Charles University in Prague
Faculty of Science

BACHELOR THESIS

Petra Gajdošová

Structural identification of protein-DNA interactions using machine learning

Supervisor of the bachelor thesis:	RNDr. David Hoksza, Ph.D.
Study programme:	Bioinformatika
Specialization:	Bioinformatika

Prague 2020

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In Prague on 13.8.2020 Petra Gajdošová

Název práce: Strukturní identifikace protein-DNA interakcí pomocí strojového učení

Autor: Petra Gajdošová

Vedoucí bakalářské práce: RNDr. David Hoksza, Ph.D.

Abstrakt: Interakcie DNA a proteínov sú dôležitou súčasťou bunky a bunečného cyklu. Aby sme mohli predikovať ich interakcie mali by sme poznať štruktúru DNA a proteínov. Pre predikciu interakcií sme zvolili strojové učenie, ktoré má adekvátne výsledky v oblasti biologickej predikcie. V tejto práci používame a upravujeme P2Rank pre predikciu DNA väzobných miest na povrchu proteínu. P2rank bol pôvodne navrhnutý pre predikciu väzobných miest ligandov. Rovnako sme pripravili popis existujúcich metód pre predikciu DNA väzobných miest. Návrhy nových vlastností pre predikciu väzobných miest je súčasťou popisu P2Rank.

Klíčová slova: bioinformatika; strukturní bioinoformatika; strojové učení

Title: Structural identification of protein-DNA interactions using machine learning

Author: Petra Gajdošová

Supervisor of the bachelor thesis: RNDr. David Hoksza, Ph.D.

Abstract: DNA-protein interactions are essential parts of cell life and cell cycle. Prediction of these interactions requires knowledge of DNA and a protein structure. Because machine learning approaches show adequate results in biological predictions, we chose to use it for the prediction of protein-DNA interactions. In this thesis, we use the machine learning tool P2Rank that was originally designed for prediction of ligand-binding sites and adapt it to predict DNA-binding sites. Apart of that, the thesis serves as a summary of existing prediction tools/methods and includes suggestions for further modifications of P2Rank.

Keywords: bioinformatics; structural bioinformatics; machine learning

Contents

1 Introduction 2

2 DNA, proteins, and their interactions 3
2.1 DNA . . . 3
2.2 Proteins . . . 5
2.3 Protein DNA interactions . . . 6

3 Existing methods 9
3.1 Sequence-based methods . . . 9
3.2 Structure-based methods . . . 10
3.2.1 DISPLAR . . . 11
3.2.2 iDBS . . . 12
3.2.3 DBSI . . . 14

4 P2Rank 17
4.1 Modifications . . . 17
4.1.1 Electrostatic potential . . . 19
4.1.2 Dipole moment . . . 20
4.1.3 P2Rank integration . . . 21
4.2 Evaluation . . . 21
4.2.1 Dataset construction . . . 21
4.2.2 Training and parameter selection . . . 22
4.2.3 Comparison with existing methods . . . 24

5 Conclusion 27

Bibliography 28

List of Figures 34

List of Tables 35

Chapter 1

Introduction

With the increasing amount of known protein structures, interest in their functions grows too. Proteins are involved in cell division, organization, development, and differentiation, but also in genome rearrangement, signal cascades, and gene regulations. The functions of proteins depend on their ability to bind other molecules. The binding interactions range from covalent bonds, through hydrogen bonds, to electrostatic interaction. Studies typically concentrate research on one type of protein interactions and their predictions, such as protein-ligand, protein-RNA, protein-protein, or protein-DNA. In this thesis we focus on DNA and their interactions with proteins.

Predictions of protein-DNA interactions can accelerate research of proteins with unknown functions, or clarify their binding properties and involved binding residues. Available tools have visible results in predicting protein-DNA interactions, but they still have room for improvement. Interactions examined in detail brought new knowledge and mechanisms that can improve the prediction ability of existing or newly created tools [1, 2, 3, 4]. A significant amount of available tools use only a protein sequence as a resource needed for prediction. Other methods use sequence information together with structure information or only structure information itself. All methods use several approaches to distinguish binding and non-binding residues/atoms, such as machine learning, linear scoring, or other statistical functions.

The similarity of binding sites between DNA and ligands lies in physicochemical properties participating in almost all molecular interactions. This lead us to think about adapting the existing machine learning tool P2Rank to predict DNA-binding sites on the protein surface. P2Rank predicts ligand-binding sites on protein surface using Random Forest classification and protein properties described as features.

In the next sections, we describe the key molecular players participating in protein-DNA interactions. Then, we introduce existing template-free prediction methods, which use machine learning algorithms. These methods guide us in choosing features for P2Rank. From all features, electrostatic potential has the most impact on predictions, and therefore we pick it as a new feature for P2Rank together with a dipole moment of proteins.

Chapter 2

DNA, proteins, and their interactions

DNA-binding sites include interactions between two macromolecules essential for all living organisms. Proteins have various forms and functions. DNA stores the information for cells, including protein forms and functions.

In order to predict DNA-binding sites, we need to get familiar with both the macromoleculer players, DNA section 2.1, and proteins section 2.2. Analysis of experimentally examined protein-DNA interactions section 2.3 brought several principles that demonstrate recognition, association, and disassembly of protein DNA interactions in biological conditions.

2.1 DNA

A molecule of DNA carries genetic information encoded in its nucleotides sequence. The exact order of nucleotides determines stored information.

DNA keeps two complementary chains of a nucleotide sequence with the same information, forming double helix DNA. Even a small change in the nucleotide sequence or nucleotide modifications may cause changes in the entire double helix. Furthermore, nucleotides composition itself is necessary knowledge for the identification and prediction of proteins-DNA interactions.

Even a small change in the nucleotide sequence or nucleotide modifications may cause changes in the entire double helix [5].Furthermore, nucleotides composition itself is necessary knowledge for the identification and prediction of proteins-DNA interactions.

Basic units of DNA are 2-deoxyribonucleotides composed of a five-carbon sugar 2-deoxyribose, a phosphate group, and a nitrogen base. A single chain of the nucleotide sequence arises in the process of DNA synthesis. During synthesis, new nucleotide connects with a phosphate group on a carbon C5 to a carbon C3 of already incorporated ribose and thus creates a phosphodiester bond. Phosphates groups and sugars form a backbone of a double helix located outside of the DNA molecule. The inner part of a DNA molecule consists of complementary nitrogen bases connected to a carbon C1 of ribose by N-glycosidic bond [6, 7].

Two complement single chains create a DNA double helix. Double helix can adopt several conformations based on water activity, surrounding ions, and

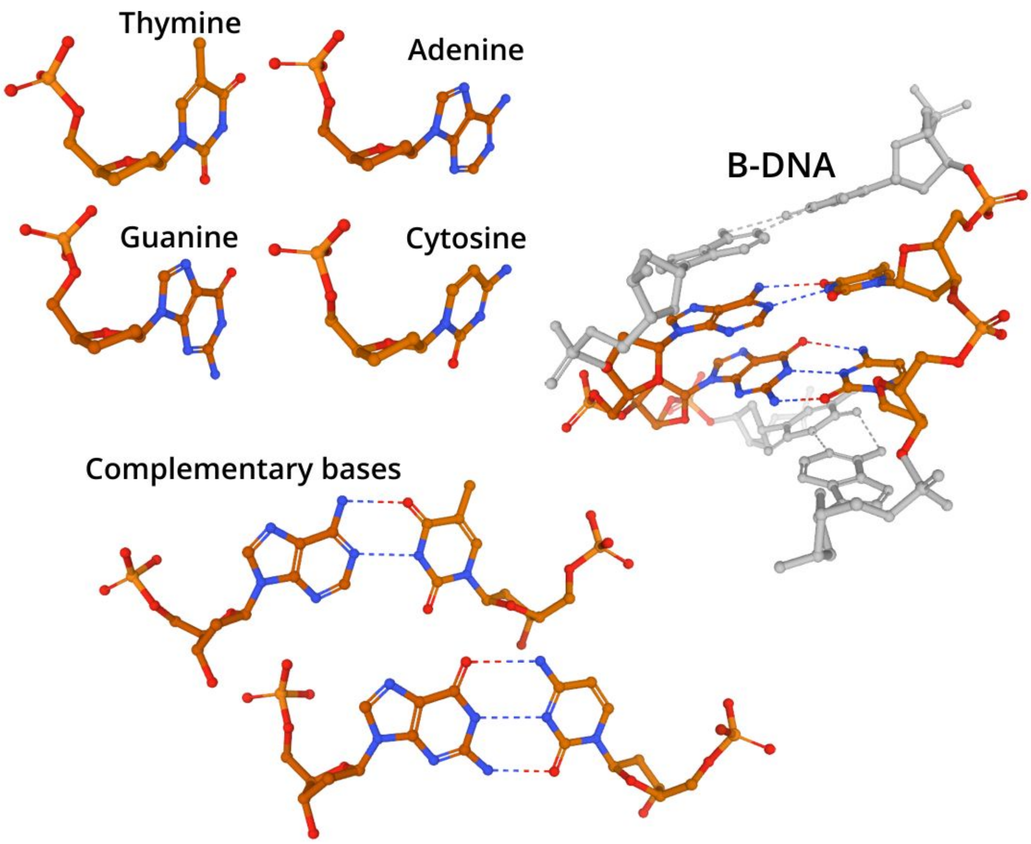

Figure 2.1: The figure shows the base units (nucleotides) of DNA: thymine, adenine, guanine, and cytosine. On the right is part of the B-DNA molecule, and under them is a representation of standard Watson-Crick complementary base pairing.

present proteins [8]. The characteristic form of DNA double helix is B-DNA. B-DNA typically occurs in physiobiological conditions.

Ribose, a nitrogenous base, and their bond have specific properties for different types of DNA double helix forms. Ribose has a carbon C2 above a sugar pentose plain that creates a spatial arrangement called C2-endo conformation. Connected base on ribose form torsion angle of N-glycosidic bond depending on base orientation. For B-DNA, it is usually anti orientation. DNA can transit its structural conformation with changes in humidity or ions concentrations. That leads to rotation of the bases around N-glycosidic bond, changes in sugar spatial arrangement, but also in a shift of base pairing complementarity. All of the mentioned properties affect the state of a DNA molecule and its other interactions.

