University of Alberta

Sequence-based prediction and characterization of disorder-to-order transitioning binding sites in proteins

by

Fatemeh Miri Disfani

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in

Software Engineering and Intelligent Systems

Department of Electrical and Computer Engineering

Spring 2012
Edmonton, Alberta

Abstract

Molecular Recognition Feature (MoRF) regions are disordered binding sites that become structured upon binding. MoRFs are implicated in important biological processes, including signaling and regulation. However, only a limited number of experimentally validated MoRFs is known, which motivates development of computational methods that predict MoRFs from protein chains.

We introduce a new MoRF predictor, MoRFpred, which identifies all MoRF types (α, β, coil, and complex). We develop a comprehensive dataset of annotated MoRFs and use it to build and empirically compare our method. Empirical evaluation shows that MoRFpred statistically significantly outperforms existing predictors by 0.07 in AUC and 10% in success rate. We show that our predicted MoRF regions have non-random sequence similarity with native MoRFs. We use this observation along with the fact that predictions with higher probability are more accurate to identify putative MoRF regions. We present case studies to analyze these putative MoRFs.

Table of Contents

List of Tables

List of Figures

Abbreviations

Abbreviations	Definition
AA	Amino Acid
AUC	Area Under the Curve
DisProt	Database of Disordered Proteins
FPR	False Positive Rate
IDP	Intrinsically Disordered Protein
MoRF	Molecular Recognition Features
NMR	Nuclear Magnetic Resonance spectroscopy
PDB	Protein Data Bank
PSSM	Position Specific Scoring Matrix
ROC	Receiver Operating Characteristic
RSA	Relative Solvent Accessibility
SVM	Support Vector Machines
TPR	True Positive Rate

1 Introduction

Protein is a polymer consisting of several dozens to thousands of amino acids. Proteins, which are among the most important biological molecules, implement a diverse range of functions. They serve as structural elements that form muscles, bones and nails, as enzymes that catalyze chemical reactions, and as hormones that trigger biological events, to name just a few of their functions. Originally it was thought that function of a protein arises from its three-dimensional (3D) shape which has a well-defined structure that is encoded in its amino acid sequence (Anfinsen, 1973). This paradigm was challenged by the discovery of unstructured proteins which lack a rigid 3D structure and are functional in their extended form (Wright & Dyson, 1999). Since the discovery of intrinsically unstructured/disordered proteins (IDP) a couple of decades ago, evidence of their participation in a variety of biological processes have been found. Many studies agree on the role of disordered proteins in processes such as transcription, transcription regulation, and signal transduction (Dunker & Kriwacki, 2011). Disordered proteins are also implicated in several diseases such as cancer, neurodegeneration and cardiovascular diseases (Midic et al., 2009; Uversky et al., 2009). The above motivates further studies on IDPs.

To have a better understanding of how IDPs function, we need to understand the nature of their interactions with other molecules. Generally speaking, proteins implement their functions through interaction/binding with other biological molecules such as other proteins, nucleic acids, and smaller ligands such as nucleotides, metals, etc. These interactions happen in specific regions of protein structure called binding sites. The prevalent paradigm says that all binding sites have a well-defined structure and would only bind to a ligand that complements their structure; by analogy, by fitting of a key (ligand) into a lock (protein) to unlock its function. However, observation of IDPs shows that some of the disordered regions also act as binding sites by going through a disorder to order transition upon binding (coupled folding and binding) to adapt their shape to the ligand's binding site (Cheng et al., 2007); by analogy, by adjusting the lock mechanism/shape to fit certain keys. Flexibility of IDPs, which is the hallmark of the disorder, gains them an advantage over their globular counterparts by allowing them to have highly specific but reversible binding and by allowing diversity in binding. This unique characteristic of disordered binding sites is especially useful in signaling and regulation processes (Oldfield et al., 2005).

To study disorder and disordered binding sites, we need a dataset of known disordered proteins with the annotated native (experimentally validated) disorder. However, due to a relatively slow progress in experimental determination of disorder, only a small number of proteins with annotated disorder is known. The DispProt database (Sickmeier et al, 2007), which is by far the largest database concerning protein disorder, contains about 600 annotated proteins. This is a small number compared to the much larger number of known protein sequences (in millions) that remain unannotated, which results in a wide annotation gap. Abundance of known protein sequences motivated the development of computational methods that predict disorder from sequences. These methods are a viable alternative to reduce the annotation gap and to investigate the disorder (He et al., 2009). So far several dozens methods have been developed to predict protein disorder from sequence. The best of these methods are able to predict disorder relatively accurately and their predictive performance is rising (Uverski & Dunker, 2010; Monastyrskyy et al., 2011).

