB-ZFPs is one of the largest families of transcription factors in higher vertebrates, and through the interaction of KRAB domain with its universal cofactor KAP1, they regulate gene repression. Here, what needs to be taken into consideration, is the role of the KRAB domains. In this study they don't seem to affect the DNA specificity of the ZF domains. On the other hand, it would be important to investigate how the various KRAB domains interact (if they do so) with the DNA or other protein complexes (e.g., KAP1).

In conclusion, our computational structure-based approach gives results that are very similar to the latest experimental-based recognition methods for ZFPs. It has the advantage to consider neighboring ZF domains and their effect on DNA binding specificity.

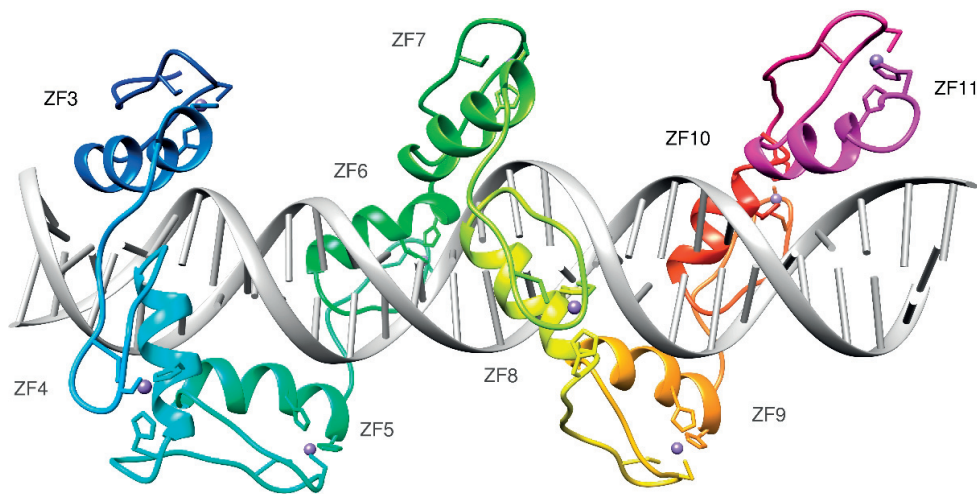


## Chapter 5      Analysis of the DNA binding landscape of human KRAB-ZFPs

### 5.1. Introduction

KRAB-ZFPs belong to one of the largest families of transcription factors in higher vertebrates (Bellefroid et al., 1991). They contain one or two KRAB domains in the N-terminal segment and a tandem of C2H2 ZF domain repeats in the C-terminal one. KRAB-ZFPs are one of the most recent and fastest growing gene families in primates. This expansion possibly enables primates to repress newly emerged transposable elements in embryonic stem cells (Jacobs et al., 2014). This was also confirmed by a recent phylogenetic and genomic study on human KRAB-ZFPs (Imbeault et al., 2017).

In the previous chapter, we proposed a recognition code between ZFPs and DNA, based on the assumption that ZF domains bind canonically on the DNA (amino acids in positions -1, 3 and 6 interact with 3 bps). Our recognition code is efficient in predicting the DNA targets for canonical binding, but fails in cases where the binding is more complicated. Likewise, it seems that the current recognition codes cannot predict accurately the DNA targets for many ZFPs. For example, in the study of Imbeault et al. the experimentally defined DNA binding motifs of ZNF248, ZNF649, ZNF765 and ZNF93 (KRAB-ZFPs that bind on primate-specific L1 elements), don't match with the motifs that are predicted by the recognition codes. This is an indication of non-canonical binding, but it needs confirmation by the solved structures of these complexes. In addition, the binding motif of ZFP568 is different from the predicted one (Patel et al., 2018). Nevertheless, in this case we have a crystal structure that confirms the non-canonical binding (Figure 19). Specifically, three out of nine binding domains bind canonically and their predicted DNA motifs match with the experimental ones. For the remaining six ZF domains, the binding mode varies. It seems that recognition codes cannot always predict the DNA targets of the ZFPs since the binding modes are very complex and we still miss exhaustive structural information. Here our aim is to recover some of this missing information about non-canonical binding. In particular, we explored other possible ways that a ZFP can bind on its DNA target, besides the canonical binding by the analyzing the DNA binding motifs from Imbeault et al. study.

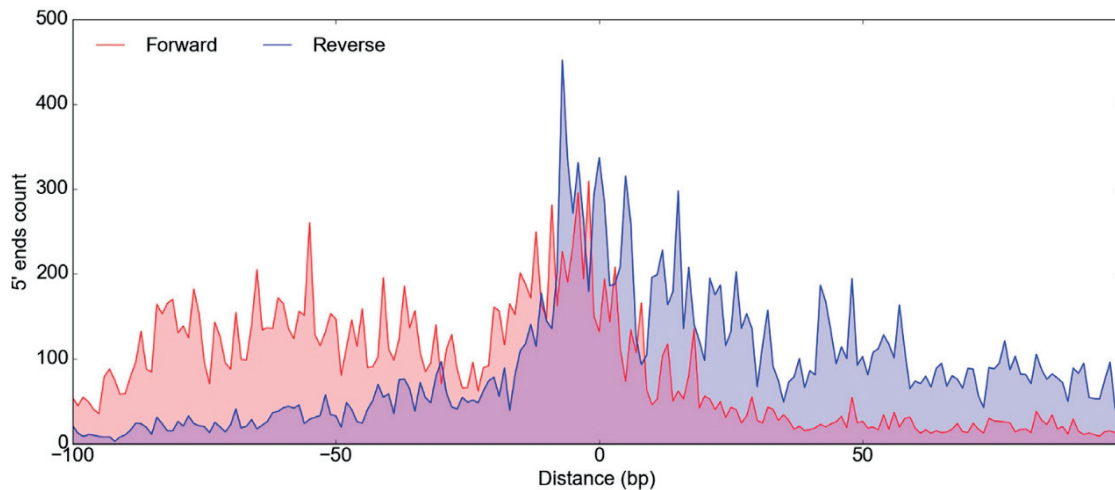


**Figure 19** - Crystal structure of the mouse ZFP568, ZF3-ZF11 (pdb ids: 5v3m and 5wjg). ZF3, ZF10 and ZF11 bind canonically to the DNA. The DNA is shown in gray, the zinc ions in purple and the tandem of ZF domains in rainbow.

## 5.2. Results

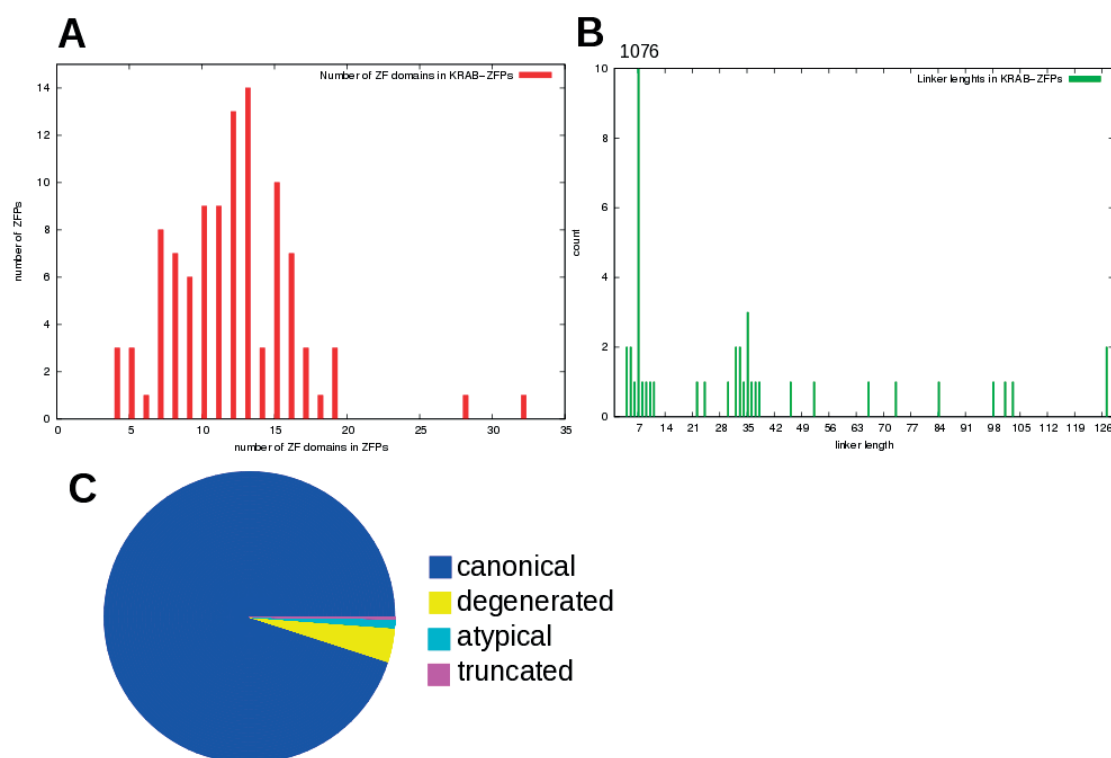
In the study of Imbeault et al. in 2017, the DNA targets of 325 human KRAB-ZFPs were determined by performing ChIP-exo (Chromatin Immunoprecipitation with exonuclease digestion). ChIP-exo is a method to identify genomic location of DNA-binding proteins at near single nucleotide accuracy (Figure 20). This method is a high-throughput sequencing approach that generates large data sets. To analyze all these data and predict DNA-binding sites, computational methods are used (peak calling methods). In the study of Imbeault they used the MACS (Model-based Analysis of ChIP-Seq) software, which gives robust and high resolution ChIP-Seq peak predictions (Zhang et al., 2008). In order to identify sets of highly significant peaks they used three different thresholds to filter out peaks: MAC scores 50, 80 and 100. Higher scores are associated with lower, more significant p-values and with more reads in the peak. For example, MACS score 80 indicates that the p-values are filtered to be less than  $1 \times 10^{-8}$ . For each of the three MACS score thresholds, we were provided with the genomic coordinates of the peaks (bed files), from which we extracted the sequence fragments.





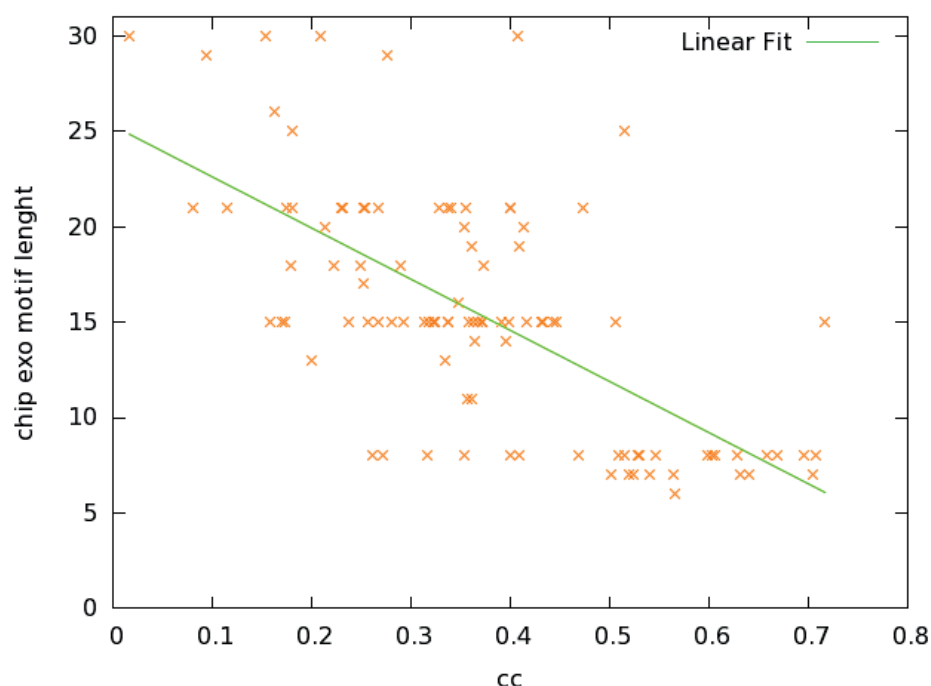
**Figure 20** - ChIP-exo reads of ZNF765. Binding regions have characteristic peak shape. The more the forward and reverse strands overlap, the higher is the motif enrichment.

