2016

Features of structural binding motifs and their predictive power

MASTER THESIS IN BIOCHEMISTRY

SUBMITTED BY B.SC. JANNIK SCHWAB
BORN 05.08.1989 IN FRIEDRICHSHAFEN

DATE OF ORAL EXAM: 07.07.2015
START OF MASTER THESIS PROJECT: 15.08.2015
WORK HANDED IN: 11.02.2016

LMU MUNICH, GENE CENTER

Statement of authorship:

I hereby certify that this thesis has been composed by me and is based on my own work, unless stated otherwise. The work was compiled at the Max-Planck-Institute of Biochemistry in Munich under the supervision and guidance of Dr. Bianca Habermann.

No other persons work has been used without due acknowledgement in this thesis. All references have been quoted, and all sources of information, including graphs and data sets, have been specifically acknowledged.

Date: Signature:

1st proof-reader: Prof. Dr. Martin

2nd proof-reader: Prof. Dr. Beckmann

Acknowledgment:

First I want to thank Dr. Bianca Habermann for the opportunity to realize this thesis in her group, for supervising me during the time of the project and scientific advice. I also want to thank all other members of the Habermann group for scientific input, help and the great atmosphere inside and outside the office.

Thanks also to Dr. Martin Steger (Mann group, MPI Munich) for letting me participate in one of their projects which turned out to fit very nicely into this thesis.

I'm also thankful to all my study colleagues for their support and the wonderful time as a student in Munich.

Last but not least I also want to thank my parents for their financial as well as moral support during my whole study time.

Abstract:

It is a common feature of proteins to bind to other ligands like nucleic acids, peptides, metals and a wide range of small molecules. The prediction and detection of these binding sites is a major step in order to guide wet-lab experiments and ultimately determine the function of a protein. While a proteins function can sometimes be identified by sequence homologs, the prediction of binding sites based on a proteins structure is another major approach in order to overcome this task. Binding sites are often closely linked to structural binding motifs, which are specific three-dimensional arrangements of amino acids within a protein. Other than sequence motifs, functional structural motifs occur in space without close proximity of participating residues on sequence level. Thus, they are extremely difficult to detect. So far we know very little about the features of functional structural motifs. Are there for instance recognizable short sequence motifs in the vicinity of motif residues? What is the typical spatial and sequential distance of motif residues? Are there any preferences of amino acid types forming structural motifs? These and other questions will be answered in this thesis in order to get a comprehensive understanding of functional binding motifs for different ligands. Following that we developed a novel prediction algorithm combining structural information with statistics evaluated from significant datasets. This algorithm is not only able to detect already known structural motifs but also has the potential to detect yet unknown binding sites formed by novel motifs.

Zusammenfassung:

Proteine sind in der Lage eine Großzahl verschiedener Liganden wie etwa Nucleinsäuren, Peptide, Metalle oder verschiedenste kleine Moleküle zu binden. Die Vorhersage und Identifizierung solcher Bindestellen ist fundamental, um weiterführenden Laboruntersuchungen die Richtung zu weisen und letztlich die Funktion eines Proteins zu ergründen. In einigen Fällen lässt sich die Funktion mithilfe eines homologen Proteins, sprich einem Protein mit ausreichend ähnlicher Aminosäuresequenz, ermitteln. Häufig lassen sich Bindestellen aber auch anhand der Proteinstruktur vorhersagen. Bindestellen werden häufig von sogenannten strukturellen Motiven gebildet, sehr spezifischen, dreidimensionalen Anordnungen von Aminosäuren innerhalb eines Proteins. Im Gegensatz zu Sequenzemotiven können Aminosäuren, welche sich in einem Strukturmotiv in räumlicher Nähe befinden, über einen großen Sequenzbereich des Proteins verteilt sein. Aus diesem Grund ist ihre Identifikation schwierig. Zum jetzigen Zeitpunkt sind nur sehr wenige Strukturmotive und die von ihnen geformten Bindestellen eingehend untersucht worden. Lassen sich zum Beispiel kurze, erkennbare Sequenzähnlichkeiten innerhalb der Strukturmotive erkennen? Befinden sich die Aminosäurereste innerhalb von typischen räumlichen oder sequenziellen Distanzen? Sind bestimmte Aminosäuren häufiger vertreten und welche Rolle kommt ihnen innerhalb der Motive zu? Dies sind nur einige Fragen, welche in dieser Arbeit beantwortet werden sollen, um ein umfassendes Verständnis von funktionalen Strukturmotiven zu erlangen. Basierend auf diesem Wissen wurde ein neuer Algorithmus, welcher strukturelle Informationen und Statistik kombiniert, entwickelt. Dieser ist nicht nur in der Lage, bereits bekannte Motive zu erkennen, sondern auch bisher unbekannte Bindestellen und die damit einhergehenden Motive zu entdecken.

Table of contents

List of figures:

List of tables:

1 Introduction

Up to this date the number of solved protein structures has by far surpassed the corresponding evaluation of their function. While structure determination by x-ray crystallography, NMR or electron microscopy have advanced rapidly, the function of many proteins remains unclear (*1*). It is now in the hand of computational biology to guide the often time-consuming and expensive wet-lab experiments into the right directions. Predicting the possible function of a protein based on sequence or structure is still a challenging task. Many different methods have been developed making use of the fact that proteins with comparable sequence or structure quite often show similarities in function (*2, 3*). By using information gathered from these so called homologues proteins, conclusions can be made to predict the function of an uncharacterised protein.

1.1 Conservation in sequence and structure

The detection of reoccurring patterns is one major approach of today's cooperative biochemistry field in order to establish relations between different genes, proteins or even whole organisms. In the cases of proteins and nucleic acids, a pattern of nucleotides or amino acids that appears very frequently and therefore has a biological significance is referred to as a motif (*4*). Genes and proteins are conserved on several levels throughout evolution. It needs to be pointed out that there are motifs in sequence and in structure, thus called "sequence motif" or "structural motif" accordingly. Sequence motifs are very common in nucleic acids and proteins, whereas structural motifs appear in proteins, although some structural motifs in RNA molecules have been also observed (*5, 6*). Sequence motifs can be identified by aligning sequences. Structural motifs are three-dimensional arrangements and are much more difficult to identify. Given the strong linkage of sequence and structure, a sequence motif of sufficient length will most likely result in a structural motif, whereas some structural motifs show no sequence conservation at all (*6-8*).

Most proteins of related function show sequence motifs as well as similarities on structural level (*3, 9*). While close sequence homologs can easily be detected by aligning their amino acid sequence with such popular methods as BLAST or FASTA, this gets

more and more challenging with decreasing sequence identity (*10, 11*). More advanced sequence based methods use predefined patterns, like PROSITE, or apply complex mathematical models like Hidden Markov models to identify sequence motifs with very weak similarities (*12, 13*). Nevertheless, due to insertions and deletions it can become impossible to find significant matches on sequence level. At this point it can often be useful to look at the proteins structure instead of its sequence, since structure is evolutionary much more conserved than sequence. Even proteins with far less than 20 % of sequence identity can share common fold elements (*14, 15*). Since the relationship between structure and function has been well established during the last few years, this approach can give good insight into a proteins function inside the so called "twilight zone" of sequence similarity (*16, 17*). One of the most impressive examples to demonstrate structural similarity in absence of sequence similarity are the proteases of the PA clan. They all share a chymotrypsin like fold and a similar mechanism of proteolysis. But their overall sequence identity can be lower than 10 % (*18, 19*). The important role of the spatial arrangements of amino acids has been demonstrated for several other biological processes besides catalytic activity (*20, 21*), namely DNA/RNA interaction (*22, 23*), ion fixation (*24*), antigen-antibody-interaction (*25*) and structural stabilization (*26*).

1.2 Structural motifs and their role in binding mechanisms

One very important property of proteins is the ability to bind a wide variety of other molecules. This feature is often closely related to a proteins function. For example, most transcription factors bind to nucleic acids while many proteins, which are associated to the membrane, can bind lipids. But these are only two of many different binding partners. Other binding sites can for example bind to metals, peptides, small molecules like ATP or sugars as well as other proteins in the formation of protein-protein complexes. This very common trait can have many different forms, from a general binding mechanism for binding partners belonging to a whole group (e.g. metals, nucleic acids) to selective binding to only one very specific molecule or nucleic acid sequence. The detection of possible binding sites is therefor on major step towards the understanding of a proteins function. The fact that most binding sites are closely linked to a structural motif can be exploited to detect possible binding sites if a proteins structure is known (*6, 7, 27*).

Before an approach can be made to predict binding motifs, it is first necessary to learn more about the features of the different binding mechanism in detail. What are the main forces involved in the binding mechanisms, how do binding sites differ and what do they have in common? Are they very diverse within a single binding type and what are their most outstanding features? These are only a few questions, which need to be addressed beforehand. In order to do so, reliable sources of protein – ligand complexes and their interfaces are required.

A wide variety of databases and datasets can be found, but a database only containing interactions establishes by already known structural motifs is not available. Instead many databases contain protein – ligand complexes based on the evaluation of the Protein Data Bank. They differ strongly in size, content, up-to-dateness, redundancy and the definition of ligand – protein interface. The simplest way to classify a residue as an interface residue is by distance. If any atom of a residue is within 3-5 Å of the ligand, most databases define it as an interface residue. In addition to distance, rule-based definitions are sometimes used to identify and also to describe the interaction in more detail as Van-der-Waals, hydrogen-bonding, hydrophobic or electrostatic interactions. This can go even further and consider water bridged interactions (*28-30*). This definitions already display the main forces involved in binding mechanisms. Many publications tried to point out the characteristics of different protein – ligand complexes. A more consistent picture comparing different mechanisms with structures and interfaces originating from one source has yet to be drawn (*31-33*).

One of the most comprehensive databases up to this date is *BioLiP*. The database is updated weekly. It contains 331,591 entries (11.12.2015), separated into DNA/RNA ligands, peptide ligands, metal ligands and other regular ligands. In addition, *BioLiP* uses a composite automated and manual procedure for examining the biological relevance of ligands in the PDB database. Each entry in *BioLiP* contains a comprehensive list of annotations on: ligand-binding residues, ligand binding affinity (from the original literature, plus Binding MOAD, PDBbind-CN, BindingDB), catalytic site residues (mapped from Catalytic Site Atlas), Enzyme Commission numbers, Gene Ontology terms, cross-links to the PDB, UniProt, PDBsum, PDBe, and PubMed databases. Interface residues defined by a cut-off which is set to be 0.5 Å plus the sum of the Van–der-Waals radius of the two atoms under investigation (*30, 34*).

Other databases worth mentioning and used in this work are: *epitome* (*35*) and *AgAbDb* for antigen-antibody complexes (*36*), *3d-footprint* for protein-DNA interactions (*29*) and

PRIDB for protein-RNA interactions (*37*). To gather data on druggable small molecule binding sites *scPDB* (*28*) can be used.

1.3 DNA/RNA binding motifs

Proteins binding to DNA is a fundamental biologic process. Nearly every mechanism involved in the regulation of gene expression can be related to a protein binding to DNA. This can be as simple as in the case of prokaryotic operons, where s single protein can repress or activate gene expression. In most eukaryotic organisms the process is much more complex and many different proteins can be involved. Nevertheless, binding of proteins to DNA is a common, very important process. While unspecific binding can be achieved with a wide variance of proteins, the binding to a specific side usually is achieved by DNA binding motifs. There is a wide range of DNA binding motifs known. The most common ones are discussed in the following section to give insight into the overall diversity, the concept of structural motifs and to stress important points regarding this topic (*38*, *39*).

Helix-turn-helix

One of the first motifs to be observed was the helix-turn-helix (HTH). Consisting of two α - helices connected by a short strand of amino acids, it appears in transcription regulatory proteins like Cro, CAP, and the λ repressor. It binds to the mayor groove of the DNA double helix through several hydrogen bonds and Van-der-Walls interactions. The recognition is achieved by one of the two helices, while the other one is required for the stabilization of the interaction (*40-42*).

The so called homeodomain proteins are a specific class of proteins containing the HTH. They were first discovered in *Drosophila* and led thereby to the important conclusion that principles of gene regulation established in bacteria are relevant to higher organisms as well. In the so called homeodomain, the HTH motif is always surrounded by the same structure. In bacteria the helix-turn-helix can be found embedded in many different structural contexts. Since then, structural studies have shown that a yeast homeodomain protein and a *Drosophila* homeodomain protein have very similar conformations and recognize DNA in almost exactly the same manner, although they are identical at only 17

of 60 amino acid positions (*43*). This fact stresses two important points. First, like mentioned before, we see that structure is much more conserved than sequence. In addition the important difference between the term “structural motif” and “domain” can be observed, although they are quite often falsely used in a likewise fashion in literature. A domain is independently stable, can fold autonomously, and a very specific function can be assigned. A structural motif is very unstable and will not or can’t fold on its own. Some structural motifs are only established after the whole protein is folded, consisting residues from several different regions of the protein. Nevertheless, the confusion about the terminology is not very surprising, since at some point the transition becomes fluid. Some zinc finger motifs for example are stable on their own and can, like a domain, be used to create fusion proteins resulting in engineered zinc finger transcription factors and zinc finger nucleases. By doing so, a desired genomic DNA sequence can be target (*44*).

Basic helix-loop-helix

The basic helix-loop-helix (HLH) motif should not be confused with the former helix-turn-helix, although it can also be found in transcription factors. While the helix-turn-helix contains two helices of equal length, the HLH motif consists a short helix connected by a very flexible loop to a longer alpha helix. Due to this flexibility, one helix can fold back and pack against the same helix in another protein. This motif therefor binds to the DNA and always against another HLH motif of a second protein to establish specific binding. Therefore, other than for the HTH motif, dimerization is required for DNA binding (*45*).

Helix-hairpin-helix

A quite similar to the helix-turn-helix, but much less common and rather newly reported motif is the helix-hairpin-helix (HhH). Like the HTH it has two helices connected by a short turn. But while the HTH binds very specific to the major groove, this is not the case for the HhH. Here the interaction with the DNA is established by a conserved loop at the N-terminal end of the second helix via hydrogen bonds between the protein backbone and the DNA phosphates. This difference then also reflects back to the functional level. While HTH establishes a very sequence specific interaction for gene regulatory proteins, HhH

motifs can for example be found in enzymes which bind to DNA, but show no sequence specificity (*46, 47*).

This fact stresses on important point. Even through the fold in the HTH and HhH can be classified as similar and is most likely found in a direct structural comparison, the function differs. Therefor also structural comparison has its limitations when similar folds are established by different compositions of amino acids.

Zinc finger

Like the name states, the zinc finger includes a zinc ion to stabilize its fold. The term was first used in relation to the *Xenopus laevis* transcription factor IIA (*48*). The zinc ion is essential for the structural integrity of the fold and thereby for the gene regulatory function of the protein. The classical zinc finger consists of an α – helix and an antiparallel β – sheet. The zinc ion is most often coordinated by two cysteines and two histidines. Besides this so called Cys_2His_2 zinc finger, many other zinc finger like motifs are known. These include the Gag-knuckle zinc finger (*49*), treble-clef zinc finger (*50*), zinc ribbon (*51*) and a Zn_2/Cys_6 which can for example be found in the Gal4 protein (*52*).

Despite this variety, proteins containing zinc finger motifs mostly bind to DNA or RNA (*53-55*). Nevertheless some zinc finger proteins binding other proteins or lipids have also been observed (*56, 57*). Given that it is not beside the point to assume that there might be very general concepts underlying even very distantly related binding mechanisms.

Leucine zipper

The leucine zipper contains one long α – helix (60-80 amino acids) with the name giving trait that it shows a leucine at every 7th amino acids. Like the HLH, two zipper find each other and form the basic leucine zipper (bZIP). The leucine zipper is very well studied and shows a high binding affinity for certain DNA sequence motifs. The helices sit in the major groove of the DNA and basic amino acids establish contact to the sugar-phosphate backbone. The array of periodic leucine residues are the one facilitating dimerization. The mechanism of dimerization to homo - or also heterodimers is a very common feature of binding motifs in general, which makes it even more difficult to identify them, due to the fact that a novel protein structure might only contain the monomer when crystalized without ligand (*58*).

DNA recognition β-sheet

So far the motifs we looked at established binding to DNA by a helical structure. But β - sheets are also able to recognize DNA. Here the interaction is maintained by residues extending from a two-stranded β-sheet. The recognized sequence depends on the amino acid composition of the sheet. An example for the β - ribbon motif is the bacterial *met* repressor (*59*).

In addition to the once mentioned above, there are some other, rare binding motifs like the HMG-box, the Wor3 domain and the OB-fold.

RNA binding

Like DNA also RNA can be bound by structural binding motifs, although less binding motifs are known. The most common one is simply called RNA-recognition motif (RRM) consisting a four stranded β-sheet and two α – helices. The main processes involving the RRM are mRNA/rRNA processing, splicing, translation regulation, RNA export, and RNA stability. Up to this date around ten different RRMs are known. Despite their different target sequences they all share common features. The interaction is established via residues of the β-sheets. Variation is achieved by the interaction between different RRM motifs, which are connected by a linker. This linker can then also be involved in the RNA binding itself.

Further a double-stranded RNA-binding motif is known. It is involved into RNA processing, RNA localization, RNA interference, RNA editing, and translational repression. This very rare motif has only been observed in up to three structures, but the feature that it binds only to dsRNA instead of ssRNA is unique (*60*).

Last but not least some zinc finger motifs can be used to bind RNA. Normally DNA binding mediated by zinc fingers is a cooperative process involving several fingers, which are combined in modular fashion. It has been discovered that binding to RNA can be achieved via zinc fingers by intermolecular hydrogen bonds and the Watson-Crick edges of the single stranded RNA bases. By this binding mode a sequence specific binding can be achieved (*61*). This shows again, that although the fold might be the same, the interaction partner can differ based on the exact amino acid composition of the motif.

1.4 Binding of peptides

Another important interaction can be found between peptides and proteins. It needs to be distinguished from large protein-protein interfaces found in protein complexes. Interactions with peptides normally involve only a short protein stretch (3-10 amino acids). This kind of binding is normally low-affinity or related to post-translational modification events like phosphorylation. The short peptide often contains a short linear motif, a short type of sequence motif, while the binding site more often can be described as a structural motif. Examples are RG-rich peptides with SMN domains, the Epstein-Barr virus LMP1 with TRADD domains, DBC1 with Sir2, and the Ago hook with the argonaute PIWI domain (*32, 62, 63*).

Antibody-Antigen interaction

One interaction closely related to the interaction between a protein and a peptide is antigen-antibody interaction. The most interesting property of antigen-antibody interaction is the fact that an antibody can quite often bind to a range of antigens while showing little cross-reactivity. The interaction is achieved by six hypervariable loops which can be very different in sequence and are very flexible. This is also one of the biggest differences to peptide interactions which involve much more stable secondary structures. It has been shown that the interaction mostly involves aromatic residues (*25*). To get further insight in the differences between protein - peptide interactions and antigen-antibody interactions we will also investigate this kind of interaction in this work.

1.5 Small molecule binding

The biggest and probably most diverse category of binding sites are those involved in small molecule binding. Proteins can bind a huge variety of small molecules. The *BioLiP* database lists over 53 000 of its over 90000 binding sites as small molecule binding sites. Since some sites are able to bind different ligands, this yields a total of 183 014 different small molecules. Small molecule binding sites are probably most difficult to detect due to their diversity and the rather small interface area, but are of major interest since they

are very often also binding sites for possible drugs. Identifying those sites thereby is one major step in drug development (*64*).

1.6 Metal binding sites

Another very important mechanism is metal binding. It is estimated that over half of all proteins contain a metal ion (*65*). Metal binding in proteins can have very different roles. Often the binding of a metal ion has mostly structural reasons. Electrostatic interactions between charged residues and metal ions can give rise to a very distinct motif, like the zinc finger. Other binding sites are more functional and are involved in processes like metal storage or enhance certain properties of a protein. The metal ions most frequently found in proteins are Na^{+}, K^{+}, Mg^{2+}, Ca^{2+}, Zn^{2+}, Mn^{2+}, Ni^{2+}, $Cu^{+/2+}$, $Fe^{2+/3+}$, and $Co^{2+/3+}$. There are several processes directly or indirectly involving metal binding like catalytic reactions, signal transduction, metal-induced protein folding and aggregation as well as heavy metal poisoning and metal-based therapy (*66-68*). The prediction of metal binding sites is therefore crucial but also very difficult since it involves only a few residues of the whole protein.

1.7 Catalytic binding sites: Enzymes

One type of binding which is very common but will not be further investigated in this work is the binding of enzymes to their substrate. Many enzyme – substrate complexes are well characterized and their specific binding pocket is well conserved through their enzyme family. If a similar pocket can be found in a protein, it is very likely be a related enzyme itself (*69*). Enzymes – substrate interactions are also the only type of binding mechanism for which a very detailed interface database is available, the *Catalytic Site Atlas* (*70*).

1.8 Characteristics of structural binding motifs

Although many methods have been developed to predict binding motifs, very little effort has been put in the evaluation of their features and the comparison of the binding to different ligands. Therefor we first need to define what the relevant features to look at are and what their role is in the binding process. Especially machine learning approaches often use a huge amount of different classifiers like charge, hydrophobicity, hydrogen-bond-tendency, bulkiness etc. Many of these values are closely related.

Logically all binding reactions involve an area on the proteins surface. But which residues are actually part of the proteins surface is a much more complicated question. In the case of macromolecules like proteins, the surface of interest is normally called solvent-accessible surface, since only the parts, which can be accessed by the solvent, are also accessible for a ligand. There are many applications to calculate the solvent accessible surface area, normally based on the *Shrake-Rupley* algorithm which simulates a ball with the size of a water molecule rolling over the surface (*71*). Although this method can tell if a residue has any contact with the surface, it is not suitable to include or exclude a residue from a possible binding site. Most of a side chain can still be buried by other residues and will most likely not be able to participate in any binding mechanisms, if only a small portion of atoms is actually located at the surface. Due to the fact that residues differ in size, the absolute SAS value for a residue needs to be normalized. Doing so results in the so-called relative solvent accessible surface area, first defined as:

$$RSAS = \frac{SAS}{SAS_{Maximum}} \qquad (1)$$

Given this formula the easiest two-state definition stated that a residue below 16 % RSAS is buried, while every residue with a RSAS equalling or above 16% is exposed to the solvent (*72*). The publication first introducing the RSAS also lists many other possible multi-state definitions. Up to this date, other publications have shown the flaws in this method with regard to the cut-off and the calculation method itself. Nevertheless it remains the used approach. The cut-off of 16 % is strongly discussed, varying from as little as 5 % to up to 32 % in literature (*73-75*). Some publications even state that with a cut-off as low as 5 %, important surface areas might get lost. Further bias is caused by the maximized SAS for the residue. It is calculated based in a simulated tripeptide containing the residue within two glycine. Recently it has been shown to be often much

higher in actual structures than in the simulated tripeptide resulting in an RSAS above 100 % (*73*).

Besides all these difficulties the RSAS is still a very interesting value regarding binding sites. Regardless which exact values or calculation method is used, it still will reveal if a specific type of binding happens more often in a cavity-like fashion or by a rather exposed area off the surface.

The most important feature are the different amino acids, which make up the interface and interact with the ligands. The types of amino acids located at the interface are closely linked to other biochemical features like charge, hydrophobicity, bulkiness, aromaticity or flexibility. In order to find areas on a proteins surface different from what would be expected, a reference is needed. The general composition of amino acids within all known proteins is quite well established and stabile, since it can be calculated simply based on sequence information. Although it is variable between proteins and also protein families, the general means are quite well established (*76, 77*). It is shown by a histogram in Figure 1.

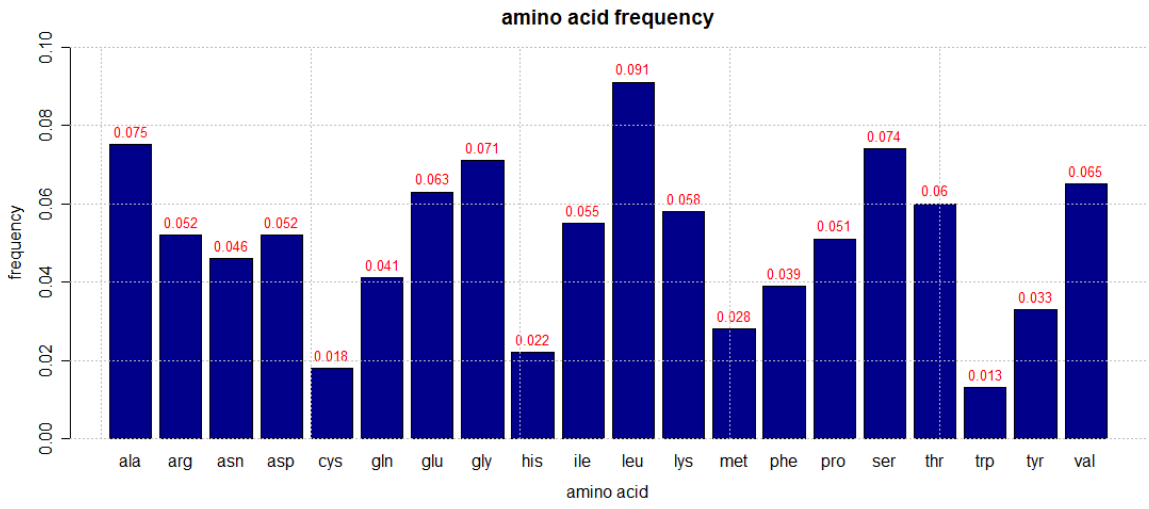


Figure 1: distribution of amino acids in known protein sequences. Exact values of each bar are shown in red. Different sources can list slightly different values.

Since we only look at the surface of the protein this doesn`t help much in order to find significant differences for the interface area. The distribution between amino acids buried or exposed differs quite strongly based on their hydrophobicity. Unluckily, mostly caused by the difficult definition of a proteins surface, there is much less information about the distribution of amino acids on protein surfaces. Most publications simply rely on a distribution based on their own definition of protein surface and the dataset in hand. This can lead to over or underestimation of the relevance of certain amino acids, since the

distribution of surface residues can also be different for a class of proteins in general (*63, 75, 78*). In addition to this, protein structures found in the Protein Data Bank often contain only a part of the protein or protein complex. Given that, the surface of a particular structure might quite well be buried *in vivo* by some other structure. The frequency of surface residues therefore will always be a rough approximation, changing based on dataset and definition of protein surface.

One of the most important structural features to look at in the analysis of structural motifs is the distance in space between interface residues. Certain arrangements of amino acids show a quite distinct pattern for the distance between residues. Some methods related to protein threading for example try to identify similar folds by calculating distances between all residues and then try to find a similar pattern of distances in other proteins (*79*). The disadvantage of this methods is the high computational effort required. In the case of binding sites, the reoccurrence of a specific distance could give a very strong value to identify certain motifs or even in more general a type of binding. For example it can easily be expected that the distance in space for metal binding might be very small, while it can be quite big for the binding of DNA (*31*).

Even if close in spatial distance, residues making up the interface in the case of structural binding motifs can be very distant in sequence or even can be located on different chains. Therefor the distribution of sequence distance might give a good indication for a certain motif. This might be particularly useful when searching for a distinct motif or substructure. Although it hast to be considered that same binding site might be achieved by residues with a very different sequence distance as well.

The secondary structure can also be a very good hint for certain binding types. The binding reactions involving antigen-antibody interactions for example are mostly established by residues inside a loop, while binding of nucleotides is involving α-helices and very rarely a β-sheet like structure.

1.9 Side chain flexibility on binding sites

One major challenge when working with structural information is the question how to handle structural flexibility. This becomes even more relevant while looking not only at the proteins backbone but also considering structural information given by the side chain

conformation. An x-ray structure is only a "snap-shot" of one possible conformation. In addition to this, the side chain conformation observed is also depended on the crystal packing of the protein. Information gathered by this method can be quite different from what could be found in solution or *in vivo (80)*. Nevertheless, most methods still are developed and evaluated based on x-ray structures, since for a long time, it was the most reliable experimental source. While big structural changes in a proteins backbone can occur, they are rather rare compared to the huge conformational space a side chain can occupy. When we consider spatial distances between side chains not only by a Cβ-Cβ distance, which will be quite steady regardless of conformational changes, but also by last-non-hydrogen to last-non-hydrogen atom, the changes can be very big. The question of how flexible side chains are is still very controversially discussed (*81, 82*). While many publications state, that for surface residues the changes in conformation happen quite often and rapidly, they are considered as fixed into on state in method development.

Since the amount of NMR structures in the Protein Data Bank has recently increased very fast, this question can now be addressed in more detail than ever before. Since NMR results contain multiple models for one structure they can display regions with high flexibility. This can be a very useful information source in the case of binding sites (*83*). It could be expected that the side chains inside a binding site are more flexible in order to bind to a ligand or sometimes need to be found in one specific conformation in order for binding to happen. An analysis of surface residues based on NMR results can therefore not only be helpful in a predictive approach, but also might lead to a better understanding of protein flexibility.