In spite of the progress in prediction of disorder, computational determination of disordered binding sites did not attract as much attention. The first step towards developing such prediction method was to identify sequence and structural characteristics of these binding sites. Oldfield *et al.* characterized a specific structural element which mediates many of disordered binding events (Oldfield et al., 2005). This short region in the protein sequence (formed by 5 to 25 consecutive amino acids), which undergoes coupled binding and folding, is referred to as Molecular Recognition Feature (MoRF) and is flanked by (regular/non-binding) disordered segments. MoRF regions are classified based on the secondary structure they take upon binding to their ligand. Secondary structure refers to local three-dimensional conformations of amino acid segments in the protein chain which are established between adjacent amino acids and are categorized into three major states: helices, sheets, and regions with irregular secondary structure. Based on this categorization, MoRFs can be divided to three subtypes: α-MoRFs which take the shape of a helix, β-MoRFs which take the form of sheets, and MoRFs that do not have a regular secondary structure.

To date, three methods have been developed to predict MoRF regions from protein sequence. The first two prediction methods, α-Morf-PredI (Oldfield et al., 2005) and α-Morf-PredII (Cheng et al., 2007), were designed to predict only α-

MoRFs. The third approach, ANCHOR (Dosztányi et al., 2009), predicts MoRF regions regardless of their subtype. These methods were trained on a relatively small datasets of annotated MoRF regions (14 regions from 12 proteins) (Cheng et al., 2007). While these approaches succeed in identifying some MoRF regions, there is a pressing need for new solutions. First of all, predictions of the two alpha MoRF methods are limited to only one MoRF subtype, while this subtype covers minority of the MoRFs. Secondly, ANCHOR, which is designed to predict all MoRF subtypes, was trained on a small dataset and its predictive quality is relatively poor.

Our hypothesis is that it is possible to build a new MoRF predictor that uses protein sequence as input and which outperforms the current methods in prediction of disordered binding sites. We devise and evaluate a new solution to address this hypothesis. Our approach consists of two steps. First, the protein sequence is converted into a numerical feature vector that represents different attributes/characteristics of the input protein. Next, this vector is inputted into a machine learning classifier that generates predictions. Three novel design and development aspects contributed to the improvements that were achieved by our solution. First, we use a larger and more comprehensive dataset of annotated MoRF regions to design our method and to empirically compare it with the existing solutions. Second, we use a more comprehensive set of features which encode previously unexplored characteristics of the protein chain and we utilize a modern and well-performing machine learning model called Support Vector Machine. Third, we extend our design by combining the machine learning-based predictions with predictions generated using sequence alignment, which exploits similarity of the predicted sequence to a dataset of annotated protein sequences.

The thesis is organized as follows. In Chapter 2 we introduce background information concerning protein structure and disorder, prediction of disorder, datasets, and evaluation protocols that are used to assess predictors. Chapter 3 explains the design of our predictor, which is evaluated and compared with the existing solutions in Chapter 4. The latter chapter also analyzes our predictions and predictive model to provide interesting insights that characterize MoRF regions and that help to interpret the predictions. Chapter 5 concludes the thesis and lists significant contributions.

2 Background and Materials

This chapter provides preliminary information required to understand the nature of protein function and interactions. We discuss certain characteristics of disordered regions and how they allow for development of computational methods that predict disordered regions. We introduce some of the top disorder prediction methods and their application in identification of disordered binding sites. Finally, we overview the available predictors of disordered binding sites and we explain protocols that are used to assess and compare predictive quality of these methods.

2.1 Definitions

2.1.1 Amino acids

Proteins are biological polymers that consist of several dozen to several thousands of amino acids (AAs). It is believed that the structure of the protein is determined by its sequence of AAs (Anfinsen 1973). There are 20 standard AA types as shown in Table 1. An amino acid is defined as $H2N\text{-}C_{\alpha}H(R)\text{-}COOH$ where the amino group (H2N) in one AA connects to the carboxyl group (COOH) of its adjacent AA to form a sequence. Together, the amino group, carboxyl group and the C_{α} atom that connects to the side chain (R) create the backbone of the protein structure, i.e., a chain that folds in space to form the 3D shape of this molecule. R is the side chain of a given AA that protrudes from the backbone and determines the type of the AA. Physiochemical differences in the side chains are responsible for different properties of individual AAs and provide a comprehensive set of building blocks to assemble unique protein structures.