We analysed the ChIP-exo results of the 325 KRAB-ZFPs for the three MACS scores. Initially, we used the MEME-ChIP server in order to discover the different binding motifs and PWMs. Next, we compared these binding motifs, for each KRAB-ZFP, with the three thresholds. We decided to continue our analysis with 102 KRAB-ZFP that had the same predicted DNA binding motif, for the thresholds of 80 and 100 MACS score. We excluded the MACS score 50 since it had high number of false positive predictions and we aimed in keeping the most significant and therefore high quality predictions. These 102 KRAB-ZFP can be a representative set of KRAB-ZFPs, since they contain 1216 ZF domains. These domains include 1157 canonical zinc finger domains (C-X<sub>2-5</sub>-C-X<sub>12-15</sub>-H-X<sub>3-5</sub>-H motif), 12 atypical zinc finger domains (the spacing between Cys or His residues is slightly altered), 46 degenerate zinc finger domains (one or two Cys or His are replaced by other amino acids), and 1 truncated zinc finger domain (half ZF domain). While 80 ZFPs contain only canonical linkers (linker length less than 15 amino acids) and 22 ZFPs contain at least one long linker (Figure 21).



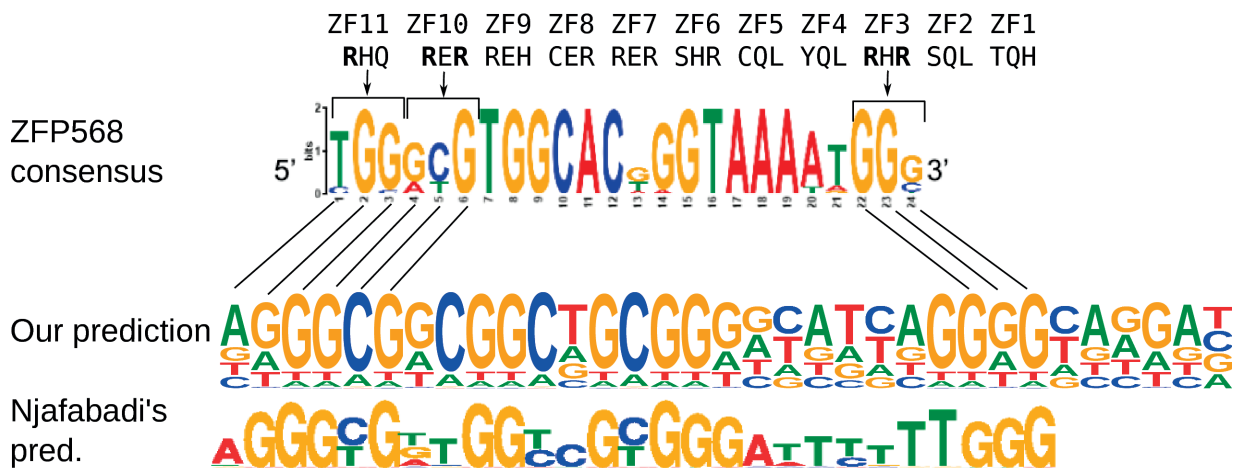
**Figure 21** - Statistics of the remaining 102 human KRAB-ZFPs. A. Number of ZF domains in ZFPs. Most ZFPs contain 12-13 ZF domain. B. Number of linkers in ZFPs. Most of the linkers between neighboring ZF domains are seven amino acids long. C. Types of ZF domains in ZFPs. 95% of the ZF domains are canonical.

For these 102 KRAB-ZFPs we predicted their DNA binding motif (PWM), using our ZFP-DNA recognition code. Next, we compared the motifs predicted by MEME with the motifs predicted by our recognition code. For the moment, we used only our recognition code, but other recognition codes could be used for further validation. However, for each KRAB-ZFP, we used multiple overlaps of the MEME PWM and our recognition's code PWM, in order to maximize their correlation coefficient and output a matrix alignment (both strands and no gaps). 25 motifs have a high correlation of over 0.5 (Figure 21), which is an indication that our recognition code can provide DNA binding motifs with high accuracy (Figure 22). Another thing to be noted is that shorter MEME motifs seem to correlate better with the predictions. This could be an indication that if fewer ZF domains bind to DNA the more canonical is the binding, and therefore easier to predict its binding mode.



**Figure 22** - Correlation between the length of the ChIP-exo motif and the correlation coefficient (cc) between our prediction and the MEME motifs. For shorter ChIP-exo motifs we have better predictions.

From the analysis above, it seems that for shorter ChIP-exo motifs we have better predictions. In addition, 25 ZFPs could possibly bind canonically, but for the remaining 77 the binding mode could be non-canonical. For example, mouse ZFP568 with 11 ZF domains makes numerous non-canonical ZF-DNA interactions (Patel et al., 2018). In particular, the third and the two last ZF domains bind canonically to the DNA (1 ZF domain binds to 3 bps), while the 6 ZF domains that are in the middle interact with 2, 3 or 4 bps per ZF domain, and few also interact with the opposite strand (Figure 23). In the case of ZFP568, ZF domains near the ends of the tandem, that contain Arg, follow the canonical binding. This could be an indication that the binding starts from the ends of the tandem and ZFs that contain Arg bind with high specificity to a G in the DNA. These ZF domains could lead the binding, while the remaining ZF domains possibly follow by making non-canonical contacts.



**Figure 23** - ZFP568 consensus DNA-binding sequence. ZF3, ZF10 and ZF11 binding canonically to the DNA. The predictions of two recognition codes agree with the consensus at these positions.

Next, we decided to truncate the ZFPs and calculate the cc between our prediction and the ChIP-exo predictions. In particular, we truncated the ZFPs in 2, 3 and 4 overlapping ZF domains and for each truncated ZFP we aligned with the predicted PWM with the ChIP-exo predicted PWM. We calculated the cc for both data sets (potential canonical and non-canonical binding) and as successful alignment where consider the ones with cc over 0.75 (Table 9):

	Potential canonical binding ZFPs	Potential non-canonical binding ZFPs
<b>Number of ZFPs</b>	25	77
<b>2-ZF domains (Number of ZFPs with cc &gt; 0.75)</b>	9 (36%)	31 (40%)
<b>3-ZF domains (Number of ZFPs with cc &gt; 0.75)</b>	3 (12%)	11 (14%)
<b>4-ZF domains (Number of ZFPs with cc &gt; 0.75)</b>	1 (4%)	5 (6%)

**Table 9** - Alignments of truncated ZFPs.

It seems that for both data sets we have similar results and there is not significant difference between the cc scores for the alignments between canonical and non-canonical binding ZFPs. Moreover, the alignments are more successful with shorter ZFPs, while when the ZFPs are long is difficult to have a proper alignment.

### Analysis of the ZF domains

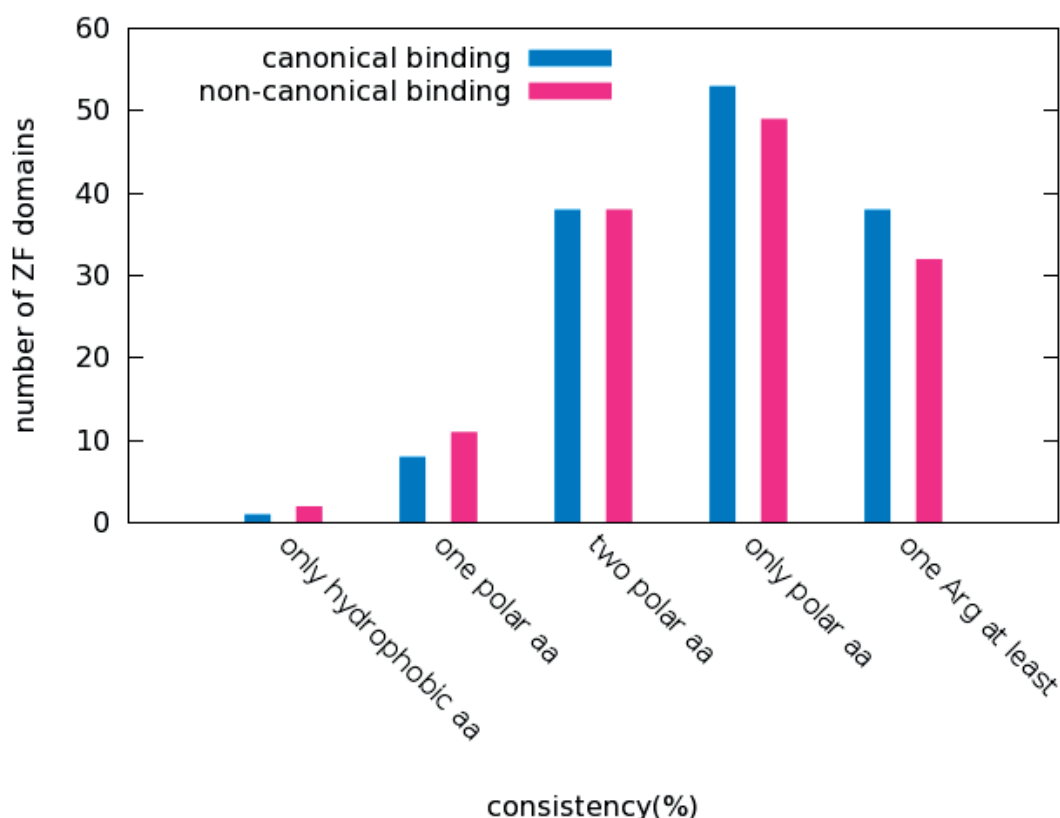
In the dataset of 77 ZFPs that potentially bind non-canonically to the DNA, we

decided to check if there are ZF domains that could bind at the edges of the MEME motif. As edges of the motif, we considered the first and last two-nucleotide triplets. In particular, we aligned the PWM for each of the 907 ZF domains (a 3x4 matrix) with 4 different PWM (3x4 matrices of the first and last two nucleotide triplets). We considered as a strong indication for binding only if the cc between the two aligned matrices is over 0.9. In our case, we have 32 out of 77 ZFPs that have at least one ZF domain that could bind at the edges of the motif (cc > 0.9). In addition, 15 out of these 32 ZFPs have at least one Arg in the binding positions of their ZF domain that is predicted to bind to the edges of the motif. This could be an indication that 40% of non-canonical binding ZF domains potentially bind at the edges of the motif, but not always Arg residues are involved in the binding.

Next, we checked if the residues in the 3 binding positions contain Arg residues, and in general, if they tend to contain more polar than hydrophobic amino acids. The results are presented in the next table (Table 9) and in Figure 24:

	<b>Potential canonical binding ZFPs</b>	<b>Potential non-canonical binding ZFPs</b>
<b>Number of ZFPs</b>	25	77
<b>Number of ZF domains</b>	308	907
<b>Number of domains containing only hydrophobic aa</b>	4 (1%)	15 (2%)
<b>Number of domains containing 1 polar aa</b>	25 (8%)	104 (11%)
<b>Number of domains containing 2 polar aa</b>	116 (38%)	342 (38%)
<b>Number of domains containing only polar aa (3 polar aa)</b>	163 (53%)	446 (49%)
<b>Number of domains containing one Arg at least</b>	117 (38%)	287 (32%)

**Table 10** - Consistency of the 3 binding positions



**Figure 24** - Consistency of the 3 binding positions

From the analysis above, it seems that there are very few ZF domains that contain only hydrophobic amino acids in their three DNA-binding positions. In addition, there are few ZF domains that contain 1 polar and 2 hydrophobic amino acids. This is compatible with the nature of ZF-DNA binding, since not only H-bonds tend to be stronger than hydrophobic interactions, but they also induce specificity between ZFPs and the DNA. In addition, the potential canonical binding ZFPs have higher consistency of Arg residues and in general of polar amino acids, than the potential non-canonical binding ZFPs. This could partially explain why the potential canonical binding ZFPs tend to bind with a short tandem of ZF domains and each ZF domain interacts with 3 bps. On the contrary, the potential non-canonical binding ZFPs have less polar amino acids, therefore more disturbed binding, and possibly lower specificity and affinity.