1.10 Prediction methods of structural motifs and binding sites

Up to this date, many different methods have already been developed to predict different structural motifs or binding sites. Some have a rather limited field of application, predicting only certain interactions or are limited by the usage of strict prior knowledge.

The most common approach to identify areas of similar structure is the so called fold matching (or motif matching). Like mentioned before, proteins sharing similar function often share a similar fold. Nevertheless, function may alter during evolution, resulting in proteins sharing a fold but showing different functions. One of the best known methods is DALI (*84*). Other popular methods are SSM, GRATH or VAST. A newer, very fast

method is FAST. They differ in the underlying algorithm and also speed but most often give rise to the same results (*8, 27, 85*). Although these methods are very reliable today, they depend on strong prior knowledge and are limited to very close matches, therefor prohibiting the discovery of new structural motifs.

While these methods consider mostly the overall fold, other methods try to identify surface clefts or specific binding pockets on the proteins surface to assign a specific function. This can either be done template-based using a known surface cleft or use a more generic approach. A well-established web server to compare binding sites to known clefts is pvSOAR or SURFACE (*86*). Since binding sites can undergo conformational changes leading to differences between the proteins and the complex structure, this kind of approach might have problems recognizing the binding site.

The strictest approach, which can be used, are residue template-based methods. Most functions are carried out by only a few amino acids (e.g. in enzymes). Often a very specific arrangement is necessary to carry out the function correctly. These arrangements are highly conserved and even if the rest of a protein might undergo severe changes during evolution, these arrangements most likely will remain stable in space. If scanning a template of a crucial arrangement of amino acids against a structure of unknown function gives a genuine match, the function of the protein is probably found. The templates can have different origins like literature searches, manual construction or can even be automatically generated. The *Catalytic Site Atlas* for example is one well-known source for catalytic centres of enzymes, which could be used as templates (*87*). Methods which use user-defined patterns are for example ASSAM or RIGOR/SPASM, which search through structures for user-defined patterns of residues or residue properties. PINTS is a newer method, which detects the largest common three-dimensional arrangement of residues between any two structures (*27, 88*). All these methods are well established. Which one to use depends on the specific task. Nevertheless all of them strongly depend on prior knowledge and structural information of different levels. This strongly limits their ability to discover distant relationships or identifying proteins with novel folds.

First approaches, which do not rely on the comparison of structures, are those based on machine learning. Like stated before, structural comparisons strongly prohibits the detection of proteins with novel folds. Therefor more general rules are necessary. Many different approaches using machine learning have been published. Nevertheless only very few of them have been pursued further. The major problem is that these approaches mostly work well on the data they are trained on, but are less successful on unseen data.

Comparing the approaches is very difficult, also due to the absence of good benchmark datasets. In addition, due to their high complexity they are difficult to adapt to other tasks. An end-user friendly integration into a user-interface is also lacking for most of the published methods (*89-92*).

Besides these two main approaches, there are those, which are more based on statistics. These approaches are in between the complex machine learning approach and the purely structure based methods. They try to use generic features, while also giving insight into the process itself. The most often used feature is the search for unusually high residue frequencies on the surface. Finding areas of a certain surface area with very unusual distribution of specific amino acids can be good predictive hints for binding sites (*33, 93*). Nevertheless no recent method so far is known to combine the abundancy of amino acids with structural information like distances and accessibility.

1.11 Goal of this project

This thesis project consists of two parts. First we wanted to understand more about the general features of structural binding motifs. Although many structural motifs and binding sites are known, very little work focuses on comparing their overall features, as well as their similarities and differences. We compared the binding of different ligands to each other, unravelling, what distinguishes them from each other and show, what they have in common. This might lead to a deeper understand of binding mechanism in general, while obvious differences are useful for a predictive approach. On the other hand, it is also important to learn about the difference between binding motifs, which bind the same ligand but show very different structural features. Only then we will be able to get insight into the overall concepts underlying these diverse mechanisms. We also demonstrate how crucial such approaches also depend on definitions of surface and interface residues.

In the second part we built a novel prediction method, based on the previously identified features. The idea was to combine structural information with statistics in order to develop a method which can not only find known motifs, but also is able to discover possible new binding sites and their motifs. The method aims for the following:

- Recognition of known motifs: given data from a set of known motifs the method should be able to identify the motif afterwards
- Prediction should be not limited by geometry but rely on more general features to identify unknown binding sites: Methods using strict structural features are good if an exact match can be found but otherwise give very little information. By using only very generic structural information it might be possible to find more unknown binding sites or distant relationships. In addition, the difficulty caused by side chain flexibility is less severe.
- High flexibility and fast calculation: especially methods bases only on machine learning approaches are only developed for one type of binding and a very specific problem in hand, making it very difficult to adapt to problems other than the one designed for. This makes the comparison to other methods very difficult. In our approach, we develop a general method, which can not only be used to find binding motifs, but can easily be transferred to other problems where a certain similarity between two proteins surfaces might be of interest.

After method development a brief assessment was performed resulting in a proof-of-concept. If successful, further work can be invested into method optimization as well as end-user fitness.

2 Methods

2.1 Datasets

A database offering protein complexes separated by already known structural binding motifs is not available. Therefore we used interaction databases to extract different datasets. The datasets available differ in size, redundancy and definition of interface residues. To guarantee comparability between the different analyses, we mostly stick to one main dataset originating from BioLiP.

Features and reduction of BioLiP dataset:

BioLiP is the main source we used in this work. BioLiP defines interface residues purely by distance. Interface residues are defined by a cut-off, which is set to be 0.5 Å plus the sum of the Van-der-Waals radius of the two atoms under investigation. Compared to other definitions, this definition might be the most restricted one resulting in lesser interface residues. For all calculations, the offered non-redundant dataset was used (ligand-protein interaction sites with binding site residues identity > 90% and receptor sequence identity >90% are removed) and separated into the four different binding types: DNA/RNA binding, peptide binding, metal binding and small-molecule binding.

To further reduce the amount of structures, the dataset was cross-referenced with the PROSITE database. For DNA/RNA binding, only structures, which are also found via the search term “dna-binding” or respectively “rna-binding” are considered. This approach can also be used to later extract structural motifs from the dataset, which are annotated in PROSITE by searching for the motif in the PROSITE database. The same reduction was done for peptide binding and metal binding. The final amount of structures in each dataset is represented in Table 1.

Table 1: size of the used BioLiP datasets after cross-referencing with PROSITE

Binding type	Number of structures
DNA/RNA	2017
Peptide	1665
Metal	2003

Since proteins in BioLiP often contain multiple structural binding motifs within the same chain, but are listed as one interface (e.g. several zinc finger motifs within on chain), it is

difficult to get reliable information on spatial distances for the separate motifs, since it will result in very long distances for interface residues located on different structural motifs. Therefor the information extracted from PROSITE was also used to calculate only the distance between residues, which are also part of the same structural motif. This was done by restricting distance calculations to residues, which are inside the same motif given in PROSITE.

Non redundant Antibody-Antigen-Dataset from literature

To evaluate a more specific case of protein peptide interaction, a non-redundant dataset was generated from two sources of literature containing 166 structures of antibody-antigen structures (see Table 10, appendix). The corresponding interface residues where extracted from two databases (*epitome* and *AgAbDB*) (*35, 36, 94, 95*).

2.2 Surface amino acid frequency

In order to get a comprehensive reference for the frequency of surface amino acids a non-redundant form of the pdb database (*nr-PDB*) was used from *NCBI/VAST*. It is filtered by a BLAST sequence alignment with a cut-off of a BLAST p-value of 10^{-7} resulting in 13467 structures. The representatives are picked by a number of priority measures (*96*). As definition for surface residues, the original definition was used, defining any residue with an RSAS above 16 % as exposed.

2.3 Statistics on binding motif characteristics

A series of python scripts based on Biopython were developed to generate statistics on any given set of pdb files and a corresponding list of interface residues (*97, 98*). The scripts can be used to evaluate the following features of the interface residues averaging over the whole dataset:

- Interface residues frequency: the frequency of each interface residue in dependency of its type is calculated. Only standard amino acids are taken into account.

- Next-in-sequence frequency: for each interface residue the type of the next and after-next residue in sequence is determined, if possible. The frequency is calculated over the whole dataset. This is also done dependent on the type of the interface residue itself. Residues further away in sequence can be also calculated if required.
- Relative surface accessibility of interface residues: for each interface residue the RSAS is calculated. To do so the residue needs to be complete (no missing atoms) in the pdb file. The RSAS calculations are done by DSSP (*99*). A density distribution is calculated by Gaussian kernel density estimation.
- Secondary structure: for each residue the secondary structure the residue is located in is determined from the structure via DSSP prediction. The possible structural elements are: G = 3-turn helix (3_{10} helix), Min length 3 residues; H = 4-turn helix (α helix). Min length 4 residues; I = 5-turn helix (π helix). Min length 5 residues; T = hydrogen bonded turn (3, 4 or 5 turn); E = extended strand in parallel and/or anti-parallel β-sheet conformation. Min length 2 residues; B = residue in isolated β-bridge (single pair β-sheet hydrogen bond formation); S = bend (the only non-hydrogen-bond based assignment); - = other (residues which are not in any of the above conformations).
- Sequence distance between interface residues: the sequence distance between each interface residue and the nearest interface residue up and downstream in the protein sequence is calculated
- Spatial distance between interface residues: for all interface residues in one structure, the spatial distance is calculated. This can either be done using Cβ to Cβ (simulated Cβ for glycine) or from the last-non-hydrogen atom to last-non-hydrogen atom. A density distribution is calculated by Gaussian kernel density estimation.
- Side chain angles χ1 and χ2: for each interface residue the side chain angles χ1 and χ2 are calculated

The output of the statistics is given as a plain text files. For visualisation purposes, "R" was used (*100*). Probability density distributions where calculated by kernel density estimation with a Gaussian kernel (Python SciPy Package).

2.4 Evaluation of side chain flexibility on binding sites

In order to evaluate the flexibility of side chains in the binding sites, information from NMR structures was obtained. From the DNA/RNA binding dataset from BioLiP (before cross-referencing), all structures based on solution NMR were extracted. Since the BioLiP data only contains the first model of an NMR structure, the corresponding full PDB files were collected from the Protein Data Bank. If only one NMR model was submitted the corresponding entries were excluded from the dataset. For the remaining proteins, a protein-protein blast search of the Protein Data Bank was performed to find homologs with 98 % or more sequence identity. All homologues structures were received and if the source was solution NMR, no DNA/RNA ligand was present and more than one model was submitted, the file was kept. The result was a dataset of 65 DNA/RNA binding proteins with a NMR structure in the bound state and one to six NMR structures in the unbound state (Table 11, see appendix). Under consideration of the different numbering of the residues in the pdb files, the side chain flexibility and rotameric state of each interface residue was evaluated based on the side chain torsion angle $\chi 1$ (for definition of $\chi 1$ for each amino acid type see appendix, Table 12).

The following methods were used to evaluate side chain flexibility:

- Distribution of angles: the rotameric state of the side chains can be displayed in the bound and the unbound state by a histogram. A distribution clustering around the preferred states of the amino acid indicates less flexibility, while an unusual or broadened distribution can either indicate higher flexibility or an enforced conformation by ligand binding.
- Average deviation from preferred state: The preferred states for the side chain torsion angles can be received from well-established rotamer libraries (see appendix, Table 13). By calculating the average deviation from the closest preferred state, a one-value measurement can be calculated for each amino acid indicating higher occurrence of unusual states. This value nevertheless doesn't give any information if, an unusual state is caused by high flexibility and represented transition states or is established by force.
- Distribution of deviation from preferred state: By creating a distribution over all recorded deviations from the preferred state, flexibility and forced states can be distinguished. A very broad distribution indicates several different states due to force, while more discrete deviation values and peaks can account for transition states between preferred rotamers.

2.5 Prediction method and training procedure

Based on the results found for the different datasets, a general prediction method was developed to identify possible interface residues based on a set of statistic files originating from a variable number of pdb files and interface residues. Following files are required:

Frequency of interface residues by amino acid type, frequency of next residue by amino acid type, frequency of after next residue by amino acid type, list of RSAS of interface residues, list of spatial distance of interface residues, frequency of secondary structure elements, and frequency of sequence distances between interface residues. The general workflow is represented in Figure 2.

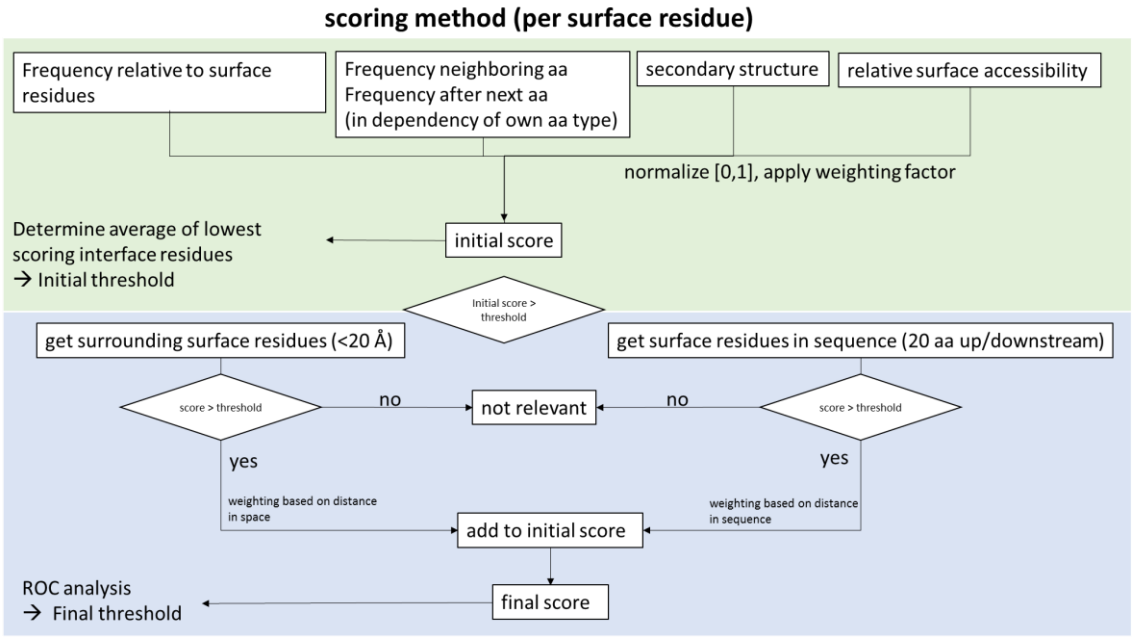


Figure 2: overview of the overall prediction method. Green area – Initial score: Each surface residue is considered separately considering its frequency relative to the frequency of surface residues, the frequency of its neighbours dependent on its own type and the secondary structure represented by a normalized value between zero and one multiplied by a weighting factor. The relative surface accessibility is given by a probability density function. Threshold after one round of training is determined by the average over all lowest scoring interface residues. Blue area – Final Score: Surface residues have to have an initial score above threshold to enter final score calculation, which considers surrounding surface residues and surface residues in sequence. Their score is weighted according to the distance probability function or distance in sequence fraction respectively and added to the initial score. Final threshold is determined by ROC analysis.

The prediction method can easily be separated into two different parts. In the first scoring cycle (shown in green) each surface residue is considered separately and scored after the following pseudo-code:

```
check if residue X is exposed (RSAS>16), standard amino acids and is not missing any atoms:

       determine residue type, secondary structure, type of next and after next amino acid up
       and downstream, relative surface accessibility.

       Initial score = normalized residue frequency [range 0-1] * 1.0

              + normalized residue frequency next residue up and downstream [range 0-1] * 1.0

              + normalized residue frequency after-next residue up and downstream [range 0-1]
              * 1.0

              + probability of RSAS(X) given by probability density function * 0.1

              + normalized frequency of secondary structure element [range 0-1] * 0.1

else:

       initial score = 0
```

Therefore, the features, which are taken into account are:

- the frequency relative to surface residue (weighting factor 1.0)
- frequency of next and after-next residue relative to general amino acid distribution (weighting factor 1.0)
- RSAS given by a density function calculated via Kernel density estimation (weighting factor 0.1)
- Secondary structure the residue is located in (weighting factor 0.1)

RSAS and secondary structure are less important, therefore a weighting factor needs to be applied. The performance seems to mostly depend on the type of the residue itself and its direct neighbours in sequence.

After this first cycle all surface residues, which surpassed a certain threshold enter a second scoring cycle, which is based on the surrounding residues, described in the following pseudo-code:

```
For each residue X if initial score residue X > threshold:
        Determine all residues Y with spatial distance of residue X below the mean given by
        the distance distribution
        For each residue Y with initial score > threshold:
                score (Y) = initial score (Y) * distance probability (distance X-Y)
                spatial distance score (X) =  spatial distance score (X) + score (Y)
        spatial distance score (X) = spatial distance score (X)/number of residues Y
        final score (X) = initial score (X) + (spatial distance score (X) * 100)

        Determine all residues (Y) 20 amino acids up and downstream of residue X.
        For each residue Y with initial score of residue Y > threshold:
                score (Y) = initial score (Y) * sequence distance fraction (X-Y)
                sequence distance score (X) = sequence distance score (X) + score (Y)
        final score (X) =final score (X) + sequence distance score (X) * 0.1
else:
        final score = initial score
```

Therefore, the features, which are taken into account in this part are:

- Clustering of interface residues in one area with a characteristic with regards to their preferred spatial distance. If a residue is surrounded by many residues with a high initial score, its own final score will increase, dependent on the distance distribution.
- Specific, characteristic sequence distances. If for a residue other high scoring residues can be found in a specific, very frequently appearing sequence distance, the final score of this residue increases.

After a first test, it quickly emerged that the sequence distance seems to be much less important, so a much lower weighting factor needs to be applied. The clustering of

residues within a certain distance is the important value, although excluding the sequence distance results in a small drop in performance.

Overall training procedure on datasets:

In order to get ready for prediction, two rounds of training on the training set are required. In the first round, only the first cycle of the scoring is needed, resulting in determination of the first threshold. This threshold is given by an average over the lowest scoring interface residue for each structure in the dataset. By choosing the threshold this way we make sure to exclude the lowest scoring residues and therefore statistically most irrelevant ones in the first round without losing too much information in the first scoring cycle. If a higher specificity is aimed for, also another approach like a ROC analysis can be used to determine a suitable threshold. Nevertheless this seems to result in a strong lose in sensitivity after the second cycle, since more interface residues will already be excluded from further evaluation.

The second, optimized threshold is then calculated by a ROC analysis (for further explanation on ROC calculations see 2.6).

2.6 Assessment of prediction method

Several calculations to evaluate the predictive power of the approach were performed. The most direct approach is to check, if the prediction method is able to identify interface residues correctly. This is also the most difficult approach since the number of interface residues is utterly small compared to the number of surface residues for most proteins. Single surface residues can also possess features, which resemble the ones found for interface residues. Second, and for our approach most relevant is the assessment by a so called patch, an area in the proteins surface. This method tries to check, if the algorithm is able to find the right area inside a protein and therefor to assign the right binding site. Last but not least, an assessment by protein can be done, checking if the method is able to assign the right function to a certain protein within a test set of functionally different proteins.

Assessment by surface residue

For assessment by residue for each structure in the dataset, the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are calculated like the following:

TP → surface residue with a score above the final threshold, which is an interface residue

TN → surface residue with a score below the final threshold, which is not an interface residue

FP → surface residue with a score above the final threshold, which is not an interface residue

FN → surface residue with a score below the final threshold, which is an interface residue

Based on this values, measurements for statistical analyses are used to assess the predictive value of the method. This is done for each structure in the dataset independently. The final value is an average over all structures.

Sensitivity (or true positive rate, TPR): Ability of the method to identify positives correctly.

$$TPR = \frac{TP}{TP + FN} \qquad (2)$$

Specificity (or true negative rate, TNR): Ability of the method to identify negatives correctly.

$$TNR = \frac{TN}{TN + FP} \qquad (3)$$

Precision (or positive predictive value): Fraction of positive identified value.

$$Precision = \frac{TP}{TP + FP} \qquad (4)$$

Accuracy: level of measurement that yields true (no systematic errors) *and* consistent (no random errors) results.

$$Acc = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \quad (5)$$

FMeasure: The FMeasure (also F1 Score) is a way to test for accuracy without considering true negatives, resulting in a value between 0 for the worst outcome and +1 for the best.

$$F1 = \frac{2 * \mathrm{TP}}{(2 * \mathrm{TP} + \mathrm{FN} + \mathrm{FP})} \quad (6)$$

Matthews correlation coefficient: The Matthews correlation coefficient is a measure for the predictive power of a method resulting in a value in-between -1 (total disagreement between prediction and observation) and +1 (perfect prediction). A value of 0 would equal a random prediction.

$$MCC = \frac{\mathrm{TP} * \mathrm{TN} - \mathrm{FP} * \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \quad (7)$$

ROC Curves:

Since our results are threshold dependent, a ROC analysis is also a very good way to assess the predictive power. Therefore many different, final thresholds (step size 0.1) are applied and the TPR and the FPR (1-TNR) are calculated for each threshold. When plotting the TPR against the FPR, the predictive value can be evaluated and illustrated by a ROC curve. Curves above the diagonal have a positive predictive power, below a negative. By calculating the so called *Youden*-Index, which is given by

$$Y - Index = \mathrm{TPF} + \mathrm{TNR} - 1 \quad (8)$$

an optimized threshold can be determined. An idealized picture of a threshold dependent ROC curve is shown in Figure 3. The more ideal the method, the more the values are located towards the upper left corner with a high TPR and a low FPR.

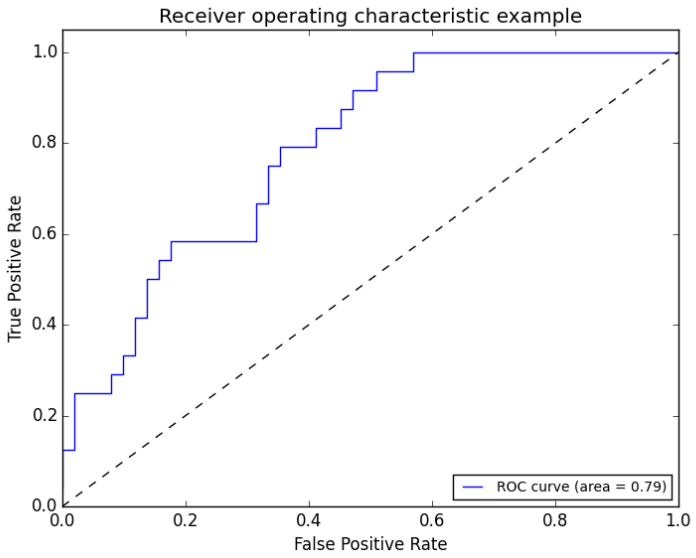


Figure 3: representation of a threshold dependent ROC analysis (101). The diagonal equals a random guess. The best threshold is found for the maximum distance from the diagonal towards the upper left corner, with a FPR of 0.0 and a TPR of 1.0.

By patch assessment:

For the assessment by patch, several surface patches are generated by scanning over the proteins surface. A surface patch contains a central residue and all the surrounding surface residues within a cut-off of 10 Å spatial distance (Cβ-Cβ). This size was chosen since most interfaces are around 20 Å in size, but other sizes could also be applied. The average of the final score for all residues within the patch is than calculated by summarizing the individual residue scores and dividing them by the number of residues contributing to the patch. All patches generated for one protein are ranked by this average to find the highest ranking one. The highest ranked patch should be the one most likely to be part of the binding site and therefor contain also many interface residues. The highest ranked patch can then be compared to the patch with the highest hits by following definitions:

Hits: Number of interface residues within a patch.

Highest hits patch: Patches, which show the highest hits possible for this structure given a certain patch size. Can be more than one.

Highest ranked patch: Patch with the highest score.

Hits percentage: Comparing the maximum amount of hits possible with the number of hits within the highest ranked patch. This is relevant for very big interfaces, where many potential patches are possible.

Overlap: Fraction of atoms which are in the highest scored patch and appear also in a patch with the highest hits. If several patches with the same amount of hits are found, the highest overlap is calculated.

Average rank: The rank after scoring at which the first patch appears reaching the maximum amount of hits possible. If the method is ideal, the highest ranked patch would always be one with the highest hits possible. Can be compared to the average number of patches.

Correlation between number of hits and score: A good predictive power can be assumed if the number of hits within a patch correlations with the score. The Pearson correlation coefficient can be used in order to evaluate for that.

2.7 Method flexibility and application

In cooperation with results from another group (Dr. Martin Steger, Group of Prof. Dr. Mann, MPI Biochemistry, Munich) the method was already used in an applicative manner to assess reliability, flexibility and end-user value. The group recently discovered human Rab proteins, a class of highly conserved GTPases, as a substrate for the kinase LRKK2 *in vitro* and *in vivo* (not yet published). LRKK2 is of huge scientific interest since mutations of the LRKK2 gene are closely related to the Parkinson disease. Surprisingly, the specificity of this reaction is limited to a subgroup of Rab proteins. All Rab proteins contain a region referred to as *switch2*. In contrast to the rest of the Rab proteins structure, *switch2* is a rather flexible region containing a short α – helix followed by a loop, shown in Figure 4.

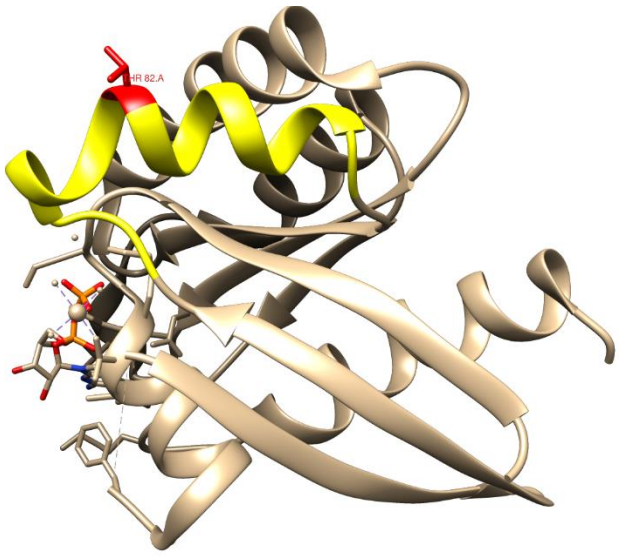


Figure 4: switch2 region of human Rab43 (yellow). The central threonine which is phosphorylated by TRKK2 is shown in red.

Switch2 is highly conserved among the different Rab proteins on sequence level, demonstrated by sequence alignment displayed in Table 2. The group observed that phosphorylation only takes place, if the sequence shows a threonine at a very specific position within the sequence of *switch2*. Although many kinases phosphorylate threonine and serine, no reaction could be observed, if this position is occupied by a serine *in vivo* and only a much weaker reaction *in vitro*. In addition, the phosphorylation is only observed in the GDP bound state, not the GTP bound state. Nevertheless, for a linear peptide of *switch2* no phosphorylation can be observed at all. This led to the conclusion that the recognition is not only sequence based, but also depends on more general surface features around the binding site. In order to find other suitable targets for substrate testing our new algorithm was adapted and used to find possible candidates.