A molecule of DNA should provide the same information through the entire cell cycle. That is partially accomplished by a DNA double helix stabilization and by packing the DNA double helix into chromosomes [9]. Hydrogen bonds between adjacent bases, hydrogen bonds between complementary bases and hydrogen bonds with surrounding water molecules assist in the stabilization of DNA double helix. On the next level, proteins interacting with the DNA double helix assist in stabilizing the packaged chromosomes.

Proteins do not serve only for DNA stabilization, but their interaction with DNA provides various functions for a cell cycle. The next sections (see Section

2.2) provide more details about these interactions.

2.2 Proteins

Smaller molecules included in the DNA-protein interactions are proteins. Proteins appear in most biological processes with a large variety of shapes and functions. Proteins have more freedom of movement compared to DNA, and their structural changes can be more pronounced. A protein molecule contains one or more polypeptide chains.

The polypeptide chain is also called the primary structure of the protein. In addition to the primary structure, proteins form secondary structures such as α-helices and β-sheets. Secondary structures group and form the tertiary structure of a protein in the process of folding. Tertiary structures can be active proteins by itself, but also they can create aggregates called quaternary structures.

Amino acids are biomolecules composed of tetrahedral carbon with four substituents. All substituents are attached to one carbon, and therefore we are talking about the C_α atom and the α-amino acids [10]. The different substituents are responsible for the asymmetry of the molecule, which can be described by chirality. Almost all amino acids in nature are L-amino acids named by chemical nomenclature guidelines. Substituents bonded to C_α are a hydrogen atom, a hydroxyl group, an amino group, and a side chain. These substituents give amino acids their properties. For example, polarity and charge of amino acids depend mainly on the attached side chain.

Amino acids are bonded together in a polypeptides chain by a covalent peptide bond. The bond arises between the nitrogen of the amyl group of new amino acid and C-terminus carbon of the carboxyl group already incorporated amino acid. These peptide bonds resonate between two states, single and double bonds, which creates a partial double-bond character. A rotation around this bond is not possible because of its double-bond nature, and therefore a peptide bond is planar and rigid. Atoms included in a planar plane are C_α carbons together with amyl nitrogen with bonded hydrogen and carboxyl carbon with bonded oxygen. Only two orientations are accessible for the planar plane considering sterical constraints of peptide bond, cis and trans conformation. Dihedral angles between $C_\alpha - N$ bond and $C_\alpha - C$ bond fully describe cis and trans conformation. For proteins is preferred trans conformation, due to sterical clashes of cis conformation.

Nevertheless, a polypeptide chain backbone has two rotatable bonds per residue. First is a bond between C_α and amyl nitrogen, and second is a bond between C_α and carboxyl carbon. Angles of rotation around these two bonds fully describe the protein backbone even if their values are restricted. The spatial arrangement of these rotations creates secondary structures of proteins. The Ramachandran plot represents allowed values of angles with two main areas of the plot appears for two main types of secondary structures in a protein [7].

The main regular shapes of secondary structures are α-helixes and β-sheets. These shapes provide hydrogen bonds to donors and acceptors in the interior of proteins. Nevertheless, secondary structures can have random shapes too. Irregular structures are loops and coils, and they often occur on the protein surface. They can form links between regular structures and are flexible.

Secondary structure elements are parts of wholly folded proteins. The folding

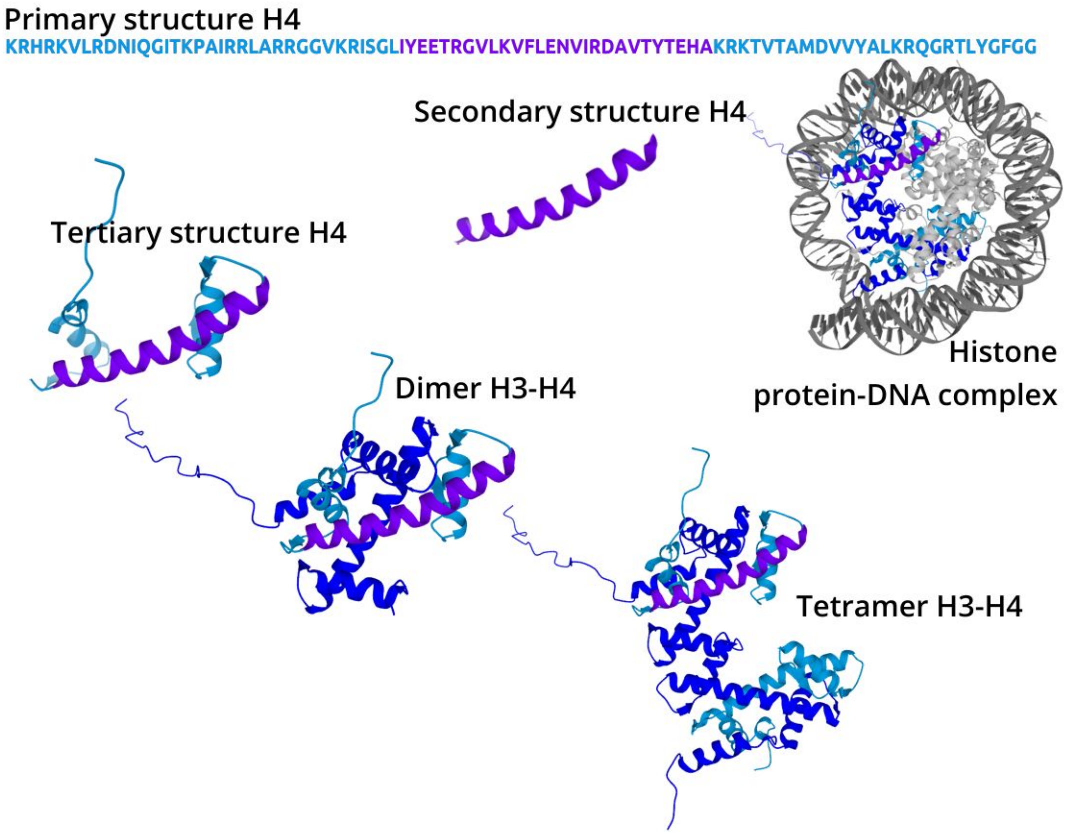

Figure 2.2: The figure shows the histone protein-DNA complex (1AOI) and histone H4 primary, secondary, tertiary structure, and its formed dimer and tetramer with histone H3.

process includes the formation of secondary structures, a hydrophobic core, and composition to the tertiary structure. Folding progress at the same time as protein grows in a process called co-translation. Amino acid sequences have many possible folds, but proteins cannot test them all to find the most energy-efficient one. Protein native conformation is a final product of folding, which is conformation with biological function. Folding is a complex process, and the amino acid sequence with kinetic of the folding process itself influences the final fold.

The alphabet of proteins is limited to twenty amino acids (+ two less common amino acids), and some of them have similar properties. Therefore, the exchange of two amino acid residues with a similar hydrophobicity or similar charge does not typically change the whole protein conformation and protein interactions.

Proteins interact with other molecules, including proteins. The interaction between protein units creates quaternary structures aggregates composed of the same type or several different types of protein units. These structures are again able to bind other molecules or macromolecules.

2.3 Protein DNA interactions

DNA binding proteins handle different types of functions. Regulatory proteins control transcriptions. Repairs of DNA require various proteins with an excising and joining activity. Topoisomerases are significant for the maintenance of an overwinding state of the DNA, and structural proteins are responsible for the state of chromatin. Processing proteins, such as polymerases, are essential proteins in DNA synthesis [11].

Different experiments of protein-DNA interaction show how proteins interact with DNA, but not all underlying principles of their interactions are clear [12, 13].Proteins can interact with DNA in one or more steps. When protein recognizes the DNA binding site, it can bind to the recognized site or make a move or rearrange its conformation. Possible moves include flip, shift, and slide [4, 1]. All of these moves could slightly change or add a protein site, which interacts with DNA.

An interaction's ability to distinguish specific sequences or shapes groups them into non-specific, sequence-specific and shape-specific interactions [11].

General properties of macromolecules determine the non-specific interactions. Features like electrostatic forces, hydrophobicity, and aromatic interactions play a role in recognizing binding sites and can influence sequence-specific and shape-specific interactions. Non-specific interactions can establish an association between macromolecules solely by the noted forces and directly or indirectly with hydrogen bonds [11]. Hydrogen bond interactions are directly between amino acids and nucleotides, or water molecules can work as a mediator for hydrogen bond.

Specific interactions consist of specific sequence recognization or specific shapes recognizations. Specific shapes recognization includes recognization of DNA motifs and recognization by specific protein binding motifs. Proteins recognize the DNA motifs, which differ from the standard form of DNA. Protein motifs binding to DNA are separated into several families by their spatial arrangement. These structures are responsible for their binding characteristics.

DNA motifs

Specific binding proteins look for specific DNA sequences or DNA shape, as the name suggests. Proteins with sequence-specific interaction to DNA read out the DNA sequence and typically bind to the major or minor groove of DNA. Proteins with the structure specificity to DNA recognize DNA shapes such as bends of the double helix, DNA kinks, Quadruplex structures, Holiday junction, another distort in B-DNA structure, and a single-strand DNA and its deformations Figure 2.3 [3].

Protein motifs

DNA binding proteins have several specific motives separated into families. Proteins in a motive family share the same or similar structural characteristics, such as bonded ion or organization of secondary structures. Families are helix-turn-helix, helix-hairpin-helix, zinc fingers, leucine zippers, β-ribbon, high mobility group and uncategorized motifs [14, 11, 12].