In order to further investigate the specific physicochemical properties of the amino acids in each position, we used the PsychoProt server (Abriata et al., 2016). PsychoProt detects, quantifies and maps on the sequence the biophysical and biochemical traits that shape amino acid preferences throughout a protein. In particular, it analyzes tolerance to substitutions at the positions of a given protein sequence, searching for correlations with specific amino acid descriptors. We used PsychoProt for analysing the ZF domains of 102 ZFPs, which are already separated

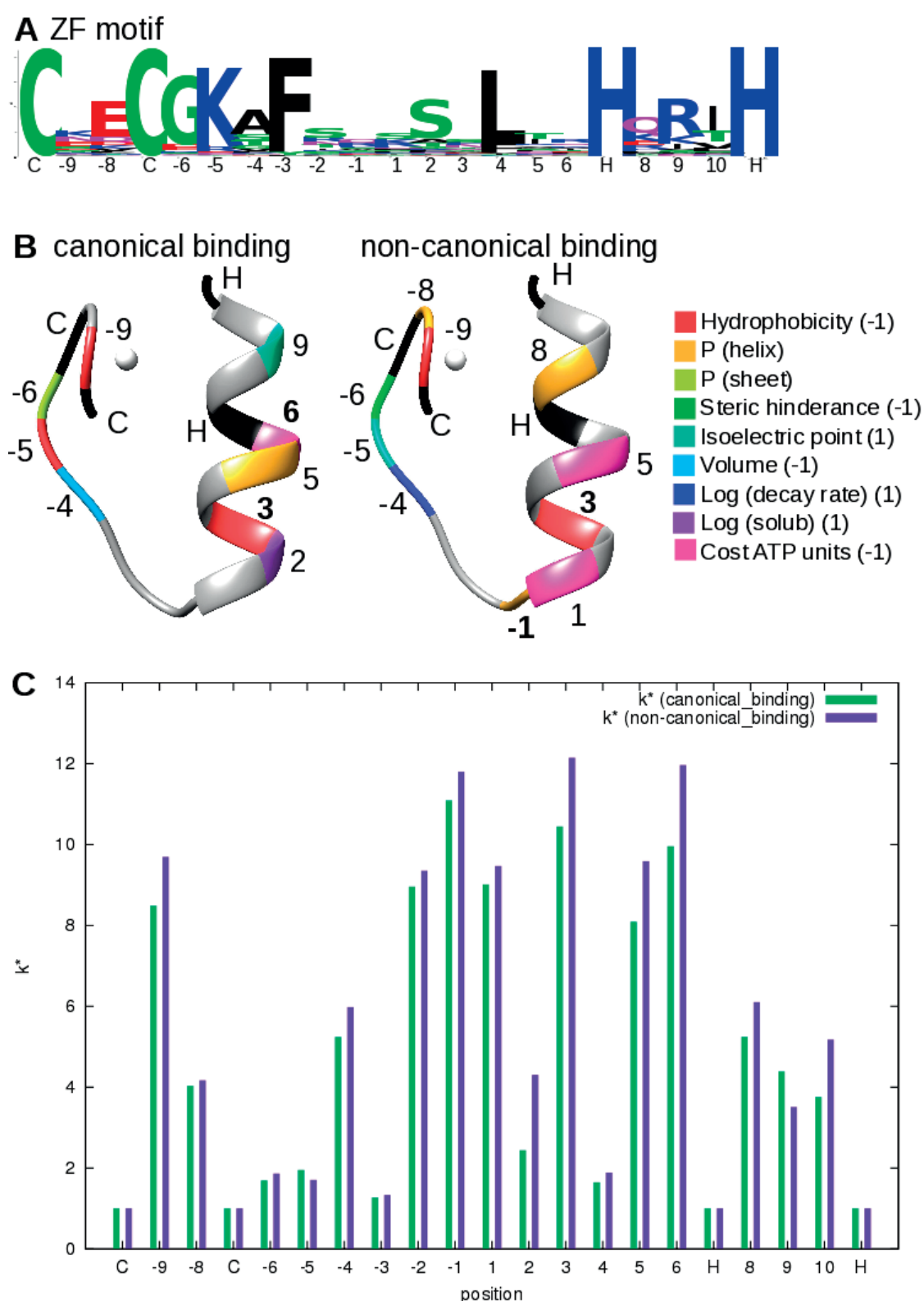
in two data sets that represent ZF domains that potentially bind canonically and non-canonically. For each of the two data sets, we aligned the 1138 ZF domains of 21 amino acids length that follow the CX<sub>2</sub>CX<sub>12</sub>HX<sub>3</sub>H motif. The data set of the potential canonical binding has 281 canonical ZF domains of 21 amino acids length, while the data set of the potential non-canonical binding has 857 canonical ZF domains of 21 amino acids length. Initially, for both data sets, we created the amino acid variability matrix from the alignment of the 21 amino acid length ZF domains. The amino acid variability matrix is the input file for PsychoProt and it consists of a number of rows equals to the length of the protein sequence and 24 columns, which include residue name, residue number. In addition, there are 20 columns with  $\Delta\Delta G$  values (distribution of each amino acid compared to a reference sequence) and a column of the  $k^*$  value (effective number of amino acids that appear at each position). The  $\Delta\Delta G$  is 0 for the reference sequence amino acid at a given position, negative for those substitutions that appear at a higher frequency than the reference amino acid and positive for those at lower frequency than the reference amino acid. The range for the  $k^*$  value is from 1 for a position fully restricted to one amino acid to 20 for a position in which any amino acid is equally represented. Next, we provided these matrices into PsychoProt's tool, to fit the amino acid variability with the physicochemical descriptors. The threshold of the p-value for fit selection was set to 0.05, in order to increase the fitness. The results are summed up in the following table and in Figure 25:

Position	Descriptor (potential canonical binding)	$k^*$ (potential canonical binding)	Descriptor (potential non-canonical binding)	$k^*$ (potential non-canonical binding)	Correlation Coefficient
Cys	-	1	-	1	1
-9	Hydrophobicity (-1)	8.49	Hydrophobicity (-1)	9.69	0.81
-8	no fit	4.03	P (helix)	4.17	0.89
Cys	-	1	-	1	1
-6	P (sheet) (-1)	1.69	Steric hindrance (-1)	1.86	0.81
-5	Hydrophobicity (-1)	1.95	Isoelectric point (1)	1.71	0.81
-4	Volume (-1)	5.24	Log (Decay rate) (-1)	5.98	0.87
-3	no fit	1.27	no fit	1.33	0.9
-2	no fit	8.96	no fit	9.35	0.93

<b>-1</b>	<b>no fit</b>	<b>11.09</b>	<b>P(helix) (1)</b>	<b>11.8</b>	<b>0.93</b>
1	no fit	9.01	Cost ATP units (-1)	9.47	0.9
2	Log (solub) (1)	2.44	no fit	4.3	0.51
<b>3</b>	<b>Hydrophobicity (-1)</b>	<b>10.44</b>	<b>Hydrophobicity (-1)</b>	<b>12.14</b>	<b>0.86</b>
4	no fit	1.65	no fit	1.89	0.9
5	P(helix) (1)	8.09	Cost ATP units (-1)	9.59	0.78
<b>6</b>	<b>Cost ATP units (-1)</b>	<b>9.95</b>	<b>no fit</b>	<b>11.96</b>	<b>0.64</b>
His	-	1	-	1	1
8	no fit	5.24	P(helix) (1)	6.1	0.88
9	Isoelectric point (1)	4.39	no fit	3.51	0.77
10	no fit	3.75	no fit	5.17	0.78
His	-	1	-	1	1

**Table 11** - PsychoProt's descriptors and  $k^*$  values. In the first column, we have the sequence of a C2H2-ZF domain, with respect to the position numbering for each amino acid. With bold we have highlighted the binding positions (-1, 3 and 6). With gray we show the coordinating 2 Cys and 2 His. In the 2<sup>nd</sup> and 4<sup>th</sup> column we have the descriptors that had the best fit for the specific position. In the 3<sup>rd</sup> and 5<sup>th</sup> columns we show the  $k^*$  values (the higher the  $k^*$  value the higher the variability). In the 6<sup>th</sup> column there is the correlation coefficient between the  $\Delta\Delta G$  values for the 20 amino acid for the data sets of the canonical and non-canonical binding.





**Figure 25** - PsychoProt's descriptors. A. Logo of the 21 amino acid length zinc finger motif and the corresponding positions. B. The physicochemical properties of the amino acids mapped on the 3D ZF structure, for the potential canonical and non-

canonical binding. C.  $k^*$  values for the 21 positions. With purple there are the  $k^*$  values for the non-canonical binding data set and with green the  $k^*$  values for the canonical binding data set.

In the cases where the  $k^*$  value is 1 is when we have the presence of only one type of amino acid in that specific position, i.e. the coordinating Cys and His. On the opposite, in the cases that we have a high  $k^*$  value ( $\sim >10$ ) we have a variety of different amino acids, like in the DNA-binding positions -1, 3 and 6. The amino acids in these positions induce binding specificity, and therefore, they should have a variability in order to bind to the different DNA motifs. Also, for both data sets, in the position -9 we have a preference for polar amino acids. For the position -8 in the potential non-canonical binding, we have a preference for amino acids that are favored to be in an  $\alpha$ -helical configuration, although there is a loop in the structure. For the position -6 in the potential canonical binding we have a preference for amino acids that are not favored to be in a  $\beta$ -strand configuration, while in the potential non-canonical binding we have a preference for amino acids with reduced steric hindrance. Moreover, in position -6 there is mainly a Gly and therefore we have reduced steric hindrance, since it's a small amino acid. In addition, Gly is favored for loops, like the one here, and not for  $\beta$ -strand configuration, since it creates kinks. For the position -5 in the potential canonical binding we have a preference for amino acids that are polar, while in the potential non-canonical binding we have a preference for amino acids that could be involved in salt bridges and hydrogen bonds on the polar surface. Specifically, we have mainly a Lys for this position. For the position -4 in the potential canonical binding we have a preference for amino acids that have small volume, while in the potential non-canonical binding we have a preference for amino acids that have low decay rate. For the positions -3 and 4 the  $k^*$  values are close to 1 and we have the presence of Phe and Leu, respectively. In particular, these two amino acids form a hydrophobic core and they help to stabilize the C2H2 zf motif. For the binding position -1 in the potential non-canonical binding we have a preference for amino acids that are favored to be in an  $\alpha$ -helical configuration. In fact, this position is the beginning of the  $\alpha$ -helix in the zf structure. For the position 1 in the potential non-canonical binding, we prefer amino acids that they have low cost in ATP units. For the position 2 in the potential canonical binding we have a preference for amino acids that are soluble. The amino acids in this position sometimes contribute to the DNA binding. Usually a Ser can form H-bonds with the DNA. For both data sets, in the binding position 3 we have a preference for polar amino acids, which induces DNA binding. For the position 5 in the potential canonical binding we have a preference for amino acids that are favored to be in an  $\alpha$ -helical configuration (there is  $\alpha$ -helix in the structure), while in the potential non-canonical binding we have a preference for amino acids that have low cost in ATP units. For the binding position 6 in the potential canonical binding, we have a preference for amino acids that have low cost in ATP units. For the position 8 in the

potential non-canonical binding, we have a preference for amino acids that are favored to be in an  $\alpha$ -helical configuration (there is  $\alpha$ -helix in the structure). For the position 9 in the potential canonical binding, we have a preference for amino acids that could be involved in salt bridges and hydrogen bonds on the polar surface, mainly Arg and Lys.

The  $k^*$  values are in general a bit higher for the non-canonical (avg  $k^* = 6.47$ ) than the canonical (avg  $k^* = 5.75$ ) binding data set. This could be due to the fact that the data set of the potential canonical binding has much less canonical ZF domains of 21 amino acids length (281) than the data set of the potential non-canonical binding (857). The above could have an impact on the variability of amino acids on the first set. Also, for the DNA-binding positions (-1, 3 and 6) where there is a variation of amino acids, we have the highest  $k^*$  values, while for the coordinating amino acids (Cys and His) we have the lowest  $k^*$  values ( $=1$ ).

Moreover, we clustered the C2H2-ZF domains of the 102 ZFPs through a phylogenetic tree. We aligned 1138 ZF domains of 21 amino acids length, which follow the  $CX_2CX_{12}HX_3H$  motif. In particular, we made an alignment for the positions -2 to 6 (8 amino acids length). To build the phylogenetic tree we used the PhyML V3.0 software, which is based on the maximum-likelihood principle (Guindon et al., 2010). The results of this analysis did not show any evolutionary relationship between the potential binding domains and the potential non-binding domains.

### Analysis of the linkers between the ZF domains

From the data set with the 102 ZFPs, 7 out of 25 (28%) of the potential canonical binding ZFPs have at least one long linker, while 14 out 77 (18%) the potential non-canonical binding ZFPs have at least one long linker. This could be an indication that since long linkers separate the ZF tandem in smaller groups of ZF domains they make it easier for shorter tandems to find their DNA-target.

### 5.3. Conclusions and perspectives

We analysed the DNA target motifs of 102 human KRAB-ZFPs from ChIP-exo experiments and we compared them the predicted DNA targets produced by our recognition code. Moreover, other recognition codes (i.e. Najafabadi et al.) could be used for further validation.

In general, for ZFPs with poly-ZF domains, the DNA-binding is not always canonical. Here, we noticed that shorter ChIP-exo DNA motifs correlate better with our predicted DNA motifs, and this could be an indication for canonical binding, since canonical binding is easier to predict. In ZFPs that bind to the DNA, the consistency of the amino acids in the DNA binding positions appears to be crucial. From our

analysis, canonical binding requires more polar amino acids (i.e. Arg) to be present in the binding positions. In addition, ZFPs that bind canonically tend to use fewer ZF domains in a tandem and therefore, bind on shorter motifs. This could partially explain the fact that there are more long linkers between ZF domains in ZFPs that bind canonically, so that their ZF domains are separated in shorter tandems. Therefore, the more ZF domains in a tandem, the more the binding is disturbed, leading to a non-canonical binding. All the above, could be used to improve the current recognition codes, in order to take into consideration non-canonical binding.