AAs are attached to each other through a peptide bond. A sequence of connected AAs is called a polypeptide. Attachment of an AA to the polypeptide always happens in 1 direction where the amino group of the unattached amino acid attaches to the carboxyl group of the polypeptide. The amino group of the first amino acid is called the N terminus, and the carboxyl group of the last amino acid is called the C terminus.

Distribution of charge in an AA side chain is a contributing factor in AA interactions. Some of the AAs have positively or negatively charged side chains. We refer to this group as charged AAs. In contrast, neutral AAs lack electric charge. The latter group can be divided to polar and non-polar categories based on the distribution of positive and negative charges inside the molecule. While polar AAs have an imbalanced distribution of charge, in non-polar amino acids charges are distributed evenly. Since charged AAs have imbalanced distribution of charges, they also belong to the polar group. Another contributing factor in AAs interactions is the hydrogen bond. As we will explain in the next section, hydrogen bonds play an important role in shaping the protein structure. A hydrogen bond occurs between an electronegative atom and a hydrogen atom that is bonded to another electronegative atom. In the case of AAs this bond occurs between the hydrogen of the H-N group and the oxygen in the C=O.

AAs can be categorized to 2 groups based on their tendency to interact with each other and with water molecules.

- Hydrophilic AAs are capable of attracting water molecules. Side chains of these AAs are polar. Due to the polarity of their side chains these AAs are capable of making hydrogen bonds with each other and with water molecules. Usually hydrogen bonds dominate the interactions of this type of AAs.

- Hydrophobic AAs are nonpolar and water "fearing". These molecules repulse/escape from water molecules and tend to pack against each other in presence of water. In protein folding (i.e., in the process that converts a polypeptide chain into a 3D protein molecule) this tendency, which is referred to as hydrophobic effect, is partly responsible for the globular shape of structured proteins (for more details refer to section 2.1.2.3).

The categorization of AAs in hydrophobic and hydrophilic is shown in Table 1. We note that various hydrophobicity scales are devised to quantify hydrophobic/hydrophilic tendencies of AAs (Carolina, 1971), so this categorization is somehow "flexible" for certain, borderline AAs.

AAs can also be characterized based on other physical and chemical attributes. A comprehensive list of amino acid indexes which quantifies biophysical and biochemical attributes of AAs such as their size, volume, acidity, charge, polarity, etc., is available in the AAindex database at http://www.genome.jp/aaindex/

(Kawashima & Kanehisa, 2000). In section 3.2 we explain how this information can be used to build a feature set for the prediction of disordered binding sites.

2.1.2 Protein structure

Protein structure is the three dimensional conformation of AAs in a protein and can be best described in the atomic level by a set of coordinates specifying position of each atom in each AA in the structure. Though highly informative, the atomic representation of a protein does not allow for characterization and classification of protein molecules. Thus, protein structure is categorized in three hierarchical levels: primary, secondary, and tertiary structures.

Amino Acid	Short	Abbrev.	Side chain	Hydrophobic	Polar	Charge
Alanine	A	Ala	-CH3	Hydrophobic	Nonpolar	Neutral
Cysteine	C	Cys	-CH2SH	Hydrophobic	Polar	Neutral
Aspartic acid	D	Asp	-CH2COOH	Hydrophilic	Polar	Negative
Glutamic acid	E	Glu	-CH2CH2COOH	Hydrophilic	Polar	Negative
Phenylalanine	F	Phe	-CH2C6H5	Hydrophobic	Nonpolar	Neutral
Glycine	G	Gly	-H	Hydrophobic	Nonpolar	Neutral
Histidine	H	His	-CH2-C3H3N2	Hydrophilic	Polar	Negative
Isoleucine	I	Ile	-CH(CH3)CH2CH3	Hydrophobic	Nonpolar	Neutral
Lysine	K	Lys	-(CH2)4NH2	Hydrophilic	Polar	Positive
Leucine	L	Leu	-CH2CH(CH3)2	Hydrophobic	Nonpolar	Neutral
Methionine	M	Met	-CH2CH2SCH3	Hydrophobic	Nonpolar	Neutral
Asparagine	N	Asn	-CH2CONH2	Hydrophilic	Polar	Neutral
Proline	P	Pro	-CH2CH2CH2-	Hydrophobic	Nonpolar	Neutral
Glutamine	Q	Gln	-CH2CH2CONH2	Hydrophilic	Polar	Neutral
Arginine	R	Arg	-(CH2)3NH-C(NH)NH2	Hydrophilic	Polar	Positive
Serine	S	Ser	-CH2OH	Hydrophilic	Polar	Neutral
Threonine	T	Thr	-CH(OH)CH3	Hydrophilic	Polar	Neutral
Valine	V	Val	-CH(CH3)2	Hydrophobic	Nonpolar	Neutral
Tryptophan	W	Trp	-CH2C8H6N	Hydrophobic	Nonpolar	Neutral
Tyrosine	Y	Tyr	-CH2-C6H4OH	Hydrophilic	Polar	Neutral