Table 2: alignment of switch2 region in human Rab proteins (Source: Mann Group, MPI Munich)

UniProt ID	Protein	Start	Sequence	Stop
P61026	Rab10	54	Q-GKKIKLQIWDTAGQERFHTITTSYYRGAMGIMLV	88
P61006	Rab8A	53	D-GKRIKLQIWDTAGQERFRTITTAYYRGAMGIMLV	87
Q92930	Rab8B	53	D-GKKIKLQIWDTAGQERFRTITTAYYRGAMGIMLV	87
P20339	Rab5A	65	D-DTTVKFEIWDTAGQERYHSLAPMYYRGAQAAIVV	99
Q6IQ22	Rab12	87	R-GKKIRLQIWDTAGQERFNSITSAYYRSAKGIILV	121
P51153	Rab13	53	E-GKKIKLQVWDTAGQERFKTITTAYYRGAMGIILV	87
P62820	Rab1A	56	D-GKTIKLQIWDTAGQERFRTITSSYYRGAHGIIVV	90
Q9H0U4	Rab1B	56	D-GKTIKLQIWDTAGQERFRTITSSYYRGAHGIIVV	90
Q92928	Rab1C	56	D-GKTIKLQIWDTAGQERFWTITSSYYRGAHGFLVV	90
Q86YS6	Rab43	63	Q-GKRVKLQIWDTAGQERFRTITQSYYRSANGAILA	97
A4D1S5	Rab19	62	D-GKKVKMQVWDTAGQERFRTITQSYYRSAHAAIIA	96
Q15286	Rab35	53	N-GEKVKLQIWDTAGQERFRTITSTYYRGTHGVIVV	87
P20336	Rab3A	67	N-DKRIKLQIWDTAGQERYRTITTAYYRGAMGFILM	101
P20337	Rab3B	67	H-EKRVKLQIWDTAGQERYRTITTAYYRGAMGFILM	101
Q96E17	Rab3C	75	N-EKRIKLQIWDTAGQERYRTITTAYYRGAMGFILM	109
O95716	Rab3D	75	H-DKRIKLQIWDTAGQERYRTITTAYYRGAMGFLLM	109
Q15771	Rab30	54	N-GEKVKLQIWDTAGQERFRSITQSYYRSANALILT	88
Q9NP72	Rab18	53	D-GNKAKLAIWDTAGQERFRTLTPSYYRGAQGVILV	87
Q14964	Rab39A	58	EPGKRIKLQLWDTAGQERFRSITRSYYRNSVGGFLV	92
P59190	Rab15	53	D-GIKVRIQIWDTAGQERYQTITKQYYRRAQGIFLV	87
Q96DA2	Rab39B	54	EPGKRIKLQIWDTAGQERFRSITRAYYRNSVGGLLL	88
P20338	Rab4A	58	G-GKYVKLQIWDTAGQERFRSVTRSYYRGAAGALLV	92
P61019	Rab2B	51	D-GKTILVDFWDTAGQERFQSMHASYYHKAHACIMV	85
Q8WUD1	Rab2A	51	D-GKQIKLQIWDTAGQESFRSITRSYYRGAAGALLV	85
Q96AX2	Rab37	75	D-GVRVKLQIWDTAGQERFRSVTHAYYRDAQALLLL	109
Q9ULW5	Rab26	109	D-GVKVKLQMWDTAGQERFRSVTHAYYRDAHALLLL	143
Q8IZ41	Rab45	586	D-GERTVLQLWDTAGQERFRSIAKSYFRKADGVLLL	620
Q13636	Rab31	50	G-NELHKFLIWDTAGQERFHSLAPMYYRGSAAAVIV	84
Q5JT25	Rab41	76	E-DQIVQLQLWDTAGQERFHSLIPSYIRDSTIAVVV	110
P51159	Rab27A	64	R-GQRIHLQLWDTAGQERFRSLTTAFFRDAMGFLLL	98
O00194	Rab27B	64	K-AFKVHLQLWDTAGQERFRSLTTAFFRDAMGFLLM	98
P61020	Rab5B	65	D-DTTVKFEIWDTAGQERYHSLAPMYYRGAQAAIVV	99
P51148	Rab5C	66	D-DTTVKFEIWDTAGQERYHSLAPMYYRGAQAAIVV	100
P51151	Rab9A	52	D-GHFVTMQIWDTAGQERFRSLRTPFYRGSDCCLLT	86
Q9NP90	Rab9B	52	D-GRFVTLQIWDTAGQERFKSLRTPFYRGADCCLLT	86
Q7Z6P3	Rab44	581	D-NKCFVLQLWDTAGQERYHSMTRQLLRKADGVVLM	615
P20340	Rab6A	58	E-DRTIRLQLWDTAGQERFRSLIPSYIRDSAAAVVV	92
Q9NRW1	Rab6B	58	E-DRTVRLQLWDTAGQERFRSLIPSYIRDSTVAVVV	92
Q9H0N0	Rab6C	58	E-DGTIGLRLWDTAGQERLRSLIPRYIRDSAAAVVV	92
P51149	Rab7A	53	D-DRLVTMQIWDTAGQERFQSLGVAFYRGADCCVLV	87
Q96AH8	Rab7B	53	G-DTTLKLQIWDTGGQERFRSMVSTFYKGSDGCILA	87
O14966	Rab29	52	SDYEIVRLQLWDIAGQERFTSMTRLYYRDASACVIM	87

Since it is unknown how exactly the kinase binds to the Rab protein as substrate, the interface and the contributing residues are unknown. The approach therefor differs slightly from the original method, demonstrating the high adaptive features of the method in practical application, a feature often lacked by the stricter machine learning based approaches. The adapted workflow is visualised in Figure 5.

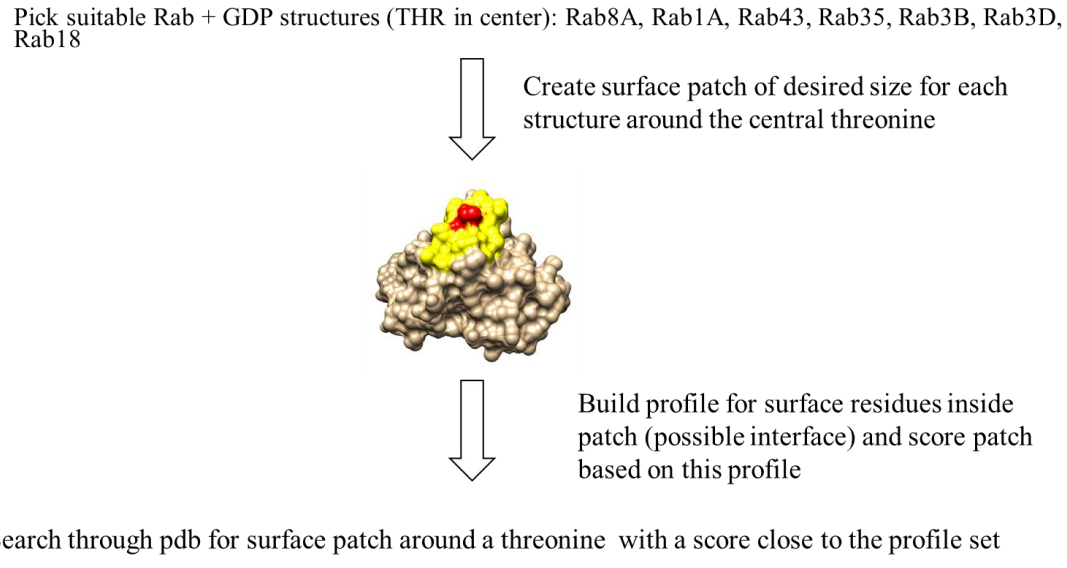


Figure 5: workflow to generate statistics without known interface. After picking a suitable set of structures a round surface patch is created for each of them (yellow) around the central residue (red). All residues inside the patch are considered interface residues.

At first, a suitable set of Rab proteins was chosen. They needed to possess the threonine at the right sequence position. A high resolution structure in the GDP bound state needed to be available with structural information in the *switch2* region present. Due to the fact that *switch2* is highly flexible, some structures actually lack this region within their x-ray structure, since they cannot be solved properly. Some complexed structures were therefore also taken into account. A set of seven Rabs with threonine were chosen (see Figure 5). As a control, also a dataset containing eleven Rab proteins with a serine at the same sequence position was constructed. For each Rab protein, a surface patch was generated containing all surface residues within a cut-off of 10 Å (Cβ-Cβ) of the serine or threonine. 10 Å was chosen the maximum distance between interface residues is around 20 Å. All surface residues within this area are considered an interface residues. In the same fashion as for the original method, a profile is generated based on this interface.

A dataset of 9733 non redundant human proteins (less than 30 % sequence identity) was retrieved from the pdb database. For each protein structure, every possible surface patch around an accessible threonine (and serine in case of the control set) was generated and the average score for the patch was calculated. In contrast to the original method, a residue is only considered in the second scoring cycle if it's part of the patch, not if they are located outside of the patch. The highest scoring patch was considered the site with most similarities to the *switch2* site for this structure. The same scoring procedure was also performed on the two datasets of Rab proteins and compared with the results from the

search. The proteins with the most similar patches should rank close to the training set and higher than the control set.

3 Results and Discussion

3.1 Surface composition compared to the normal composition of proteins

Before we discuss the different features for the binding mechanisms, the distribution of surface amino acids in the non-redundant pdb set will be analysed and compared to the well established distribution of amino acids within proteins in general. A histogram of the distribution of surface amino acids compared to the average composition of proteins is shown in Figure 6.

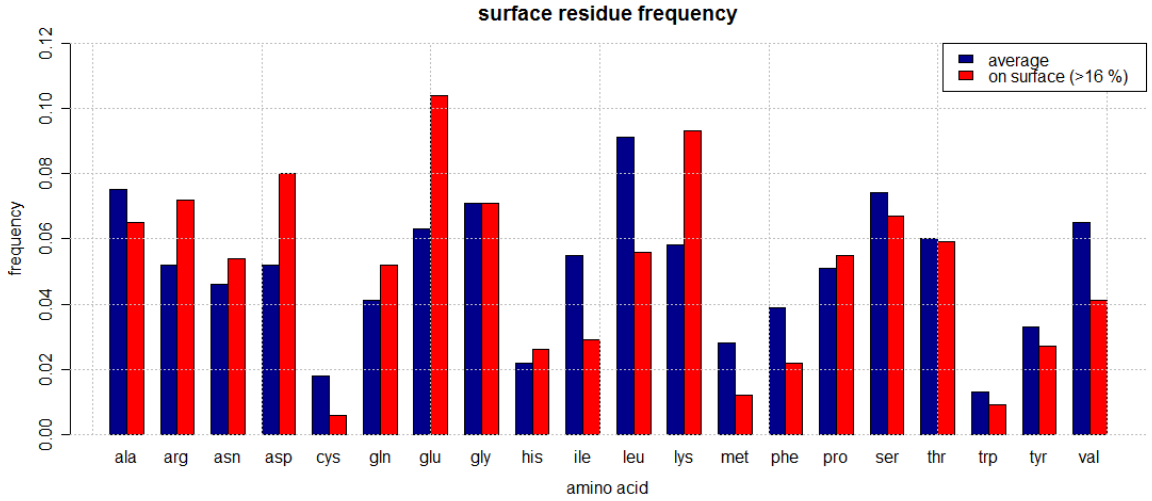


Figure 6. Histogram of the composition of the proteins surface (red) calculated based on nr-PDB compared to the general composition of proteins (blue). All residues with an RSAS > 16 % are considered to be surface residues.

Like expected the hydrophobic amino acids isoleucine, leucine, valine and phenylalanine are underrepresented on the surface and most likely buried in the hydrophobic core. Although alanine, proline and glycine also belong to the group of hydrophobic amino acids, the difference between the surface and the average value is insignificant. This can very easily be explained by their small size which makes it quite easy to shield them from the proteins surface, resulting in a RSAS close or below the cut-off of 16 %. The surface of a protein is dominated by the charged amino acids lysine, glutamic acid, aspartic acid and arginine. Most polar amino acids show no tendency towards the surface, although cysteine is strongly underrepresented. This is due to its role in disulphide bridges. In this specific structural arrangement they will normally not be surface accessible.

Interesting facts can be observed when comparing these results with related calculations for surface residues in literature, demonstrating how crucial this kind of statistics depends on the chosen definitions of surface residues. Several publications have created statistics of surface residues and their corresponding RSAS distribution (*73*). For some residues like tryptophan or tyrosine the major portion is located within a window of a RSAS of 5-20 %, indicating that the cut-off of 16 % will miss many of these residues despite the fact that they might still be relevant. We also calculated this statistics for our interface residues in particular and can observe similar situations for several residues, like demonstrated in Figure 7, showing the results for three amino acids from the DNA/RNA binding dataset. Especially if the binding site is not an exposed area of the protein but a cavity, this might lead to misinterpretation. For example, tryptophan, which, like we will later show plays a significant role in DNA binding, quite often shows an RSAS < 20 %. Most of these residues will therefore not be taken into account when using the 16 % cut-off value, although they might be quite relevant in the binding mechanism. Lysine and arginine on the other hand show a very broad distribution and a higher accessibility in general.

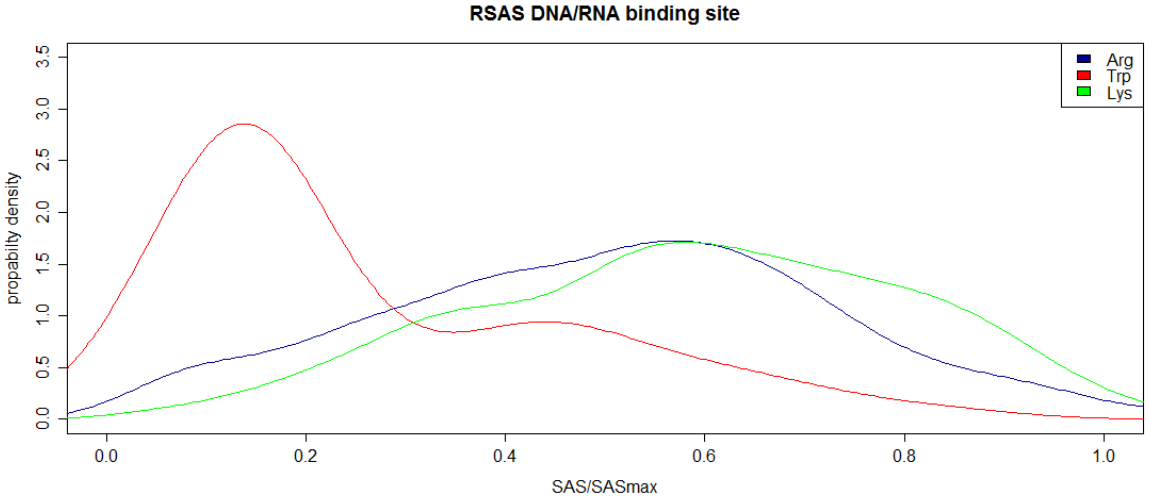


Figure 7: RSAS of the amino acids arginine (blue), tryptophan (red) and lysine (green) when present in a DNA/RNA binding site.

Nevertheless we decided to stick with the cut-off of 16 % for most calculations since there is no reliable information yet available clarifying, which cut-off would be most suitable for binding site residues. Further work could evaluate the performance change of our method in dependency of the definition of surface residues, therefore not only improving the prediction method, but also demonstrating, which cut-off would be most suitable for binding sites. The amino acid type specific RSAS distributions are not yet included in the calculation, mostly because for rare interface residues, there are too few values to calculate a statistical significant distribution. Nevertheless a type-specific distribution

could be used for the most abundant residues in future. Therefore we first would need to check, if the accessibility of this specific interface residues shows distinct differences from surface residues in general.

All frequencies regarding surface residues from here on will be represented normalized to the frequencies shown above if not stated otherwise.

3.2 Statistics on DNA/RNA binding

Since most, well defined structural motifs are known for DNA/RNA binding mechanism, they will be discussed first. We will have a look at the BioLiP data for nucleotide binding proteins, which contains several structural motifs but also some other unusual DNA/RNA binding interfaces, resulting in a very diverse set. Following that, we will investigate different motifs on their own in order to demonstrate the high diversity within this one group.

Distribution of surface amino acids

Figure 9 shows a histogram of the distribution of interface residues in the DNA/RNA binding dataset based on BioLiP with absolute frequencies before normalization based on the surface composition. To check if the calculated data is statistical relevant and stable, the average standard deviation was calculated by a bootstrap procedure (random 10 % of the dataset, calculated 50 times). The values are very stable and only a very small standard deviation can be reported. On first glance arginine and lysine seem to be the most important participants. But like mentioned before, this statistics needs to be normalized by the distribution for surface amino acids shown in Figure 6. The result after normalization is shown in Figure 9, revealing which amino acids are enriched compared to what would be expected to be found on the surface.

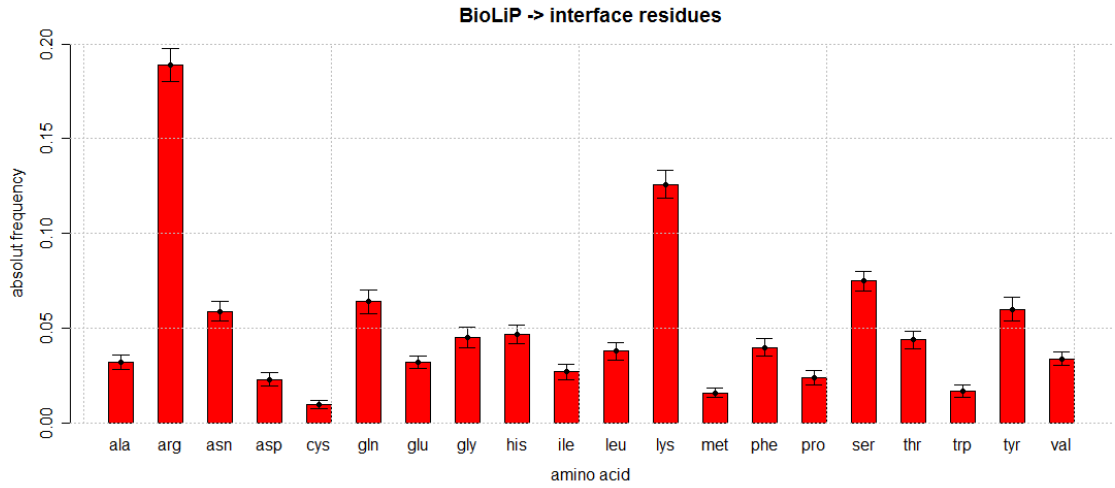


Figure 8: distribution of interface residues in the DNA/RNA binding dataset based on BioLiP with absolute frequencies before normalization based on the surface composition. The error bars for standard deviation were calculated by a bootstrap procedure (random 10 % of the dataset, calculated 50 times).

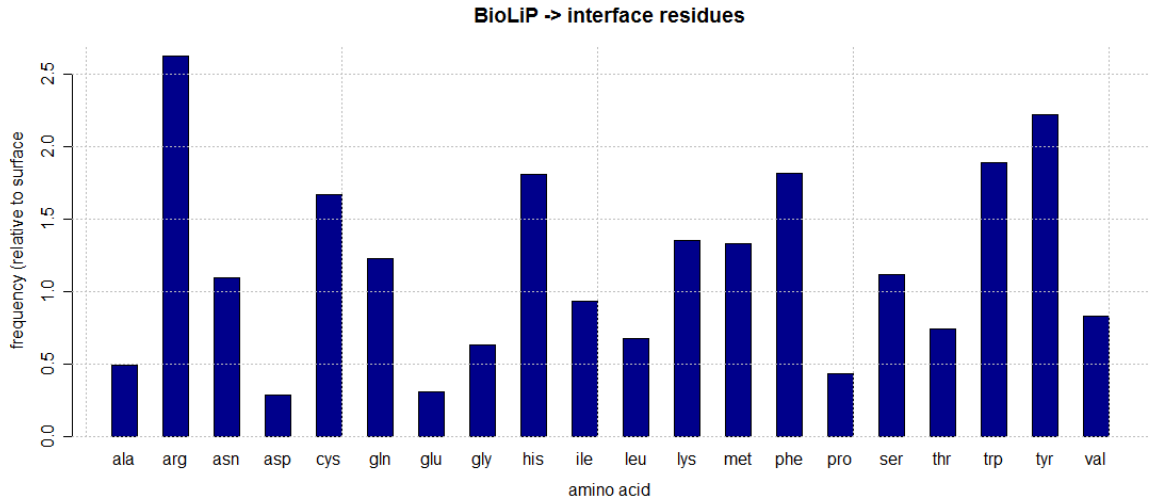


Figure 9: composition of the interface residues of DNA/RNA binding sites extracted from BioLiP. The frequency is relative to the surface residue frequency calculated on the nr-PDB dataset.

The main interactions in this type of binding seems to be established by arginine, which appears over 2.5 times more often at an interface than it should appear on the surface, and tyrosine, which appears more than twice as often on an interface. Other important amino acids are tryptophan, lysine, phenylalanine and histidine. This result is not very surprising. The main force establishing interaction between nucleic acids and proteins is charge, mainly delivered by arginine. The preference of arginine above lysine is most likely caused by the longer side chain and thereby higher flexibility of arginine. The second important interaction type is a so called stacking interaction between aromatic amino acids and the nucleic acids bases. It can be established by tyrosine and tryptophan.

It is less often caused by phenylalanine and histidine. In addition to charge interactions and stacking interactions, hydrogen bonding is the third force establishing the interaction between nucleic acids and proteins. Interesting results can be observed when looking at the amino acids surrounding the surface residues. Figure 10 shows a histogram of the next and the after next amino acids up and downstream of an interface residue in sequence.

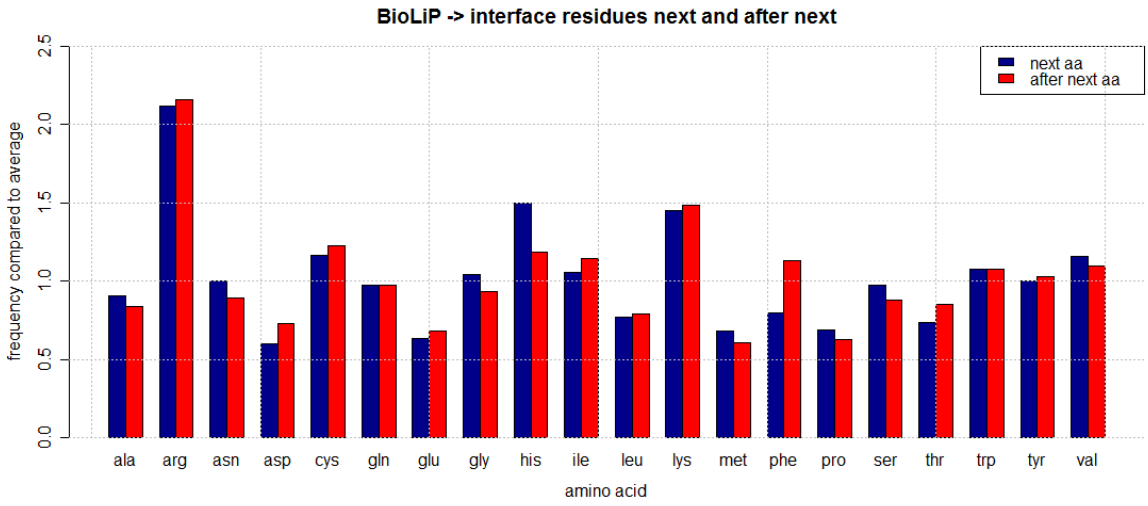


Figure 10: next and after next amino acid in nucleic acid binding proteins. The next and after next amino acid are evaluated in both directions, up and downstream the protein sequence.

Arginine and lysine again play a major role, indicating that most charged residues participating are also very close together in sequence. The aromatic residues need more space and will not be next to another interface residue in sequence. This can be easily demonstrated in detail, when separating this kind of statistics by the interface residue

itself, creating neighbour statistics for each amino acid type. This is shown for arginine and tryptophan in Figure 11 and Figure 12.

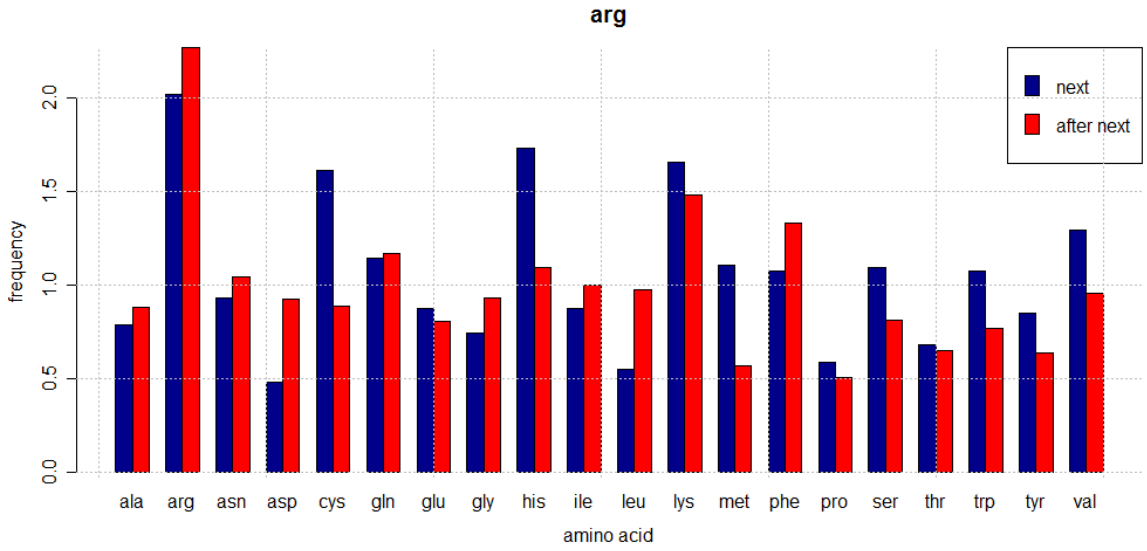


Figure 11: next (blue) and after next (red) amino acid of arginine. An interface arginine is very likely neighboured by another charges residue.

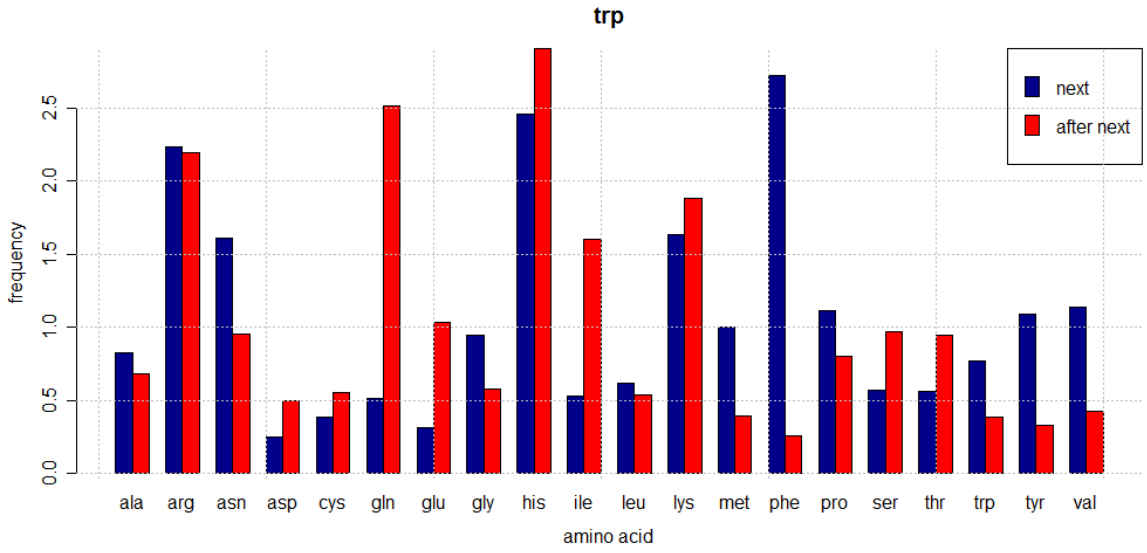


Figure 12: next (blue) and after next (red) amino acid of tryptophan.

Arginine is quite likely followed by another arginine or a lysine. The high frequency of cysteine and histidine is probably caused by zinc fingers in the dataset, which will bind the zinc ion with histidine and/or cysteine. For tryptophan, we observe a very different profile. The probability to find a tryptophan on the interface followed by another one is very low, but it will most likely be neighboured by a phenylalanine. For easier comparison this is shown again in Figure 13 for the next amino acid in sequence for arginine and tryptophan in one histogram combined.

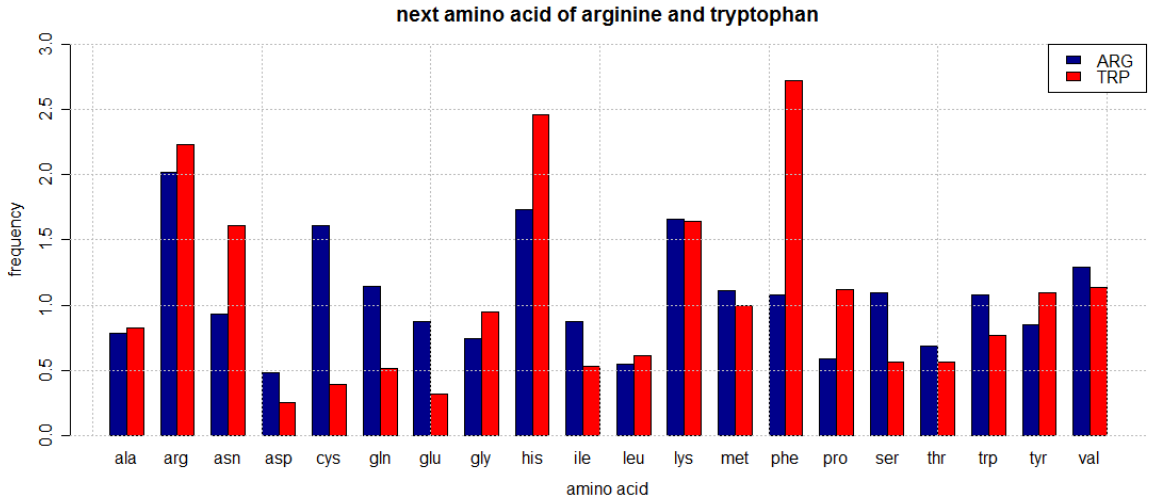


Figure 13: next amino acid for arginine (blue) and tryptophan (red) in DNA/RNA binding proteins combined on histogram.