To investigate more other binding modes, more data from ChIP analysis and solved structures would be useful. Consequently, in order to discover new binding modes, we should prioritize the crystallization of ZFP complexes with poly-ZF domains, no long linkers in between and low consistency of polar amino acids in the binding positions. In addition, the improvement of motif discovery algorithms is important, in order to get more significant and accurate predictions on the DNA binding motifs.

## Chapter 6      Conclusions and perspectives

Protein-DNA interactions are involved in many biological functions; there are proteins that recognize specific DNA bases (i.e. ZFPs) and proteins that recognize a specific DNA shape (Rohs et al., 2010). Therefore, in order to understand these functions, it is important to know the structures of the protein-DNA complexes that are involved. Since there are not so many experimentally solved protein-DNA structures, computational and experimental methods are combined through Integrative Modeling, in order to build atomistic models of protein-DNA complexes and study their interactions. To predict the binding residues on the protein-DNA interface sequence-based or structure-based approaches are used (Si et al., 2015). Structure-based methods make more accurate predictions than the sequence-based methods, since structure is more conserved than the sequence. Many of these methods follow machine learning approaches, but either they are limited to predict if a protein can bind to the DNA, either they predict the protein binding interface with low accuracy.

In this work, I followed an Integrative Modeling approach to predict the interface of various protein-DNA complexes (structure-based approach). In particular, we used a machine learning approach (ANN) to predict the correct interface among different orientations of a protein on the DNA, of known complexes. The predictions of the binding-interfaces were of high accuracy. This method can be used to discriminate the correct complexes from the rest, in an unknown set of complexes, given a DNA-binding protein and a DNA model. Therefore, it can predict the interacting interface of a DNA-binding protein, without knowing the structure of the complex. Unfortunately, our method cannot distinguish between proteins that recognize specific DNA bases and proteins that recognize a specific DNA shape. In addition, more descriptors of the interface, derived from structural information, could help to improve our predictions. Therefore, a combination of various descriptors and the state-of-art machine learning methods, could improve the accuracy of our predictions. Furthermore, not only the static properties of protein-DNA complexes should be considered, but also the conformational changes that occur during the formation of the complex.

An important family of DNA-binding proteins are the ZFPs that act as TFs and participate in gene silencing. In the last years, various recognition codes between ZF domains and the DNA have been proposed. Although there are different experimental and computational approaches followed, these recognition codes cannot capture the complexity of the DNA binding landscape.

Here, we defined a computational structure-based method, which gives results that are very similar to the latest experimental-based recognition methods for ZFPs. Our recognition code between ZFPs and the DNA is used to predict which ZF domains of multi-domain ZFPs bind to their DNA target. This recognition code is also used to predict the most probable DNA target sequence of any ZFP. Our method

showed that for a ZF domain that recognize 3 nucleotides in a row (canonical binding), the binding specificity can be predicted. It also includes the effect of the neighboring ZF domains on binding specificity, which is not always taken into consideration from other recognition codes. Moreover, our recognition code was tested on a data set of 17 solved ZFP-DNA complexes and 2 ZFPs with unsolved structures. It predicted successfully the DNA binding targets in both data sets (where the binding is canonical). However, in the future we should test our recognition code with more unsolved complexes. Other limitation is that our code can predict only the canonical binding mode. Similarly, the other recognition codes model only the canonical binding. Therefore, more solved ZFP-DNA complexes are needed, in order to have more information about the other non-canonical binding modes. Since most of the ZFPs in this study contain a KRAB domain, a further extension to these recognition codes could be the investigation of the KRAB domain and its role on DNA specificity.

In some ZFPs with poly-ZF domains, the DNA-binding seems to be disturbed and follow non-canonical patterns. Here, we tried to investigate these other non-canonical binding modes of ZFPs. For this reason, we analysed the DNA target motifs of 102 human KRAB-ZFPs from ChIP-exo experiments and compared them with the predicted DNA target motifs produced by our recognition code.

In particular, we noticed that shorter ChIP-exo DNA motifs correlate better with our predicted DNA motifs and this could be an indication for canonical binding, since canonical binding is easier to predict. In addition, the type of the amino acids in the DNA binding positions seems to affect the binding mode. For example, canonical binding requires more polar amino acids for the binding positions. It appears that ZFPs that bind canonically tend to use fewer ZF domains in a row tandem and bind on shorter motifs. Also, the higher is the number of ZF domains in a poly-ZF ZFP, most probable the binding will be non-canonical. These conclusions should be validated by comparing the ChIP-exo predicted motifs with the predicted motifs from other recognition codes. Thus, our analysis could be used to extend the current recognition codes. To further investigate, it would be useful to have more ChIP data results and solved ZFP-DNA complexes. Nevertheless, the determination of which ZF domains and how do they bind to the DNA is a challenge.

To conclude, the prediction of the protein-DNA binding sites and of the DNA target sequences is very important. In particular, by knowing more about protein-DNA complexes, we can have an insight into transcription, chromosome packaging and other functions that protein-DNA interactions are involved. The last years there have been many approaches, for the various DNA-binding protein families, and specifically for the KRAB-ZFP family. Nevertheless, the robustness of the predictions of these approaches is average. One reason is that the DNA-binding domains of the proteins have many degrees of freedom, in contrast to the DNA that has limited degrees of freedom. To improve these predictions, both structure and sequence should be taken into consideration. In addition, there are many different types of DNA-binding

domains and it is challenging to predict how they will bind to the DNA. Therefore, it would be better to develop prediction methods for different types of DNA-binding proteins (i.e. recognition code between ZFPs and DNA). Also, more descriptors of the interface should be included in these methods (i.e. H-bonds).

Although, studying protein-DNA interactions is challenging, in the next years the increase of the computational power and the number of solved protein-DNA complexes, will lead to better predictions on what concerns the field of the protein-DNA interactions. This will give an insight into their structures and therefore it will lead to better understanding of their functions.

## Chapter 7      Appendix – Related works addressing protein-DNA systems

### 7.1. Molecular modeling of the KAP1 complex

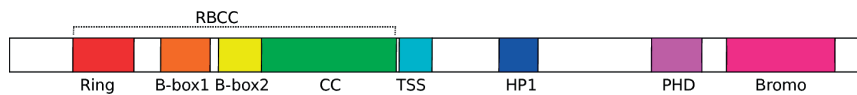
(Fonti G., Marcaida M.J, Bryan L.C., Traeger S., Kalantzi A.S., Helleboid P.Y.J.L., Demurtas D., Tully M., Grudinin S., Trono D., Fierz B., Dal Peraro M. KAP1 is an antiparallel dimer with a natively functional asymmetry – submitted, preprint available on at *bioRxiv* 553511; doi: <https://doi.org/10.1101/553511>)

The KRAB-ZFP/KAP1 complex is an important regulator of the genome. Unfortunately, the structures and functions of the proteins that participate in this complex are partially unknown. Our goal was to provide biomolecular models of the proteins that participate in the KRAB-ZFP/KAP1 complex in order to investigate its functions. Here, we followed a computational approach and propose an atomistic model for the KAP1 protein, in order to reveal its structure and its stoichiometry. The aim of these *in silico* model is not only to have a better insight into the KRAB-ZFP/KAP1 complex, but also to be integrated with experimental data. Eventually, our model for KAP1 was integrated with Cryo-EM and SAXS experimental data. I contributed to this project by making the initial model for KAP1 and investigating its stoichiometry, through homology modeling.

#### The KRAB-ZFP/KAP1 complex

KAP1 (KRAB Associated Protein 1) or TRIM28 is a member of the TRIM family (Tripartite motif) and regulates transcriptional repression. Its structure is modular: at the N-terminal there are three zinc-finger domains (one RING (Really Interesting New Gene) followed by two B-boxes), and a CC (Coiled Coils) region, constituting the so-called RBCC domain (Figure 27). Experiments have revealed that the RBCC domain binds as a trimer to a single KRAB domain (Peng et al., 2000). The central region of KAP1 includes the TSS domain (TFF1 Signature Sequence) and the HP1-binding domain (PXVXL motif), while the remaining central region has not a well-defined structure. The C-terminal region contains a PHD domain (Plant Homeodomain) and a bromodomain that interact with chromatin-remodeling proteins, i.e. the PHD domain has E3 SUMO-protein ligase activity.



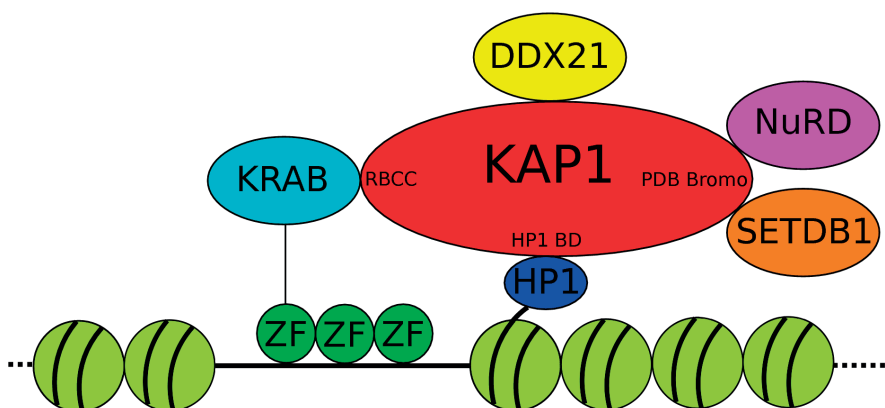


**Figure 27** - KAP1 domains. RBCC, TSS, HP1 and PHD-Bromodomain.

The KRAB-ZFP/KAP1 complex is a transcriptional network that participates in the control of the homeostasis of hematopoietic stem cells in higher vertebrates. It has been demonstrated that KRAB-ZFPs act as site-specific transcriptional repressors, which together with KAP1 mediate the silencing of ERVs and of murine leukemia virus in embryonic stem cells (Aktas et al., 2010; Wolf & Goff, 2007). As well, the highly conserved KRAB-ZFP family member ZFP57 is responsible for the maintenance of imprinting marks, including DNA methylation, during early embryogenesis, through its binding to a methylated hexa-nucleotide found in all murine and human imprinting control regions (ICR), where it and KAP1 recruit SETDB1, Dnmt1, Dnmt3a, Dnmt3b and NP95 (Li et al., 2009; Mackay et al., 2008). In addition, recently it has been shown that the KRAB/KAP1-mediated sequence-specific recognition of thousands of genomic loci, including ERVs, in embryonic stem cells leads to their DNA methylation, thus ensuring the genome-wide establishment of site-specific epigenetic marks that are subsequently maintained during development (Quenneville et al., 2013). Finally, KAP1 and/or specific KRAB-ZFPs have indeed been demonstrated to play an important role in tumorigenesis, in the control of behavioral stress and Parkinson disease, in the differentiation and activation of both the humoral and cellular arms of the immune system, and in the sex-specific metabolic control of hormones, drugs, and xenobiotics in the liver (Cammass et al., 2000; Jakobsson et al., 2008; Sio et al., 2013). Despite this emerging importance in biological processes and diseases, little is known about the specific interactions of the various proteins complexes that participate in this system.