Helix-turn-helix [2] motifs occur in different protein structures with various functions and origin. Typically, two helices are linked together by turn at 120° to each other. One of the helixes is binding and associates only with B-DNA's

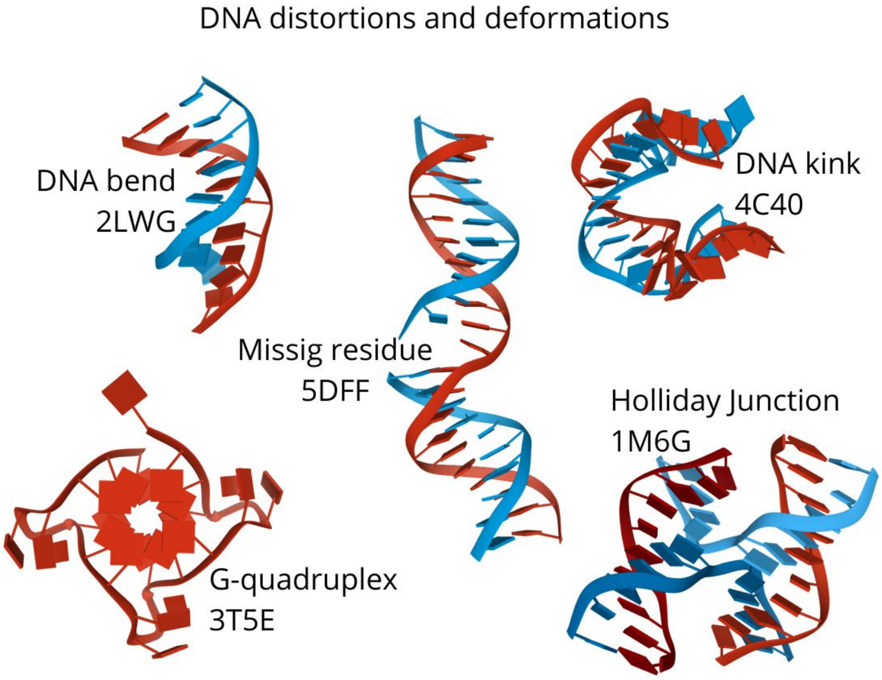

Figure 2.3: The figure shows typical distortions and deformations, which occur in DNA molecules, and which are recognized by proteins.

major groove because of its size. This motif can be part of the larger structures containing more helices, flexible turns, and small $\beta-$sheets.

Helix-hairpin-helix motifs [15] are non-specific binding motifs. Motifs recall of a helix-turn-helix with two helices linked together. A short turn between helices consists of strongly conserved amino acids and forms angle between helices, which do not allow the association of one helix with the major groove of the DNA. Functions of proteins with this motif also differ from helix-turn-helix. While helix turn-helix generally occurs in regulatory proteins, a helix-hairpin-helix frequently occurs in enzymatic proteins.

Zinc fingers [16] are specific binding motifs with a wide of folds. They generally bond in the major groove of DNA and contain one or more zinc ions. These motifs vary in secondary structures, binding zinc ions, and binding to DNA.

Leucine zippers [2] are motifs with regular leucine residues in long double helices. Helices interact directly or by connecting domains and form coiled-coil conformation. Associations with the major groove of DNA consists of specific and non-specific interactions.

The family of the motifs binding by their β-structures are β-ribbons [17]. They interact with the major groove of the DNA but also with the minor groove. A quite large amount of them can bend DNA or create kinks.

High mobility groups of motifs are diversified. Their common feature is an irregular arrangement of the three helices, but they differ in the specificity of recognizations. They also differ in contact with the major and minor grooves of DNA or interactions with double-helix or single-straight DNA.

Last but not least, atypical motifs or newly observed motifs are part of the miscellaneous family.

Chapter 3

Existing methods

Several different approaches to predict DNA binding protein sites demonstrate the complexity of the task. Methods use different algorithms and protein properties with several types of output results [13, 12]. We can classify methods by their used protein properties, applied types of algorithms, or output. Used protein properties depend on algorithm inputs and divide methods into two main groups: sequence-based methods and structure-based methods. Statistical methods or neural network are an example of used algorithms to analyze protein properties. The output has several forms. Most used is a list of binding atoms or residues, then binding sites, and simple binary classification of the whole protein.

A significant amount of methods is in the form of protocols and reports, and we do not have an implementation for existing software to test them.

3.1 Sequence-based methods

Sequence-based methods use the primary protein structure, which is easily accessible compared to tertiary structures. That is why the number of methods using only protein sequences is larger than the number of those using structure information [18].

The existing sequence-based method uses different features and different classifications to distinguish binding and non-binding residues or sites.

Methods use different protein properties calculated from a sequence
for method features. Commonly used property is sequence conservation [19, 20, 21, 22, 23]. Other properties are hydrophobicity, side chains pKa, the molecular mass of amino acids, a dipole of side chains, side chains volume [23, 24, 25]. Prediction method can also use charge of residues, sites, or proteins [26]. Pseudo amino acid compositions or the amino acid distributions on protein sites are also properties suitable for prediction [27, 28]. Two methods use sequence homologies of know DNA binding proteins, domains or motifs to predict binding sites.[29, 30]. Due to known DNA binding motifs and their secondary structures, predictions of secondary structures or their descriptions are additional properties for predictions [31, 25, 32].

Many methods use DNA-binding proteins and their properties to train machine learning classifiers including support vector machine (SVM) [33, 24, 32, 21], Random Forest (RF) [23, 25, 27], neural networks [22] or a various combinations of them [34]. But not all methods use machine learning. The methods can use linear-score [31], Hidden Markov model [20] or Gaussian network model [35].

3.2 Structure-based methods

With the increasing number of available tertiary protein structures, the number of methods using protein structures has grown as well. First methods had available only a small amount of protein structures, often with low structural resolution. The current experimental methods not only produce more protein structures with higher resolutions, but also predictions of the protein structures are more accurate. This results in more structural data for analysis and study.

Structure-based methods use different protein properties for the prediction of binding sites. Methods can use known binding motifs, the spatial arrangement of structure, and properties of protein residues. Some methods use protein properties to label proteins only as binding or non-binding.

Identifications of DNA-binding motifs set in motion the development of methods looking for structural motifs [36, 37, 38, 39, 40, 41]. Structural motifs have rather well-conserved shapes within the motif family. Known DNA binding motifs collected in databases or datasets serve as templates for the comparison of studied proteins. Comparisons return the root mean square deviation (RMSD), which shows how much proteins or parts are similar in the structural arrangement. The structural properties of proteins are used in studies to split binding and non-binding proteins often by their RMSD. Binding proteins do not create a connection to DNA with their full surface, but only with the specific binding site. Therefore methods use local curvatures and shapes to separate binding and non-binding residues [42, 43].

Several methods use properties of residues on the protein surface to predict binding sites, including electrostatic potential, accessible surface area, charge, local curvature, hydrophobicity, residue propensity, amino acid composition, and residue conservation. [44, 45, 46, 47, 38, 42, 48, 49, 50].

Types and amounts of the secondary structures, net charge, dipole moment, and quadrupole moment of entry protein are properties used in studies that separate binding and non-binding proteins without looking for specific binding sites [49, 51, 52].

Similar to the sequence-based method, many of structure-based methods use protein properties to train machine learning classifiers.Considering the number of implementations, the methods in the machine learning group use mostly neural networks [49, 53, 45, 47] followed by SVM [50, 44, 54], and Random Forest [55]. We can also find methods, which use linear-score [38] or statistical evaluation functions [42, 43] to discriminate binding and non-binding residues.

The next sections describe three methods that use residue properties on the protein surface and use a machine learning classifier to discriminate binding and non-binding residues. These parameters correspond with the P2Rank approach, which we want to modify to predict DNA binding sites. Methods references are available on the omictools.com by searching for the ”protein DNA interaction prediction.”

3.2.1 DISPLAR

DISPLAR [56] is a structure-based method designed for the prediction of DNA binding sites on a protein's surface. Binding site prediction uses a neural network classifier, and it is a modified version of previously published method PPISP [57, 58].PPISP predict protein-protein interactions using protein residue properties. Modifications of PPISP consisted of the newly implemented features that use information form the protein structure and the composition of the amino acids in the protein.

Data used for the classification consist of three parts. First is a sequence profile of the protein, which is produced by PSI-BLAST [59]. Second is solvent accessibility for each residue on the protein surface, and third is fourteen neighboring residues for the investigated residue.

Neural network training uses two feed-forward, back-propagation neural networks. Both neural networks have different numbers of nodes. The first has 15x21 nodes, where 15 represent considered residue and 14 spatial neighbor residues, and 21 represent input variables, including solvent accessibility of considered residue and sequence profile. The first neural network output is a layer of two nodes, one for predicting non-binding and predicting binding residue. This output is part of the input for the second neural network together with solvent accessibility for residues. The second neural network designed for the window with 15 residues has two nodes with the binding state on the output. The training outcome is a weight matrix initialized with random values and modified each round in the training process.

As suggested in PPISP, the classification uses a consensus approach of the collecting weight matrices for cross-training. The consensus approach includes two steps. The first step was to cluster all positive predictions from weight matrices by their consensus score. The consensus score is the sum of positives predictions in different weight matrices. In the second step, the consensus score and cluster size determine an optimal collection of the weight matrices.

The neural training runs with the collected dataset and separately on the dataset with partially trimmed non-binding residues to improve the accuracy of predictions.

Evaluation dataset

The dataset used for the neural network training was obtained from the Protein Data Bank where the filter was set so that the structures needed to contain at least one DNA chain and one protein chain. All of the obtained protein chains were processed by PSI-BLAST afterward to achieve the sequence identity among chains less or equal to 50%. The removal of similar structures increased the non-redundancy of the dataset, typically the entries with the highest structure resolution were retained. The next step of pre-processing was to remove chains with sequence length forty or less because of their inability to obtain a position-specific scoring matrix from PSI-BLAST. The neural network algorithm operates on residue level, and for that reason, residues of training sequences has to have a label determining their binding state. Pair of heavy atoms have to exist across the protein-DNA interface within a distance of 5Å to label a residue as binding.

Division	Accuracy	Coverage
three-tier	64.8	80.2
two-tier	63.9	84.2

Table 3.1: The cross-training results for the three-tier and the two-tier divisions of the dataset using DISPLAR.

Results

Predictions assessment is rated for DISPLAR by coverage 3.1 and accuracy 3.2. The three-tier and the two-tier are division used for the training, cross-training, and testing. Table 3.1 presents their coverage and accuracy.

$$coverage = \frac{n_{tp}}{N_{dc}} \tag{3.1}$$

$$\begin{aligned} n_{tp} &- \text{number of the predicted true positive residues} \\ N_{dc} &- \text{number of real DNA binding residues} \end{aligned}$$

$$accuracy = \frac{n'_{tp}}{N_{pr}} \tag{3.2}$$

$$\begin{aligned} n'_{tp} &- \text{addapted number of predicted true possitive residues} \\ &\text{(counting as positive four nearest neighbors)} \\ N_{pr} &- \text{number of the predicted binding residues} \end{aligned}$$

Protein-DNA interactions have an impact on the structural conformations of proteins and DNA. Fourteen of twenty-five proteins in the two-tier division had been available with their unbound structures and thus were used to test the prediction ability of the DISPLAR for this scenario. All of these proteins, coverage, and accuracy are computed together with the root-mean-square deviation (RMSD) of $C\alpha$ atoms. Results of cross-training for bound and for unbound structures are comparable, but with notable deterioration.