This kind of statistics can be very useful to predict structural motifs with a very weak underlying sequence conservation. In more general it can also be interesting to just look at the sequence distance of the interface residues. Since most of the interface residues are still close together, it is closely related to the next and after next amino acid statistics. Figure 14 shows a histogram of the sequence distance to the next interface amino acid.

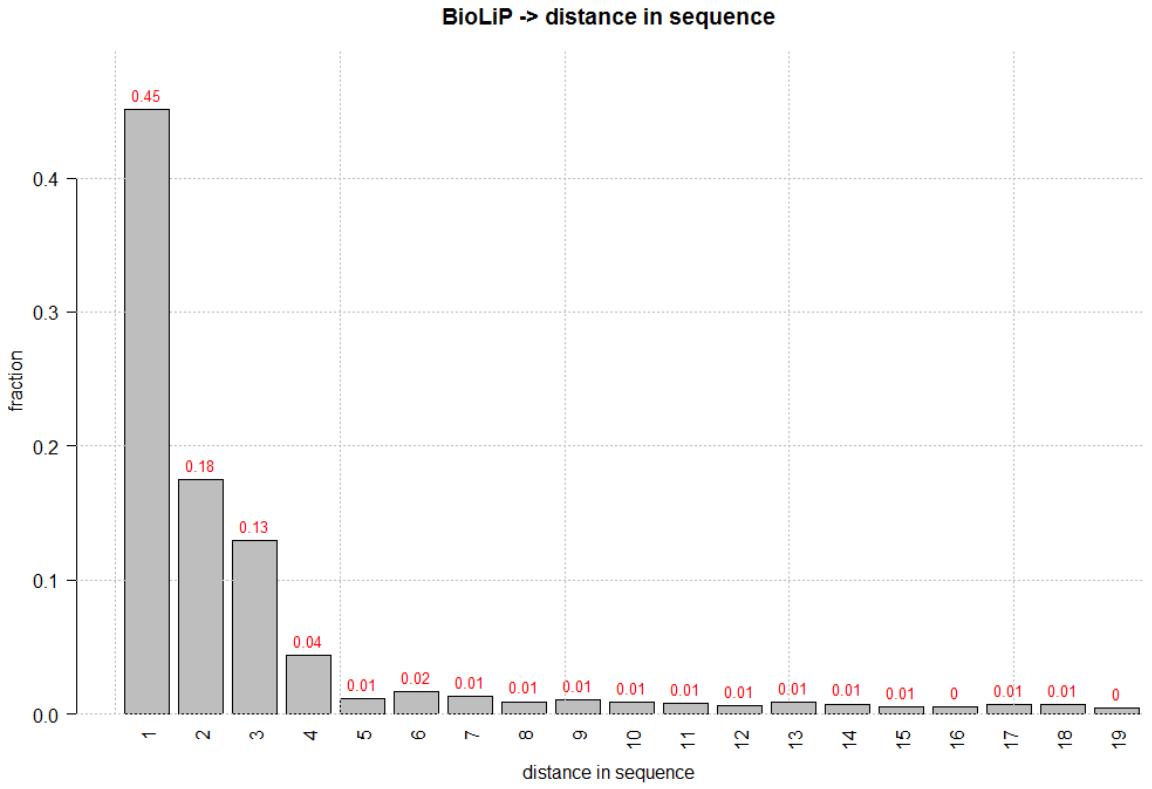


Figure 14: distance in sequence for interface residues in nucleic acid binding proteins. The absolute values are shown in read. A distance of one equals neighbouring amino acids.

Quite obviously, most interface residues are still very close in in sequence with a distances between one (neighbouring) and 18 amino acids distance. For such a big dataset a lot of noise above 20 amino acids distance is retrieved, resulting in values of up to 177 amino acids sequence distance. This shows already how far apart the interface residues can be in sequence.

The next logical step is to look at what defines in particular structural motifs moving away from sequence related information to the spatial arrangement. Figure 15 shows a density distribution based in the BioLiP dataset for the Cβ – Cβ distance.

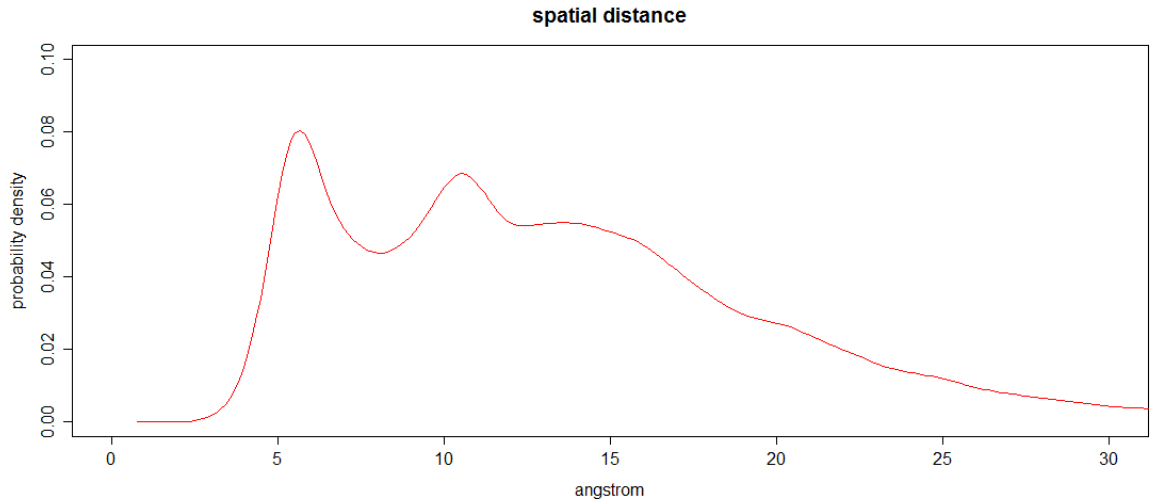


Figure 15: spatial distance distribution for nucleic acid binding proteins calculated by Kernel density estimation.

Most structural motifs for DNA/RNA binding cover an interface area of around 25 Å. The two peaks above five and ten angstrom are caused by interface residues which are next or after next in sequence in an α-helix, resulting in a very specific distance for their Cβ distance. Besides this a broad distribution of different distances can be found. In general they will not come each other closer to each other than five angstrom, which is not surprising, given the fact that most of the residues will be charged or aromatic.

A similar distribution can be calculated for the surface accessibility (shown in Figure 16). It results in a very broad distribution although a median accessibility seems to be most favoured. A very low accessibility would support the idea of a cavity, which is normally not the case for nucleic acid binding, while an accessibility close to the maximum of 1.0 would be caused by highly isolated side chains emerging from the surface. Since the charged side chains are packed against each other, but the overall motif is exposed to create a charged surface patch, a medium accessibility of broad variety is exactly what

would be expected. On the other hand, such a broad distribution doesn't contain much informative value in sense of predictive power.

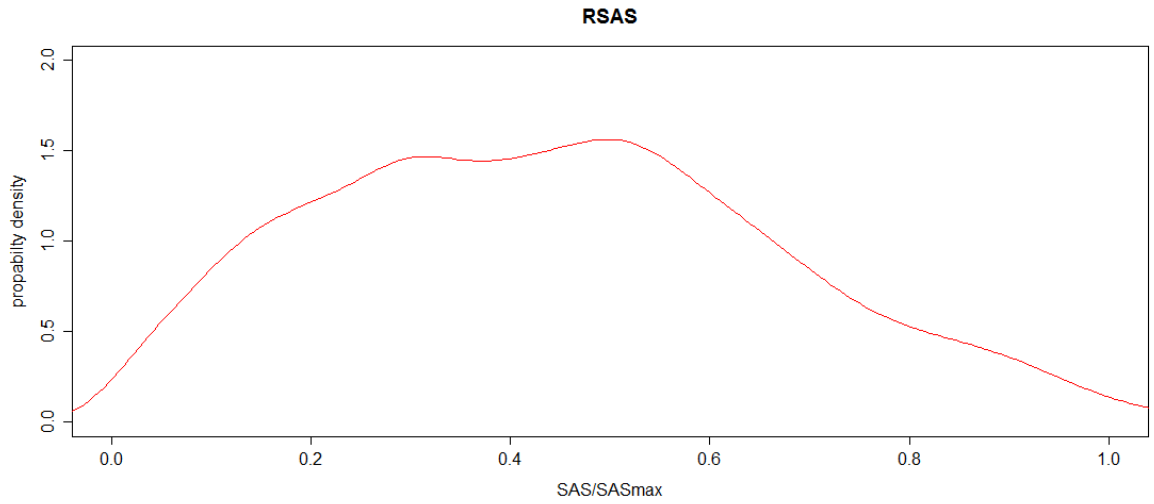


Figure 16: surface accessibility distribution for interface residues in nucleic acid binding proteins

Last but not least we can have a look at the secondary structure elements which are involved in the binding mechanism.

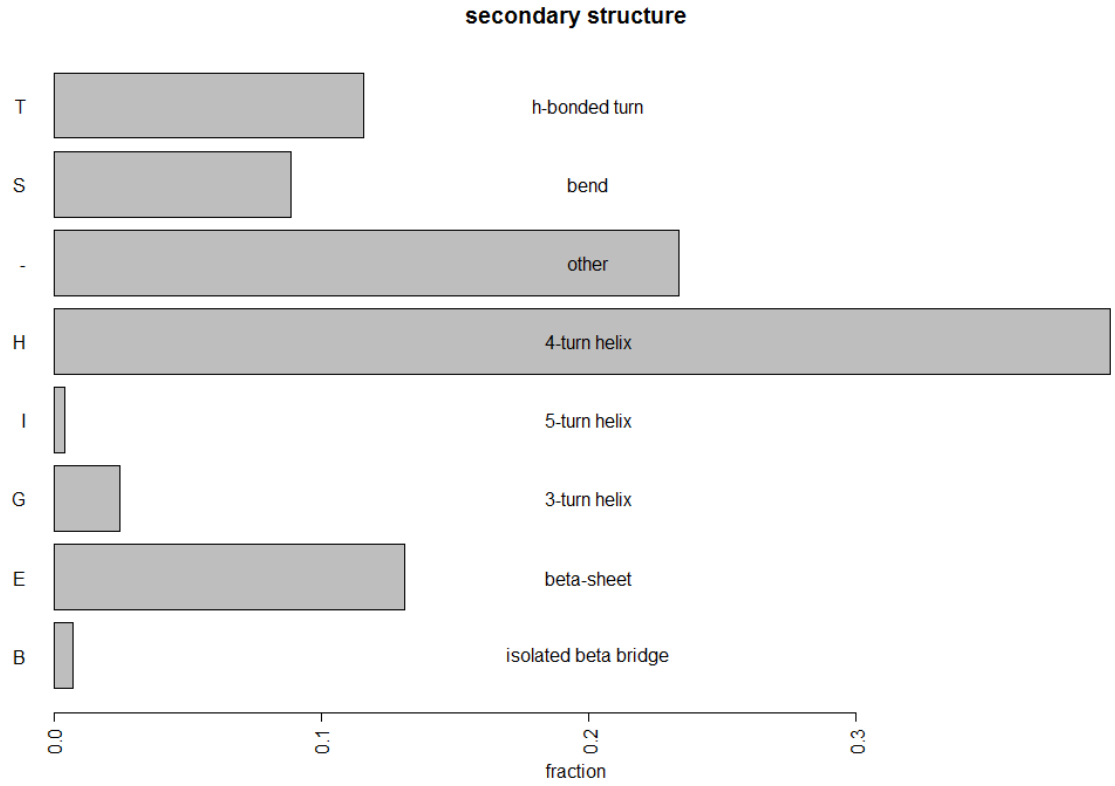


Figure 17: secondary structure elements involved in nucleic acid binding

Quite obviously the 4-turn helix is the most common structural element among all the elements. This can easily be explained by the fact that most DNA binding motifs consist helical structures. Nevertheless, also β-sheets and bends/turns occur as structural elements.

Diversity of structural binding motifs on the example of DNA/RNA binding

The statistics on the huge dataset shown above can easily be separated into smaller datasets containing only one specific structural motif if the information is available (see methods). By doing so, we can demonstrate the diversity of structural motifs even within one binding type, in this case DNA/RNA binding. This is also the main reason why it's very difficult to predict new structural motifs *de novo*. Due to their diversity it is difficult to define their most important feature. Even if they could be defined it would not be sufficient to predict them correctly. In the following we will show the statistics for three different structural DNA binding motifs to underline this. The helix-loop-helix motif and the helix-turn-helix motif, which both consist of two helices, but still show distinct differences in the interacting amino acids, and the leucine zipper motif.

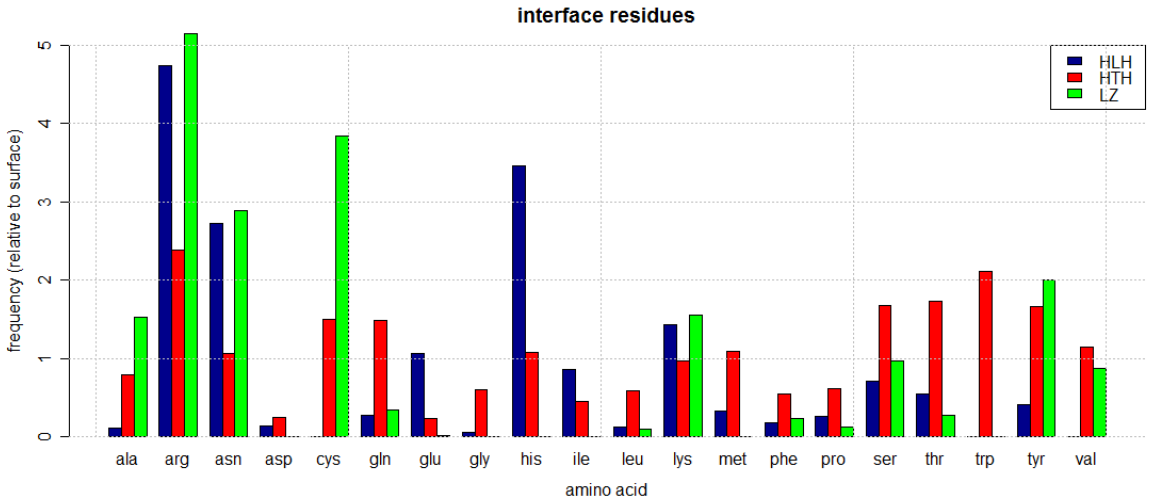


Figure 18: Interface residue frequency for the helix-loop-helix motif (blue), the helix-turn-helix motif (red) and the leucine zipper motif (green). All frequencies are normalized against the general surface residue frequency.

Figure 18 shows the frequency of interface residues. Again arginine is the most abundant amino acid for the HLH motif and the LZ motif, but is much less important in the HTH motif. For the HTH motif the so called stacking interactions by aromatic residues seem to be much more important. In addition, the HLH motif and the LZ show a high

occurrence of asparagine. This seems illogical first, since asparagine carries no charge und should therefore should interact weakly with the negative charge of the DNA backbone. Closer investigation shows that these residues are located at the outskirts of the motifs and the interface area, forming hydrogen bonds and, therefore basically locking the ligand within the positively charged area in the desired position. The high amount of cysteine in the interface of the leucine zipper on the other hand plays an important role during dimerization of the two leucine zipper motifs. Again located at the edge of the interface area, this cysteines promote a stabile dimerization of the two leucine zippers upon interaction.

In case of spatial distances the motifs demonstrate the strong informational value of distance distributions in order to identify structural motifs.

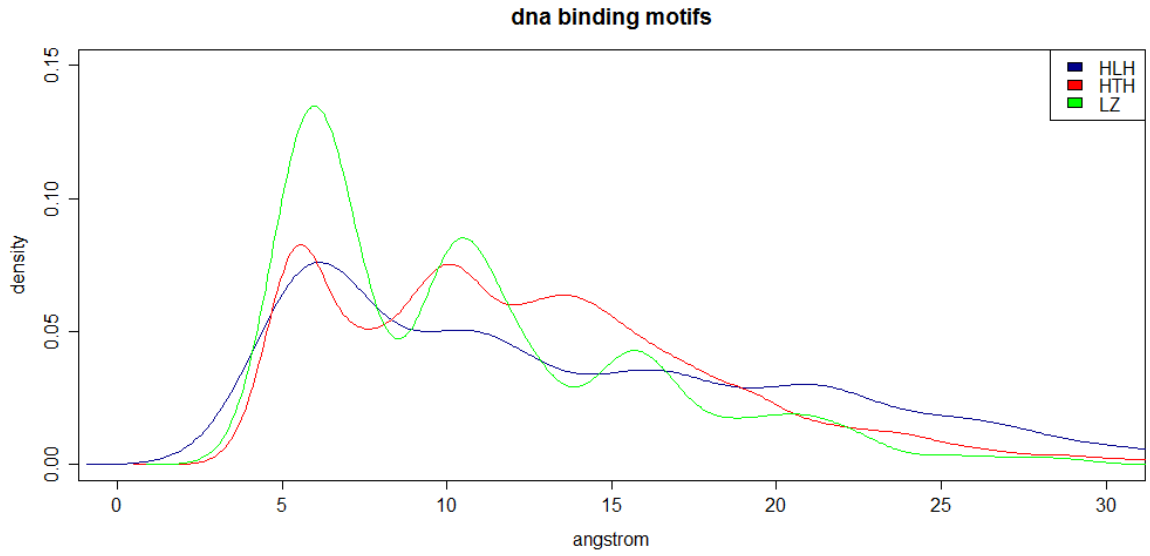


Figure 19: spatial distribution between interface residues in three DNA binding motifs, the helix-loop-helix motif (blue), the helix-turn-helix motif (red) and the leucine zipper motif (green).

As is shown in Figure 19, the leucine zipper shows a very distinct distribution of distances. This is due to the fact that it is basically a very straight helix with the interface residues in an even distance to each other. Such a distinct pattern of distances in combination with residue frequencies could therefor already be a good indicator for the motif itself or motifs with another underlying secondary structure but the same spatial arrangement

The same can be done for the surface accessibility. Again, the HLH motif shows more similarity with the LZ motif than with the HTH.

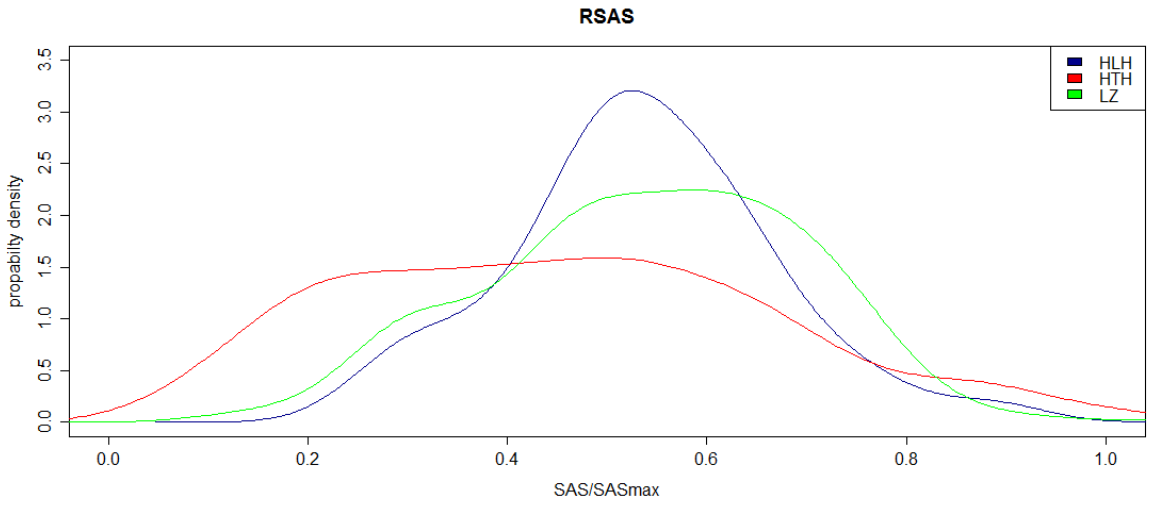


Figure 20: RSAS distribution for three DNA binding motifs, the helix-loop-helix motif (blue), the helix-turn-helix motif (red) and the leucine zipper motif (green).

Although the HLH and HTH might seem more related when it comes to their structural elements, their binding mechanism is quite different. The binding mechanism of the HLH resembles the one of the leucine zipper, often also including dimerization of two HLH motifs to form a scissor like arrangement. In this HLH and LZ show very similar features, although structurally different. This already shows that direct structural comparison not always will give good results, since the same situation could apply to yet unknown motifs.

Our results towards the binding of nucleic acids are conform to the results reported previously in literature, claiming arginine as the major participant. This is true in general, but other motifs or unknown binding sites might only rely very little on arginine or charge in general, but focus much more on stacking interactions. Therefore, searching only for charged patches on a proteins surface might be able to find some binding sites but will still miss those where charge is not the main force involved. In addition, very little information was gathered about the size of binding sites. We now know that most of these interactions take place within a window of 30 angstrom, with a mean around 15 angstrom. The area has to be highly exposed, but the single residues show a median exposure to the surface, since the charged residues seem to cluster together while stacking interactions need more space.

3.3 Statistics of peptide binding

Next we will have a look at the interaction of proteins with small peptides. Figure 21 shows the distribution of interface residues as well as the next and after next amino acid in sequence. In contrast to nucleic acid binding, the main interaction seems to be established by Van-der-Waals forces of aromatic residues. Although an interaction between a protein and a peptide could also be established by charge if complementary residues are used on the peptide, it makes more sense to establish this kind of interaction by a different force in order to limit cross-reactivity.

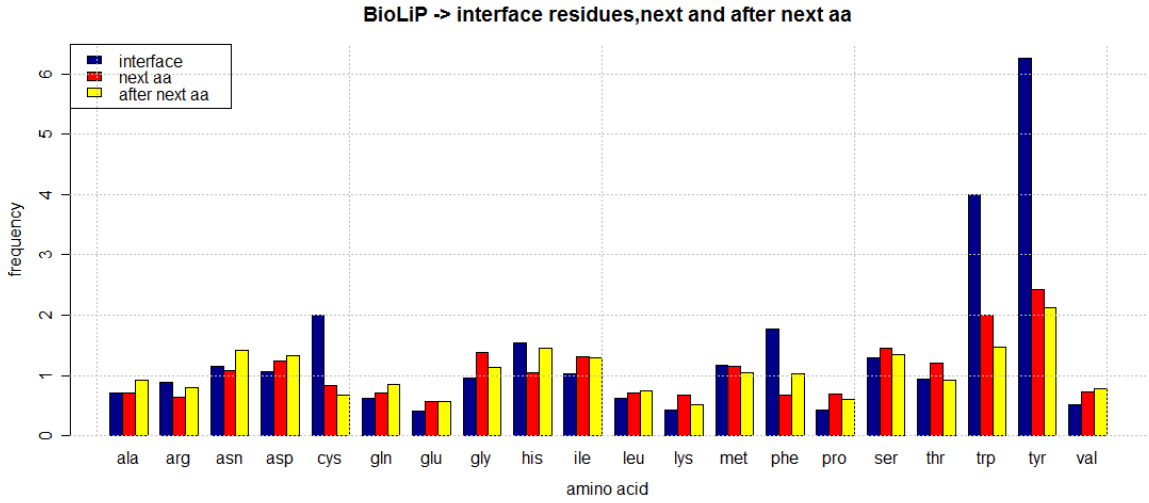


Figure 21: interface residue frequency (blue), next (red) and after next (yellow) residue frequency of peptide binding proteins

Tyrosine and tryptophan are the main amino acids involved in this reaction with a frequency six respectively four times higher on the interface than on the surface on average. Cysteine, probably due to its metal binding or its ability to build disulphide bridges and phenylalanine also seem to be relevant in this interaction type, but are negligible compared to the overwhelming enrichment of aromatic residues. Charge coupled mechanism seem to be totally irrelevant, the interfaces even seems to be charge reduced compared to a normal surface. In addition, the amino acids which are relevant don't seem to bulk together as is the case for nucleic acid interactions. This is not very surprising, since the aromatic amino acids need more space, especially when they pair with an interaction partner. This can be explained in more detail by Figure 22 and Figure 23. Although tryptophan is still quite often followed by another tryptophan, the amino acids tyrosine, methionine, isoleucine and aspartic acid are also quite frequent neighbours. Interestingly the after next amino acid of tryptophan is statistically never

another tryptophan so they never appear in more than a pair before another amino acid follows. This is most likely due to steric reasons. There would be simply not enough space for another tryptophan in between the two aromatic residues. Nevertheless histidine and aspartic acid as well as aspartate itself are very frequent as after next amino acids.

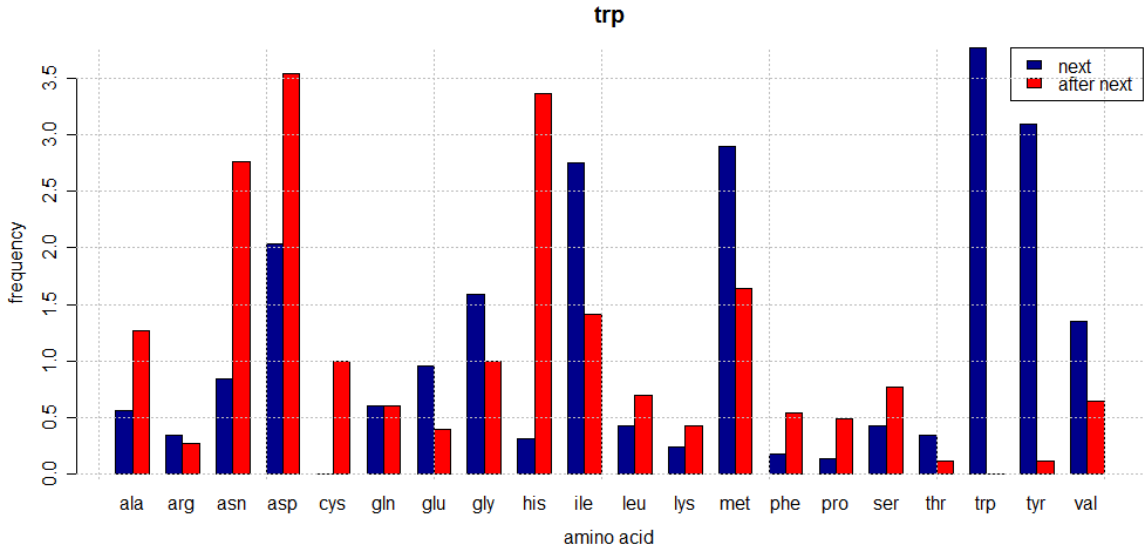


Figure 22: next (blue) and after next (red) amino acid of tryptophan involved in peptide binding reactions.

For tyrosine the situation is quite similar.

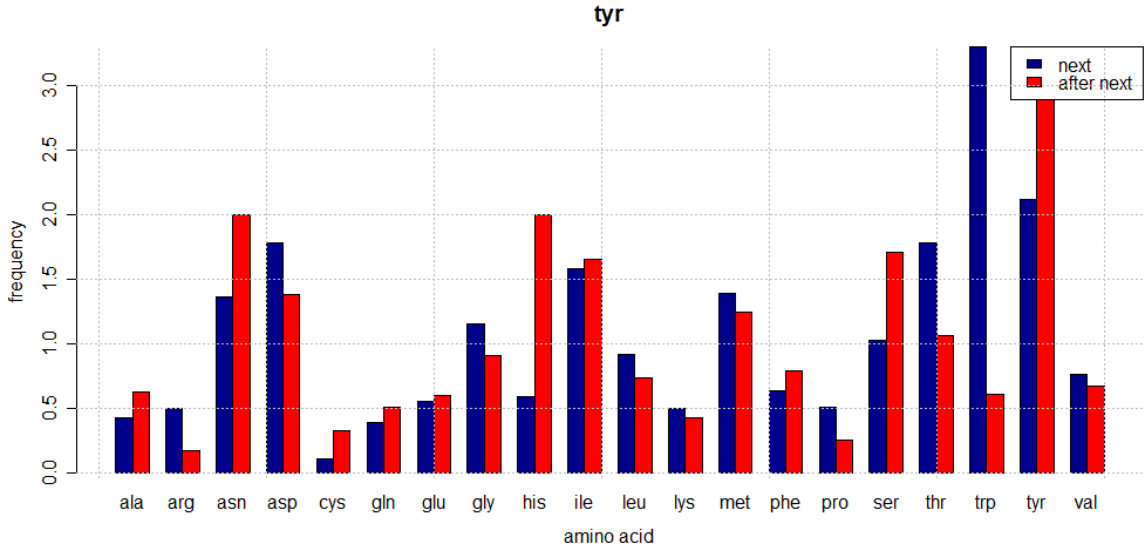


Figure 23: next (blue) and after next (red) amino acid of tyrosine involved in peptide binding reactions.

The distribution of amino acids can be complemented by the distance in sequence for the interface residues, pictured in Figure 24. While half of the residues are direct neighbours, several other sequence distances are more present and more frequent than for a

DNA/RNA binding motif. An addition, the highest observed sequence distance in the whole dataset was 203 amino acids.