KRAB-Zinc Finger Proteins and KAP1 can mediate long-range transcriptional repression through heterochromatin spreading (Groner et al., 2010). Literature suggests that KAP1 performs diverse functions by forming a complex assembly with chromatin-remodeling proteins (Mi2a, SETDB1, HP1 etc.), which allow it to regulate a heterochromatic microenvironment (Sripathy et al., 2006) (Figure 28). KAP1 recruitment to the genome is made by its RBCC (N-terminal RING finger/B-box/Coiled Coil) domain that can bind with the KRAB domain of KRAB-ZFPs, which target KAP1 to specific DNA sequences through their DNA-binding zinc fingers. Silencing by the KRAB-ZNF/KAP1 complex requires post-translational modification of KAP1 by SUMO on specific lysines in the KAP1 bromodomain. The adjacent PHD domain serves as an intramolecular E3 ligase for bromodomain sumoylation. The histone deacetylase complex NuRD and the histone methyltransferase SETDB1, leading to histone deacetylation, deposition of H3K9me3 and binding of HP1 recognize the attached SUMO moiety (Friedman et al., 1996). These proteins place chromatin repressive marks on histones and help to establish silenced state of KRAB-ZNF target genes. The SUMO modification of proteins in

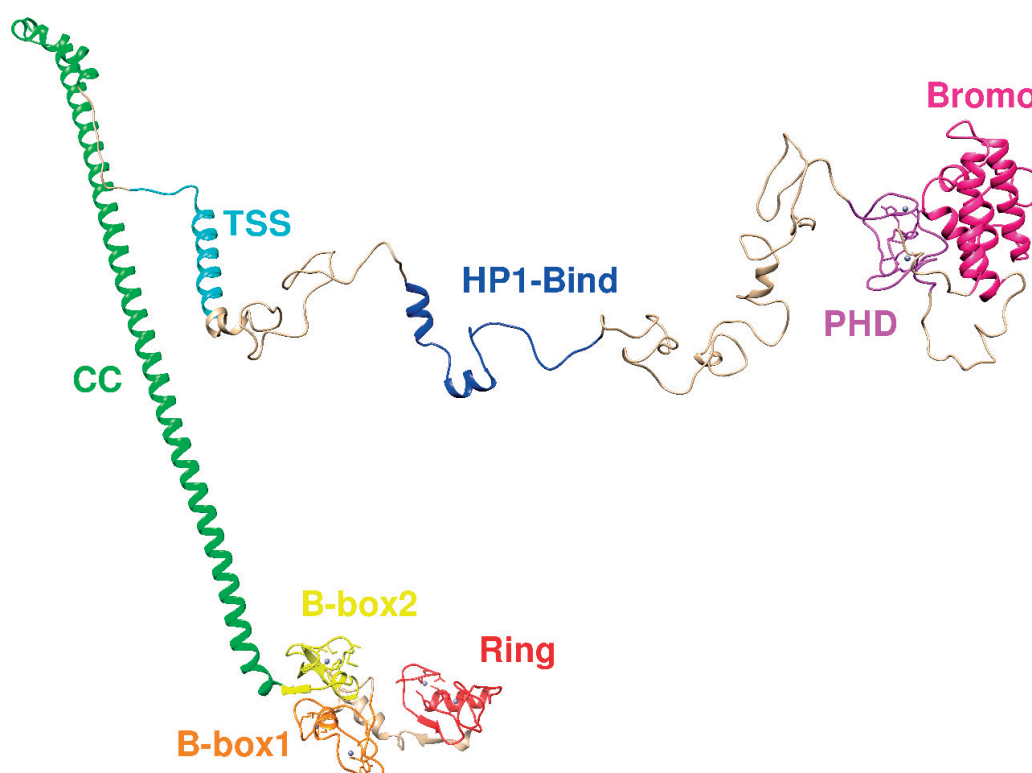
general is required for normal chromosome condensation and mitosis. Also, KAP1 mediates localized compaction of euchromatin to heterochromatin that is necessary for suppression of specific gene transcription, and that is associated with chromatin modifications (Iyengar & Farnham, 2011).



**Figure 28** - The KRAB-ZFP/KAP1 complex and its cofactors.

#### A molecular model for KAP1

KAP1 model is designed based on existing, homologous structures of RING, B-box and potential coiled  $\alpha$ -helical domains. The TSS and HP1-binding domains occurred by the combination of predicted and real structures, while the PHD-bromodomain is the only part that was previously solved. By combining different modeling techniques, we have built a structural model for the KAP1 (Uniprot: Q13263) (Figure 29). To construct the RBCC domain we used SWISS Model server (Schwede et al., 2003) for homologous modeling. In particular, the RING domain was based on the RING-RING dimer of Rad18 (pdb: 2Y43) (Huang et al., 2011), the B-Box1 was based on the B-Box from MuRF1 (pdb: 3DDT) (Mrosek et al., 2008), and the B-Box2 was based on the B-Box domain of Trim54 (pdb: 3Q1D). The CC domain was based on the CC domain of the TRIM69 (pdb: 4NQJ) (Li et al., 2014). For the central part of KAP1, Robetta server was used to model the TSS domain. While Modeller v9.14 (Eswar et al., 2006) was used for the HP1-binding domain, using as a multiple template the structure of the EMSY-HP1 complex (pdb: 2FMM\_E) (Huang et al., 2006) and the structure of the small subunit of the mammalian mitoribosome (pdb: 5AJ3\_a) (Greber et al., 2014). For the PHD finger-bromodomain there are already NMR structures deposited in PDB (pdb: 2RO1) (Zeng et al., 2008). All the missing parts and loops were constructed by Rosetta's loop modeling application v3.5 (Wang et al., 2007; Leaver-fay et al., 2011).



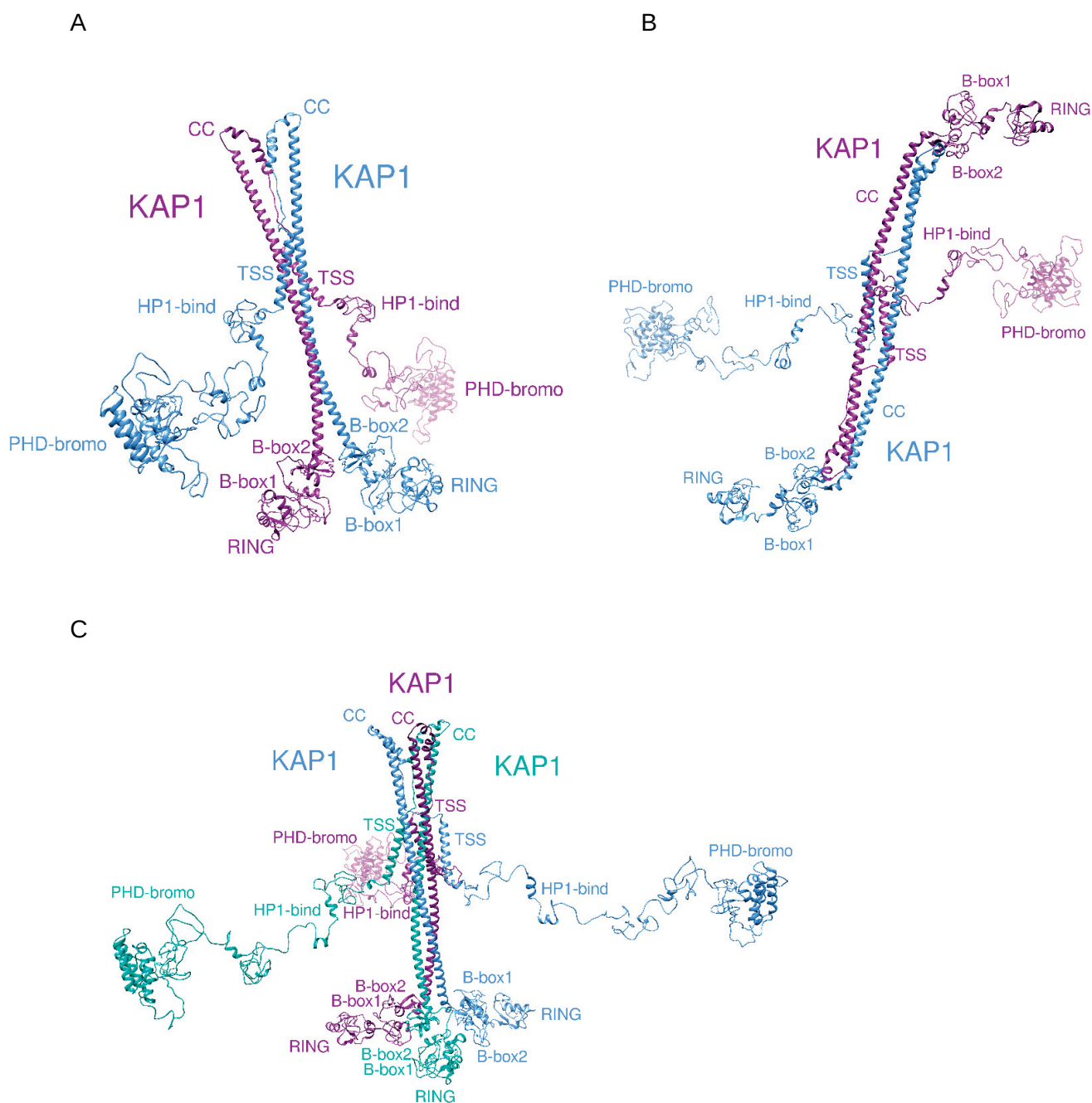
**Figure 29** – Initial molecular model for KAP1

The models of the RING domain and the second B-box domain of the RBCC were built with high accuracy since there were solved structures with homology 30% and 32%, respectively. The model of the first B-box domain was also based on a homologous domain (25%), but since it's an atypical B-box domain I expect that the accuracy can be lower possibly due to a partially incorrect sequence alignment. In addition, the Ring and the two B-box domain models had some missing parts that were reconstructed and this might have introduced inaccuracies. The CC domain model was based on modeling from a  $\alpha$ -helical domain of the same family, which provide a quite accurate model. The TSS and HP1-Bind domains are not based on homologous proteins; consequently, these are the domains that can suffer from higher inaccuracies. In particular, TSS domain model was based on a structural prediction, while HP1-Bind domain was built by combining a low homologous peptide (21%) with the structure of another HP1-binding domain. The loops between the domains are flexible, therefore this preliminary model does not allow at the moment to have insights into the specific interactions of the various domains with the complex itself and other cofactors. Functional and structural data that will be collected in the future will aid to improve the model in this respect. The KAP1 model that was developed here is a monomer and therefore it does not represent the overall image of the active polymer that interacts with the KRAB domain, which was proposed in the literature to act as a trimer (Peng et al., 2000). However, this preliminary model

has so far been functional to aid the experiments in our lab, by providing structural information about the boundaries of each domain.

### Possible oligomerization states of the RBCC domain

In past experiments, it was shown that three RBCC domains can form a homo-trimer and interact with one KRAB domain (Peng et al., 2000). Also, according to the MultiCoil program (a program for predicting two- and three-stranded coiled coils in an alignment) (Wolf et al., 1997), the probability in average of the CC domain to form a trimer is two times higher than the probability to form a dimer. Another indication for a trimer, according to the CCBUILDER tool v1.0 (Wood et al., 2014), is that Rosetta's energy score for the CC homo-trimer molecular model is negative indicating a stable complex formation, while for the (parallel or anti-parallel) homo-dimer model is positive. Nevertheless, experiments in our lab using an array of different techniques clearly indicate that the RBCC domain forms a dimer (of molecular weight 92 KDa). In particular, after cloning and purifying a construct of the RBCC domain, our lab performed experiments for the biophysical characterization of the oligomerization state of this construct in solution (using static light scattering and analytical ultracentrifugation), and concluded that it forms a dimer (in solution). An additional evidence that RBCC can have a dimer form in solution is brought by the crystal structure of the B-Box-Coiled-coil region of Rhesus Trim5alpha which belongs to the TRIM family and forms itself an anti-parallel dimer (Goldstone et al., 2014). These opposing indications lead us to develop a third hypothesis in which a dimer of the RBCC domain forms eventually a trimer coiled coil complex with one of the helices of the KRAB domain of a ZFP. To sort and evaluate these three cases we will build using CCBUILDER v1.0 all the three possible models of the CC domains. We have built molecular models of the KAP1 complex for an anti-parallel dimer based on the crystal structure of Rhesus Trim5alpha, and a parallel dimer and a trimer based on software predictions and constructions (Figure 30).



**Figure 30** - Possible oligomerization states of the RBCC domain: A. Anti-parallel dimer. B. Parallel dimer. C. Trimer

These models can give an insight into the potential structural complexes that the RBCC domains, and in extent the KAP1 oligomers, could form between them. Yet the fluctuations of the domains and the mechanics of the overall complexes due the flexibility of the loops, are not observable. The trimer model comes in agreement with literature's suggestion that three KAP1 proteins form a complex to interact with a KRAB domain, but it needs more experimental data to be confirmed. The parallel

dimer model also confirms the experiments in our lab, but the packing of the two CC domains is not favorable, making this model unrealistic. The anti-parallel dimer model could be a realistic representation of the KAP1 polymer, since it is based on a crystal structure of a protein complex of the same family (TRIM family) and it confirms the preliminary observations of the experiments from our lab. These facts make the anti-parallel dimer model of KAP1, the most consistent model between the three that are proposed here. Eventually, both CryoEM and SAXS results from lab experiments confirm the anti-parallel dimer model as the most accurate model, leading to further refinement and improvement (see submitted paper).

In this work, we showed that KAP1 is an elongated antiparallel dimer with a native asymmetry at the C-terminus domain. This conformation supports our finding that the RING domain influences KAP1 autoSUMOylation. This intrinsic asymmetry has key functional implications for the KAP1 network of interactions, as the heterochromatin protein 1 (HP1) occupies only one of the two putative HP1 binding sites on the KAP1 dimer, resulting in an unexpected stoichiometry, even in the context of chromatin fibers. Similarly, this functional asymmetry in recruiting partners with controlled stoichiometry can be present for other KAP1 interacting partners, such as the MAGE (melanoma antigen genes) protein of KRAB-ZFPs, that interact with the CC domain (Moonsmann et al., 1996; Yang et al., 2007; Schultz et al., 2002). For the latter in fact it has been recently shown that KAP1 FL binds only one KRAB domain of the ZNF93 at around the CC domain (to be submitted).