Only two of fourteen protein structures have RMSD more than 2.5Å. These two proteins represent two types of conformation changes in proteins, the global distortion and proteins domain rearrangement. Despite modifications in structural conformation, DISPALR can determine the part of binding residues.

3.2.2 iDBS

iDBS [60, 55], is a random forest model for detection of functional regions on the surface of DNA-binding proteins. Core of iDBS is another method PatchFinder [61, 62], created to identify conserved function regions on the surface of proteins. PatchFinder, together with local and global physicochemical features, is used to get required specificity for the model of DNA-binding proteins. A most prominent feature is a positive electrostatic charge of binding regions, and hydrogen bonds have an important role, especially in recognition.

Due to common helix-related DNA-binding motifs, residues' helical conformation is one of the features used for classification. The next ten features added from the work of Szilágyi, A. & Skolnick [52] are dipole moment of the full molecule, percentage of Arginine, Alanine, Glycine, Lysine and Aspartic acids and spatial asymmetry of Arginine, Glycine, Asparagine, and Serine.

Because the model's training includes results from PatchFinder, a brief description of this algorithm is fitting. PatchFinder uses 3D structures of proteins, multiple sequence alignment of a query protein MSA for input. PatchFinder output is a cluster of residues on the protein surface, which PatchFinder assigned as a functional site of a protein. These functional sites are called ML-patches.

The final classifier is composed of two separately trained models. One uses 16 features and is suitable for prediction on protein chains, which have too few sequence homologs to predict ML-patches. The second classifier uses all 33 defined features, including features based on the ML-patches.

The models use 138 binding protein chains and 110 non-binding protein chains for training. Evaluation of the model is determined using 10-fold cross-validation runs on the same dataset.

Evaluation dataset

Model training uses the dataset from the previously published method [52]. Workflow for creating the dataset requires to retrieve all PDB structures with x-ray resolution less than 3.0Å from a set of protein-DNA complexes containing double-helix DNA.

From retrieved structures are removed all protein chains, which have less than 41 residues and less than 5 DNA binding residues. If the PDB structure has less than five pairs of DNA-bases, than they remove protein chains of structure from the dataset. From retrieved structures are removed all protein chains, which have less than 41 residues and less than 5 DNA binding residues. If the PDB structure has less than five pairs of DNA-bases, than they remove protein chains of structure from the dataset. They also remove protein chains with pairwise sequence identities more than 35% to create a non-redundant dataset from protein chains. The process returns 138 proteins chains.

A dataset with 110 non-binding protein chains is created by the same redundancy criteria to provide real data for training the model on binding and non-binding proteins.

The same redundancy criteria create two more datasets for testing purposes.

A small independent dataset formed from eleven additional structures found in the Protein Data Bank satisfies the 35% identity. This dataset is prepared to test the trained model on bond and unbound structures.

An extended dataset or "real-data" dataset is created by adding the non-redundant structures, satisfying 35% identity to simulate the real data experience. Furthermore, the extended dataset consists of the original dataset enriched by 733 non-binding proteins. The dataset represents the ration of the binding and non-binding proteins in the proteome established by observed data in the database.

Results

Results of the classifier rated by MCC, Matthew's correlation coefficient, for dataset reaches value 0.80. Specificity and sensitivity measure the performance of the model too. (see Table 3.2).

Tests running on additional eleven proteins available with their unbound structure identify binding protein correctly with specificity and sensitivity 0.9.

Tests run on the "real life" dataset get the drop in specificity to 0.72, but sensitivity stays at 0.9. However, an increase in specificity reached by a suitable cutoff of the evolutionary score is possible, but in this case, sensitivity decrease.

Dataset	Specificity	Sensitivity	MCC	AUC
training dataset	0.9	0.9	0.8	0.96
extended dataset	0.72	0.9	.	.
extended dataset with score cutoff	0.85	0.82	.	.

Table 3.2: The cross-training results of the iDBS using 10-fold cross-validation on the dataset of 138 bindings and 110 non-binding protein chains.

Used features groups are ranked to show their significance. From seven features groups, electrostatic potential has the highest impact on the model. Other categories have a smaller effect: amino acid conservation pattern, secondary structure, amino acid content, hydrogen donor/acceptor bonds, dipole moment, and amino acid asymmetry in exact order.

Method tested the hypothesis of the high conservation of functional regions. Results suggest that at least a small part of residues is conserved, and at least half of the residues found in ML-patch are in contact with DNA.

3.2.3 DBSI

DBSI [63] uses for training is an SVM model, precisely SVM light program [64], which classify binding and non-binding residues. To avoid over-fitting, the selection of used features runs through an iterative process. Two smaller datasets are applied to accelerate the model's training during iterative feature selection and SVM parameter selection. Datasets collected from previously published articles are processed to get rid of similarity biases. Nevertheless, method bias lies in datasets due to the imbalance between binding and non-binding residues. This imbalance favors accuracy in non-binding residues predictions because the number of non-binding residues exceeds the number of binding residues.

The article presents a set of features, but only a small part of them pass the iterative selection. Features use standard residue properties such as hydrophobicity, size, charge, but also secondary structure assignment by DSSP [65], SASA by NACCESS [66], availability of the hydrogen bonds, electrostatic properties calculated by PBEQ-solver [67, 68], surface curvature by SURFCV [69] and local atom density by FADE [70]. Features describe residue level properties or atom level electrostatic properties. Nevertheless, also describe the microenvironment of a residue using neighboring residues within several distances.

Evaluation dataset

Training and testing of the SVM model apply datasets from previously published articles. Training dataset, TRAIN-263 use PDB structures collected for DISPLAR [56] method without the 1MJE structure, witch misses too many fragments. Testing dataset, TEST-206 use available PDB structures for proteins used in metaDBsite [34], which is a compilation of the methods using the sequence-based approach. Binding residues have to be well defined to obtain separation of bound and unbound residues. For training, if an amino acid residue is a surface residue and has any heavy atom within 5Å distance from any DNA heavy atom, than the amino acid residue is binding.

The training dataset went through the modification process to prepare it for cross-validation and decrease the bias against sequence and structure similarity. Training data create ten groups. Structures belong to the same group if their sequence similarity is more than 25% or if they share common fold or if their TM-score [71] is more than 0.3. The TM-score was computed by TM-align using the RMSD for comparison of the structural similarity.

Structures with bound and unbound form add information about the prediction ability of DBSI in biological conditions, where proteins can dramatically change their tertiary structure depending on the bound components. Obtained HOLO and APO structures from articles of Ozbek [35] and Xiong [48, 54] were also processed and sorted for testing. Processing included removing the structures with too many residues or structures with only $C\alpha$ atoms, and then homology was calculated between HOLO and APO structures and the training data. 29 APO and 30 HOLO forms obtained as outcome had applied TM-align to determine values of structural changes between bound and unbound structures.

Model training

The article presents 480 features for model training to describe protein properties. Not all of the features provide necessary information for training, and to avoid over-fitting, only few of them are used. Training is composed of three separate steps: iterative feature selection, SVM parameter selection, final model training.

Iterative feature selection uses 1000 randomly chosen residues, which respond by the ration 18% to 82% of a bind and non-bind residues to the entire training dataset. Features are applied one by one to train the model on a reduced dataset. The best scoring model features combine to retrain the model.

The iterative selection process adds the next features to the combined model. Features are added to the mode one by one. In each iteration, they select a feature from the best scoring model for the next retraining. Iterations repeat until the model score is not improving anymore. The score used to determine how good is a trained model is the F1 score. The best values in single feature training reached electrostatic features and PSSM feature groups.

The final model with specific parameters trained on the entire dataset uses 10-cross fold validation for evaluation.

Results

The prediction ability of the DBSI is comparable to other methods with the accuracy, specificity, and sensitivity with values 0.83, 0.85, and 0.74, respectively. Different features contribute differently to the prediction ability of the final model.

Dataset	MCC	AUC	Specificity	Sensitivity	Accuracy	Precision	F1	Strength
TRAIN-263	0.48	0.86	0.85	0.70	0.82	0.5	0.58	0.77
TEST-206	0.51	0.88	0.85	0.74	0.83	0.49	0.59	0.80
HOLO-30	0.44	0.85	0.89	0.6	0.85	0.45	0.52	0.75
APO-29	0.41	0.83	0.89	0.58	0.86	0.42	0.48	0.73

Table 3.3: The table shows the cross-validation results of the model for datasets used by DBSI.

Iterative selection allowed test individual features' prediction ability and prevented the over-fitting of the model. The best scoring features were electrostatic features followed by the PSSM features and features of the residue environment.

Electrostatic features computed for atom level use sphere with a radius of 1.4Å, which corresponds to the value of the solvent probe radius. The original approach of the DBSI shows that omitting the properties within the inner part of the spere can lead to filtering the false positives and increase the sensitivity of the model. Ignored the spere's inner radius, which achieved the best result was 0.5Å from the van der Waals molecule surface.

PPSM features show a promising result individually with no significant difference with the use of the different window sizes for neighboring residues. The best scoring electrostatic feature model and best PPSM features model serve as the combined model for further iterative additions.

In the iteration process, the features of the residue microenvironment added to the trained model do not show separately better results but increased the model's predictive ability. Features included in the residue microenvironment include the local shape, electrostatic features, and distribution of the amino acids for the residue and its neighboring residue.

Testing on the dataset of HOLO and APO structures indicate low sensitivity to spatial changes of protein conformation.

Chapter 4

P2Rank

As stated before section 2.2, proteins have a wide diversity of shapes but also handle various functions. The active state of a protein often depends on bonded ligands, which are small molecules able to bind on the protein surface. This dependency led to the implementation of tools/software able to detect ligand binding sites on a protein surface. P2Rank [72, 73, 74, 75] based on machine learning using Random Forest classifier is one of them.

P2Rank predicts ligand-binding pockets from a protein structure. At first, P2Rank generates Connoly's point on the solvent-accessible surface of the protein. These points get assigned averaged feature vectors of all neighboring exposed atoms in the selected radius. All exposed atoms have a feature vector that describes properties suitable to identify ligand-binding interactions. Based on the averaged feature vectors, the P2Rank classifier assign sore to each Connoly point. Connoly points scoring above the threshold are clustered together and create pockets. Then P2Rank ranks pockets by their cumulative score for output.

Our problem, prediction of DNA-binding sites on proteins surface, has similar characteristics to the prediction of ligand-binding pockets. We have a protein structure on the input, and we want to find DNA binding sites.