Table 1. The names, abbreviations, and side chain formulas for the 20 AAs. The last two columns indicate hydrophobic and hydrophilic classes of AAs. This table was borrowed from http://en.wikipedia.org/wiki/Proteinogenic_amino_acid#Side_chain_properties.

2.1.2.1 Primary structure

The sequence of AAs in a polypeptide chain is referred to as the primary structure. Primary structure shows the order of AAs in the chain and is written and read in the direction that a given protein is created: from N-terminus to C-terminus. The primary structure of a protein determines how this protein folds into the secondary and tertiary structures.

2.1.2.2 Secondary structure

Recurring patterns of local three dimensional conformations in a protein chain are referred to as secondary structure. They are established through hydrogen bonds between N-H and C=O groups of adjacent (in space) AAs. Two common forms of secondary structures are α-helices and β-sheets.

An α-helix is a cylindrical structure made by a peptide chain where the carbonyl oxygen atom of each AA forms a hydrogen bond with the amide nitrogen four residues (we use AA and residue as synonyms) further along in the sequence. The backbones of the AAs form the wall of the cylinder where the side chains are protruded from the structure. Side chains are determining factors of the interactions that occur between the helix and the other parts of the protein. Figure 1a depicts the structure of a helix.

A β-sheet is a structure consisting of two or more strands of AAs that are taking extended conformations and are bonded to each other through backbone hydrogen bonds. If the two bonded strands are oriented in the same direction they are called parallel sheets. In contrast, strands that are in opposite directions form antiparallel sheets. Figure 1b depicts the conformation of parallel and antiparallel β-sheets.

While these two types of secondary structure are the most common, other forms of secondary structure such as other forms of helixes and strands (called bridges) has also been observed. These structures will be discussed in section 2.1.3.

Segments of the protein which lack a regular secondary structure are referred to as coils, and they include turns, bends, and other less regular shapes. These coils serve as “connectors” between helices and strands.

2.1.2.3 Tertiary structure

Tertiary structure refers to the three-dimensional (3D) conformation of a protein, where secondary structure elements fold into a compact globular molecule, see Figure 2. This fold is the most energetically stable state of the molecule and (often) the only functional conformation. This structure is stabilized by weak intermolecular interactions between polar and nonpolar groups. The structured state of the protein is referred to as the native state and the process of reaching this state is called protein folding. Generally, when the term “protein structure” is used, it refers to the tertiary structure.