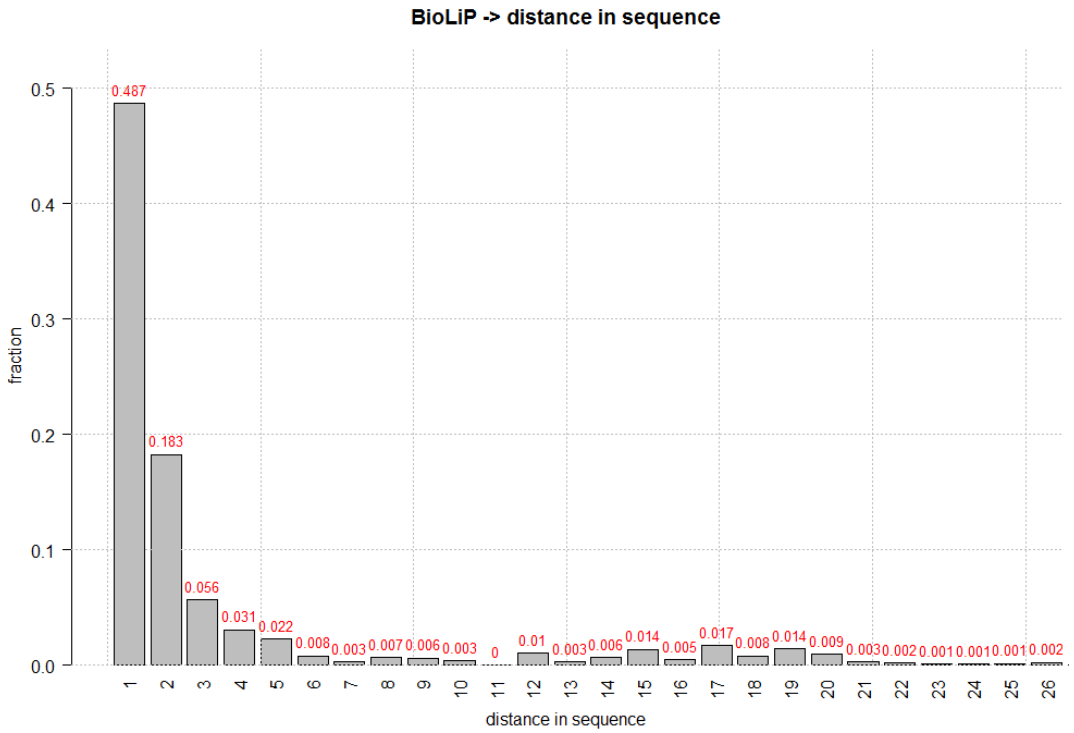


Figure 24: distance in sequence for interface residues involved in peptide binding. Absolute values are shown in red.

Again we can estimate the size of the interface by the distance between the interface residues, like shown in Figure 25. The interfaces seem to be smaller than for DNA binding, which is also to be expected, since the peptides are small compared to the big nucleic acid molecules some motifs bind to. In addition, the peaks for the residues in a specific sequence distance are much less prominent. This underlines that the variation in sequence distance is much higher in peptide binding motifs than in the ones that bind to DNA. But this is not the only reason for the smoother distribution of distances. Figure 26 shows the secondary structure elements involved in peptide binding. While DNA binding is established by very stable helixes, peptide binding involves mostly flexible loops and sheets.

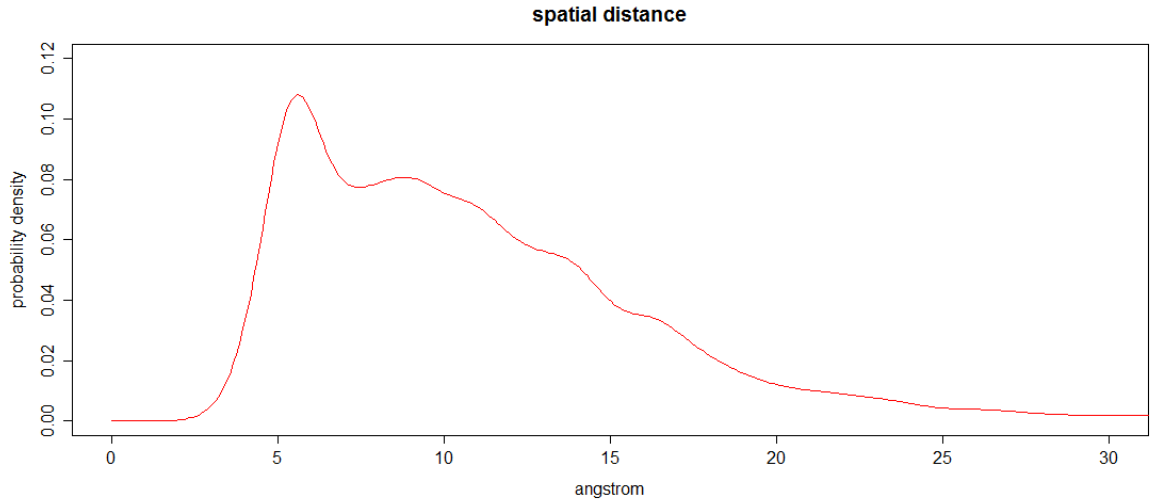


Figure 25: spatial distance between interface residues involved in peptide binding.

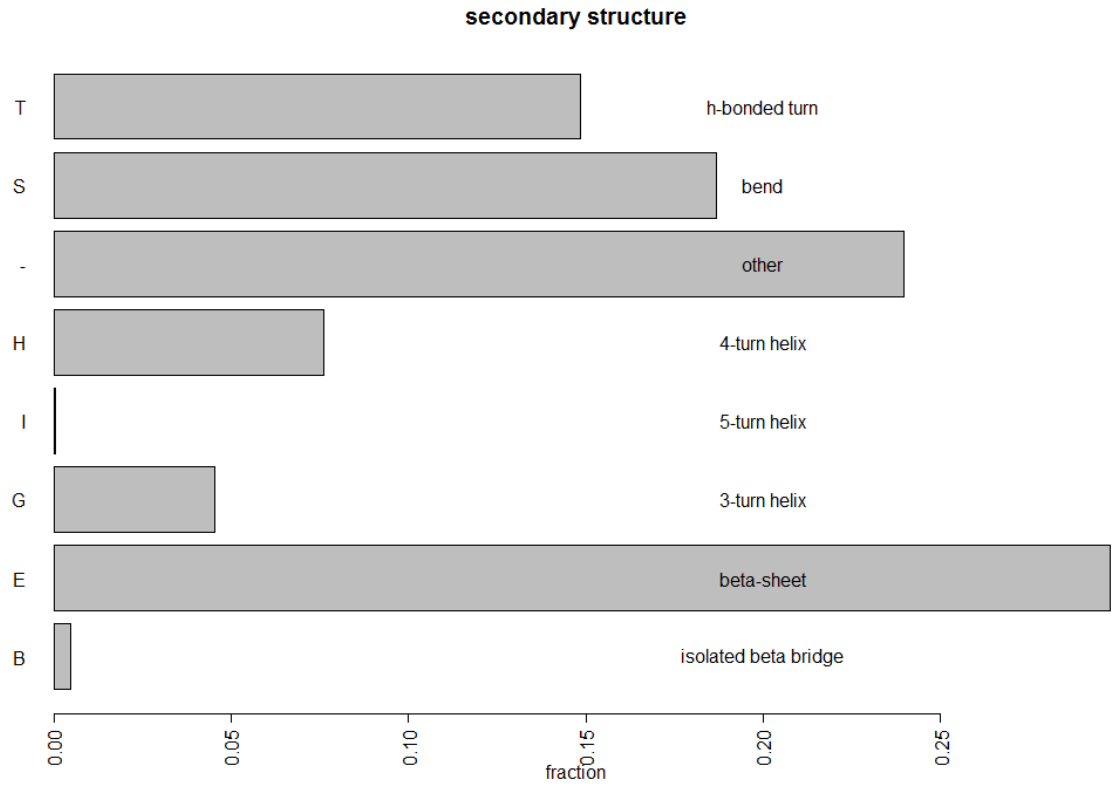


Figure 26: secondary structure elements involved in peptide binding reactions

Helices are irrelevant in the process of peptide binding. This leads to the conclusion that peptide binding involves more protein flexibility. This could raise the question if peptide binding in general is less specific than nucleic acid binding. We will confirm this observation in the next subchapter on the example of antigen-antibody interaction.

Last but not least we also want to examine the surface with respect to accessibility. The distribution shown in Figure 27 is comparable to the distribution for DNA/RNA binding

proteins. Although a broad range of RSAS can be observed, more low RSAS values are reported, supporting a more cavity-like arrangement of residues. This is in fact true. Peptide fragments, stretching out from a protein, often bind inside a pocket.

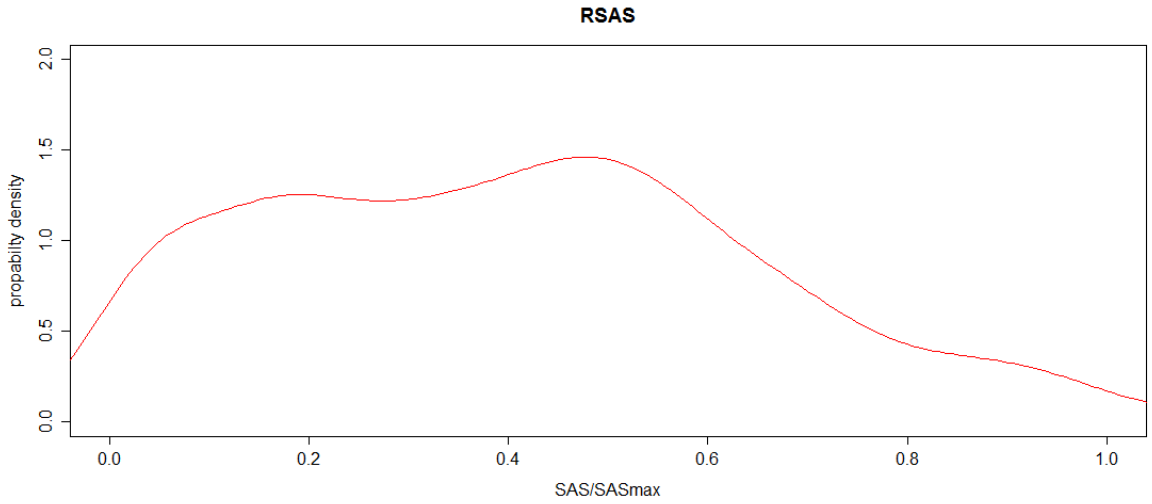


Figure 27: solvent accessibility in the case of peptide binding proteins

Antigen-Antibody interaction:

One specific kind of complex formation we will now look at is Antigen-Antibody (AG-AB) interaction. Figure 28 shows the results for the amino acid frequencies in a non-redundant AG-AB dataset. The distribution is quite similar to the one for peptide interactions in general. Differences can be found for cysteine, which is in fact logical since these interactions are non-covalent and the formation of disulphide bonds is unusual. Besides that, serine seems to be involved in the process of AG-AB interaction,

which is not the case for peptide binding in general. This interaction could be hydrogen-bond related.

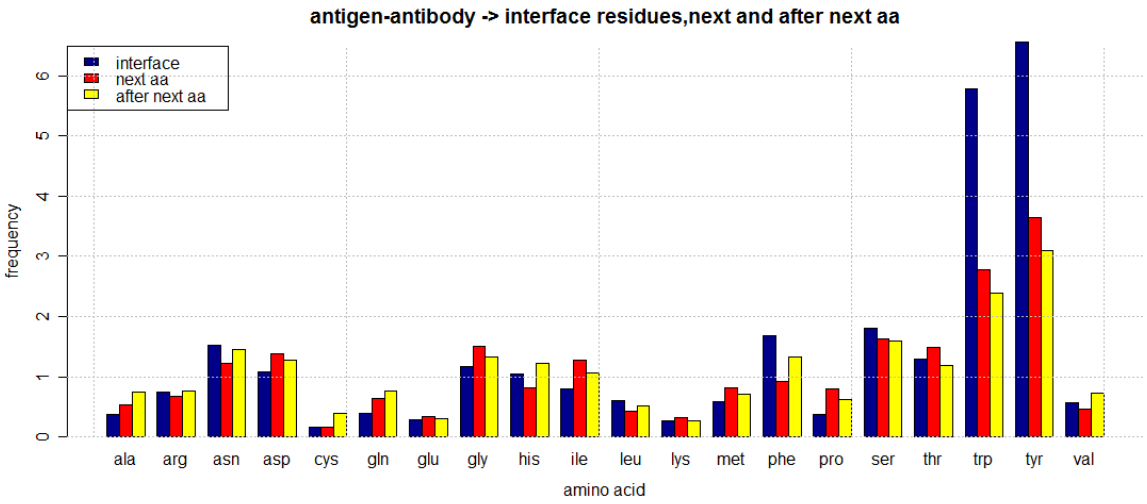


Figure 28: interface residue frequency (blue), next (red) and after next (yellow) residue frequency of AG-AB interactions

The spatial distances get even higher for AG-AB interactions and are more evenly distribuited. The peak for neighbouring amino acids nearly dissapears and has changed into a flat tray, represented in Figure 29

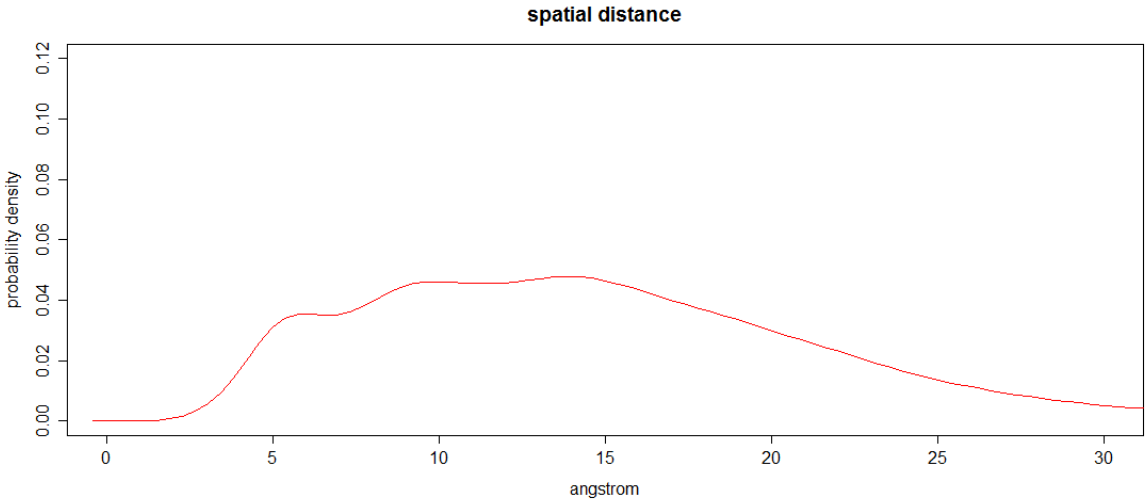


Figure 29: spatial distance between interface residues involved in AG-AB interaction

Interstingly this change can not be explained by the distance in sequence since most interface residues seem still to be neighbouring each other, as is shown in Figure 30. There are two other possible reasons. First, the secondary structure involved in this kind

of interaction is very flexible. Figure 31 shows this statistic, identifieyng turns, bends and sheets as the main elements, resulting in a much less even arrangement of amino acids.

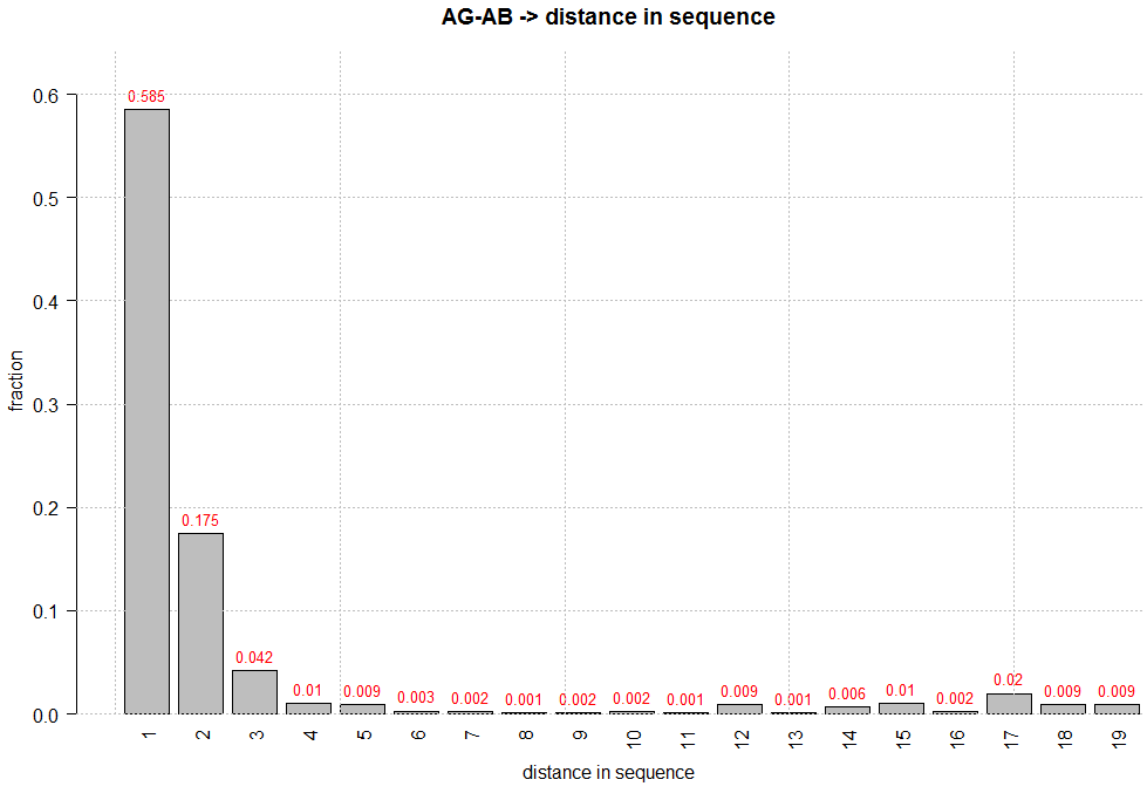


Figure 30: distance in sequence between interface residues involved in AG-AB interaction

The even more important factor cannot be evaluated from the statistics. While most previous motifs where limited to a single chain of amino acids, the motifs of antibodies are formed by two different chains. For these residues, a distance in sequence can of course not be calculated, while their spatial distance is still part of our statistics. This is one very important feature also our prediction method will incorporate. Many methods are limited to single chains, while our approach is not restricted by the number of chains but includes all residues regardless of the chain they are located on.

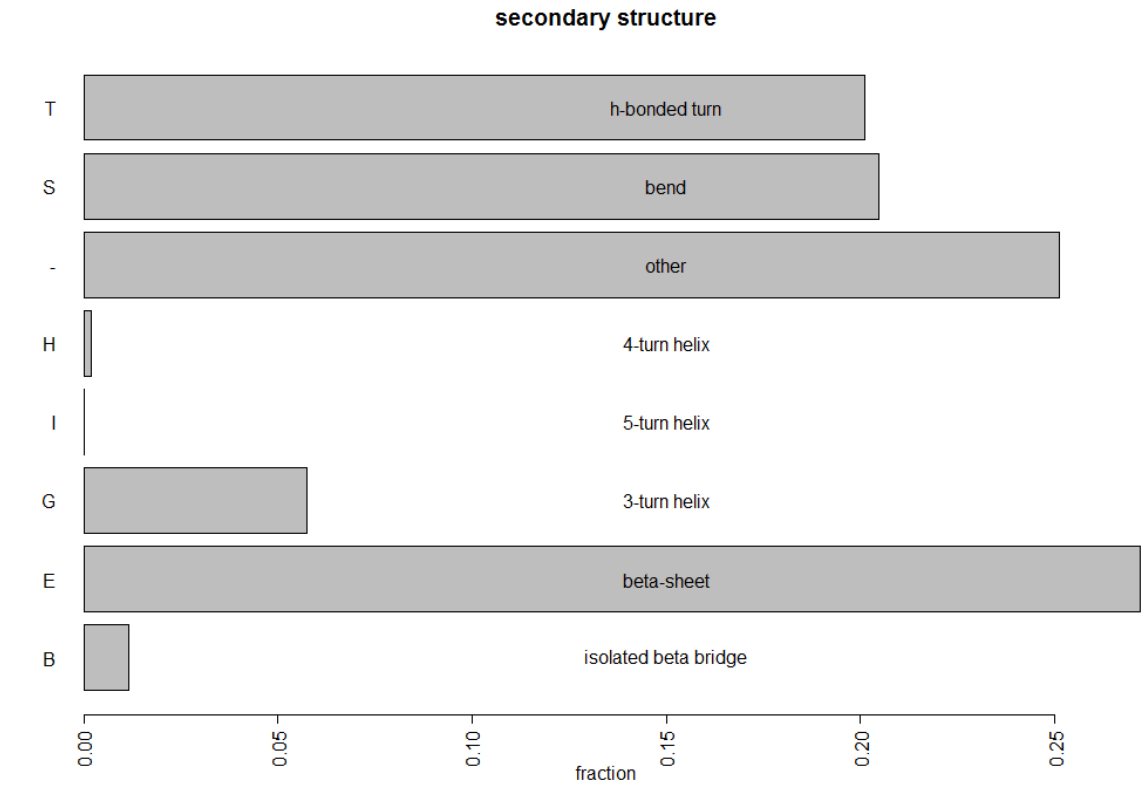


Figure 31: secondary structure contributing to AG-AB interaction

For completion we can examine the RSAS, which is not considerable different from the general peptide interaction.

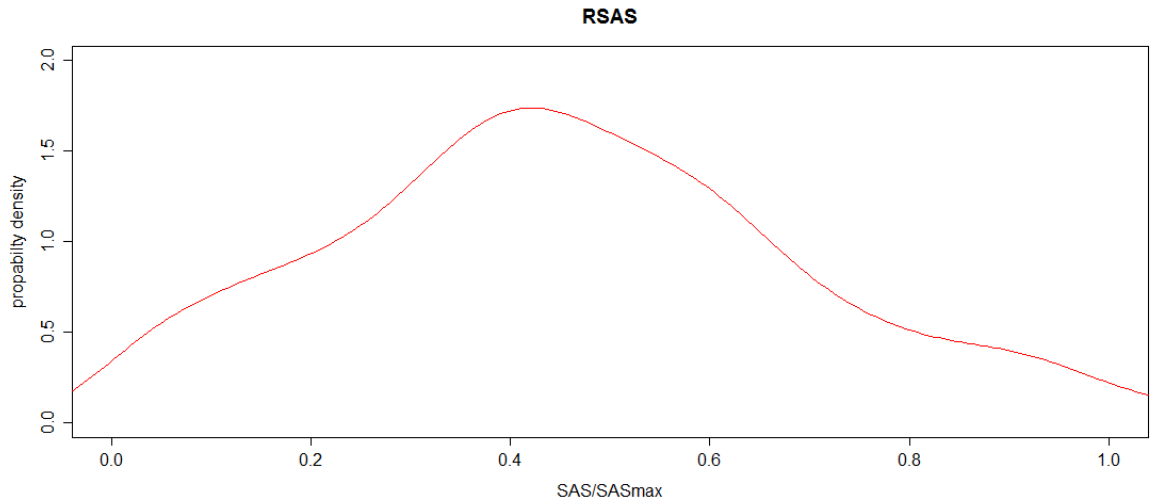


Figure 32: solvent accessibility in the case of AG-AB interaction

3.4 Statistics of Calcium and Magnesium binding

The last statistical analysis will point out the differences and similarities in two very similar mechanism, the binding of Magnesium and Calcium. Both metals are very frequent ligands and identifying metal binding sites inside proteins can be very useful.

The main amino acid involved in binding of calcium as well as magnesium is aspartic acid. The histogram of the amino acid frequency is shown in Figure 33. Other participating amino acids can be glutamic acid, aspartate and cysteine. At first glance the distribution for both metals seems to be nearly the same, which would make it difficult to distinguish between those two.

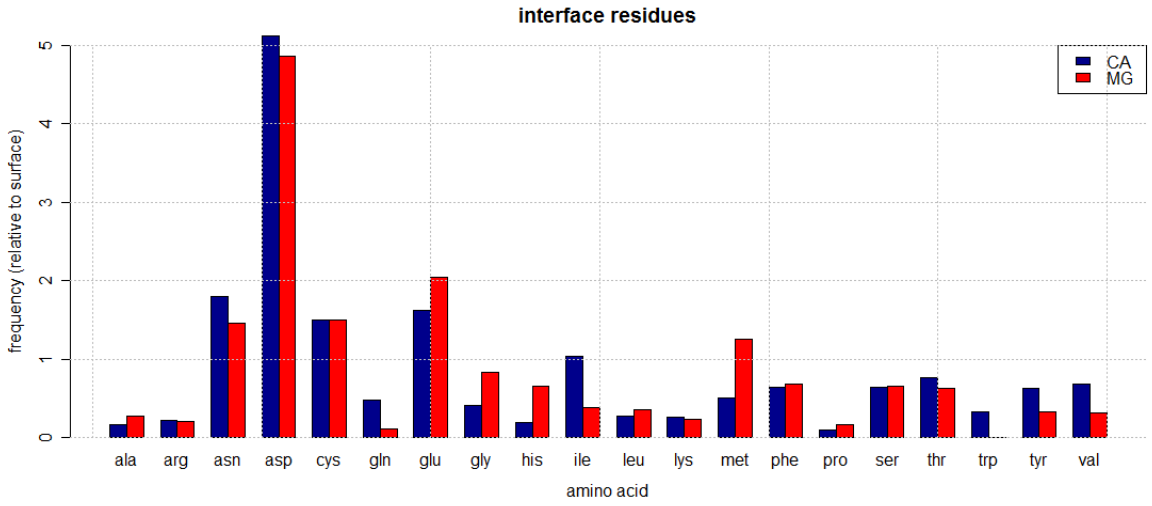


Figure 33: interface residue frequency in proteins binding calcium (blue) and magnesium (red)

This problem can be tackled when looking at the distance between the amino acids in sequence, as well as the distance in space between the residues. In case of magnesium binding, the interface residues, quite often neighbour each other, but can also be up to 22 amino acids apart. For calcium binding, the next interface residue often appears with one other residue in between (see Figure 34 and Figure 35). This fact is also observed in the distance distribution, with the distribution of calcium shifted to higher values, like visualized in Figure 36. Interestingly, although for calcium binding some residues do not seem to be direct neighbours, the characteristic peak in the distance distribution can't be reported. This indicates that the residues, even though they are not direct neighbours, come quite close to each other. Since the calcium ion is a single interaction point, this is logical. The reason for this differences between magnesium binding and calcium binding are obviously the size of the metal ion, with calcium being bigger than magnesium and therefore requiring a bigger cavity. In this case we can quite certainly speak of a cavity like illustrated by the RSAS distribution (see Figure 37 and Figure 38). Here we also see a major problem with the definition of surface residues. For metal binding the

accessibility falls below 16 % for a major portion of the amino acids. Following this, a much smaller cut-off value has to be chosen, otherwise many relevant residues will get lost during the calculations. For further calculations regarding metal binding, we therefor apply a cut-off value of 3 %.