Interaction between proteins and ligands uses the physicochemical and geometrical properties of amino acids exposed atoms. The differences in interaction are binding molecules and values of properties. P2Rank can be used for prediction of ligand-binding pockets and also for prediction of DNA-binding sites, but thresholds of used properties can differ. Therefore, we have to retrain a model with appropriate thresholds to get correctly defined binding properties. Also, we want to expand P2Rank with new features that are not available yet. These features should help with DNA-binding site predictions but also with predictions of ligand-binding sites.

4.1 Modifications

P2Rank shows better results than alternative methods for structure-based detection of ligand-binding sites. One of its advantages is the ability to find binding sites formed by multiple protein chains. P2Rank is also independent of additional bioinformatics tools or databases. For these reasons, we use P2Rank as the basis for the prediction of DNA-binding sites.

To properly modify the P2Rank, we have to analyze how DNA-binding differs from the ligand-binding and describe their differences. These differences then determine features which can be implemented into P2Rank for it to be better suited for the DNA-binding detection task.

If we compare ligand-binding sites to DNA-binding sites of proteins, then some differences are noticeable. At first sight, the size and shape of the sites clearly differs. DNA-binding sites have shallower clefts than ligan-binding sites, and the size of their binding sites can be more extensive. Feature vectors of P2Rank already include these properties, such as the number of exposed atoms and protein surface protrusion, which reflect this. Therefore, adjusting their values by retraining the model on as suitable DNA-protein binding dataset is sufficient.

The difference in the electrostatic potential on a protein surface is another contrast between DNA-binding and ligand-binding sites. An electrostatic potential map (Figure 4.1) shows charge distributions on a protein surface for ligand-binding (2SRC) and DNA-binding proteins(4O5K). DNA-binding protein shows a more compact distribution of electrostatic potential with a wider positive cluster, part of DNA-binding pocket.

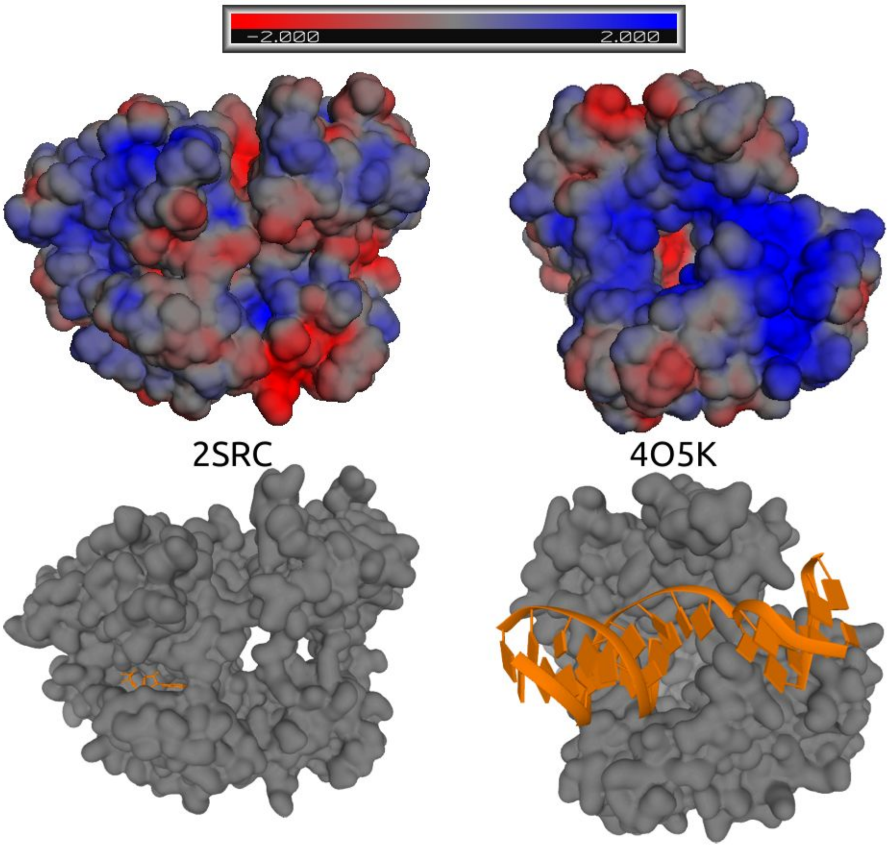

Figure 4.1: Electrostatic potential maps for ligand-binding structure (2SRC) and DNA-binding structure (4O5K) and their bonded forms with ligand and DNA show a difference in electrostatic potential of binding pockets.

Implementation of electrostatic potential as a feature for P2Rank requires a more accurate description. The subsection 4.1.1 explains in more detail electrostatic potential, its calculation, and obtaining an electrostatic potential map using existing tools.

4.1.1 Electrostatic potential

Previously published studies showed that electrostatic potential had a significant impact on the prediction of DNA-binding sites. This and differences between DNA-binding and ligand-binding pockets were the main reasons to implement electrostatic potential into P2Rank.

The molecule's electrostatic potential is the interaction energy between the molecule charge distribution and the positive charge of a unit [76]. The electrostatic potential depends on an electric charge of atoms and their distance from each other, but it is also necessary to take a solvent charge into account. The irregular shape of proteins also adds variables for the calculation of electrostatic potential. The Poisson-Boltzmann equation incorporates all of the variables and solves the problem of the electrostatic potential of proteins. The Poisson-Boltzmann equation is a non-linear differential equation that usually approximated as a series expansion with just the first term retained [77].

Several teams work on the problem of generating an electrostatic potential map for macromolecules. They use the Poisson-Boltzmann equation to solve the electrostatic potential and generate a map (see section 4.1.1).

Electrostatic potential implementation

Before we assign values to feature vectors to atoms, we have to generate an electrostatic potential map of proteins. For that purpose, we used previously published work and tools to create a script that uses the APBS [78] and DelPhi [79, 80] tools for calculation of protein electrostatic potential.

Electrostatic potential:

```
$file$ ⇐ path to PDB file
$library$ ⇐ chosen library
$forcefield$ ⇐ chosen force field for library
if $library$ is DelPhi then
    prepare PRM file with $file$ and $forcefield$
    run DelPhi with PRM file
    save CUBE file and LOG file
end if
if $library$ is DelPhi then
    run PDB2PQR with $forcefield$ and $file$
    set apbs-input file
    run APBS with apbs-input file
    run DX2CUBE
    save CUBE file and LOG file
end if
```

Both these tools can generate an electrostatic potential map based on the chosen force field, and this option is allowed for our script too. Both of the tools need PDB structure files and different input data for calculation.

The DelPhi needs input files, which contain atoms charges and atoms sizes. Several pre-prepared files containing this data are available by DelPhi. Paths to these files are included in the PRM format file to describe other parameters available for calculation. The returned electrostatic potential map use CUBE format.

The APBS needs for input pre-prepared PDB structure containing again sizes and charges of atoms and specific input files with other parameters.The PDB2PQR [78] is a tool, which prepares this PDB structure and generates an input file for APBS. Unlike the DelPhi, APBS returns electrostatic potential map in DX format, and therefore available python script converts the format SX to CUBE.

Our script needs only structure in the PDB format and returns the LOG file and CUBE file, simplifying the workflow.

4.1.2 Dipole moment

Calculation of electrostatic potential is not the only way how to use the charge of atoms in proteins. The dipole moment of proteins describes the electrical polarity of proteins using bonded atoms in a protein. Bonded atoms can differ in their electronegativity, and hence arise a separation between positive and negative charge. For protein, the dipole moment composes of partial dipole moments of individual bonds in a protein (see section 4.1.2).

The article [51] uses net charge, dipole moment, and quadrupole moment for predicting DNA-binding proteins. The study has shown that these properties can distinguish DNA-binding proteins with high accuracy from non-binding proteins. Furthermore, we can consider the dipole moment as a new feature for P2Rank. We calculate one dipole moment for the entire protein, and therefore it is a global feature. To use the global feature in P2Rank, we can add for each vector the same value.

Dipole moment calculation

We used the formula (4.1) to express the dipole moment of a molecule from [51]. The reference point (4.2) set on the geometric center of protein gives the best classification between binding and non-binding proteins, and the formula shows its estimation.

Dipole moment:

```
$file \Leftarrow$ path to PDB file
$forcefield \Leftarrow$ chosen force field
$dipolemoment \Leftarrow [0, 0, 0]$
run PDB2PQR with $forcefield$ and $file$
$R_0 \Leftarrow$ geometric center of PDB strucutre
for each atom in PDB file do
    $dipolemoment += $ (distance of atom and $R_0$) $*$ atom charge
end for
```

Before we calculate dipole moment, we need to assign charges and coordinations in space for all atoms. Coordinations are available from PDB structure files. The PDB2PQR tool, used for electrostatic potential script, can generate the charge for all atoms in a PDB structure.

$$\sum (R_i - R_0) q_i \tag{4.1}$$

$$R_i - \text{position vector of i-th atom}$$
$$R_0 - \text{geometric center of protein}$$
$$q_i - \text{charge of i-th atom}$$

$$R_0 = \sum \frac{R_i}{N} \tag{4.2}$$

$$R_0 - \text{geometric center of protein}$$
$$R_i - \text{position vector of i-th atom}$$
$$N - \text{number of atoms in a protein}$$

4.1.3 P2Rank integration

The integration of the P2Rank includes changes in source code, which is a complex task. Therefore integration of new features will be prepared by a third party, which develops features for P2Rank. The integration of features is still in progress (see subsection 4.2.2).

4.2 Evaluation

In order to evaluate adapted P2Rank, we took a few preprocessing steps. We created the new dataset for P2Rank, which contains protein-DNA complexes. On this dataset are trained new models with different parameters. For testing, we have chosen the trained model with the best results. Newly designed features are in the process of implementation, and therefore we present test results only for models trained without them. Finally, we compare P2Rank with existing methods.

4.2.1 Dataset construction

The dataset used in the above introduced methods does not represent the available protein-DNA complexes to this date. The number of available structures increased over time, and X-ray diffraction resolutions have improved too. Therefore, we created a new dataset for training, evaluating, and testing P2Rank.

Process of creating the dataset

- We obtained list of PDB entries (entries.idx) from RCSB database summaries.