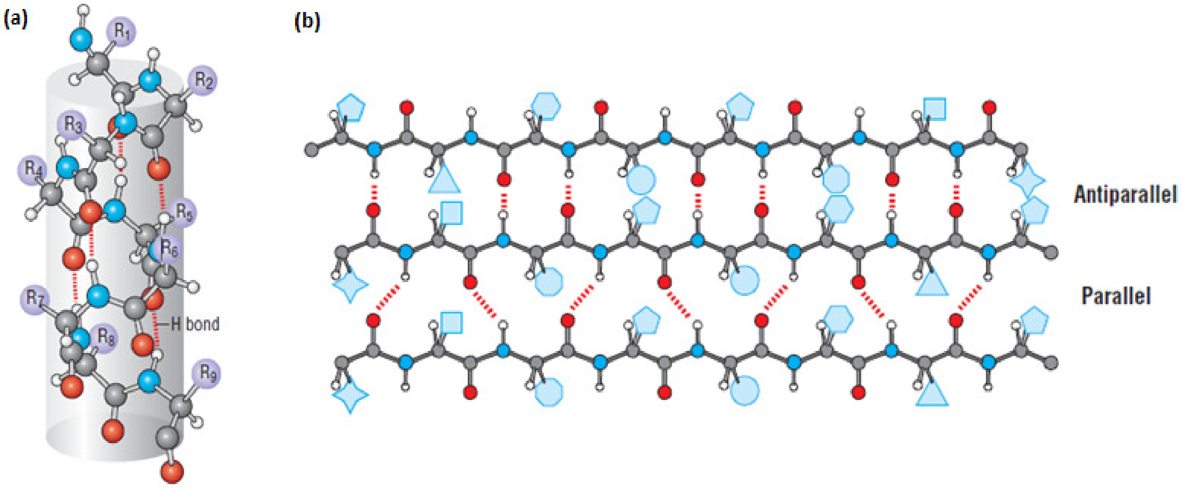

Figure 1. α-helix (panel a) and β-sheet (panel b) structures. The carbon atoms are depicted by gray circles, oxygen atoms by red circles and nitrogen atoms by blue circles. Hydrogen atoms are depicted by small white circles and hydrogen bonds are shown by red dashed lines. (This figure was taken from the book “Protein structure and function”, Petsko, G.A. & Ringe, D., 2004, Sinauer Associates).

A protein is a combination of different AAs with polar and non-polar side chains. The order and composition of AAs differ from one protein to another. Here we discuss the effects of presence of different AA types in protein folding. As we mentioned earlier, polar and charged AAs tend to make hydrogen bonds with water molecules in their surrounding area. This is what happens to polar residues of a protein in an extended unstructured form. However, the presence of nonpolar residues that cannot form hydrogen bonds would disrupt the network of hydrogen bonded water molecules in the solution. This disruption is energetically unfavorable for the protein and would cause nonpolar residues to escape from water molecules and aggregate together (the hydrophobic effect). Finally the tendency of polar residues to make bonds with water molecules and non-polar residues to cluster together would drive the protein to fold into a compact globular structure.

Protein structure is described in terms of atomic coordinates and can be determined using different methods. Among these methods, the X-ray crystallography is the most common way to determine the structure; this method can provide high resolution information. It uses protein crystals to capture the well-structured portions of proteins and provides information about flexibility of individual AAs. The Nuclear magnetic resonance spectroscopy (NMR) is the other common way to find protein structure. This approach can capture changes of protein structure on a time scale. This method works in (more) native environment in solution (in contrast to a non-native crystallized state), but it can only be used for relatively small proteins and provides a lower resolution.

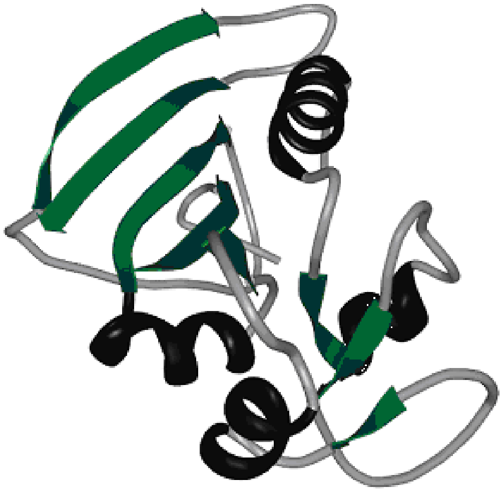

Figure 2. Tertiary structure of DHFR which is an important enzyme in nucleotide metabolism. The structure is shown in cartoon format where helices are in black, strands in green (dark gray), and coils in light gray.

2.1.2.4 Sequence alignment

Sequence alignment arranges two or more sequences against each other to find similar regions/stretches of AAs. Sequence similarity can be used to identify structural and functional similarity across a diverse set of protein sequences. The alignment is a valuable tool to find structural and functional annotations of a protein sequence that lack annotations by inferring them from the annotated regions on the similar sequences. This method works best for unannotated sequences for which we can find similar sequences (usually with at least 30% similarity) with known structure and function. In some cases only parts of the sequence are aligned (matched) and this could be sufficient to transfer the annotations.

Alignment algorithms can be categorized into two groups: local and global, see Figure 3. Global alignment tries to align every residue in both sequences and is useful when the sequences are highly similar and of nearly the same size/length.

Local sequence alignment approaches are more suitable for sequences which have similar regions but are not entirely similar and may have different size.

Multiple sequence alignment is an approach to align more than two sequences at the same time. Multiple sequence alignment is computationally expensive. These methods usually use a heuristics approach rather than the global optimization to find approximately best alignment in order to reduce the computational cost.