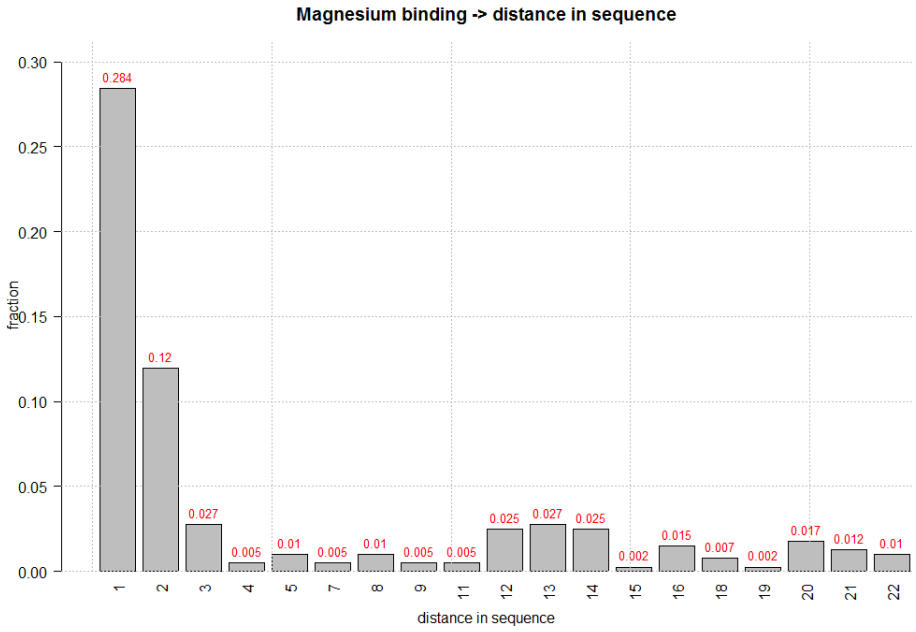


Figure 34: distance in sequence for magnesium binding sites

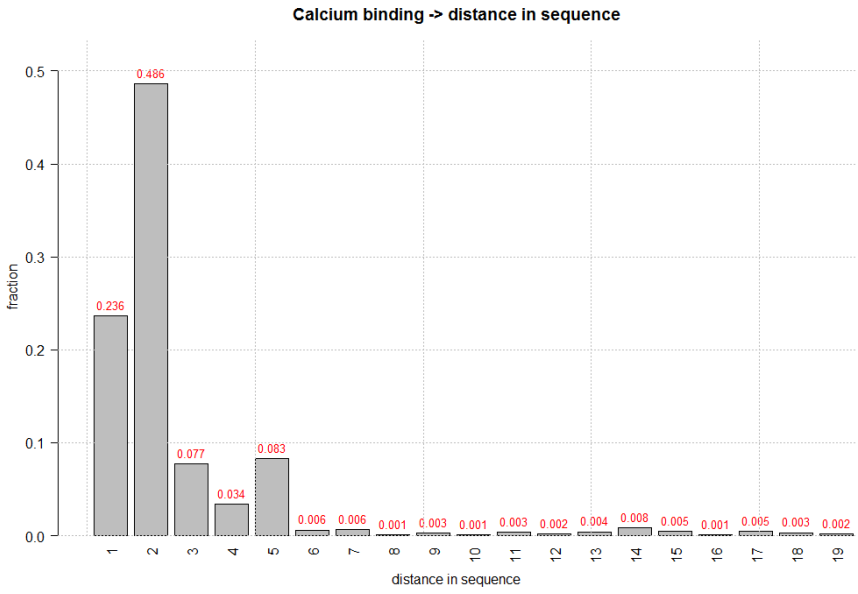


Figure 35: distance in sequence for calcium binding sites

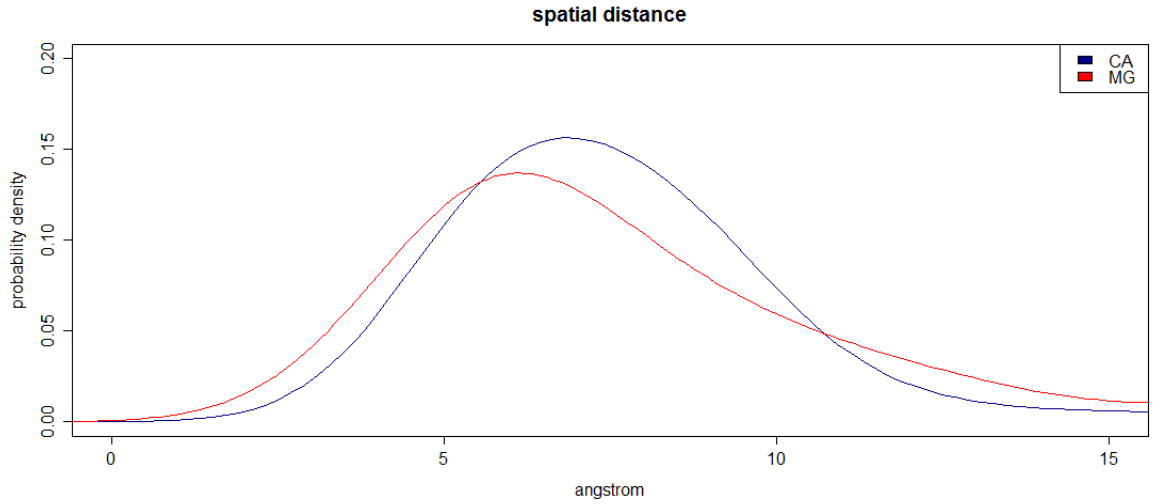


Figure 36: spatial distribution of interface amino acids in calcium binding (blue) proteins and magnesium binding proteins (red)

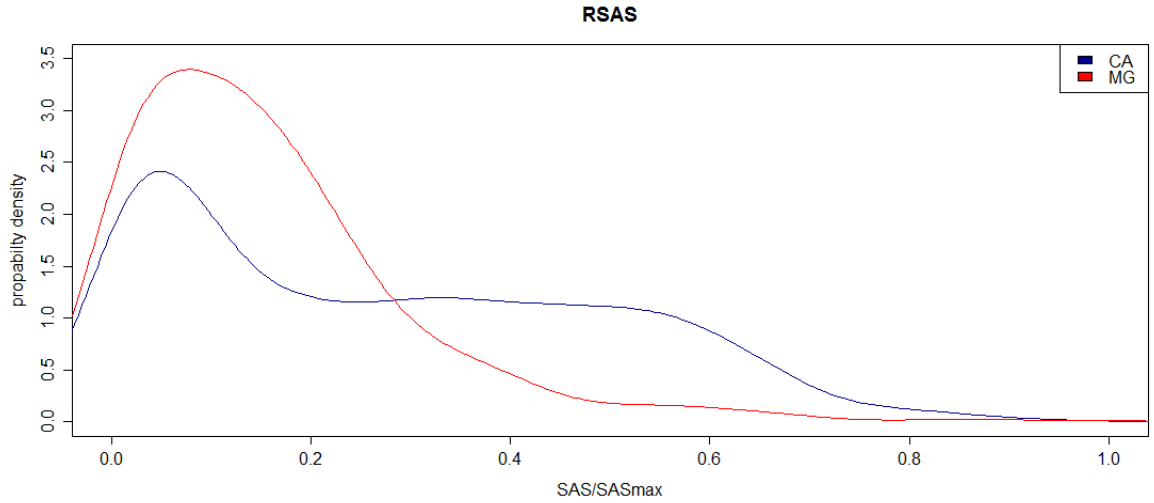


Figure 37: surface accessibility for calcium binding sites (blue) and magnesium binding sited (red)

Calcium and magnesium binding also differ quite strongly in the participating secondary structure. While bends are the major structural element in involved in calcium binding, most residues for magnesium binding are located within a β-sheet.

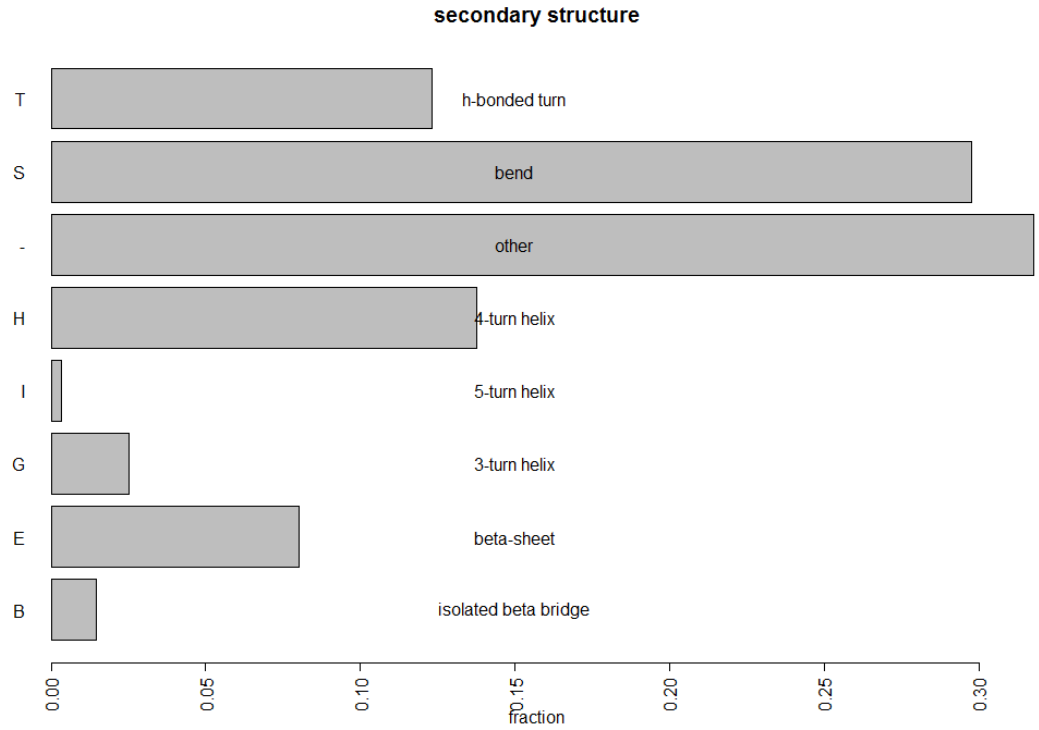


Figure 38: secondary structure involved in calcium binding

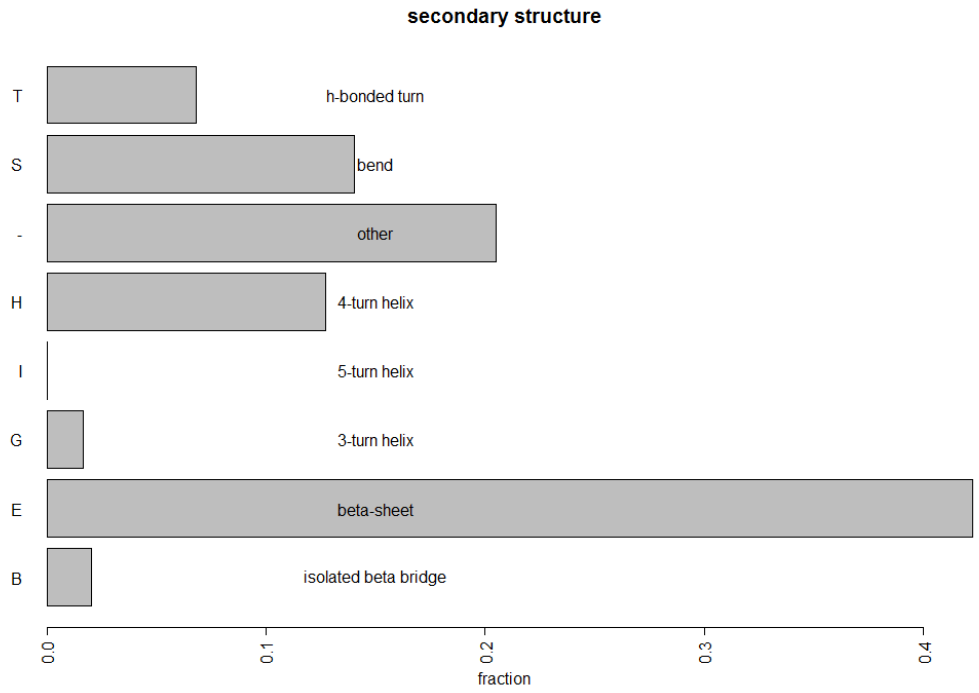
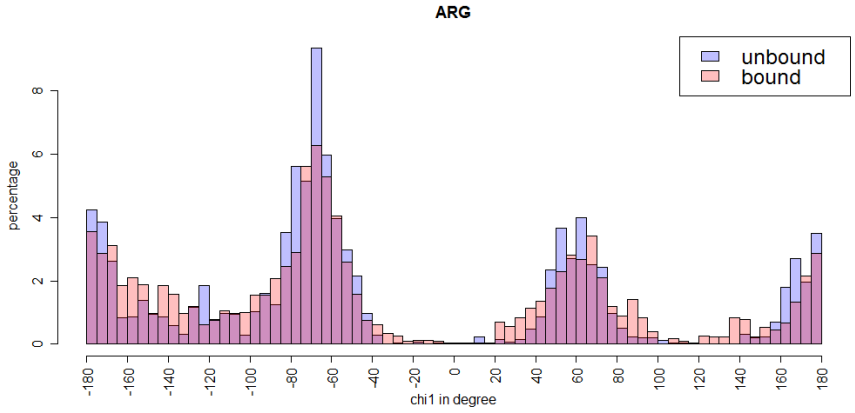


Figure 39: secondary structure involved in magnesium binding

3.5 Excursion: Side chain flexibility in DNA/RNA binding sites evaluated by NMR structures

In this excursion the flexibility of side chains and changes in the distribution of rotamers in the binding site of nucleic acid binding proteins will be discussed. Many methods using structural information consider side chains to be fixed in one conformation which is in fact not true. Side chains have many preferred conformations, between which they can switch very rapidly. This is especially true for unbound surface residues. Considering the atomic information based on the side chains can therefore be misleading.

This fact can easily be demonstrated by the three histograms shown in Figure 40, representing the three cases we can observe for most residues on the binding sites after calculation of all χ1 angles. For the very relevant, charged residue arginine the differences between the bound and unbound state are rather small, displayed in first histogram. In the bound state, the distribution is a little more spread towards unusual states, nevertheless in general the distribution is in general quite similar. Since the side chains are long and quite flexible in the unbound state, the conformational space is quite big. The occupation of more unusual states in the bound state is most likely caused by the attraction between the charged residue and the interaction with the DNA backbone, forcing the side chain into these states. The same was observed for lysine.



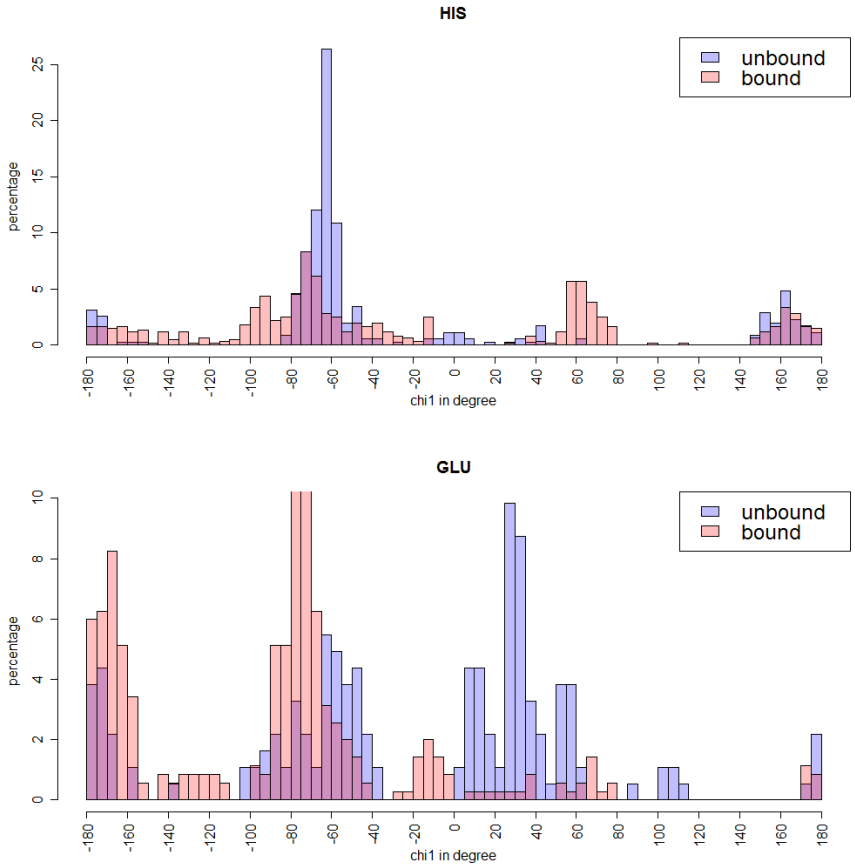


Figure 40: side chain angle distribution. Top: distribution of rotamers for arginine in the bound (red) and unbound state (blue). The preferred states of arginine are 62°, -177°, -67°, and -62°. Middle: distribution of rotamers for histidine in the bound (red) and unbound state (blue). The preferred states of histidine are 62°, -177°, and -65°. Bottom: distribution of rotamers for glutamic acid in the bound (red) and unbound state (blue). The preferred states of glutamic acid are 62°, 70°, -177, -65°, and -67°.

Due to the high flexibility of the long arginine side chains this is rather unsurprising to occur. Nevertheless it indicates structural changes during binding. For aromatic residues involved in the binding, like histidine or tryptophan, we can report quite the same but in a much severe fashion. The unbound and bound state differ strongly. While the unbound residue shows a rather limited flexibility, therefore occupying mainly the preferred rotameric states, the bound states shows a broad distribution. Moreover, it contains residues in states, which are not reported at all for the unbound mode. This might quite likely be due to the stacking interactions. For this kind of interaction to occur the aromatic residues has to show a certain spatial arrangement towards the base pairs (*102*). Since the

base pairs themselves are inflexible, the only way to establish the interaction is due to changes in the amino acid conformation, as is reported here.

The opposite behaviour can be observed for most residues, which are rather unlikely to be present in the binding site like glutamic acid. In an unbound state, many different states can be reported, most likely indicating a high amount of flexibility. Interestingly also a high amount of uncommon states can be reported. Glutamic acid shows a cluster of states around 40° for χ1. This might be caused by reciprocal action with residues of the opposite charge on the binding site. Upon binding, the states shift to more discrete values and often preferred values. For example in the case of Glutamic acid, which carries a negative charge, this might be due to the repulsive forces, resulting in a more limited conformational space, basically locking the residues in one or two preferred conformations. This could even be a possible binding concept. In the unbound state parts of the charged patch might be neutralized by opposite charged residues in a less preferred conformational state. If ligand and protein come in close contact, the repulsive forces can easily force these residues into a more preferred conformational state, exposing the binding site fully in order to establish contact. A concept quite similar to the so-called induced fit in the case of enzymes.

The result can be confirmed by the average deviation from the preferred side chain angles shown in Table 3. Lysine and Arginine show higher deviation in the bound state than in the unbound state although the difference is not that severe, while for tryptophan and histidine very high values can be reported. For cysteine and, as mentioned before, glutamic acid, the deviation from the preferred states is much higher in the unbound state than in the bound state.

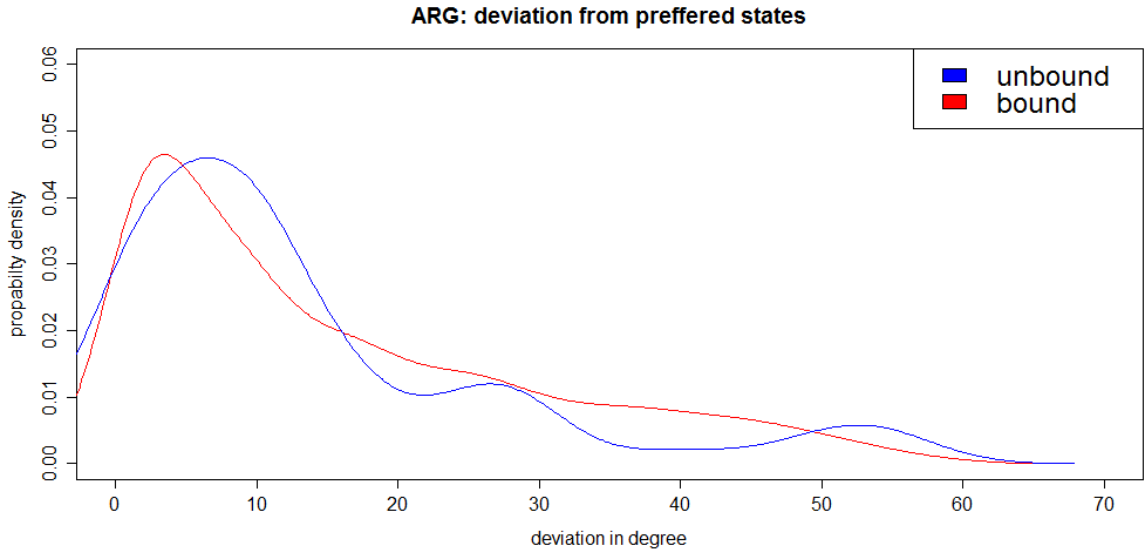
Table 3: deviation from preferred state for χ1

	Average deviation in degree from closest preferred state	
amino acid	**bound**	**unbound**
ARG	15.60	13.54
ASN	15.41	13.11
ASP	13.06	11.87
CYS	12.12	23.30
GLN	15.15	9.40
GLU	14.94	27.17
HIS	16.36	10.61
ILE	12.91	13.76
LEU	13.46	7.52
LYS	13.52	9.42
MET	13.80	11.66

PHE	14.23	15.27
PRO	7.27	8.07
SER	16.52	10.73
THR	20.76	20.05
TRP	19.54	13.48
TYR	14.59	19.75
VAL	15.09	12.35

By creating a distribution over the deviation instead of calculation an average, the behaviour can be described in more detail. This type of distribution is shown in Figure 41. A peak close to zero is caused by all rotamers in a conformation close to a preferred state.

A broader peak in this area in the unbound state might indicate more flexibility in general, while a broader distribution over several values quite likely can be interpreted as residues, which are forced into unfavourable states. It needs to be considered that in an NMR structure, the models submitted are still filtered, containing only these with the lowest energies.



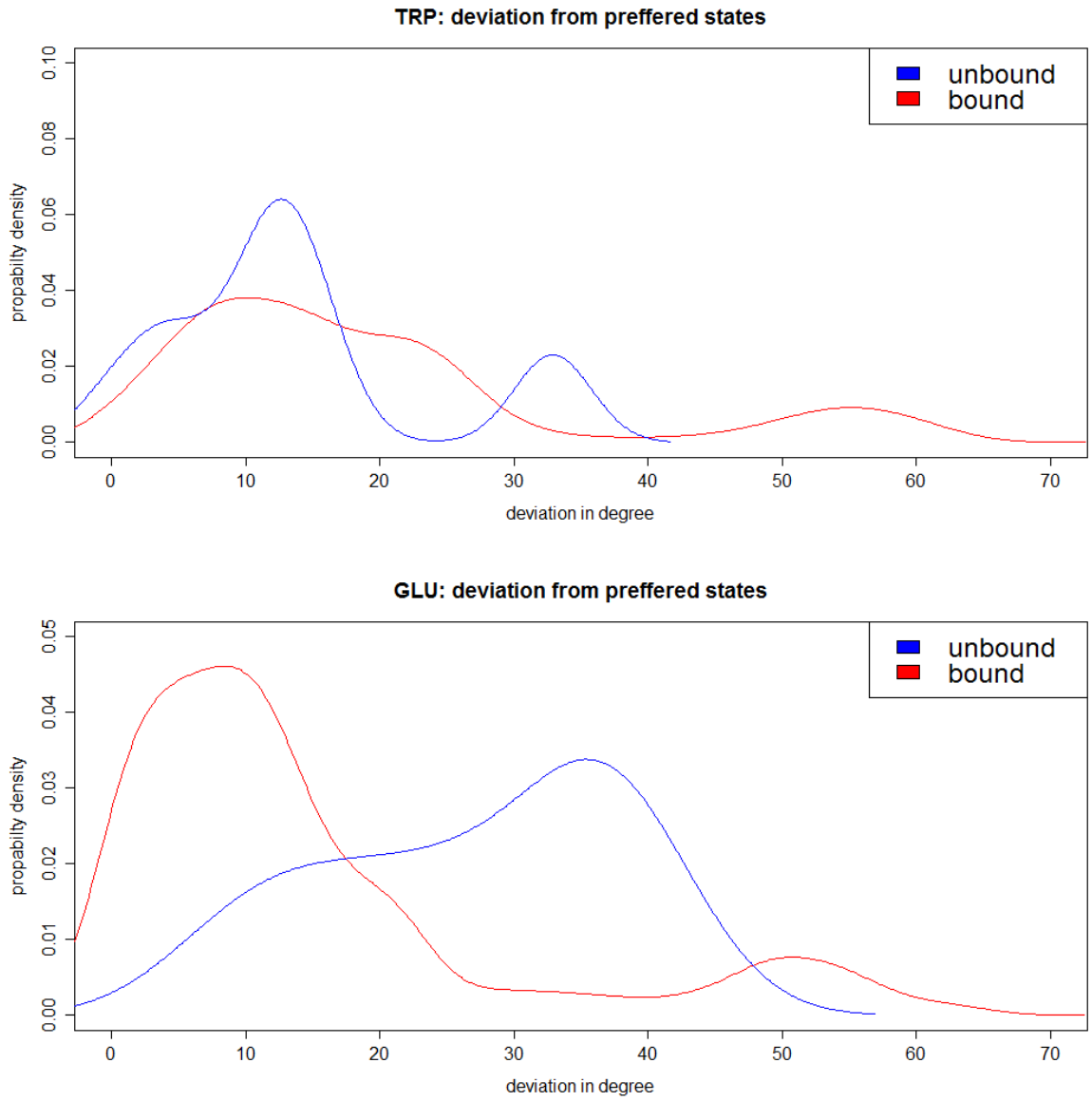


Figure 41: Top: distribution over the deviation from the preferred states for arginine. Middle: distribution over the deviation from the preferred states for tryptophan. Bottom: distribution over the deviation from the preferred states for glutamic acid.

For arginine the bound and unbound state don't differ much and most rotamers are close to the preferred states. Nevertheless in the bound state the distribution shows several peaks. The first broad peak close to zero can be assigned to residues in the preferred state. The two smaller peaks around 30° and 55° are caused by transition states of one preferred rotamer into another. This peaks disappear in the broader distribution of the bound state. Here many more states seem to be present due to interaction with the ligand. The same can be reported for tryptophan, where the bound residue shows a very broad distribution with states deviating up to 30° from the preferred state in order to establish stacking interactions. We see quite the opposite we see for glutamic acid. The unbound residue

very often occupies unusual rotamer states, while in the bound state it shifts strongly towards the preferred states.

In summary, this small excursion shows that the consideration of atomic coordinates of the side chains is very challenging. The changes in conformation are quite frequent, especially for residues involved in the interaction. A previously rather flexible residue can be found in a rather unusual conformation upon binding. These changes show the difficulty of purely geometry based approaches. A certain arrangement of amino acids might not be present in the unbound protein. Since the prior knowledge is mostly based on bound structures, many possible binding sites might be missed. This disadvantage is not present in methods which rely on more general features. In addition, it also needs to be mentioned, that most assessment and benchmarking is done on datasets containing already bound structures. A much more accurate approach would be to use data as a benchmark, which is available in an unbound and a bound state. Only then the true predictive power of a method can be assessed. Our results are in line with the very few results given in literature, also reporting changes in structure and side chains upon binding, although they are based an X-Ray structures and account only for protein – protein interactions (*103*).

Given our results, it might be the best approach to only consider low structural information like the Cβ – orientation for accessible side chains, which will still give a good impression of the orientation of the chain regardless of changes in the rotameric state and which should not change much upon ligand binding. With the rapidly growing amount of NMR structures available the demonstrated approach can be used for much more detailed analysis of side chain flexibility, also covering the $\chi 2$ angle or a direct structural comparison of atomic coordinates including changes in backbone conformation. Knowledge about this topic is not only relevant in the shown context of prediction methods, but also plays a major role in protein modelling and docking approaches.

3.6 By residue assessment

ROC curves:

To determine score performance and the optimal threshold ROC curves where calculated. For the for complete BioLiP datasets the ROC curves are shown in Figure 42.

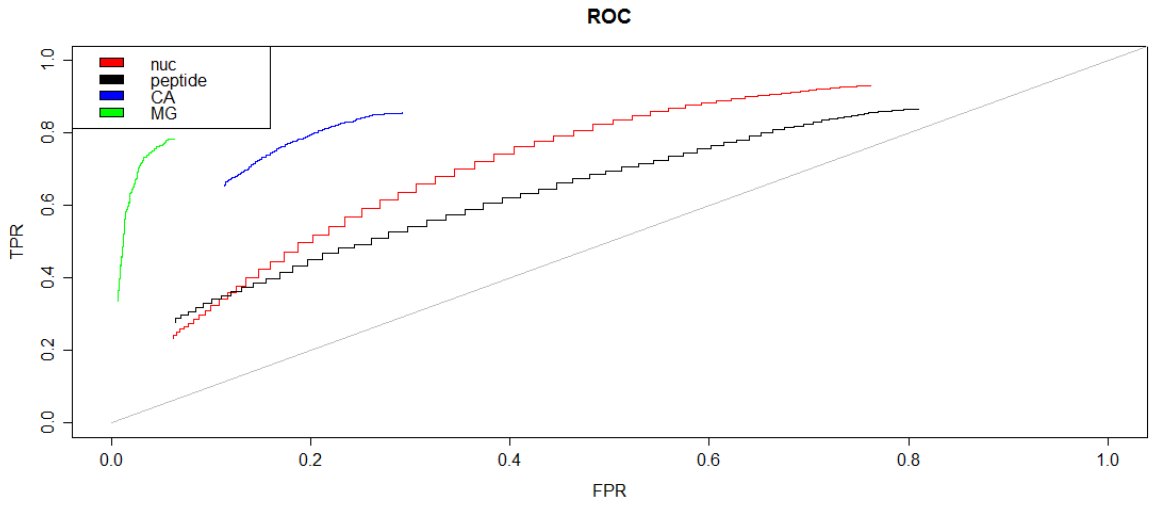


Figure 42: ROC analysis for nucleic acid binding (red), peptide binding (black), calcium binding (blue) and magnesium binding (green). Since the ROC analysis follows the first scoring cycle, the curves never reach an FPR of 1.0. The threshold was increased in steps of 0.1.