- We used the RCSB Search API, to select only structures containing protein-DNA complexes. The query to API had the ”Polymer Entity Type” parameter set to protein and DNA.
- The query returned a list of PDB IDs, which contains protein and DNA structure.
- We merged both lists by the PDB ID of entries.
- To represent each protein by one structure, we downloaded unitprot_pdb.csv from ftp.ebi.ac.uk/pub/databases/msd/sifts/csv/. File contain list of UnitProt entries with mapped corresponding PDB IDs.
- In the next step, we removed all UniProt entries, which do not have at least one protein-DNA structure.
- Retained UniProt entries have a list of PDB IDs. PDB IDs are for structures with protein-DNA structures and also for simple protein structures. We removed all simple structures to get only complexes for training. In case when we will want to test models on unbound structures, PDB IDs for simple structures are saved separately for each relevant UniProt entry.
- For each UniProt entry, we want only one structure for representation. Therefore we sort the list of PDB IDs by their accession date and their resolution if the used experimental method was X-Ray diffraction.
- We retained the first structure from the sorted list as represent for each UniProt entry.
- We got 1032 protein-DNA structures in our dataset.
- Next, we randomly divided the dataset into three parts: the training set (516 structures), the evaluation set (170 structures), and the testing set (346 structures).

We produce a script for this process, which allows us to obtain a resent representation of database data if the new structures appear. Datasets are available on https://github.com/gajdosp2/p2rank-data-dna.

Radius	Binding	Non-binding	Ratio
4.0	562717	4171513	0.135
4.5	637756	4096477	0.156
5.0	704672	4029558	0.175
5.5	769246	3964984	0.194
6.0	842311	3891919	0.216

Table 4.1: The number of residues labeled binding and non-binding in the dataset for different radius.

4.2.2 Training and parameter selection

We trained models applying newly created datasets without implemented features described in section 3 because their integration is still in process. Before we could train models, we had to prepare datasets. P2Rank is prepared for recognization binding ligands in structure, but not for recognization of binding

DNA. Nevertheless, P2Rank is capable of training, evaluating, and testing with additional data, which contains labeled residues by their binding state.

We need two additional files for training, evaluation, and testing dataset. The first file **dataset_chains** should contain a path to PDB structure and protein chain name (**../eval/4BQA.pdb.gz G**). We need structured data in three lines for the second file **dataset_lables**. The first line has a PDB ID, followed by a protein chain name (**>4BQAG**). The second line has a protein chain sequence using the one-letter name (**GPIQLWQF...**), and the third line has a protein chain sequence using binary labels for their binding state (**00011000...**).

Dataset labeling:

```
for each protien chains in the PDB strucutre do
    add PDB ID and chain name to dataset_chains
    for for each residue in protien chain  do
        if residue is in radius R from DNA  then
            set residue label to 1
        else
            set residue label to 0
        end if
        add to dataset_lables: PDB ID and chain name
        add to dataset_lables: protein chain residue sequence
        add to dataset_lables: protein chain residue labels
    end for
end for
```

Described methods in chapter 3 use different definitions for binding residue. Definitions differ in the distance of a residue from the DNA molecule. To select an optimal radius for our model, we trained several models using a radius of 4.0, 4.5, 5.0, 5.5, and 6.0Å (see Table 4.2).

We wanted to know how good are newly trained models in predictions compared to P2Rank trained for ligand-binding proteins (P2Rank_ligand). Therefore we evaluate P2Rank our new test dataset for all radii. Then we evaluated our new models (P2Rank_DNA models) using the same test dataset.

Each model trained on the new dataset shows better results than the model trained for ligands. MCC value of P2Rank_DNA models increases from 4.0 to 5.0 bind radius, and for radius 5.5 and 6.0 value decrease. AUC values, accuracy, and specificity slightly decrease with a growing radius. On the other hand, sensitivity, precious, and F1 values increase with a growing radius. After taking into account the evaluation result and the radii used in previous methods, we chose the P2Rank_DNA model using bind radius 5Å for the next demonstration.

The table shows us statistical data of P2Rank_ligand and P2Rank_DNA models, but we would like to demonstrate different outputs in protein structure predictions. For that purpose, we choose structure 6PAX, which is DNA binding.

The Figure 4.2 illustrates predictions for the P2Rank_ligand model, the DBSI web server tool, and the P2Rank_DNA model with a 5Å bind radius. DBSI recognized the binding place on the surface in large measures as one binding place. P2Rank_ligand was not able to identify all binding residues, but the identified part looks correctly. The P2Rank_DNA found three binding sites, which are, in

Model	Radius	MCC	AUC	Accuracy	Specificity	Sensitivity	Precision	F1
Ligand	4.0	0.1634	0.6849	0.6437	0.6479	0.6083	0.1719	0.2881
DNA		0.4844	0.8977	0.8734	0.8956	0.6881	0.4421	0.5383
Ligand	4.5	0.171	0.6815	0.6447	0.6509	0.6011	0.1957	0.2953
DNA		0.4912	0.8922	0.8603	0.8823	0.7036	0.4569	0.554
Ligand	5.0	0.1717	0.6734	0.644	0.6534	0.5864	0.217	0.3167
DNA		0.4933	0.8857	0.8423	0.8608	0.7301	0.4606	0.5648
Ligand	5.5	0.1746	0.6682	0.6436	0.6564	0.5758	0.2396	0.3384
DNA		0.489	0.8785	0.824	0.8389	0.7446	0.4638	0.5716
Ligand	6.0	0.1717	0.6565	0.6407	0.6586	0.5588	0.2631	0.3577
DNA		0.4905	0.8732	0.8066	0.8159	0.7641	0.474	0.585

Table 4.2: The table shows the evaluation of our test dataset on the P2Rank_ligand model and the all trained models P2Rank_DNA for different bind radius.

fact, one binding site. These show us that the P2Rank_DNA model can found binding residues, but clustering them to binding sites is not optimized for DNA-binding sites.

We created a demonstration of structure 6PAX for all newly trained models to compare them with each other .

The Figure 4.3 demonstrate residues defined as binding by our dataset preparation (see section 4.2.2) for structure 6PAX. On top of the figure, the color scale distinguishes the binding residues by their radius. Under that, we have predictions generated by P2Rank_DNA models by the corresponding radius. Each model predicts at least two binding sites for protein structure. Overall, they predict residues partially correctly, but we can see a problem identifying the binding site as a whole.

4.2.3 Comparison with existing methods

We use 10-fold cross-validation for comparison of P2Rank with existing methods. Table 4.3 shows validation results using the same dataset for all methods, except for the iDBS. Data for DISPLAR and DBSI is from the article [63], and for P2Rank, we run 10-fold cross-validation using the same dataset TRAIN-263. The P2Rank shows comparable results to other methods. Our specificity surpassed other methods, but our sensitivity is lower than the values of the method.

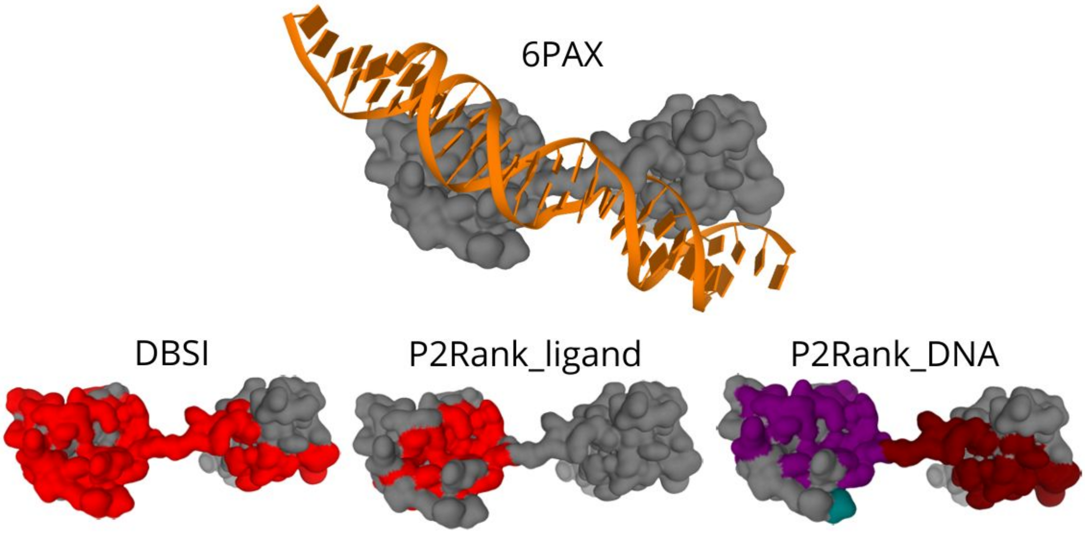

Figure 4.2: The figure shows protein Pax-6 (structure 6PAX) with bonded DNA. The structures under 6PAX present predicted residues by P2Rank and DSBI. P2Rank_ligand is the default model for ligand-binding classification. Used DSBI is a web server tool of the previously described method. P2Rank_DNA was trained on newly created datasets using a bind radius of 5Å.

Method	MCC	Accuracy	Specificity	Sensitivity	Precision	F1
P2Rank_DNA	0.41	0.86	0.93	0.45	0.54	0.49
DBSI	0.48	0.82	0.85	0.70	0.50	0.58
DISPLAR	0.34	0.76	0.79	0.60	0.39	0.47
iDBS	0.8	.	0.9	0.9	.	.

Table 4.3: The table shows 10-fold cross-validation on the dataset used for training DISPLAR and DBSI. We run cross-validation on P2Rank with the same dataset. The model used for cross-validation is P2Rank_DNA with a 5Å bind radius, which corresponds to the radius used for DISPLAR and DBSI. Underline is cross-validation of iDBS, which uses a different dataset.