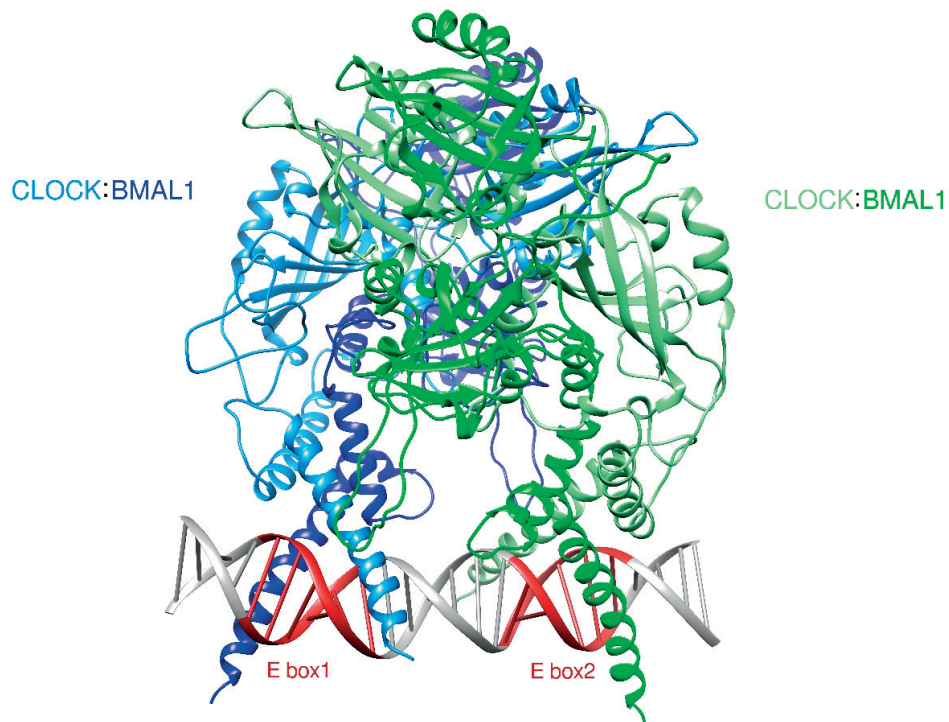
## 7.2. Molecular modeling of the CLOCK:BMAL1 complex bound to DNA

(Sobel J.A., Krier I., Andersin T., Raghav S., Canella D., Gilardi F., Kalantzi A.S., Rey G., Weger B., Gachon F., Dal Peraro M., Hernandez N., Schibler U., Deplancke B. and Naef F. Transcriptional regulatory logic of the diurnal cycle in the mouse liver. PLoS Biol. 15:e2001069, 2017)

The CLOCK:BMAL1 complex consists of two proteins arranged in a heterodimeric transcriptional activator, a key component of the circadian clock in mammals. It regulates gene expression by interacting with a palindromic promoter element termed E-box (CACGTG), while it has been proposed that a repeat of two E-box sequences with a distance of 6 or 7 base pairs between them is required (Nakahata, et al. 2008). Also, BMAL1 footprint reveals temporally distinct protein-DNA complexes, consistent with a hetero-tetramer DNA-binding model. Based on the above, within the manuscript of Sobel et al. I proposed two models of the two CLOCK:BMAL1 complexes binding on the two E-Boxes with a distance of few nucleotides between them. In the first model the spacing between the two E-boxes is 6 base pairs (sp6) (Figure 31) and in the second model the spacing is 7 base pairs (sp7) (not shown). For the model of the single CLOCK:BMAL1 complex we used the crystal structure of the heterodimeric CLOCK:BMAL1 (pdb id: 4F3L) (Huang et al., 2012), in which we build the missing parts of the flexible loops. In order to bind the single CLOCK:BMAL1 model on the E-box we used the complex crystal structure of the CLOCK:BMAL1 basic helix-loop-helix domains bonded on the E-box (CACGTG) (pdb id:4H10) (Wang et al., 2013). Eventually, we superimpose the two single CLOCK:BMAL1:E-box models, with the sp6 DNA and the sp7 DNA, forming the respective symmetric dimer models.

The models give an insight into the molecular mechanisms of the circadian clock, in respect with the literature and the footprint results. An analysis of the interactions between the two CLOCK:BMAL1 complexes could indicate the amino acids that maintain the stability of the overall complex, although experiments *in vitro* will be still necessary to confirm it.





**Figure 31** - A model for the CLOCK:BMAL1 heterotetramer (with 6 bps spacing)

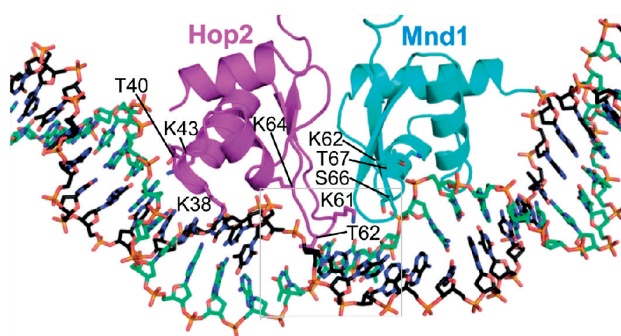
Eventually, the analysis of DNase I cuts at nucleotide resolution showed dynamically changing footprints consistent with dynamic binding of CLOCK:BMAL1 complexes. Our structural model suggested that these footprints are driven by a transient heterotetramer binding configuration at peak activity. Together, our temporal DNase I mappings allowed us to decipher the global regulation of diurnal transcription rhythms in the mouse liver.



### 7.3. Molecular modeling and molecular dynamics simulations of the Hop2-Mnd1-DNA complex

(Kang H.A., Shin H.C., Kalantzi A.S., Toseland C.P., Kim H.M., Gruber S., Peraro M.D., Oh B.H. Crystal structure of Hop2-Mnd1 and mechanistic insights into its role in meiotic recombination. *Nucleic Acids Research*. 43:3841-56, 2015)

In meiotic DNA recombination, the Hop2-Mnd1 complex promotes Dmc1-mediated single-stranded DNA (ssDNA) invasion into homologous chromosomes to form a synaptic complex by a yet-unclear mechanism. I used available structural information to understand how the WHD pair of Hop2-Mnd1 might bind dsDNA. The WHDs of both Hop2 and Mnd1 are structurally most similar to the WHD of transcription regulator TtgV among the known structures of the WHDs in complex with the DNA, according to the Dali server (Homl et al., 2010). Structural superposition of the TtgV:dsDNA complex onto both WHDs of Hop2 and Mnd1 indicated that binding of the juxtaposed WHDs to a continuous DNA is likely to require severe distortion of the DNA. To investigate further, we modeled dsDNA bound to the WHDs based on the TtgV:dsDNA structure (Figure 32), and after geometry optimization, we could confirm indeed that dsDNA is highly perturbed in the model. Based on this initial model, we performed MD simulations to further test the stability of the complex and the structural changes produced upon dsDNA binding. MD simulations revealed a distortion in the base pairing in between the WHDs.



**Figure 32** - A model for dsDNA binding to the WHD pair of Hop2-Mnd1. The model aims to explain how the Hop2-Mnd1 complex binds to dsDNA and distorts its structure, promoting the Dmc1-mediated ssDNA strand invasion. This model can be used to build a complex with the Dmc1 filament and give an insight into the mechanism of meiotic recombination.

It seems that the winged-helix domains are juxtaposed at fixed relative orientation, and according to our molecular simulations, their binding to DNA is likely

to perturb the base pairing. These findings allow us to propose a model explaining how Hop2-Mnd1 juxtaposes Dmc1-bound ssDNA with distorted recipient double-stranded DNA and thus facilitates strand invasion.

## References

- Abriata L.A., Bovigny C. and Dal Peraro M. Detection and sequence/structure mapping of biophysical constraints to protein variation in saturated mutational libraries and protein sequence alignments with a dedicated server. *BMC Bioinformatics*. 17:242, 2016.
- Bailey T.L., Boden M., Buske F.A., Frith M., Grant C.E., Clementi L., Ren J., Li W.W. and Noble W.S. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208, 2009.
- Bailey T.L. and Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol.* 2:28-36, 1994.
- Barazandeh M., Lambert S.A., Albu M. and Hughes T.R. Comparison of ChIP-Seq Data and a Reference Motif Set for Human KRAB C2H2 Zinc Finger Proteins. *G3: Genes, Genomes, Genetics*. 8:219-229., 2018.
- Bellefroid E.J., Poncelet D., Lecocq P.J., Revelant O. and Martial J. The evolutionary conserved Krüppel-associated box domain defines a subfamily of eukaryotic multifingered proteins. *Proceedings of the National Academy of Sciences of the United States of America*. 88:3608–3612, 1991.
- Biasini M., Bienert S., Waterhouse A., Arnold K., Studer G., Schmidt T., Kiefer F., Cassarino T.G., Bertoni M., Bordoli L. and Schwede T. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 42: W252-W258, 2014.
- Bulyk M. Discovering DNA regulatory elements with bacteria. *Nature Biotechnology*. 23:942–944, 2005.
- Cai Y.H., Huang H. Advances in the study of protein–DNA interaction. *Amino Acids*. 43:1141-1146, 2012.
- Camero S., Benítez M.J., Jiménez J.S. Anomalous protein-DNA interactions behind neurological disorders. *Adv Protein Chem Struct Biol*. 91:37-63, 2013.
- Cammas F., Mark M., Dollé P., Dierich A., Chambon P. and Losson, R. Mice lacking the transcriptional corepressor TIF1  $\beta$  are defective in early postimplantation development. *Development*. 127:2955-63, 2000.
- Case D.A., Cheatham T.E., Darden T., Gohlke H., Luo R., Merz K.M., Onufriev A., Simmerling C., Wang B. and Woods R.J. The Amber biomolecular simulation

programs. *Journal Computational Chemistry*, 26, 1668-88, 2005.

Ching Ang H., Joerger A.C, Mayer S. and Fersht A.R. Effects of Common Cancer Mutations on Stability and DNA Binding of Full-length p53 Compared with Isolated Core Domains. *The Journal of Biological Chemistry*. 281,21934-21941, 2006.

Chu W.Y., Huang Y.F., Huang C.C., Cheng Y.S., Huang C.K. and Oyang Y.J. ProteDNA: A sequence-based predictor of sequence-specific DNA-binding residues in transcription factors. *Nucleic Acids Res.* 37: W396–W401, 2009.

Collins T. and Sander T.L. The Superfamily of SCAN Domain Containing Zinc Finger Transcription Factors. *Zinc Finger Proteins*. 156-167, 2005.

D'haeseleer P. How does DNA sequence motif discovery work? *Nat Biotechnol.* 24:959-61, 2006.

Degiacomi M.T. and Dal Peraro M. Macromolecular Symmetric Assembly Prediction Using Swarm Intelligence Dynamic Modeling. *Structure*. 21:1097, 2013.

Eswar, N., Webb, B., Marti-renom, M. A., Madhusudhan, M. S., Shen, M., Pieper, U. and Sali, A. Comparative Protein Structure Modeling Using Modeller. *Curr Protoc Bioinformatics*. Chapter 5:5.6, 2006.

Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*. 27:861-874, 2006.

Fiser A., Do R.K., and Sali A. Modeling of loops in protein structures. *Protein Science*, 9, 1753-1773, 2000.

Frenkel D. and Smit B. Understanding Molecular Simulation: From Algorithms to Applications (2nd Edition). ISBN:9780122673511, 2001.

Fried M., Crothers D.M. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis". *Nucleic Acids Res.* 9: 6505–25, 1981.

Friedman J.R., Fredericks W.J., Jensen D.E., Speicher D.W., Huang X.P., Neilson E.G., Rauscher F.J. 3rd. KAP-1, a novel corepressor for the highly conserved KRAB repression domain. *Genes Dev.* 10:2067–2078, 1996.

Galas D.J., Schmitz A. DNase footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Research*. 5:3157–70, 1978.

Garner M.M., Revzin A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli

lactose operon regulatory system. *Nucleic Acids Res.* 9:3047–60, 1981.

Garton M., Najafabadi H.S., Schmitges F.W., Radovani E., Hughes T.R. and Kim P.M. A structural approach reveals how neighbouring C<sub>2</sub>H<sub>2</sub> zinc fingers influence DNA binding specificity. *Nucleic Acids Research.* 43:9147–9157, 2015.

Gilmour D.S., Lis J.T. Detecting protein-DNA interactions in vivo: Distribution of RNA polymerase on specific bacterial genes. *PNAS.* 81:4275–4279, 1984.

Goldstone D.C., Walker P.A., Calder L.J., Coombs P.J., Kirkpatrick J., Ball N.J., Hilditch L., Yap M.W., Rosenthal P.B., Stoye J.P. and Taylor, I.A. Structural studies of postentry restriction factors reveal antiparallel dimers that enable avid binding to the HIV-1 capsid lattice. *Proceedings of the National Academy of Sciences.* 111:9609-14, 2014.