Global	Input sequence	FTFTALILLAVAV
	Aligned sequence	F––TAL–LLA–AV
Local	Input sequence	FTFTALILLA–VAV
	Aligned sequence	––FTAL–LLAAV––

Figure 3. Results of global and local alignment when aligning FTFTALILLAVAV sequence to FTALLLAAV sequence. In both cases inserting gap in the alignment is allowed. Gaps are empty spaces inserted between sequences to allow for a better alignment and are represented by –.

2.1.3 **1D descriptors**

The last few decades observed development of a number of lower-level descriptors of protein structure that provide an alternative and somehow complementary way to describe, analyze, and predict protein structure and function when compared with the structure defined using atomic coordinates. These descriptors quantify certain structural properties of AAs, such as secondary structure, their position with respect to the protein surface, and their flexibility. We refer to these descriptors as 1-dimensional (1D) descriptors since they project 3D structural features onto 1D strings of residue-wise structural assignments. Among available descriptors, in this section we discuss secondary structure, solvent accessibility and flexibility descriptors, which are utilized to design our predictor of disordered binding sites.

2.1.3.1 Secondary structure

Secondary structures are determined based on the patterns of hydrogen bonds in protein structure and they are categorized into three major states: helices, sheets, and regions with irregular secondary structure. The DSSP method (Kabsch and Sander, 1983) assigns one of the following eight secondary structure types for each of the structured residues (residues that have three-dimensional coordinates) in the protein sequence:

- G: 3-turn helix (also referred to as 3_{10} helix). In this secondary structure the carboxyl group of a given AA forms a hydrogen bond with amid group of the AA three positions down in the sequence forming a tight, right-handed helical structure with three residues per turn.
- H: 4-turn helix (also referred to as α-helix). This structure is similar to the 3-turn helix, however, the hydrogen bonds are formed between consecutive AAs that are four positions away in the protein chain. This is the most prevalent helix type.
- I: 5-turn helix (also referred to as π-helix). In this type of the helix the hydrogen bonding occurs between residues spaced five positions away from each other and which also results in a right-handed helical structure; left-handed π-helices are relatively rare.
- E: extended strand in parallel or anti-parallel sheet conformation. Two or more strands are connected laterally by at least two hydrogen bonds forming a pleated sheet.

- B: residue in an isolated beta-bridge, which is a single residue pair sheet formed based on the hydrogen bond.
- T: hydrogen bonded turn. A turn in the protein chain in which a single hydrogen bond is formed between residues spaced 3, 4, or 5 positions away in the protein chain.
- S: bend, which denotes a fragment of protein chain with high curvature where the angle between the vector from $C_{\alpha i}$ to $C_{\alpha i+2}$ (C_{α} atoms at i^{th} and $i+2^{th}$ positions) and the vector from $C_{\alpha i-2}$ to $C_{\alpha i}$ is < 70°; this is the only non-hydrogen bond-based regular secondary structure type.
- – : irregular secondary structure (also referred to as loops and random coils), which corresponds to the remaining conformations.

The above eight types are often mapped into three states as follows

- H: α-helix. This secondary structure state encompasses right or left handed cylindrical/helical conformations that include G, H, and I types.
- E: β-strand. This state corresponds to pleated sheet structures and it includes E and B secondary structure types.
- C: coil. This state represent the remaining types of the local confirmations and it includes S, T, and – types.

2.1.3.2 Solvent accessibility

Solvent-accessible area of a protein molecule was first defined by Lee and Richards in early 1970s (Lee and Richards, 1971) as the area traced out by the center of a virtual probe sphere representing a solvent molecule as it is rolled over the protein surface. In the follow up definition (Richards, 1977), the solvent-accessible area consists of the part of the Van der Waals surface of the atoms that are accessible to the probe sphere. The accessible surfaces of atoms are connected to each other by a network of concave and saddle-shaped surfaces that smoothes over the crevices and pits between the atoms. The 1D descriptor of the solvent accessibility (also referred to as the relative solvent accessibility) is defined as the ratio between the solvent exposed surface area of a given residue observed in a given protein structure (i.e., the corresponding part of the solvent-accessible area of this protein) and the maximum obtainable value of the solvent-exposed surface area for this AA (Adamczak et al., 2004). The ratio is used to normalize between different AA types. The values for the accessible surface area are often calculated using the DSSP program. The maximum obtainable values of the solvent exposed surface area correspond to the surface exposed area of a given residue type observed in an extended tripeptide conformation flanked with either glycine or alanine residues. The relative solvent accessibility ranges between 0%, for fully buried residues, and 100%, for fully solvent accessible residues.