For all four datasets a positive predictive value can be observed in the ROC curve. The prediction of the Magnesium and Calcium binding residues seems to work much better than for the DNA/RNA and peptide data. This is due to the high diversity in the latter two, which makes it more difficult to achieve a more accurate prediction.

Table 4: ROC analysis of the four main datasets DNA/RNA, peptide, magnesium and calcium binding

Dataset	Y-Index	Threshold
DNA/RNA	0.337	9.1
Peptide	0.240	10.5
Magnesium	0.724	2.2
Calcium	0.596	7.0

Table 5: statistical measurements for the four main datasets DNA/RNA, peptide, magnesium and calcium binding

Dataset	Sensitivity	Specificity	Precision	Accuracy	FMeasue	MCC
DNA/RNA	0.742 (0.197)	0.596 (0.252)	0.261 (0.142)	0.632 (0.142)	0.363 (0.162)	0.244 (0.151)
peptide	0.451 (0.373)	0.789 (0.122)	0.108 (0.114)	0.761 (0.097)	0.151 (0.121)	0.117 (0.153)
MG	0.788 (0.300)	0.911 (0.023)	0.103 (0.082)	0.909 (0.0.23)	0.174 (0.121)	0.251 (0.137)
CA	0.790 (0.100)	0.806 (0.261)	0.223 (0.139)	0.794 (0.089)	0.330 (0.173)	0.340 (0.166)

All four datasets perform very well in case of sensitivity and specificity. Nevertheless, precision is a huge problem. Due to the overwhelming amount of residues compared to the very few interface residues, the number of false positives can be quite high. Nevertheless this is not an unexpected problem. Other methods based on machine learning approaches report values in a comparable range for a residue based assessment, although direct comparison is difficult due to the different datasets and definitions (*89, 104, 105*). In addition, a residue based assessment is not the main goal of the method, which does not aim to identify discrete residues, but a specific area in the protein.

Isolated DNA motifs:

To check how performance changes when working with more distinct, smaller datasets the similar assessment method was performed on the DNA binding motifs with the corresponding profiles. The ROC curves can be found in Figure 43.

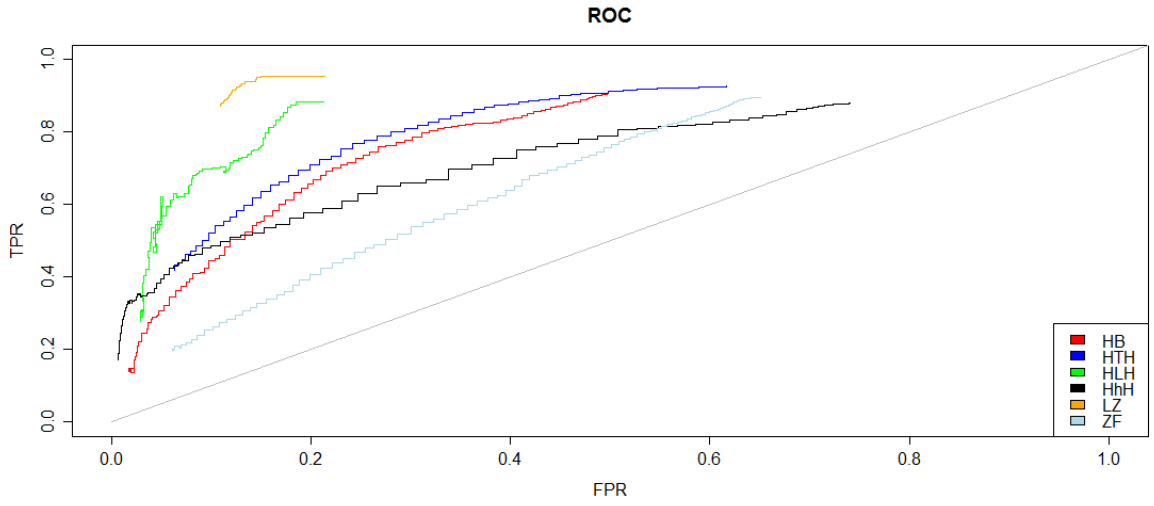


Figure 43: Roc analysis for the six different DNA binding motifs. Since the ROC analysis follows the first scoring cycle, the curves never reach an FPR of 1.0. The threshold was increased in steps of 0.1.

Table 6: ROC analysis for isolated DNA binding motif datasets

Dataset	Y-Index	Threshold
HLH	0.693	3.0
HTH	0.513	9.5
HhH	0.380	7.2
LZ	0.804	7.6
ZF	0.260	5.7
HB	0.484	6.1

The performance on the motif datasets is much better than on the unspecific datasets. Especially the LZ, the HLH and HTH motif dataset perfom very well. This might be due to the fact that these sets are the most distinct ones with very little differences within the structures, creating a very significant profile. This can also be confirmed when locking at statistical measurements

Table 7: statistical measurements for isolated DNA binding motif datasets (standard deviation)

Dataset	Sensitivity	Specificity	Precision	Accuracy	FMeasure	MCC
HLH	0.881 (0.145)	0.811 (0.070)	0.438 (0.158)	0.818 (0.063)	0.564 (0.157)	0.528 (0.149)
HTH	0.767 (0.236)	0.746 (0.137)	0.342 (0.170)	0.758 (0.100)	0.443 (0.172)	0.378 (0.159)
HhH	0.509 (0.285)	0.871 (0.033)	0.160 (0.100)	0.853 (0.028)	0.235 (0.138)	0.219 (0.158)
LZ	0.954 (0.062)	0.849 (0.046)	0.496 (0.130)	0.863 (0.039)	0.642 (0.118)	0.622 (0.101)
ZF	0.816 (0.213)	0.443 (0.221)	0.242 (0.128)	0.526 (0.160)	0.358 (0150)	0.205 (0.131)
HB	0.759 (0.169)	0.725 (0.262)	0.489 (0.154)	0.742 (0.100)	0.562 (0.154)	0.446 (0.168)

For the specific motifs, the method achieves very good performance values. This confirms what we already expected from the analysis. Structural binding motifs are very diverse and prediction from a general set of features is challenging. Assuming that most interactions are established by either a known motif or a structure with very similar features to one of the known ones, it could be useful to rely on smaller, more precise datasets for training instead.

3.7 By patch assessment

The assessment by patch is in fact much more relevant to our approach, but also more difficult to realize. In our case, we use a very rough approximation in form of a circle shaped patch of a fixed size. This is of course not accurate, since the real interface can be shaped quite differently. The difficulty of patch assignment has already been discussed in literature and it was shown that finding the right patch size and shape can very strongly influence performance (*33*). Given that it would be worthwhile to develop a method which assigns the patches more dynamically and also is able to recognize clustering of high scoring residues. First calculations regarding this process have already been

performed during this work was written. Nevertheless, the approximation of a round patch gives as a first impression if our method is able to identify the right binding area.

Table 8: assessment values per patch (standard deviation)

Dataset	Hits percentage	overlap	Rank	correlation
RNA/DNA	41 % (33 %)	30 % (33 %)	28/129	0.29 (0.24)
Peptide	27 % (30 %)	29 % (39 %)	28/155	0.22 (0.18)
Calcium	42 % (37 %)	45 % (36 %)	11/153	0.51 (0.24)
Magnesium	46 % (46 %)	45 % (47 %)	15/299	0.250 (0.17)
HLH	84 % (20 %)	86 % (15 %)	2/65	0.77 (0.11)
HTH	53 % (30 %)	46 % (30 %)	12/116	0.40 (0.20)
HhH	48 % (34%)	30 % (33 %)	65/322	0.31 (0.25)
LZ	79 % (19 %)	72 % (23 %)	3/65	0.81 (0.11)
ZF	37 % (32 %)	31 % (32 %)	19/90	0.27 (0.29)
HB	61 % (27 %)	41 % (37 %)	19/67	0.48 (0.23)

Like we can see, the average values are promising, although we report a rather high standard deviation, indicating that for some structures, a faulty patch is assigned, most likely due to a very different shape. Correlation between the hits and the score of a patch seems to be good, as is displayed in Figure 44 for the HLH motif.

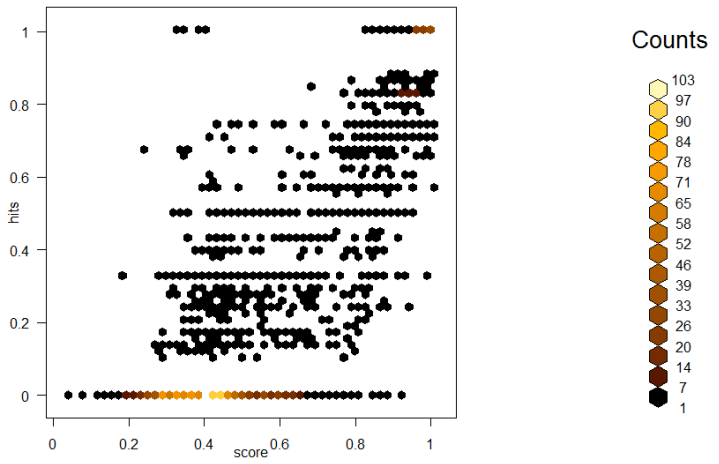


Figure 44: Correlation between number of hits and the score for the HLH motif, both values normalized between zero and one. Low quality patches should be located in the left lower corner, high quality patches in the upper right.

The assessment of the developed method shows that the approach in general works, although it differs strongly between the datasets. What exactly influences this is up to discussion and could be further investigated by using smaller datasets and analysing individual cases. In summary the more specific approach using small datasets seems to

work much better. This is not surprising given the fact that the big datasets result in a quite broad average over all containing binding sites. Further assessment is required especially by independent datasets, which were not part of the training procedure. Up to this date only for DNA and RNA binding some benchmark datasets are available. Other publications have shown that the comparison of such approaches is very difficult, mostly caused by the many different ways to define interface residues (*89*). The most interesting assessment way is a more advanced patch analysis, assigning non spherical patches. The only method using a rather similar approach in order to predict protein-protein binding sites also states patch assignment as one main difficulty and reports comparable values for overlap of patches (*33*).

3.8 Method flexibility and application

In this last chapter of the results we will demonstrate our methods flexibility and reliability based on actual application. Like discussed in the methods, we adapted our algorithm towards a more generic surface comparison in order to identify proteins with a surface patch similar to the *switch2* region in Rab proteins. Table 9 shows the top results of the calculations done on a non-redundant dataset of 9733 human protein structures, as well as on a dataset of seven Rab proteins with a threonine in the centre (Rab_T – set) of *switch2* and 11 Rab proteins with a serine in the centre of *switch2* (Rab_S – set).

Table 9: top results of the patch search for 9733 human proteins. The proteins of the Rab_T – Set in green, the *Rab_S – set in red and the Rab_S – set with inclusion of serine as patch centre in orange.*

#	pdbID	found residue	centre	score	ΔÅ	Short description
0	4lhw	T72.C		4.85		rab8 in its active gppnhp-bound form
1	4LHY	T72.A	T72.A	4.17	0.0	gdp-bound rab8:rabin8
2	2HUP	T82.A	T82.A	3.96	0.0	rab43 in complex with gdp
3	3DZ8	T86.A	T86.A	3.8	0.0	rab3b gtpase bound with gdp
4	3NKV	T72.A	T72.A	3.73	0.0	rab1b covalently modified with amp at y77
5	2GF9	T86.A	T86.A	3.7	0.0	rab3d in complex with gdp
6	3TW8	T74.B	T72.B	3.13	6.2	dennd 1b in complex with rab gtpase rab35
8	1X3S	T72.A	T72.A	2.87	0.0	rab18 in complex with gppnhp
9	1qmn	T366.A		2.67		alpha1-antichymotrypsin serpin
10	5dj4	T480.A		2.66		leucine-bound sestrin2 from homo sapiens
11	2yd0	T886.A		2.41		aminopeptidase 1 erap1
12	3rjo	T886.A		2.36		erap1 peptide binding domain
13	4pa0	T786.B		2.36		human beta-cardiac myosin motor domain
14	4a82	T303.A		2.24		transmembrane conductance regulator

15	4pbx	T371.A		2.18		receptor protein tyrosine phosphatase sigma
16	1zzj	T15.C		2.12		kh domain of hnrnp k in complex with ssdna
17	4x9r	T517.A		2.12		plk-1 polo-box domain
18	2yd9	T151.A		2.04		receptor protein tyrosine phosphatase sigma
19	4dur	T581.B		2.02		full-length type ii human plasminogen
20	4ri0	T342.C		1.99		serine protease htra3, mutationally inactivated
21	4cim	T55.A		1.96		complex of a bcl-w bh3 mutant with a bh3 domain
22	1w9e	T198.B		1.95		pdz tandem of syntenin
23	4px9	T201.B		1.93		dead-box rna helicase ddx3x
24	4acq	T346.A		1.92		alpha-2 macroglobulin
25	2xsz	T248.E		1.91		dodecameric human ruvbl1:ruvbl2
26	4urj	T156.D		1.9		crystal structure of human bj-tsa-9
27	2F7S	S83.A		1.9	0.0	human rab27b bound to gdp
28	1hy7	T230.A		1.89		inhibitor in complex with mmp3
29	2xb2	T276.A		1.88		core mago-y14-eif4aiii-barentsz- upf3b
30	2y0n	T173.B		1.88		dosage compensation factors msl1 and msl3
31	1hap	T74.H		1.86		alpha-thrombin
32	4ft2	T340.B		1.86		mays zmet2 in complex h3(1-15)k9me2 peptide
33	2a55	T19.A		1.85		n-terminal ccp modules of c4b- binding protein
34	2yd6	T140.A		1.85		receptor protein tyrosine phosphatase delta
35	1yrv	T142.A		1.84		novel ubiquitin-conjugating enzyme
36	4rdu	T142.D		1.84		homeobox protein 5 (dlx5) from homo sapiens
37	2da3	T23.A		1.83		homeobox of atbf1
38	4cca	T572.A		1.83		structure of human munc18-2
39	3ex7	T276.C		1.82		ejc in its transition state
40	3n7q	T133.A		1.82		mitochondrial mterf fragment
41	4j1v	T76.A		1.82		mobkl1b
42	4uml	T139.A		1.82		gdap2 macro domain
43	2F7S	T40.A		1.82		rab27b bound to gdp
44	2FE4	SS77.A	S77.A	1.82	0.0	neuronal rab6b in its inactive gdp-bound form
45	1upk	T112.A		1.81		mo25 in complex with c-terminal peptide of strad
46	1khm	T15.A		1.8		c-terminal kh domain of hnrnp k (kh3)
47	1wlj	T54.A		1.8		human isg20
48	4jj7	T390.A		1.8		caspase-3 amino acid-based peptides
49	1mgx	T38.A		1.78		coagulation factor, mg(ii)
50	1r6j	T198.A		1.77		syntenin pdz2
51	1rgo	T152.A		1.77		tandem zinc finger domain of tis11d
52	4u1r	T103.A		1.77		platelet phosphofructokinase
53	1qtn	T390.B		1.76		complex of caspase-8 with tetrapeptide inhibitor
54	1tr2	T8.B		1.76		full-length vinculin (residues 1- 1066)
55	1xf7	T24.A		1.76		wilms' tumor suppressor protein (wt1) finger 3
56	2dal	T24.A		1.76		human fas associated factor 1 protein
57	5dot	T1214.B		1.75		human carbamoyl phosphate synthetase i (cps1)
58	4zg7	T608.A		1.74		human autotaxin
59	4iyp	T78.C		1.73		structure of the npp2ac-alpha4 complex
60	1hao	T74.H		1.71		lpha-thrombin
61	3pcv	T145.A		1.71		leukotriene c4 synthase

62	4ifs	T385.A		1.7		hssrp1 middle domain
63	4r2q	T404.A		1.7		wilms tumor protein (wt1) zinc fingers
64	1r74	T7.A		1.69		human glycine n-methyltransferase
65	1zzi	T15.A		1.69		the third kh domain of hnrnp k
66	3fk2	T1416.D		1.69		glucocorticoid receptor dna-binding factor 1
67	3qu6	T32.B		1.69		irf-3 dbd free form
68	4gv1	T448.A		1.69		pkb alpha in complex with azd5363
69	4nzq	T389.A		1.69		prothrombin deletion mutant residues 146-167
70	3r8q	T151.A		1.68		structure of fibronectin domain 12-14
71	3trt	T317.A		1.68		stabilised vimentin coil2 fragment
72	4i7y	T74.H		1.68		human alpha thrombin
73	3dd2	T74.H		1.67		rna aptamer bound to human thrombin
74	3g07	T630.F		1.67		human bicoid-interacting protein 3
75	3lru	T1970.B		1.67		hprp8 non-native subdomain
76	4bq6	T942.A		1.67		rgmb-neo1 complex form 1
77	4lt6	T411.B		1.67		human poly(a) polymerase gamma
78	4uz1	T157.A		1.67		wnt deacylase notum - crystal form iii - 1.4a
79	1qub	T168.A		1.66		human beta2- glycoprotein i
80	3l4g	T183.M		1.66		cytoplasmic phenylalanyl-trna synthetase
81	4eut	T333.A		1.66		human tbk1 kinase- uld domain
82	4oo6	T28.A		1.66		kap-beta2 bound to the nls of hcc1
83	1j5k	T15.A		1.65		hnrnp k
84	1zzk	T15.A		1.65		hnrnp k at 0.95a resolution
86	3BC1	S83.A	S83.A	1.61	0.0	complex rab27a-slp2a
87	1S8F	S1071.A	S1071.A	1.51	0.0	rab9 complexed to gdp
88	4DKX	S77.A	S77.A	1.3	0.0	rab 6a'(q72l)
89	4DKX	T27.A	S77.A	1.26	22.4	rab 6a'(q72l)
90	3BC1	T40.A	S83.A	1.24	17.2	rab27a-slp2a
91	2A5J	S75.A	S75.A	1.19	0.0	human rab2b
93	2BMD	T76.A	S74.A	1.16	6.8	gdp-bound human rab4a
94	2FE4	T54.A	S77.A	1.15	32.7	neuronal rab6b in its inactive gdp-bound form
95	2A5J	T77.A	S75.A	1.14	7.2	human rab2b
97	1Z0A	T72.A	S70.A	1.09	7.6	gdp-bound rab2a gtpase
98	1S8F	T1039.A	S1071.A	1.05	14.5	rab9 complexed to gdp
99	1Z0D	T53.A	S85.A	1.04	19.9	gdp-bound rab5c gtpase
101	2HEI	S84.A	S84.A	0.94	0.0	human rab5b in complex with gdp
102	1TU4	T52.A	S84.A	0.87	16.9	structure of rab5-gdp complex
104	2HEI	T166.A	S84.A	0.74	35.8	human rab5b in complex with gdp

Several remarkable points can be observed regarding our method and the adaption to this task. First of all, only very few proteins even score close to the values achieved for the Rab proteins from which the statistics were generated on (green). As expected, among the high scoring proteins several Rab related structures can be found, but also other proteins. For all of the structures in the Rab_T– set except for one, we can identify the

centre of *switch2* and the corresponding surface patch. The one faulty result is another threonine close to *switch2.* This might be caused by the fact that it is not a single protein structure but a complex with a GEF domain. Nevertheless we still have identified the correct site. For the Rab proteins with a serine at *switch2,* we mostly identify threonines close to the *switch2* region. If none is available there, a more or less random threonine on the surface which is accessible will be the highest scoring one, but with a very low score (red). If we do not restrict the method to a threonine but instead allow to use a serine as the centre of the patch, we will in many cases actually identify the serine in the centre of *switch2*, even thorough the profile was generated on the Rab_T– set. Nevertheless the scores of the Rab_S– set stay low compared to the Rab_T – set. This is a good indication that the method actually scales with the similarity of the surface area. The *switch2* region of the Rab_S– set might be the most similar one to the site in the Rab_T – set, but still different enough to result in a low score. In addition, since the *switch2* sequence is very well conserved, the low scores of the Rab_S – set indicates, even though the method incorporates some sequence information, the main information is contributed by surface structure similarity. This can also be shown by looking at the structures directly, comparing their biochemical features. Since the interaction of kinases is often related to hydrophobicity and charge, we will analyse the proteins in case of their hydrophobicity (based on Kyte and Doolittle scale, see appendix, Table 14 (*106*)) and their charge. Demonstrated in Figure 45, the central threonine in Rab8 is highly exposed, emerging from the proteins surface. The area seems to be only slightly charged, with a rather positive charged area on the one side and a more negative area on the other side of the central threonine, resulting in an area of low charge. In case of hydrophobicity, we can observe a small but very hydrophilic charged patch on one side of the central threonine, while most other areas are hydrophobic. For Rab2A, which contains a serine in the centre, we observe a very different surface configuration. First of all the serine is much less accessible, located in a sink like part of the protein. The whole area is negatively charged in general. Only the hydrophobicity arrangement resembles slightly the one of Rab8.

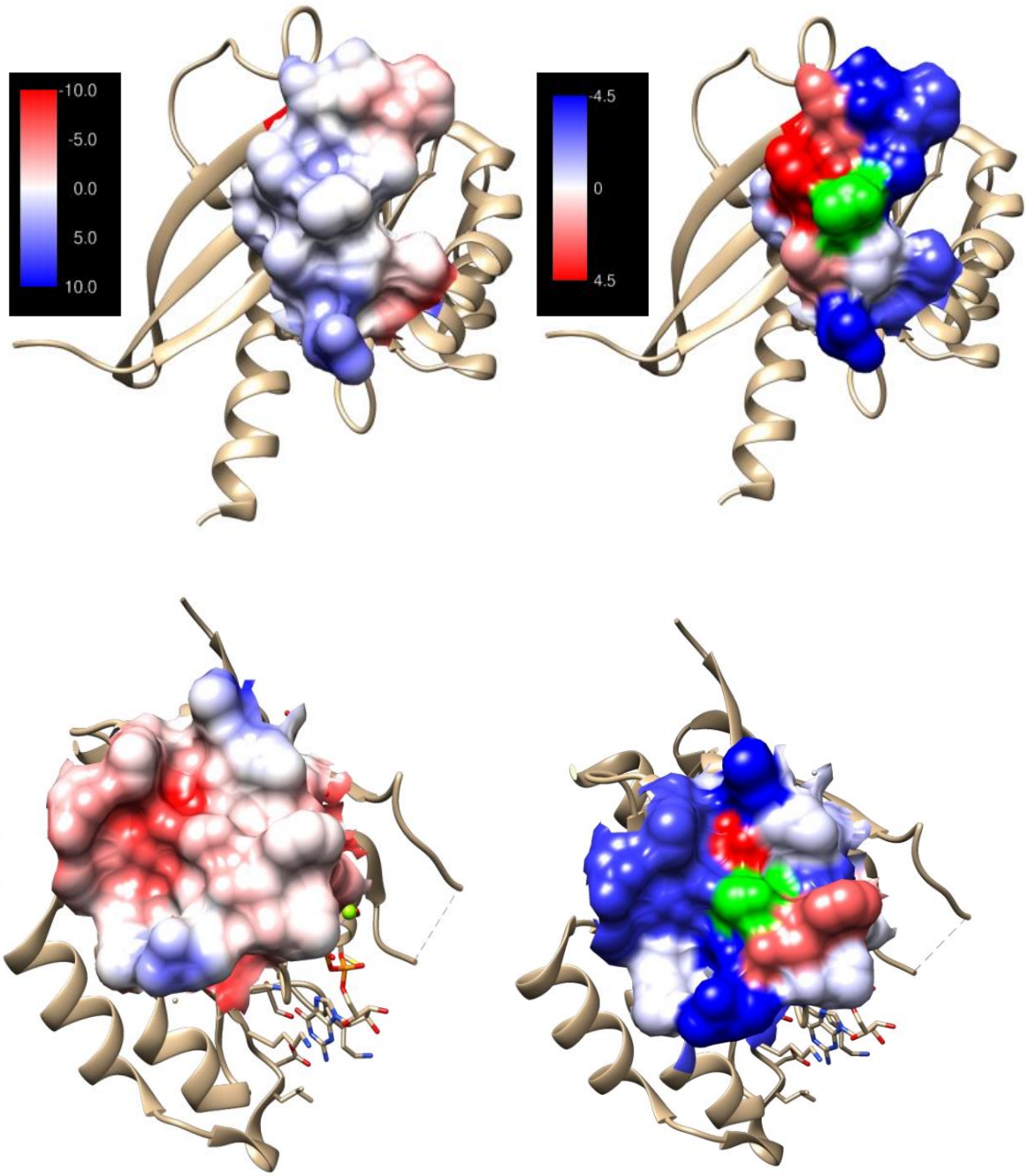


Figure 45: Upper left: Charge distribution (based on Coulomb's law) of the surface patch around the central threonine in switch2 of Rab8. Upper right: Hydrophobicity of the surface patch around the central threonine (in green) in switch2 of Rab8. Lower left: Charge distribution of the surface patch around the central serine in switch2 of Rab2A. Lower right: Hydrophobicity of the surface patch around the central threonine (in green) in switch2 of Rab2A. Images were created using the software Chimera (107).

The same analysis can and should be done, if a protein is picked from the results given in Table 9. A good candidate might be α-thrombin (PDB-ID: 4I7Y, Figure 46), appearing several times in the medium range of the upper field, always with the same central residue although it is in fact represented by slightly different structures from different publications.

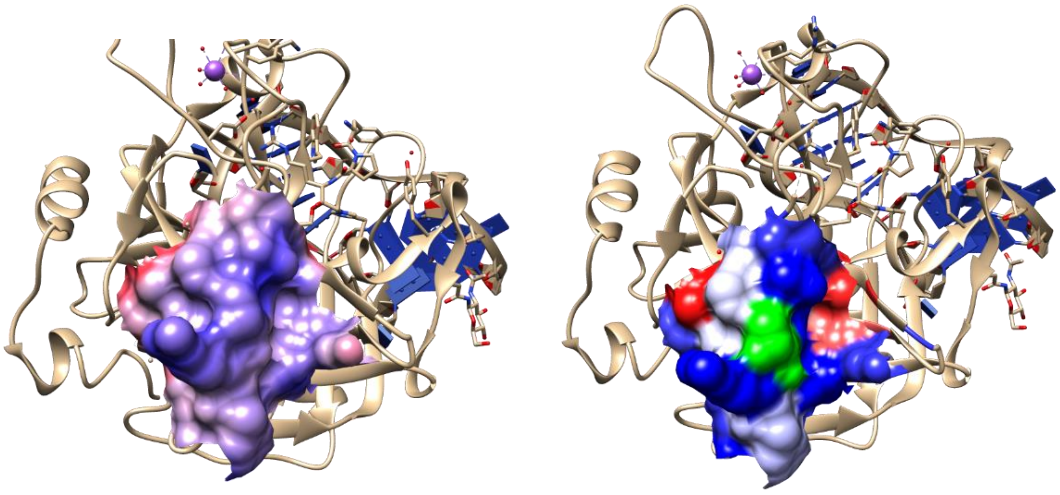


Figure 46: Left: Charge distribution (based on Coulomb's law) of the surface patch around a threonine in α-thrombin. Right: Hydrophobicity of the surface patch around the central threonine (in green) in α-thrombin. Colour scales are the same as in Figure 45.

Again we can observe a highly exposed threonine (green). The area is more positively than negatively charged and evenly hydrophilic and hydrophobic. A similar arrangement could be found also in other proteins picked from the list. In addition to the surface similarity, also the underlying protein structure shows similarities. The main residues are in general located on a loop or helical structure, while residues located at a β-sheet underneath contribute to the distal areas of the patch.

Although these results are very promising in terms of method performance and reliability, they might not give relevant results towards the actual problem of finding possible kinase substrates. Kinases often have several mechanisms to achieve specificity. Among these are also very distal binding sites (*21*). In addition we also used the approximation of a round surface patch of a fixed size and assumed that the threonine is located in the centre of the interface, which might also not be the case. An analysis on known kinase interfaces similar to the one performed on binding motifs in this work might lead to much better insight. For now, since the exact interface is unknown, the predictive value of this results is rather low. But the structural analysis and investigation of literature gave some other insight into the possible underlying mechanisms. The fact that the *switch2* region is rather disordered in the GDP-bound state and stabilizes upon binding of a factor or GDP-GTP exchange, supports the concept of an induced fit for the phosphorylation reaction. The flexibility therefor could be a requirement for the reaction to occur. In addition the fact that some phosphorylation occurs for Rab proteins with serine in *switch2 in vitro* but not *in vivo*, supports a competition between the two substrates in the organism. This situation is quite common among kinases (*21*). Given this, a distal binding site in the Rab proteins

leading to a preference towards one type of Rab proteins as substrate seems to be quite likely. This idea is also in agreement with the fact that linear peptides are not phosphorylated, lacking any distal sites. Discovering this binding site could be one major step in substrate identification and could be achieved by the shown approach. In the next step, many differently sized or even differently shaped surface patches could be used, preforming the same calculations many times. Proteins which appear in the results more often or disappear for a certain shaped patch might be more suitable for further investigation. Although this will not be part of this thesis, the application towards this problem shows that the method itself is reliable. Even more, this approach could be one major field for such applications to be used in.