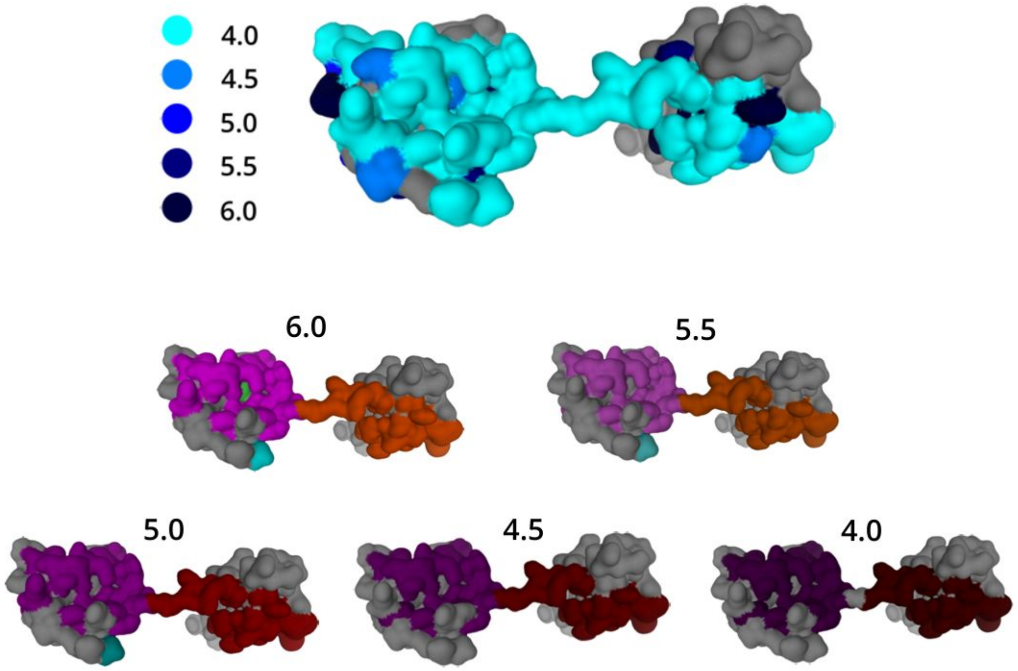

Figure 4.3: The figure shows on the structure 6PAX residues defined as binding for different radii by color scale. Under scale is a set of structures representing predictions for each newly trained model.

Chapter 5

Conclusion

Protein-DNA interactions are significant parts of cell life and cell process in all living organisms. Uncover involved principles could lead us to design treatment for different conditions in regular expressions and other cell processes.

Even the best tools do not have perfect results, but their improvement can help understand the principles involved in interactions. Vice versa, a better understanding can lead us to more reliable predictions results.

Existing methods show all sorts of prediction approaches. They are looking at the problem from a sequence view or structural view, and results are better with each new tool. Macromolecules participating in interactions have various properties with different weights for prediction protein-DNA interactions. Electrostatic potential shows the best efficiency for disguising binding and non-binding DNA residues on a protein surface.

P2Rank predictions show comparable results for the newly trained models and existing methods. They compared model shows a better result in specificity, but its sensitivity was a bit lower. Classifiers can distinguish at least the part of the binding residues in binding sites correctly. Implementation of new features should increase the prediction ability of P2Rank for protein-DNA binding sites.

Bibliography

[1] S. E. Halford. How do site-specific dna-binding proteins find their targets? *Nucleic Acids Research*, 32(10):3040–3052, Jun 2004.

[2] George P. Rédei. *Helix-Turn-Helix Motif*, pages 850–850. Springer Netherlands, Dordrecht, 2008.

[3] Sabrina Harteis and Sabine Schneider. Making the bend: Dna tertiary structure and protein-dna interactions. *International Journal of Molecular Sciences*, 15(7):12335–12363, Jul 2014.

[4] Mahipal Ganji, Margreet Docter, Stuart F.J. Le Grice, and Elio A. Abbondanzieri. Dna binding proteins explore multiple local configurations during docking via rapid rebinding. *Nucleic Acids Research*, 44(17):8376–8384, Jul 2016.

[5] Nicholas M. Luscombe and Janet M. Thornton. Protein–dna interactions: Amino acid conservation and the effects of mutations on binding specificity. *Journal of Molecular Biology*, 320(5):991–1009, Jul 2002.

[6] Garrett Soukup. *Nucleic Acids: General Properties.* 05 2003.

[7] Anders Liljas. *Textbook of structural biology.* World Scientific, New Jersey, 2017.

[8] Dave Ussery. *DNA Structure: A-, B- and Z-DNA Helix Families.* 05 2002.

[9] Burke Judd. *Nucleic Acids as Genetic Material.* 04 2001.

[10] Stanley Maloy. *Brenner's encyclopedia of genetics.* Academic Press, San Diego, CA, 2013.

[11] Stephen Neidle. *Principles of Nucleic Acid Structure.* Elsevier, 2008.

[12] Abbasali Emamjomeh, Darush Choobineh, Behzad Hajieghrari, Nafiseh MahdiNezhad, and Amir Khodavirdipour. Dna–protein interaction: identification, prediction and data analysis. *Molecular Biology Reports*, 46(3):3571–3596, Mar 2019.

[13] Jingna Si, Rui Zhao, and Rongling Wu. An overview of the prediction of protein dna-binding sites. *International Journal of Molecular Sciences*, 16(12):5194–5215, Mar 2015.

[14] Marianne Rooman and René Wintjens. Protein-dna interactions, Mar 2015.

[15] X. Shao. Common fold in helix-hairpin-helix proteins. *Nucleic Acids Research*, 28(14):2643–2650, Jul 2000.

[16] M. Isalan. *Zinc Fingers*, page 575–579. Elsevier, 2013.

[17] S. Kim. Beta ribbon: a new dna recognition motif. *Science*, 255(5049):1217–1218, Mar 1992.

[18] Yanay Ofran, Venkatesh Mysore, and Burkhard Rost. Prediction of dna-binding residues from sequence. *Bioinformatics*, 23(13):i347–i353, Jul 2007.

[19] Bruno Contreras-Moreira, Pierre-Alain Branger, and Julio Collado-Vides. Tfmodeller: comparative modelling of protein–dna complexes. *Bioinformatics*, 23(13):1694–1696, Apr 2007.

[20] S. K. Kummerfeld. Dbd: a transcription factor prediction database. *Nucleic Acids Research*, 34(90001):D74–D81, Jan 2006.

[21] S. Hwang, Z. Gou, and I. B. Kuznetsov. Dp-bind: a web server for sequence-based prediction of dna-binding residues in dna-binding proteins. *Bioinformatics*, 23(5):634–636, Jan 2007.

[22] Shandar Ahmad and Akinori Sarai. Pssm-based prediction of dna binding sites in proteins. *BMC Bioinformatics*, 6(1):33, 2005.

[23] Liangjiang Wang, Mary Yang, and Jack Y Yang. Prediction of dna-binding residues from protein sequence information using random forests. *BMC Genomics*, 10(Suppl 1):S1, 2009.

[24] L. Wang and S. J. Brown. Bindn: a web-based tool for efficient prediction of dna and rna binding sites in amino acid sequences. *Nucleic Acids Research*, 34(Web Server):W243–W248, Jul 2006.

[25] Jiansheng Wu, Hongde Liu, Xueye Duan, Yan Ding, Hongtao Wu, Yunfei Bai, and Xiao Sun. Prediction of dna-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics*, 25(1):30–35, Nov 2008.

[26] Matthew B. Carson, Robert Langlois, and Hui Lu. Naps: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Research*, 38, 05 2010.

[27] Wei-Zhong Lin, Jian-An Fang, Xuan Xiao, and Kuo-Chen Chou. idna-prot: Identification of dna binding proteins using random forest with grey model. *PLoS ONE*, 6(9):e24756, Sep 2011.

[28] Bin Liu, Jinghao Xu, Xun Lan, Ruifeng Xu, Jiyun Zhou, Xiaolong Wang, and Kuo-Chen Chou. idna-prot—dis: Identifying dna-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. *PLoS ONE*, 9(9):e106691, Sep 2014.

[29] Mu Gao and Jeffrey Skolnick. A threading-based method for the prediction of dna-binding proteins with application to the human genome. *PLoS Computational Biology*, 5(11):e1000567, Nov 2009.

[30] S. Mahony and P. V. Benos. Stamp: a web tool for exploring dna-binding motif similarities. *Nucleic Acids Research*, 35(Web Server):W253–W258, May 2007.

[31] J. D. Fischer, C. E. Mayer, and J. Söding. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, 24(5):613–620, Jan 2008.

[32] Yanping Zhang, Jun Xu, Wei Zheng, Chen Zhang, Xingye Qiu, Ke Chen, and Jishou Ruan. newdna-prot: Prediction of dna-binding proteins by employing support vector machine and a comprehensive sequence representation. *Computational Biology and Chemistry*, 52:51–59, Oct 2014.

[33] Liangjiang Wang, Caiyan Huang, Mary Qu Yang, and Jack Y Yang. Bindn+ for accurate prediction of dna and rna-binding residues from protein sequence features. *BMC Systems Biology*, 4(S1), May 2010.

[34] Jingna Si, Zengming Zhang, Biaoyang Lin, Michael Schroeder, and Bingding Huang. Metadbsite: a meta approach to improve protein dna-binding sites prediction. *BMC Systems Biology*, 5(Suppl 1):S7, 2011.

[35] Pemra Ozbek, Seren Soner, Burak Erman, and Turkan Haliloglu. Dnabindprot: fluctuation-based predictor of dna-binding residues within a network of interacting residues. *Nucleic Acids Research*, 38, 05 2010.

[36] S. Jones. Using structural motif templates to identify proteins with dna binding function. *Nucleic Acids Research*, 31(11):2811–2823, Jun 2003.

[37] H. P. Shanahan. Identifying dna-binding proteins using structural motifs and the electrostatic potential. *Nucleic Acids Research*, 32(16):4732–4741, Sep 2004.

[38] C. Ferrer-Costa, H. P. Shanahan, S. Jones, and J. M. Thornton. Hthquery: a method for detecting dna-binding proteins with a helix-turn-helix structural motif. *Bioinformatics*, 21(18):3679–3680, Jul 2005.

[39] Mu Gao and Jeffrey Skolnick. Dbd-hunter: a knowledge-based method for the prediction of dna–protein interactions. *Nucleic Acids Research*, 36(12):3978–3992, May 2008.

[40] Mu Gao and Jeffrey Skolnick. From nonspecific dna–protein encounter complexes to the prediction of dna–protein interactions. *PLoS Computational Biology*, 5(3):e1000341, Apr 2009.

[41] Andrea Zen, Cesira de Chiara, Annalisa Pastore, and Cristian Micheletti. Using dynamics-based comparisons to predict nucleic acid binding sites in proteins: an application to ob-fold domains. *Bioinformatics*, 25(15):1876–1883, May 2009.

[42] Yuko Tsuchiya, Kengo Kinoshita, and Haruki Nakamura. Structure-based prediction of dna-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins: Structure, Function, and Bioinformatics*, 55(4):885–894, Apr 2004.

[43] Y. Tsuchiya, K. Kinoshita, and H. Nakamura. Preds: a server for predicting dsdna-binding site on protein molecular surfaces. *Bioinformatics*, 21(8):1721–1723, Dec 2004.