Goodfellow I., Bengio Y., and Courville A. Deep learning. MIT Press, 2016.

Greber B.J., Bieri P., Leibundgut M., Leitner A., Aebersold R., Boehringer D. and Ban N. Ribosome. The complete structure of the 55S mammalian mitochondrial ribosome. *Science.* 348:303-8, 2015.

Gromiha M.M., Yugandhar K., Jemimah S. Protein-protein interactions: scoring schemes and binding affinity. *Current opinion in structural biology.* 44:31-38, 2016.

Groner A.C., Meylan S., Ciuffi A., Zangger N., Ambrosini G., Dénervaud N., Bucher P. and Trono D. KRAB-zinc finger proteins and KAP1 can mediate long-range transcriptional repression through heterochromatin spreading. *PLoS Genet.* 6: e1000869, 2010.

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology.* 59:307-21, 2010.

Gupta A., Christensen R.G., Bell H.A., Goodwin M., Patel R.Y., Pandey M., Enuameh M., Rayla A.L., Zhu C., Thibodeau-Beganny S., Brodsky M.H., Joung J.K., Wolfe S.A. and Stormo G.D. An improved predictive recognition model for Cys<sup>2</sup>-His<sup>2</sup> zinc finger proteins. *Nucleic Acids Research.* 42:4800-12, 2014.

Han B.Y., Foo C.S., Wu S. and Cyster J.G. The C<sub>2</sub>H<sub>2</sub>-ZF transcription factor Zfp335 recognizes two consensus motifs using separate zinc finger arrays. *Genes & Development.* 30:1509-14, 2016.

Hård T., Lundbäck T. Thermodynamics of sequence-specific protein-DNA interactions. *Biophys Chem.* 62:121-39, 1996

Harteis S., Schneider S. Making the bend: DNA tertiary structure and protein-DNA interactions. *Int J Mol Sci.* 15:12335-63, 2014.

Holm L., Rosenström P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* 38:545-9, 2010.

Huang A., Hibbert R.G., De Jong R.N., Das D., Sixma T.K. and Boelens R. Symmetry and asymmetry of the RING-RING dimer of Rad18. *Journal of Molecular Biology.* 410: 424–435, 2011.

Huang N., Chelliah Y., Shan Y., Taylor C.A., Yoo S.H., Partch C., Green C.B., Zhang H., Takahashi J.S. Crystal structure of the heterodimeric CLOCK:BMAL1 transcriptional activator complex. *Science.* 337:189-94, 2012.

Huang Y., Myers M.P., and Xu R.M. Crystal Structure of the HP1-EMSY Complex Reveals an Unusual Mode of HP1 Binding. *Structure.* 14: 703–712, 2006.

Imbeault M., Helleboid P.Y. and Trono,D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature.* 543:550-554, 2017.

Isakova A., Groux R., Imbeault M., Rainer P., Alpern D., Dainese R., Ambrosini G., Trono D., Bucher P. and Deplancke,B. SMiLE-seq identifies binding motifs of single and dimeric transcription factors. *Nature Methods.* 14:316–322, 2017.

Isakova A., Berset Y., Hatzimanikatis V. and Deplancke B. Quantification of cooperativity in heterodimer-DNA binding improves the accuracy of binding specificity models. *J. Biol. Chem.* 291:10293–10306, 2016.

Iyengar S. and Farnham, P.J. KAP1 protein: An enigmatic master regulator of the genome. *Journal of Biological Chemistry.* 286:26267–26276, 2011.

Jacobs F.M.J., Greenberg D., Nguyen N., Haeussler M., Ewing A.D., Katzman S. and Haussler D. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature.* 516:242–245, 2014.

Jakobsson J., Cordero M.I., Bisaz R., Groner A.C., Busskamp V., Bensadoun J.C., Cammas F., Losson R., Mansuy I.M., Sandi C. and Trono D. KAP1-mediated epigenetic repression in the forebrain modulates behavioral vulnerability to stress. *Neuron.* 60:818-31, 2008.

Jamee N.C., Rajaiya J. and Webb C.F. Mutations in the DNA-binding Domain of the Transcription Factor Bright Act as Dominant Negative Proteins and Interfere with Immunoglobulin Transactivation. *The Journal of Biological Chemistry.* 279: 52465-52472, 2004.

Jordan M.I. and Bishop C.M. Neural Networks – 2<sup>nd</sup> edition. 2004.

Kang H.A., Shin H.C., Kalantzi A.S., Toseland C.P., Kim H.M., Gruber S., Peraro M.D., Oh B.H. Crystal structure of Hop2-Mnd1 and mechanistic insights into its role in meiotic recombination. *Nucleic Acids Research*. 43:3841-56, 2015.

Karplus M. and McCammon J.A. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*. 9:646–652, 2002.

Kennedy J. and Eberhart R. Particle Swarm Optimization. *Proceedings of IEEE International Conference on Neural Networks*. :1942–1948, 1995.

Kim D.E., Chivian D. and Baker D Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*. 32: W526–31, 2004.

Latchman D.S. Transcription factors: an overview. *Int. J. Exp. Path.* 74:417-422, 1993.

Luscombe N.M. and Thornton J.M. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J Mol Biol*. 320:991-1009, 2002.

Leaver-Fay A., ..., Bradley P. Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. *Methods Enzymol*. 487:545–574, 2011.

Lensink M.F. and Wodak S.J. Docking, scoring, and affinity prediction in CAPRI. *Proteins*. 81:2082–2095, 2013.

Letunic I. and Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*. 23:127-8, 2006.

Letunic I. and Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res*. 39:W475-8, 2011.

Letunic I. and Bork P. Interactive Tree Of Life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res*. 44:W242-5, 2016.

Li Y., Wu H., Wu W., Zhuo W., Liu W., Zhang Y., Cheng M., Chen Y.G., Gao N., Yu H., Wang L., Li W. and Yang M. Structural insights into the TRIM family of ubiquitin E3 ligases. *Cell Res*. 24:762-5, 2014.

Li X., Ito M., Zhou F., Youngson N., Zuo X., Leder P. and Ferguson-smith, A.C. (2009). paternal imprints. A maternal-zygotic effect gene Zfp57 maintains both maternal and paternal imprints. *Dev Cell*. 15: 547–557, 2008.

Liu J. and Stormo G.D. Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors. *Bioinformatics*, 24, 1850–1857, 2008.

Looman C., Abrink M., Mark C. and Hellman L. KRAB zinc finger proteins: an analysis of the molecular mechanisms governing their increase in numbers and complexity during evolution. *Molecular Biology and Evolution*. 19:2118-2130, 2002.

Lu Y., Wang X., Chen X. and Zhao G. Computational methods for DNA-binding protein and binding residue prediction. *Protein & Peptide Letters*. 20:346–351, 2013.

Lu X.J. and Olson W.K. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res*. 31:5108-21, 2003.

Lupo A., Cesaro E., Montano G., Zurlo D., Izzo P. and Costanzo P. KRAB-Zinc Finger Proteins: A Repressor Family Displaying Multiple Biological Functions, *Current Genomics*. 14:268-278, 2013.

Luscombe N.M., Austin S.E., Berman H.M. and Thornton J.M. An overview of the structures of protein–DNA complexes. *Genome Biology*. 1:1-37, 2000.

Machanick P. and Bailey T.L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*. 27:1696–1697, 2011.

Mackay D.J., Callaway J.L., Marks S.M., White H.E., Acerini C.L., Boonen S.E., Dayanikli P., Firth H.V., Goodship J.A., Haemers A.P., Hahnemann J.M., Kordonouri O., Masoud A.F., Oestergaard E., Storr J., Ellard S., Hattersley A.T., Robinson D.O., and Temple I.K. Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57. *Nat Genet*. 40:949-51, 2008.

Maerkl S.J. and Quake S.R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*. 315:233-7, 2007.

Maesani A., Iacca G., and Floreano D. Memetic Viability Evolution for Constrained Optimization. *IEEE Transactions on Evolutionary Computation*. 20, 2015.

Margolin J.F., Friedman J.R., Meyer W.K., Vissing H., Thiesen H.J. and Rauscher F.J. Krüppel-associated boxes are potent transcriptional repression domains. *Proceedings of the National Academy of Sciences of the United States of America*. 91:4509–4513, 1994.

Miller J., McLachlan A.D. and Klug,A. Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *The EMBO Journal*. 4:1609–1614,



1985.

Molparia B., Goyal K., Sarkar A., Kumar S. and Sundar D. ZiF-Predict: A Web Tool for Predicting DNA-Binding Specificity in C2H2 Zinc Finger Proteins. *Genomics Proteomics Bioinformatics*. 8:122–126, 2010.

Mrosek M., Meier S., Ucurum-Fotiadis Z., von Castelmur E., Hedbom E., Lustig A., Grzesiek S., Labeit D., Labeit S. and Mayans, O. Structural analysis of B-Box 2 from MuRF1: identification of a novel self-association pattern in a RING-like fold. *Biochemistry*. 47:10722-10730, 2008.

Najafabadi H.S., Mnaimneh S., Schmitges F.W., Garton M., Lam K.N., Yang A., Albu M., Weirauch M.T., Radovani E., Kim P.M., Greenblatt J., Frey B.J. and Hughes T.R. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature Biotechnology*. 33: 555–562, 2015.

Najafabadi H.S., Albu M. and Hughes.T.R. Identification of C2H2-ZF binding preferences from ChIP-seq data using RCADE. *Bioinformatics*. 31:2879-81, 2015.

Najafabadi H.S., Garton M., Weirauch M.T., Mnaimneh S., Yang A., Kim P.M. and Hughes T.R. Non-base-contacting residues enable kaleidoscopic evolution of metazoan C2H2 zinc finger DNA binding. *Genome Biology*. 18:167, 2017.

Nakahata Y., Kaluzova M., Grimaldi B., Sahar S., Hirayama J., Chen D., Guarente L.P. and Sassone-Corsi P. The NAD<sup>+</sup>-dependent deacetylase SIRT1 modulates CLOCK-mediated chromatin remodeling and circadian control. *Cell*. 134:329-40, 2008.

Patel A., Horton J.R., Wilson G.G., Zhang X. and Cheng X. Structural basis for human PRDM9 action at recombination hot spots. *Genes & Development*. 30:257-65, 2016.

Patel A., Yang P., Tinkham M., Pradhan M., Sun M.A., Wang Y., Hoang D., Wolf G., Horton J.R., Zhang X., Macfarlan T. and Cheng X. Evolutionary and structural basis for binding of placental-specific Igf2-P0 by ZFP568. *Cell*. 2018.

Peng H., Begg G.E., Schultz D.C., Friedman J.R., Jensen D.E., Speicher, D.W. and Rauscher F.J. Reconstitution of the KRAB-KAP-1 repressor complex: a model system for defining the molecular anatomy of RING-B box-coiled-coil domain-mediated protein-protein interactions. *Journal of Molecular Biology*. 295, 1139–1162, 2000.

Persikov A.V., Wetzel J.L., Rowland E.F., Oakes B.L., Xu D.J., Singh M. and Noyes M.B. A systematic survey of the Cys2His2 zinc finger DNA-binding landscape. *Nucleic Acids Research*. 43:1965–1984, 2015.

Peters M.B., Yang Y., Wang B., Füsti-Molnár L., Weaver M.N. and Merz K.M. Structural Survey of Zinc Containing Proteins and the Development of the Zinc AMBER Force Field (ZAFF). *Journal of Chemical Theory and Computation*, 6, 2935-2947, 2010.

Phillips J.C., Braun R., Wang W., Gumbart J., Tajkhorshid E., Villa E., Chipot C., Skeel R.D., Kale L. and Schulten K. *Scalable molecular dynamics with NAMD*. *Journal of Computational Chemistry*, 26:1781-1802, 2005.

Phillips S.E. The beta-ribbon DNA recognition motif. *Annu Rev Biophys Biomol Struct.* 23:671-701, 1994.

Quenneville S., Turelli P., Bojkowska K., and Raclot C. (2013). The KRAB-ZFP/KAP1 System Contributes to the Early Embryonic Establishment of Site-Specific DNA Methylation Patterns Maintained during Development. *Cell Rep.* 2:766–773.

Rhee H.S. and Pugh F. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell*. 147:1408-1419, 2011.

Rockel S., Geertz M. and Maerkl S.J. MITOMI: a microfluidic platform for in vitro characterization of transcription factor-DNA interaction. *Methods Mol Biol.* 786:97-114, 2012.

Rodrigues J.P and Bonvin A.M. Integrative computational modeling of protein interactions. *The FEBS journal*. 281:1988-2003,2014.

Rohs R., Jin X., West S.M., Joshi R., Honig B. and Mann R.S. Origins of specificity in protein-DNA recognition. *Annu Rev Biochem.* 79:233-69, 2010.