4 Conclusions

This work can be used as a basis for further development in many different directions. One of the most promising applications might be to use general surface features instead of limiting results by a structural arrangement to compare surfaces. Since the number of spatial arrangements of amino acids, which could achieve a characteristic surface area is virtually endless, this might lead at some point to another level of conservation within proteins (sequence → structure → surface).

Before this can be achieved, a more detailed understanding of the proteins surface is necessary. This is a very difficult task, which could also be supported in further work with the same approach. By comparing different definitions and cut-offs while using our approach, it might be possible to identify a more reliable value, which a surface residue has to pass to be not only accessible but able to participate in a functional manner. This could also help to identify, which residues are actually crucial for binding and which are not. Mutation experiments determining crucial residues as well as energy calculations are very time consuming. By analysing the performance of a prediction method in dependency of different cut-off values, a more detailed picture of interface residue relevance could be drawn.

For the prediction method itself it was shown that, even though we are in an early stage of optimization, we could demonstrate reliability and user-orientated value. Nevertheless we expect that much improvement can be achieved by further assessment and

optimization. At the moment, the weighting parameters underlying the different parts of the scoring method are chosen manually. By optimizing those using for example artificial neural networks, much better and more consistent results could be expected. Further assessment is necessary in order to understand the performance dependencies. Therefore, reliable benchmark datasets need to be created and individual cases of good and bad performing benchmark structures need to be evaluated. Also a non-static patch assessment method could improve the results quite strongly, since a round shaped patch of fixed size is a rough approximation.

Ultimately it would be constructive to integrate this approach into a user friendly interface. It might even be useful to include basic tools for statistical surface analysis, which is missing in many structural biology applications.

By exploiting the methods high flexibility many different final applications are imaginable such as protein classification and function discovery on several levels, protein surface analysis, comparison of user-defined surface areas or the discovery of yet unknown structural motifs.

5 Literature

1. H. M. Berman *et al.*, The Protein Data Bank. *Nucleic Acids Res.* **28**, 235 (January 1, 2000, 2000).
2. R. Kolodny, P. Koehl, M. Levitt, Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.* **346**, 1173 (Mar 4, 2005).
3. M. F. Perutz *et al.*, Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature* **185**, 416 (Feb 13, 1960).
4. P. Bork, E. V. Koonin, Protein sequence motifs. *Curr. Opin. Struct. Biol.* **6**, 366 (6//, 1996).
5. P. B. Moore, Structural Motifs in RNA. *Annu. Rev. Biochem.* **68**, 287 (1999/06/01, 1999).
6. C. A. Orengo, T. P. Flores, D. T. Jones, W. R. Taylor, J. M. Thornton, Recurring structural motifs in proteins with different functions. *Curr. Biol.* **3**, 131 (Mar, 1993).
7. F. Kaiser, A. Eisold, D. Labudde, A Novel Algorithm for Enhanced Structural Motif Matching in Proteins. *J. Comput. Biol.* **22**, 698 (Jul, 2015).
8. M. N. Nguyen, M. S. Madhusudhan, Biological insights from topology independent comparison of protein 3D structures. *Nucleic Acids Res.* **39**, e94 (Aug, 2011).
9. A. Ingles-Prieto *et al.*, Conservation of Protein Structure over Four Billion Years. *Structure (London, England : 1993)* **21**, 1690.
10. W. R. Pearson, Effective protein sequence comparison. *Methods Enzymol.* **266**, 227 (1996).
11. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389 (1997).
12. K. Karplus, C. Barrett, R. Hughey, Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846 (January 1, 1998, 1998).
13. C. J. Sigrist *et al.*, New and continuing developments at PROSITE. *Nucleic Acids Res.* **41**, D344 (Jan, 2013).
14. A. M. Lesk, C. Chothia, How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins. *J. Mol. Biol.* **136**, 225 (1980/01/25, 1980).
15. O. B. Ptitsyn, A. V. Finkelstein, Similarities of protein topologies: evolutionary divergence, functional convergence or principles of folding? *Q. Rev. Biophys.* **13**, 339 (1980).
16. L. Xie, L. Xie, P. E. Bourne, A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics* **25**, i305 (05/27, 2009).

17. D. D. Paul, C. Yu-Dong, J. S. Benjamin, J. D. Andrew, Prediction of Protein Function in the Absence of Significant Sequence Similarity. *Curr. Med. Chem.* **11**, 2135 (2004).
18. A. E. Gorbalenya, V. M. Blinov, A. P. Donchenko, Poliovirus-encoded proteinase 3C: a possible evolutionary link between cellular serine and cysteine proteinase families. *FEBS Lett.* **194**, 253.
19. J. F. Bazan, R. J. Fletterick, Viral cysteine proteases are homologous to the trypsin-like family of serine proteases: structural and functional implications. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 7872 (1988).
20. L. Hedstrom, Serine protease mechanism and specificity. *Chem Rev* **102**, 4501 (Dec, 2002).
21. J. A. Ubersax, J. E. Ferrell, Jr., Mechanisms of specificity in protein phosphorylation. *Nat. Rev. Mol. Cell Biol.* **8**, 530 (Jul, 2007).
22. J. Miller, A. D. McLachlan, A. Klug, Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J.* **4**, 1609 (Jun, 1985).
23. R. B. Darnell, Developing global insight into RNA regulation. *Cold Spring Harb Symp Quant Biol* **71**, 321 (2006).
24. G. K. W. Kong *et al.*, Copper binding to the Alzheimer's disease amyloid precursor protein. *European Biophysics Journal* **37**, 269 (11/21

08/30/received

10/24/revised

10/26/accepted, 2008).
25. I. S. Mian, A. R. Bradwell, A. J. Olson, Structure, function and properties of antibody binding sites. *J. Mol. Biol.* **217**, 133 (Jan 5, 1991).
26. A. Koutsotoli, A. G. Tzakos, Host-pathogen crosstalking: the mastery of taking the helm of the host. *Structure (London, England : 1993)* **20**, 1613 (Oct 10, 2012).
27. M. Moll, L. E. Kavraki, Matching of structural motifs using hashing on residue labels and geometric filtering for protein function prediction. *Computational systems bioinformatics / Life Sciences Society. Computational Systems Bioinformatics Conference* **7**, 157 (2008).
28. J. Desaphy, G. Bret, D. Rognan, E. Kellenberger, sc-PDB: a 3D-database of ligandable binding sites—10 years on. *Nucleic Acids Res.*, (October 9, 2014, 2014).
29. B. Contreras-Moreira, 3D-footprint: a database for the structural analysis of protein–DNA complexes. *Nucleic Acids Res.* **38**, D91 (09/18

08/14/received

09/02/revised

09/03/accepted, 2010).
30. J. Yang, A. Roy, Y. Zhang, BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res.* **41**, D1096 (January 1, 2013, 2013).

31. H. Deng, G. Chen, W. Yang, J. J. Yang, Predicting calcium-binding sites in proteins - a graph theory and geometry approach. *Proteins* **64**, 34 (Jul 1, 2006).
32. A. Amos-Binks *et al.*, Binding Site Prediction for Protein-Protein Interactions and Novel Motif Discovery using Re-occurring Polypeptide Sequences. *BMC Bioinformatics* **12**, 1 (2011).
33. S. Jones, J. M. Thornton, Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol.* **272**, 133 (Sep 12, 1997).
34. T. Schmidt, J. Haas, T. Gallo Cassarino, T. Schwede, Assessment of ligand-binding residue predictions in CASP9. *Proteins* **79 Suppl 10**, 126 (2011).
35. A. Schlessinger, Y. Ofran, G. Yachdav, B. Rost, Epitome: database of structure-inferred antigenic epitopes. *Nucleic Acids Res.* **34**, D777 (Jan 1, 2006).
36. U. Kulkarni-Kale, S. Raskar-Renuse, G. Natekar-Kalantre, S. A. Saxena, Antigen-Antibody Interaction Database (AgAbDb): a compendium of antigen-antibody interactions. *Methods in molecular biology (Clifton, N.J.)* **1184**, 149 (2014).
37. B. A. Lewis *et al.*, PRIDB: a protein–RNA interface database. *Nucleic Acids Res.* **39**, D277 (January 1, 2011, 2011).
38. T. Siggers, R. Gordân, Protein–DNA binding: complexities and multi-protein codes. *Nucleic Acids Res.*, (November 16, 2013, 2013).
39. G. D. Stormo, D. S. Fields, Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.* **23**, 109 (Mar, 1998).
40. W. F. Anderson, D. H. Ohlendorf, Y. Takeda, B. W. Matthews, Structure of the cro repressor from bacteriophage [lambda] and its interaction with DNA. *Nature* **290**, 754 (04/30/print, 1981).
41. B. W. Matthews, D. H. Ohlendorf, W. F. Anderson, Y. Takeda, Structure of the DNA-binding region of lac repressor inferred from its homology with cro repressor. *Proceedings of the National Academy of Sciences* **79**, 1428 (March 1, 1982, 1982).
42. R. Wintjens, M. Rooman, Structural Classification of HTH DNA-binding Domains and Protein – DNA Interaction Modes. *J. Mol. Biol.* **262**, 294 (9/20/, 1996).
43. T. R. Burglin, Homeodomain subtypes and functional diversity. *Sub-cellular biochemistry* **52**, 95 (2011).
44. F. D. Urnov, E. J. Rebar, M. C. Holmes, H. S. Zhang, P. D. Gregory, Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.* **11**, 636 (09//print, 2010).
45. C. Murre *et al.*, Structure and function of helix-loop-helix proteins. *Biochim. Biophys. Acta* **1218**, 129 (Jun 21, 1994).
46. A. J. Doherty, L. C. Serpell, C. P. Ponting, The helix-hairpin-helix DNA-binding motif: a structural basis for non-sequence-specific recognition of DNA. *Nucleic Acids Res.* **24**, 2488 (1996/07//, 1996).
47. M. M. Thayer, H. Ahern, D. Xing, R. P. Cunningham, J. A. Tainer, Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure. *The EMBO journal* **14**, 4108 (1995/08//, 1995).

48. J. Miller, A. D. McLachlan, A. Klug, Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *The EMBO Journal* **4**, 1609 (1985).
49. R. A. Katz, J. E. Jentoft, What is the role of the cys-his motif in retroviral nucleocapsid (NC) proteins? *BioEssays* **11**, 176 (Dec, 1989).
50. N. V. Grishin, Treble clef finger—a functionally diverse zinc-binding structural motif. *Nucleic Acids Res.* **29**, 1703 (April 15, 2001, 2001).
51. S. Buratowski, H. Zhou, Functional domains of transcription factor TFIIB. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 5633 (Jun 15, 1993).
52. T. Pan, J. E. Coleman, GAL4 transcription factor is not a "zinc finger" but forms a Zn(II)2Cys6 binuclear cluster. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 2077 (Mar, 1990).
53. A. Klug, D. Rhodes, Zinc Fingers: A Novel Protein Fold for Nucleic Acid Recognition. *Cold Spring Harbor Symp. Quant. Biol.* **52**, 473 (January 1, 1987, 1987).
54. C. O. Pabo, E. Peisach, R. A. Grant, Design and Selection of Novel Cys2His2 Zinc Finger Proteins. *Annu. Rev. Biochem.* **70**, 313 (2001).
55. S. A. Wolfe, L. N. and, C. O. Pabo, DNA Recognition by Cys2His2 Zinc Finger Proteins. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 183 (2000).
56. J. M. Matthews, M. Sunde, Zinc Fingers--Folds for Many Occasions. *IUBMB Life* **54**, 351 (2002).
57. R. Gamsjaeger, C. K. Liew, F. E. Loughlin, M. Crossley, J. P. Mackay, Sticky fingers: zinc-fingers as protein-recognition motifs. *Trends Biochem. Sci.* **32**, 63.
58. W. H. Landschulz, P. F. Johnson, S. L. McKnight, The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science* **240**, 1759 (Jun 24, 1988).
59. M. D. Allen, K. Yamasaki, M. Ohme-Takagi, M. Tateno, M. Suzuki, A novel mode of DNA recognition by a beta-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA. *The EMBO Journal* **17**, 5484 (1998).
60. B. M. Lunde, C. Moore, G. Varani, RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8**, 479 (Jun, 2007).
61. R. Stefl, L. Skrisovska, F. H. T. Allain, RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. *EMBO Rep.* **6**, 33 (09/22/received

11/26/accepted, 2005).
62. G. Waksman *et al.*, Crystal structure of the phosphotyrosine recognition domain SH2 of v-src complexed with tyrosine-phosphorylated peptides. *Nature* **358**, 646 (Aug 20, 1992).
63. E. Petsalaki, A. Stark, E. García-Urdiales, R. B. Russell, Accurate Prediction of Peptide Binding Sites on Protein Surfaces. *PLoS Comput. Biol.* **5**, e1000335 (2009).
64. M. Gao, J. Skolnick, A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. *PLoS Comput. Biol.* **9**, e1003302 (2013).
65. A. J. Thomson, H. B. Gray, Bio-inorganic chemistry. *Curr. Opin. Chem. Biol.* **2**, 155 (4//, 1998).

66. M. M. Yamashita, L. Wesson, G. Eisenman, D. Eisenberg, Where metal ions bind in proteins. *Proc. Natl. Acad. Sci. U. S. A.* **87**, 5648 (Aug, 1990).
67. K. J. Waldron, N. J. Robinson, How do bacterial cells ensure that metalloproteins get the correct metal? *Nat Rev Micro* **7**, 25 (01//print, 2009).
68. T. Dudev, C. Lim, Metal binding affinity and selectivity in metalloproteins: insights from computational studies. *Annual review of biophysics* **37**, 97 (2008).
69. P. D. Dobson, A. J. Doig, Distinguishing enzyme structures from non-enzymes without alignments. *J. Mol. Biol.* **330**, 771 (Jul 18, 2003).
70. N. Furnham *et al.*, The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* **42**, D485 (Jan, 2014).
71. A. Shrake, J. A. Rupley, Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* **79**, 351 (Sep 15, 1973).
72. B. Rost, C. Sander, Conservation and prediction of solvent accessibility in protein families. *Proteins* **20**, 216 (Nov, 1994).
73. M. Z. Tien, A. G. Meyer, D. K. Sydykova, S. J. Spielman, C. O. Wilke, Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLoS One* **8**, e80635 (2013).
74. H. Chen, X. Hu, I. Yoo, H.-X. Zhou, in *Second Asia-Pacific Bioinformatics Conference (APBC2004)* Y.-P. P. Chen, Ed. (ACS, Dunedin, New Zealand, 2004), vol. 29, pp. 333-338.
75. S. Miller, J. Janin, A. M. Lesk, C. Chothia, Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641 (8/5/, 1987).
76. G. Trinquier, Y. H. Sanejouand, Which effective property of amino acids is best preserved by the genetic code? *Protein Eng.* **11**, 153 (Mar, 1998).
77. M. H. Smith, The amino acid composition of proteins. *J. Theor. Biol.* **13**, 261 (12//, 1966).
78. O. Lichtarge, M. E. Sowa, Evolutionary predictions of binding surfaces and interactions. *Curr Opin Struc Biol* **12**, (2002).
79. B. K. Dukka, E. Tomita, J. Suzuki, K. Horimoto, T. Akutsu, Protein threading with profiles and distance constraints using clique based algorithms. *Journal of bioinformatics and computational biology* **4**, 19 (Feb, 2006).
80. M. P. Jacobson, R. A. Friesner, Z. Xiang, B. Honig, On the role of the crystal environment in determining protein side-chain conformations. *J. Mol. Biol.* **320**, 597 (Jul 12, 2002).
81. F. Gaudreault, M. Chartier, R. Najmanovich, Side-chain rotamer changes upon ligand binding: common, crucial, correlate with entropy and rearrange hydrogen bonding. *Bioinformatics* **28**, i423 (Sep 15, 2012).
82. A. M. Weljie, A. P. Yamniuk, H. Yoshino, Y. Izumi, H. J. Vogel, Protein conformational changes studied by diffusion NMR spectroscopy: application to helix-loop-helix calcium binding proteins. *Protein Sci.* **12**, 228 (Feb, 2003).
83. F. A. A. Mulder, Leucine Side-Chain Conformation and Dynamics in Proteins from 13C NMR Chemical Shifts. *ChemBioChem* **10**, 1477 (2009).
84. L. Holm, P. Rosenström, Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38**, W545 (July 1, 2010, 2010).

85. P. D. Dobson, Y. D. Cai, B. J. Stapley, A. J. Doig, Prediction of protein function in the absence of significant sequence similarity. *Curr. Med. Chem.* **11**, 2135 (Aug, 2004).
86. T. A. Binkowski, P. Freeman, J. Liang, pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. *Nucleic Acids Res.* **32**, W555 (Jul 1, 2004).
87. D. M. Kristensen *et al.*, Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics* **9**, 17 (01/11

06/11/received

01/11/accepted, 2008).
88. N. Nadzirin, E. J. Gardiner, P. Willett, P. J. Artymiuk, M. Firdaus-Raih, SPRITE and ASSAM: web servers for side chain 3D-motif searching in protein structures. *Nucleic Acids Res.* **40**, W380 (Jul, 2012).
89. R. R. Walia *et al.*, Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. *BMC Bioinformatics* **13**, 1 (2012).
90. Y. Guo, L. Yu, Z. Wen, M. Li, Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res.* **36**, (2008).
91. L. Wang, S. Brown, in *Proc of the 26th IEEE EMBS Ann Int Conf*. (2006).
92. I. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. (Morgan Kaufmann, San Francisco, 2005).
93. O. Kim, K. Yura, N. Go, Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.* **34**, (2006).
94. G. Robin *et al.*, Restricted diversity of antigen binding residues of antibodies revealed by computational alanine scanning of 227 antibody-antigen complexes. *J. Mol. Biol.* **426**, 3729 (Nov 11, 2014).
95. T. Ramaraj, T. Angel, E. A. Dratz, A. J. Jesaitis, B. Mumey, Antigen-antibody interface properties: composition, residue interactions, and features of 53 non-redundant structures. *Biochim. Biophys. Acta* **1824**, 520 (Mar, 2012).
96. T. Madej *et al.*, MMDB and VAST+: tracking structural similarities between macromolecular complexes. *Nucleic Acids Res.* **42**, D297 (Jan, 2014).
97. T. Hamelryck, B. Manderick, PDB file parser and structure class implemented in Python. *Bioinformatics* **19**, 2308 (November 22, 2003, 2003).
98. P. J. A. Cock *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422 (June 1, 2009, 2009).
99. W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577 (Dec, 1983).
100. R Development Core Team. (2010).
101. scikit learn developers. (2015).

102. G. B. McGaughey, M. Gagné, A. K. Rappé, π-Stacking Interactions: ALIVE AND WELL IN PROTEINS. *J. Biol. Chem.* **273**, 15458 (June 19, 1998, 1998).
103. K. Koch, F. Zollner, S. Neumann, F. Kummert, G. Sagerer, Comparing bound and unbound protein structures using energy calculation and rotamer statistics. *In silico biology* **2**, 351 (2002).
104. K. Wang *et al.*, An Accurate Method for Prediction of Protein-Ligand Binding Site on Protein Surface Using SVM and Statistical Depth Function. *BioMed Research International* **2013**, 7 (2013).
105. O. T. Kim, K. Yura, N. Go, Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.* **34**, 6450 (2006).
106. J. Kyte, R. F. Doolittle, A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105 (May 5, 1982).
107. E. F. Pettersen *et al.*, UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605 (Oct, 2004).
108. S. C. Lovell, J. M. Word, J. S. Richardson, D. C. Richardson, The penultimate rotamer library. *Proteins* **40**, 389 (Aug 15, 2000).

6 Appendix

Table 10: structures in the non-redundant antigen-antibody dataset

2NY7	**2IGF**	**1FNS**	**1EO8**	**1DEE**	**1HI6**	**2JEL**	**3LEY**	**1FJ1**
1KTR	**1NAK**	**1OAZ**	**2BRR**	**3MNZ**	**1C08**	**1YQV**	**1LK3**	**1I9R**
1JPS	**3G5Y**	**3L5W**	**1IQD**	**3CXD**	**1W72**	**3NGB**	**1KCS**	**1OB1**
2WUC	**2ZPK**	**3NH7**	**1BVK**	**1SM3**	**1PZ5**	**2QHR**	**1KCR**	**3I50**
2HRP	**1OTS**	**2H1P**	**1KC5**	**1QFU**	**1NMC**	**1FRG**	**1FBI**	**3LIZ**
3FFD	**1MPA**	**3HR5**	**1H0D**	**1FE8**	**2BDN**	**1JRH**	**1ZEA**	**3MXW**
3KR3	**3NFP**	**1ORS**	**1MVF**	**1QFW**	**2OTW**	**1AI1**	**3GO1**	
1FPT	**1KXQ**	**1OSP**	**2B1H**	**1JTO**	**3D9A**	**1IKF**	**3GI8**	
1NSN	**1A14**	**2VIR**	**3LEX**	**3IDG**	**3BKY**	**2XRA**	**1TJI**	
3DVG	**1F90**	**1ZTX**	**1UWX**	**1TZH**	**1WEJ**	**3MLY**	**1A2Y**	
2DD8	**2CK0**	**1DQJ**	**1MVU**	**3MLR**	**1NCA**	**1E6J**	**1EJO**	
1KEN	**2W9E**	**2B2X**	**1P2C**	**1YY9**	**1S78**	**1N8Z**	**2AEP**	
1E4W	**1RJL**	**2XTJ**	**3IFO**	**1AHW**	**1CU4**	**2VXT**	**3O41**	
2J4W	**2UZI**	**1UJ3**	**1QKZ**	**3AB0**	**1KXT**	**1A3R**	**3IET**	
1RVF	**3GHE**	**1V7M**	**3IFL**	**2XQY**	**1F58**	**1EZV**	**1BJ1**	
1E4X	**1GGI**	**1TQB**	**2OSL**	**1I8K**	**1NL0**	**1DZB**	**1NBY**	
3BGF	**3IDX**	**1G6V**	**3O0R**	**1TZG**	**2A6I**	**1KXV**	**3L95**	
1JHL	**1EGJ**	**3LD8**	**3GBN**	**3IFP**	**3LZF**	**2VXS**	**3C2A**	
3KJ4	**3L5X**	**1IFH**	**3G6D**	**2XQB**	**1P4B**	**1CE1**	**3IXT**	
3N85	**3MLW**	**3IU3**	**1MLC**	**1FDL**	**1NDG**	**1BQL**	**1N0X**	

Table 11: bound and unbound NMR structures with nucleic acid binding sites

Bound structure	**Unbound homologs**					
2lbsB	2luqA	1t4nA				
1f4sP	3alcA					
1oslB	1lqcA					
1t4lB	2luqA	1t4nA				
2exfA	1eskA	2l44A				
2lebA	2leaA	2kn4A				
2kejA	1lqcA					
2jzwA	1eskA	2l44A				
2eseA	2es6A	2fe9A				
2jp9A	1xf7A					
1cjgB	1lqcA					
1rcsA	1wrtR	1wrtS	1wrsR	1wrsS	2xdiA	2xdiB
1rcsB	1wrtR	1wrtS	1wrsR	1wrsS	2xdiA	2xdiB
1cjgA	1lqcA					
1co0B	1wrtR	1wrtS	1wrsR	1wrsS	2xdiA	2xdiB

2k1nA	1z0rA	1z0rB	2ro4A	2ro4B	1ysfA	1ysfB
1co0A	1wrtR	1wrtS	1wrsR	1wrsS	2xdiA	2xdiB
2lupB	2luqA	1t4nA				
1msfC	1idzA	1mbhA	1mbkA			
1l1mB	1lqcA					
2sttA	1r36A					
2kn7A	1z00B	2aq0A	2aq0B	2mutB		
2k1nC	1z0rA	1z0rB	2ro4A	2ro4B	1ysfA	1ysfB
2l3jA	2b7tA	2b7vA				
2kn7D	1z00B	2aq0A	2aq0B	2mutB		
1t2sA	1vynA					
1bj6A	1eskA	2l44A				
1wtbA	1iqtA					
1l1mA	1lqcA					
1dz5A	1fhtA					
1dz5B	1fhtA					
2lecA	2leaA	2kn4A				
2jpaA	1xf7A					
2k1nB	1z0rA	1z0rB	2ro4A	2ro4B	1ysfA	1ysfB
1a1tA	2m3zA	1mfsA	2l44A			
2o9lA	2ahqA					
2kxnB	2rrbA	2cqcA				
2lexA	1wj2A					
2kekB	1lqcA					
1audA	1fhtA					
2kekA	1lqcA					
2li8A	2cqfA					
2l3cA	2b7tA					
2keiA	1lqcA					
2keiB	1lqcA					
1x0fA	1iqtA					
2l5dA	2l5cA					
2l4lA	1eskA	2l44A				
2jx1A	2jydA					
2k1nD	1z0rA	1z0rB	2ro4A	2ro4B	1ysfA	1ysfB
1nk2P	1vndA	1qryA				
2lttB	2ltdA	2ltdB				
2lttA	2ltdA	2ltdB				
2ko0A	2jtgA					
1f6uA	2m3zA	1mfsA	2l44A			
1oslA	1lqcA					
1t2rA	1vynA					
2o8kA	2ahqA					
2kejB	1lqcA					
2rraA	2cqcA					
2l2kB	2b7vA					
2xfmA	2l5cA					

1tn9A	1bb8A					
2l1gA	2jtgA					
1lcdA	1lqcA					

Table 12: definition of χ1 side chain angles based on amino acid type

Side chain	Axis	Atoms used to define angle	Zero value	Formal range
ARG	CA-CB	N-CA-CB-CG	CG cis to N	from -180 to +180 deg
ASN	CA-CB	N-CA-CB-CG	CG cis to N	from -180 to +180 deg
ASP	CA-CB	N-CA-CB-CG	CG cis to N	from -180 to +180 deg
CYS	CA-CB	N-CA-CB-SG	SG cis to N	from -180 to +180 deg
GLN	CA-CB	N-CA-CB-CG	CG cis to N	from -180 to +180 deg
GLU	CA-CB	N-CA-CB-CG	CG cis to N	from -180 to +180 deg
HIS	CA-CB	N-CA-CB-CG	CG cis to N	from -180 to +180 deg
ILE	CA-CB	N-CA-CB-CG1	CG1 cis to N	from -180 to +180 deg
LEU	CA-CB	N-CA-CB-CG	CG cis to N	from -180 to +180 deg
LYS	CA-CB	N-CA-CB-CG	CG cis to N	from -180 to +180 deg
MET	CA-CB	N-CA-CB-CG	CG cis to N	from -180 to +180 deg
PHE	CA-CB	N-CA-CB-CG	CG cis to N	from -180 to +180 deg
PRO	CA-CB	N-CA-CB-CG	CG cis to N	CA-CB is part of ring
SER	CA-CB	N-CA-CB-OG	OG cis to N	from -180 to +180 deg
THR	CA-CB	N-CA-CB-OG1	OG1 cis to N	from -180 to +180 deg
TRP	CA-CB	N-CA-CB-CG	CG cis to N	from -180 to +180 deg
TYR	CA-CB	N-CA-CB-CG	CG cis to N	from -180 to +180 deg
VAL	CA-CB	N-CA-CB-CG1	CG1 cis to N	from -180 to +180 deg

Table 13: preferred rotamers for χ1 as listed in the Penultimate Rotamer Library (108)

Side chain	Preferred rotamers
ARG	62°, -177°, -62, -67°
ASN	62°, -177°, -70°
ASP	62, -174, -177
CYS	62°, -177°, -65°
GLN	62°, 70°, -177°, -65°, -67°
GLU	62°, 70°, -177°, -65°, -67°
HIS	63°, -177°, -65°
ILE	62°, -177°, -65°,-57°
LEU	62°, -177°, -172°, -85°, -65°
LYS	62°, -177°, -90°, -67°, -62°
MET	62°, -177°, -67°, -65°
PHE	62°, -177°, -65°
PRO	30°, -30°
SER	62°, -177°, -65°
THR	62°, -175°, -65°

TRP	62°, -177°, -65°
TYR	62°, -177°, -65°
VAL	63°, 175, -60°

Table 14: Hydrophobicity based on Kyte and Doolittle scale (106)

R	**K**	**N**	**D**	**Q**	**E**	**H**	**P**	**Y**	**W**
-4.5	-3.9	-3.5	-3.5	-3.5	-3.5	-3.2	-1.6	-1.3	-0.9
S	**T**	**G**	**A**	**M**	**C**	**F**	**L**	**V**	**I**
-0.8	-0.7	-0.4	1.8	1.9	2.5	2.8	3.8	4.2	4.5