[44] N. Bhardwaj. Kernel-based machine learning protocol for predicting dna-binding proteins. *Nucleic Acids Research*, 33(20):6486–6493, Nov 2005.

[45] Eric W. Stawiski, Lydia M. Gregoret, and Yael Mandel-Gutfreund. Annotating nucleic acid-binding function based on protein structure. *Journal of Molecular Biology*, 326(4):1065–1079, Feb 2003.

[46] S. Jones. Using electrostatic potentials to predict dna-binding sites on dna-binding proteins. *Nucleic Acids Research*, 31(24):7189–7198, Dec 2003.

[47] Matthias Keil, Thomas E. Exner, and Jürgen Brickmann. Pattern recognition strategies for molecular surfaces: Iii. binding site prediction with a neural network. *Journal of Computational Chemistry*, 25(6):779–789, 2004.

[48] Yi Xiong, Junfeng Xia, Wen Zhang, and Juan Liu. Exploiting a reduced set of weighted average features to improve prediction of dna-binding residues from 3d structures. *PLoS ONE*, 6(12):e28440, Dec 2011.

[49] S. Ahmad, M. M. Gromiha, and A. Sarai. Analysis and prediction of dna-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, 20(4):477–486, Jan 2004.

[50] Igor B. Kuznetsov, Zhenkun Gou, Run Li, and Seungwoo Hwang. Using evolutionary and structural information to predict dna-binding sites on dna-binding proteins. *Proteins: Structure, Function, and Bioinformatics*, 64(1):19–27, Mar 2006.

[51] Shandar Ahmad and Akinori Sarai. Moment-based prediction of dna-binding proteins. *Journal of Molecular Biology*, 341(1):65–71, Jul 2004.

[52] András Szilágyi and Jeffrey Skolnick. Efficient prediction of nucleic acid binding function from low-resolution protein structures. *Journal of Molecular Biology*, 358(3):922–933, May 2006.

[53] Munazah Andrabi, Kenji Mizuguchi, Akinori Sarai, and Shandar Ahmad. Prediction of mono- and di-nucleotide-specific dna-binding sites in proteins using neural networks. *BMC Structural Biology*, 9(1):30, 2009.

[54] Yi Xiong, Juan Liu, and Dong-Qing Wei. An accurate feature-based method for identifying dna-binding residues on protein surfaces. *Proteins: Structure, Function, and Bioinformatics*, 79(2):509–517, Nov 2010.

[55] Guy Nimrod, Maya Schushan, András Szilágyi, Christina Leslie, and Nir Ben-Tal. idbps: a web server for the identification of dna binding proteins. *Bioinformatics*, 26(5):692–693, Jan 2010.

[56] Harianto Tjong and Huan-Xiang Zhou. Displar: an accurate method for predicting dna-binding sites on protein surfaces. *Nucleic Acids Research*, 35(5):1465–1477, Feb 2007.

[57] Huiling Chen and Huan-Xiang Zhou. Prediction of interface residues in protein-protein complexes by a consensus neural network method: Test against nmr data. *Proteins: Structure, Function, and Bioinformatics*, 61(1):21–35, Aug 2005.

[58] Huan-Xiang Zhou and Yibing Shan. Prediction of protein interaction sites from sequence profile and residue neighbor list. *Proteins: Structure, Function, and Genetics*, 44(3):336–343, 2001.

[59] S. Altschul. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, Sep 1997.

[60] Guy Nimrod, András Szilágyi, Christina Leslie, and Nir Ben-Tal. Identification of dna-binding proteins using structural, electrostatic and evolutionary features. *Journal of Molecular Biology*, 387(4):1040–1053, Apr 2009.

[61] G. Nimrod, F. Glaser, D. Steinberg, N. Ben-Tal, and T. Pupko. In silico identification of functional regions in proteins. *Bioinformatics*, 21(Suppl 1):i328–i337, Jun 2005.

[62] Guy Nimrod, Maya Schushan, David M. Steinberg, and Nir Ben-Tal. Detection of functionally important regions in "hypothetical proteins" of known structure. *Structure*, 16(12):1755–1763, Dec 2008.

[63] Xiaolei Zhu, Spencer S. Ericksen, and Julie C. Mitchell. Dbsi: Dna-binding site identifier. *Nucleic Acids Research*, 41(16):e160–e160, Jul 2013.

[64] Thorsten Joachims. *Making large scale SVM learning practical.* Oct 1999.

[65] R. P. Joosten, T. A. H. te Beek, E. Krieger, M. L. Hekkelman, R. W. W. Hooft, R. Schneider, C. Sander, and G. Vriend. A series of pdb related databases for everyday needs. *Nucleic Acids Research*, 39(Database):D411–D419, Nov 2010.

[66] Thornton JM Hubbard SJ. misc.

[67] Sunhwan Jo, Miklos Vargyas, Judit Vasko-Szedlar, Benoît Roux, and Wonpil Im. Pbeq-solver for online visualization of electrostatic potential of biomolecules. *Nucleic Acids Research*, 36:W270–W275, May 2008.

[68] Wonpil Im, Dmitrii Beglov, and Benoît Roux. Continuum solvation model: Computation of electrostatic forces from numerical solutions to the poisson-boltzmann equation. *Computer Physics Communications*, 111(1–3):59–75, Jun 1998.

[69] Anthony Nicholls, Kim A. Sharp, and Barry Honig. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Structure, Function, and Genetics*, 11(4):281–296, Dec 1991.

[70] Julie C. Mitchell, Rex Kerr, and Lynn F. Ten Eyck. Rapid atomic density methods for molecular shape characterization. *Journal of Molecular Graphics and Modelling*, 19(3–4):325–330, Jun 2001.

[71] Y. Zhang. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic Acids Research*, 33(7):2302–2309, Apr 2005.

[72] Radoslav Krivák and David Hoksza. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of Cheminformatics*, 10(1), Aug 2018.

[73] Lukas Jendele, Radoslav Krivak, Petr Skoda, Marian Novotny, and David Hoksza. Prankweb: a web server for ligand binding site prediction and visualization. *Nucleic Acids Research*, 47:W345–W349, May 2019.

[74] Radoslav Krivák and David Hoksza. *P2RANK: Knowledge-Based Ligand Binding Site Prediction Using Aggregated Local Features*, page 41–52. Springer International Publishing, 2015.

[75] Radoslav Krivák and David Hoksza. Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features. *Journal of Cheminformatics*, 7(1), Apr 2015.

[76] Modesto Orozco and Francisco J. Luque. Generalization of the molecular electrostatic potential for the study of noncovalent interactions. In Jane S. Murray and Kalidas Sen, editors, *Molecular Electrostatic Potentials*, volume 3 of *Theoretical and Computational Chemistry*, pages 181 – 218. Elsevier, 1996.

[77] A.R. Leach. 4.05 - ligand-based approaches: Core molecular modeling. In John B. Taylor and David J. Triggle, editors, *Comprehensive Medicinal Chemistry II*, pages 87 – 118. Elsevier, Oxford, 2007.

[78] Elizabeth Jurrus, Dave Engel, Keith Star, Kyle Monson, Juan Brandi, Lisa E. Felberg, David H. Brookes, Leighton Wilson, Jiahui Chen, Karina Liles, Minju Chun, Peter Li, David W. Gohara, Todd Dolinsky, Robert Konecny, David R. Koes, Jens Erik Nielsen, Teresa Head-Gordon, Weihua Geng, Robert Krasny, Guo-Wei Wei, Michael J. Holst, J. Andrew McCammon, and Nathan A. Baker. Improvements to the apbs biomolecular solvation software suite. *Protein Science*, 27(1):112–128, 2018.

[79] Chuan Li, Zhe Jia, Arghya Chakravorty, Swagata Pahari, Yunhui Peng, Sankar Basu, Mahesh Koirala, Shailesh Kumar Panday, Marharyta Petukh, Lin Li, and et al. Delphi suite: New developments and review of functionalities. *Journal of Computational Chemistry*, 40(28):2502–2508, Jun 2019.

[80] Lin Li, Chuan Li, Subhra Sarkar, Jie Zhang, Shawn Witham, Zhe Zhang, Lin Wang, Nicholas Smith, Marharyta Petukh, and Emil Alexov. Delphi: a comprehensive suite for delphi software and associated resources. *BMC Biophysics*, 5(1):9, 2012.

List of Figures

2.1 The figure shows the base units (nucleotides) of DNA: thymine, adenine, guanine, and cytosine. On the right is part of the B-DNA molecule, and under them is a representation of standard Watson-Crick complementary base pairing. 4

2.2 The figure shows the histone protein-DNA complex (1AOI) and histone H4 primary, secondary, tertiary structure, and its formed dimer and tetramer with histone H3. 6

2.3 The figure shows typical distortions and deformations, which occur in DNA molecules, and which are recognized by proteins. 8

4.1 Electrostatic potential maps for ligand-binding structure (2SRC) and DNA-binding structure (4O5K) and their bonded forms with ligand and DNA show a difference in electrostatic potential of binding pockets. 18

4.2 The figure shows protein Pax-6 (structure 6PAX) with bonded DNA. The structures under 6PAX present predicted residues by P2Rank and DSBI. P2Rank_ligand is the default model for ligand-binding classification. Used DSBI is a web server tool of the previously described method. P2Rank_DNA was trained on newly created datasets using a bind radius of 5Å. 25

4.3 The figure shows on the structure 6PAX residues defined as binding for different radii by color scale. Under scale is a set of structures representing predictions for each newly trained model. 26

List of Tables

3.1 The cross-training results for the three-tier and the two-tier divisions of the dataset using DISPLAR. 12
3.2 The cross-training results of the iDBS using 10-fold cross-validation on the dataset of 138 bindings and 110 non-binding protein chains. 14
3.3 The table shows the cross-validation results of the model for datasets used by DBSI. 16

4.1 The number of residues labeled binding and non-binding in the dataset for different radius. 22
4.2 The table shows the evaluation of our test dataset on the P2Rank_ligand model and the all trained models P2Rank_DNA for different bind radius. 24
4.3 The table shows 10-fold cross-validation on the dataset used for training DISPLAR and DBSI. We run cross-validation on P2Rank with the same dataset. The model used for cross-validation is P2Rank_DNA with a 5Å bind radius, which corresponds to the radius used for DISPLAR and DBSI. Underline is cross-validation of iDBS, which uses a different dataset. 25