Rowe H.M., Jakobsson J., Mesnard D., Rougemont J., Reynard S., Aktas T., Maillard P.V., Layard-Liesching H., Verp S., Marquis J., Spitz F., Constam D.B., Trono D. KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature*. 463:237–240, 2010.

Rumelhart D., Zipser D. and McClelland J.L. Parallel Distributed Processing. *MIT Press*. 1:151–193, 1986.

Russell and Ingrid. The Delta Rule. University of Hartford, 2012.

Saleem R.A., Banerjee-Basu S., Berry F.B., Baxevanis A.D. and Walter M.A. Analyses of the effects that disease-causing missense mutations have on the structure and function of the winged-helix protein FOXC1. *Cell*. 68: 627-641, 2001.

Šali A. and Blundell T.L. Comparative protein modelling by satisfaction of spatial

restraints. *J. Mol. Biol.* 234:779-815, 1993.

Santoni de Sio F.R., Massacand J., Barde I., Offner S., Corsinotti A., Kapopoulou A., Bojkowska K., Dagklis A., Fernandez M., Ghia P., Thomas J.H., Pinschewer D., Harris N. and Trono D. KAP1 regulates gene networks controlling mouse B-lymphoid cell differentiation and function. *Blood*. 119:4675-85, 2012.

Schmitges F.W., Radovani E., Najafabadi H.S., Barazandeh M., Campitelli L.F., Yin Y., Jolma A., Zhong G., Guo H., Kanagalingam T., Dai W.F., Taipale J., Emili A., Greenblatt J.F. and Hughes T.R. Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Research*. 26:1742-1752, 2016.

Schwede T., Kopp J., Guex N. and Peitsch M.C. WISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* 31:3381–3385, 2003.

Shi Y. and Eberhart R.C. A modified particle swarm optimizer. *Proceedings of IEEE International Conference on Evolutionary Computation*. :69–73, 1998.

Si J., Zhao R. and Wu R. An overview of the prediction of protein DNA-binding sites. *Int J Mol Sci*. 16:5194-215, 2015.

Simons K., Kooperberg C., Huang E., Baker D. Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences using Simulated Annealing and Bayesian Scoring Functions. *J. Mol. Biol.* 268:209-225, 1997.

Sobel J.A., Krier I., Andersin T., Raghav S., Canella D., Gilardi F., Kalantzi A.S., Rey G., Weger B., Gachon F., Dal Peraro M., Hernandez N., Schibler U., Deplancke B. and Naef F. Transcriptional regulatory logic of the diurnal cycle in the mouse liver. *PLoS Biol.* 15:e2001069, 2017.

Sripathy S.P., Stevens J. and Schultz D.C. The KAP1 corepressor functions to coordinate the assembly of de novo HP1-demarcated microenvironments of heterochromatin required for KRAB zinc finger protein-mediated transcriptional repression. *Molecular and Cellular Biology*. 26: 8623–8638, 2006.

Stawiski E.W., Gregoret L.M., Mandel-Gutfreund Y. Annotating nucleic acid-binding function based on protein structure. *J Mol Biol.* 326:1065-79, 2003.

Stephen C.H. A structural taxonomy of DNA-binding domains. *Nature*. 353:715–719, 1991.

Storn R. and Price K. Differential Evolution - a Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *Journal of Global Optimization*, 11, 341–359, 1997.

Stubbs L., Sun Y. and Caetano-Anolles D. Function and evolution of C2H2 zinc finger arrays. *Subcellular Biochemistry*. 52:75–94, 2011.

Southall N.T., Dill K.A. and Haymet A.D.J. A view of the hydrophobic effect. *J. Phys. Chem.*, 106, 521–533, 2001.

Suzuki M. and Yagi N. DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Biophysics*. 91:12357–12361, 1994.

Russell S.J. and Norvig P. Artificial Intelligence: A Modern Approach - 3rd Edition, 2009.

Tamò G., Maesani A., Träger S., Degiacomi M.T., Floreano D. and Dal Peraro M. *Scientific Reports*. 7:235, 2017.

Tan S., Guschin D., Davalos A., Lee Y.L., Snowden A.W., Jouvenot Y., Zhang H.S., Howes K., McNamara A.R., Lai A., Ullman C., Reynolds L., Moore M., Isalan M., Berg L.P., Campos B., Qi H., Spratt S.K., Case C.C., Pabo C.O., Campisi J. and Gregory P.D. Zinc-finger protein-targeted gene regulation: Genomewide single-gene specificity. *PNAS*. 10: 11997–12002, 2003.

Thalassinos K., Pandurangan A.P., Xu M., Alber F. and Topf M. Conformational States of Macromolecular Assemblies Explored by Integrative Structure Calculation. *Structure*. 21:1500-1508, 2013.

Tjong H. and Zhou H. DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Res*. 35:1465-77, 2007.

Urrutia R. KRAB-containing zinc-finger repressor proteins. *Genome Biol*. 4:231, 2003.

Van Dijk M. and Bonvin A.M.J.J. A protein–DNA docking benchmark. *Nucleic Acids Res*. 36: e88, 2008.

Vandevenne M., Jacques D.A., Artuz C., Dinh Nguyen C., Kwan A.H.Y., Segal D.J., Matthews J.M., Crossley M., Mitchell Guss J. and Mackay J.P. New Insights into DNA Recognition by Zinc Fingers Revealed by Structural Analysis of the Oncoprotein ZNF217. *Journal of Biological Chemistry*, 288, 10616–10627, 2013.

Vlahavas I., Kefalas P., Bassiliades N., Kokkoras F. and Sakellariou I. Artificial Intelligence - 3rd Edition. ISBN: 978-960-8396-64-7. Publisher: University of Macedonia Press/Greece, 2011.

Walter M.C., Rattei T., Arnold R., Guldener U., Munsterkotter M., Nenova K., Kastenmuller G., Tischler P., Wolling A., Volz A. PEDANT covers all complete RefSeq genomes. *Nucleic Acids Res.* 37: D408–D411, 2009.

Wang C., Bradley P. and Baker D. Protein-protein docking with backbone flexibility. *J. Mol. Biol.* 373:503-19, 2007.

Wang L., Huang C., Yang M.Q. and Yang J.Y. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst Biol.* 1: S3, 2010.

Wang Z., Wu Y., Li L., Su X.D. Intermolecular recognition revealed by the complex structure of human CLOCK-BMAL1 basic helix-loop-helix domains with E-box DNA. *Cell Res.* 23:213-24, 2013.

Witzgall R., O'Leary E., Leaf A., Onaldi D. and Bonventre J.V. The Krüppel-associated box-A (KRAB-A) domain of zinc finger proteins mediates transcriptional repression. *Proceedings of the National Academy of Sciences of the United States of America.* 91: 4514–4518, 1994.

Wolf D., and Goff S.P. (2007). TRIM28 Mediates Primer Binding Site-Targeted Silencing of Murine Leukemia Virus in Embryonic Cells. *Cell.* 131:46–57. 2007.

Wolf E., Kim P.S. and Berger B. MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Science.* 6, 1179–1189, 1997.

Wolfe S.A., Nekludova L. and Pabo C.O. DNA recognition by Cys2His2 zinc finger proteins. *Annual review of biophysics and biomolecular structure.* 3:183–212, 1999.

Wood C.W., Bruning M., Ibarra A. Á., Bartlett G.J., Thomson, A.R., Sessions R.B., Brady L. and Woolfson D.N. CCBuilder: an interactive web-based tool for building, designing and assessing coiled-coil protein assemblies. *Bioinformatics.* 30: 3029–3035, 2014.

Wu J., Liu H., Duan X., Ding Y., Wu H., Bai Y. and Sun X. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics.* 25:30–35, 2009.

Yan C., Terribilini M., Wu F., Jernigan R.L., Dobbs D. and Honavar V. Predicting DNA-binding sites of proteins from amino acid sequence. *BMC Bioinform.* 7:262, 2006.

Zeng L., Yap K.L., Ivanov A.V., Wang X., Mujtaba S., Plotnikova O., Rauscher F.J. 3rd and Zhou M.M. Structural insights into human KAP1 PHD finger-bromodomain and its role in gene silencing. *Nat Struct Mol Biol.* 15:626-33, 2008.

Zhang Y., Liu T., Meyer C.A., Eeckhoute J., Johnson D.S., Bernstein B.E., Nusbaum C., Myers R.M., Brown M., Li W. and Liu X.S. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9:R137, 2008.

## Curriculum Vitae

### Alexandra-Styliani Kalantzi

Home address: Chemin de Champ Fleuri 20,  
1022 Chavannes-pres-Renens, Switzerland

Telephone: +41789723224

e-mail: [a.kalantzi@gmail.com](mailto:a.kalantzi@gmail.com)

**Birth date and Place** 21/10/1983 – Athens, Greece

**Nationality** Greek

**Residence Permit in Switzerland** C – Valid until 31/04/2022

**Working Experience**

- o 1/4/2012–31/3/2014 (2 years): Technician-Researcher in the Swiss Institute of Bioinformatics/University of Lausanne, Switzerland
  - Developing and testing models of protein-coding gene evolution in C++, C and Perl
  - Maintenance and debugging of the FastCodeML (C++ parallel processing software)
  - Installing and maintenance of a queuing system (SLURM) on a server
  - Comparison of various software that detect positive selection on phylogenetic trees
  - Prediction of positive selection sites in the proteins and peptides of the melanocortin system, in Perl
- o 2008–2010 (2 years): Software Developer in Printec Group, Greece
  - Developing software (debug and production versions) for VeriFone's automatic transaction terminals (EFTPOS) in C++
  - Maintenance and debugging the existing applications
  - Developing a connection between EFTPOS and cashier machine
  - Migrating software in different EFTPOS embedded systems
- o 2009 (3 months): University of Athens, Greece
  - Teaching assistance of Perl language and bioinformatics
- o 2007 (3 months): Voula Municipal Development Company, Greece
  - Teaching ECDL (Microsoft Office Word, Excel and Power Point)
- o 2001 (3 months): Pegasus Stock Exchange Company, Greece
  - Office and Secretarial Support

**Studies**

- o 2014–2019: PhD in Bioengineering (Laboratory of Biomolecular modeling), Swiss Federal institute of Technology in Lausanne (EPFL), Switzerland  
Thesis: Studies on Protein-DNA Interactions (Supervisor Prof. Matteo Dal Peraro)
  - Developing of the Zinc Fingers-DNA Recognition Code Prediction tool, in Python, BioPython libraries, PHP and HTML
  - Artificial Neural Networks for predicting protein-DNA interactions, in Python and Tensorflow libraries
  - Molecular dynamics simulations with NAMD and GROMACS

- Molecular modeling with Rosetta/Robetta, Modeller, Swiss-Model and Chimera UCFS.
- Motif discovery with MEME suit
- Teaching assistance in the following courses: C++, Bioinformatics, and Biomolecular structure and mechanics
- o 2007–2011: MSc in Bioinformatics (grade: 8.6/10), University of Athens, Greece  
Thesis: Computational Studies of  $\alpha$ -Helix-Lipid Interactions and  $\alpha$ -Helix- $\alpha$ -Helix Interactions, in  $\alpha$ -Helical Transmembrane Proteins (grade: 10/10)
- o 2001–2007: BSc in Informatics (grade: 7/10), University of Piraeus, Greece  
Thesis: Designing of UML diagrams for an e-shop (grade: 10/10)
- o 1997–2001: High School (grade: 18.1/20)

#### Programming languages

Python, Perl, C, C++, Java, MatLab, R, Bash (Unix), SQL, HTML, PHP

#### Languages

- o Greek: Mother tongue
- o English: Proficient in writing and speaking, Michigan Proficiency (E.C.P.E.)
- o Italian: Beginner level
- o French: Beginner level

#### Publications

- o Kang H.A., Shin H.C., Kalantzi A.S., Toseland C.P., Kim H.M., Gruber S., Peraro M.D., and Oh B.H. Crystal structure of Hop2-Mnd1 and mechanistic insights into its role in meiotic recombination. *Nucleic Acids Research*. 43:3841-56, 2015.
- o Sobel J.A., Krier I., Andersin T., Raghav S., Canella D., Gilardi F., Kalantzi A.S., Rey G., Weger B., Gachon F., Dal Peraro M., Hernandez N., Schibler U., Deplancke B. and Naef F. Transcriptional regulatory logic of the diurnal cycle in the mouse liver. *PLoS Biol*. 15: e2001069, 2017.
- o Fonti G., Marcaida M.J, Bryan L.C., Traeger S., Kalantzi A.S., Helleboid P.Y.J.L., Demurtas D., Tully M., Grudinin S., Trono D., Fierz B., Dal Peraro M. KAP1 is an antiparallel dimer with a natively functional asymmetry (Submitted).
- o Kalantzi A.S., Isakova A., Deplancke B., and Dal Peraro M. A computational method to predict the DNA specificity of zinc finger proteins (To be submitted).

#### Hobbies

Making beeswax ointments, dancing