2.1.3.3 Flexibility

The B-factor (also called temperature-factor or Debye-Waller factor) describes the degree to which the electron density of a given atom (or a group of atoms) in the X-ray scattering of the crystal structure of a protein is spread out. The B-factor values quantify mobility of an atom and they are computed as

$$\mathrm{B}_{\mathrm{factor}} = 8\pi^2\mathrm{U}_\mathrm{i}^2$$

where U_i^2 is the mean square displacement of the i^{th} atom which is averaged over the lattice. Since B-factors depend on several characteristics of the structure determination protocol, such as experimental resolution, crystal contacts, and refinement procedures, they should be normalized to allow comparisons between different structures. Following (Parthasarathy and Murthy, 1997), the B-factors of a given AA are expressed using the B-factors of C_α atoms that are normalized using average and standard deviation of the B-factors in a given chain as follows

$$\text{normalized\_B}_{\text{factor}} = (\text{B}_{\text{factor}} - \text{mean\_B}_{\text{factor}}) \,/\, \text{standard\_deviation\_of\_B}_{\text{factor}}$$

The values of the abovementioned 1D structural descriptors can be either computed from the known protein structures or predicted from the knowledge of the input protein sequence. An overview of the existing sequence-based predictors of the 1D structural descriptors that compares selected secondary structure, disorder, and solvent accessibility predictors can be found in (Kurgan & Disfani, 2011).

2.2 Protein disorder and its prediction

Disorder in a protein is characterized by lack of a well-defined 3D structure in parts or all of the protein. Disordered regions are flexible polypeptides which do not establish a stable conformation and can fluctuate between different conformations. Proteins which include disordered segments are called intrinsically disordered proteins (IDP). From an experimental point of view, a disordered region is defined as residues that lack coordinates in structures solved by X-ray crystallography and as residues that exhibit high variability within structure ensemble or are annotated as disordered in REMARK 465 by experimentalists for the structures solved by NMR (Kurgan & Disfani, 2011).

Studies reveal that about 40% of all human proteins contain at least one intrinsically disordered segment of 30 AAs or more, and that some 25% are likely to be disordered from beginning to end (Uverski & Dunker, 2010). Disorder has been observed to be involved in a variety of biological processes, such as protein-RNA and protein-DNA binding, transcription, translation, regulation, and signaling. Disorder also plays a part in processes associated with certain diseases such as cancer, neurodegenerative diseases, and cardiac disorders (Midic et al., 2009; Uversky et al., 2009).

As we mentioned earlier, the structure of a protein is determined by its AA sequence. This is also true for the disordered proteins. Studying sequences of the disordered proteins reveals specific features/characteristics of these sequences. Disordered sequences have low complexity (they are built from a less diverse set of AA types when compared with ordered chains) and are characterized by high net charge due to inclusion of polar AAs and low content of hydrophobic AAs. Both of these characteristics are contributing factors in disorder. Firstly, high net charge in a sequence leads to same charge-charge repulsion. Secondly, due to scarcity of hydrophobic residues one of the driving forces of protein folding, the hydrophobic effect, is weakened. As we discuss section 2.2.2, these characteristics can be used in prediction of the disordered segments.

2.2.1 **Disordered binding sites and MoRF regions**

Observation of disordered proteins defies the classical structure-to-function paradigm. This paradigm states that a unique 3D conformation of a given protein determines its interactions with other molecules. While this paradigm is true for many (majority) proteins, disordered proteins can also interact without having a defined structure in isolation (before the interaction occurs).

There are several examples of protein–protein and protein–nucleic acid interactions that involve coupled folding and binding (Uverski & Dunker, 2010). The significance of these interactions is due to two factors: (1) these interactions can be very specific due to the flexibility of the disordered binding sites; and (2) the transition from disorder to order results in a substantial loss in the system entropy, which in turn affects the binding strength. These two factors are especially beneficial in signaling and regulation where highly specific yet dispensable/weak interactions are needed.

Efforts to characterize the disordered binding sites resulted in identification of a specific structural element which mediates many of disordered binding events (Oldfield et al., 2005). This short region (which includes between 5 and 25 AAs) is referred to as Molecular Recognition Feature (MoRF) and is placed between two segments of disorder. MoRF regions can be categorized into three types based on the secondary structure they take upon binding to their ligand: α-MoRFs which take the shape of a helix, β-MoRFs which take the form of sheets, and MoRFs that do not have a regular secondary structure.

Since disordered regions have an extended structure as opposed to globular shape of structured proteins, disordered binding sites are relatively easy to locate in the primary structure, i.e., they usually form long stretches of consecutive AAs. These regions are observed to be enriched in hydrophobic residues compared to general disordered regions (Meszaros et al. 2009). Identifying these characteristics helped researchers build computational methods to find MoRF residues in sequences. More details on these prediction methods are provided in the next section.

2.2.2 Prediction of disordered and MoRF regions

Identification of disorder as one of the defining attributes of MoRF regions is essential to prediction of MoRFs. Therefore, in this section we briefly discuss disorder prediction methods and then we provide details about MoRF prediction methods.

Several studies have shown that disordered regions are characterized by relatively unique sequence signatures. As mentioned in the previous section, they often have a low content of bulky hydrophobic AAs and a high proportion of polar and charged AAs, a low content of (predicted) secondary structure, low complexity, and unique evolutionary and solvent accessibility profiles (Mizianty et al., 2010). This implies that disorder is predictable from the protein sequence. The development of computational predictors was further motivated by the fact that the disorder prediction was introduced into the Critical Assessment of protein Structure Prediction CASP experiments since 2002 (Monastyrskyy et al., 2011).

Predictors of disorder can be categorized into 4 groups based on their design (Mizianty et al., 2010). For each group we introduce a (representative) method that we used in prediction of MoRF regions:

1. *propensity-based methods* based on relative propensity of AAs to form disorder/ordered regions: IUPred (Dosztanyi et al. 2005).

2. *machine learning-based* predictors that use machine learning classifiers to perform predictions: DISOPRED2(Ward et al., 2004).

3. *consensus-based* methods that combine predictions from multiple disorder predictors. MFDp (Mizianty et al. 2010).

4. *structural models-based* approaches that make use of predicted tertiary structure models. DISOCLUST (McGuffin, 2008).

More detailed discussion of the disorder predictors can be found in (He et al., 2009)

The prediction of MoRF regions enjoys less interest, likely because it was initiated recently and is more challenging. Currently, only two predictors are available: α-MoRF-PredII (Cheng et al., 2007), which supersedes α-MoRF-PredI, and ANCHOR (Dosztányi et al., 2009).

α-MoRF-PredII is a neural network based predictor that uses disorder predictions, secondary structure predictions, and amino acid indices as its input attributes. Output of this method is a binary number specifying whether a residue is a MoRF residue (1) or a non Morf residue (0). This method concentrates on predicting α-MoRFs that forms helical structures upon binding, which limits its applications.

ANCHOR is developed to predict all classes of MoRFs. This method uses three quantities to identify a MoRF region: 1) tendency of a residue to be disordered; 2) tendency of a residue to interact with its neighbors and form structure; and 3) tendency of a residue to form favorable interactions with other globular proteins.

Using these three parameters this predictor calculates a score indicating the probability of a residue to be in a disordered binding site. This score is a real value in [0, 1] interval. To binarize the scores, residues with scores above 0.5 are considered to be in a disordered binding site. ANCHOR's predictions are characterized by a relatively weak predictive performance, which is supported by the empirical tests presented in this thesis.

2.3 Materials and Methods

2.3.1 Databases

Several online resources are available to researchers to gather information on structural and functional aspects of a protein. One of the comprehensive protein related databases is UniProt (Jain, 2009) which contains sequence information, functional annotations, and cross references to other protein related tools and databases. We use this database to extract protein sequences that are used to build a dataset to develop and evaluate our predictor.

Another valuable resource is Protein Data Bank (PDB http://www.wwpdb.org/) (Berman et al., 2007) which provides structural information for ordered/structured proteins. Each structure in PDB is represented by an ID number and contains the atomic coordinates of a protein. PDB files are used to assign secondary structures and solvent accessibility of a protein using programs such as DSSP. Segments with missing coordinates in a PDB structure is identified as disordered segments. This database served as a source to find MoRF segments.

The Database of Protein Disorder (DisProt) (Sickmeier et al, 2007) is a manually curated database of intrinsically disordered proteins containing disorder annotation of more than 600 proteins with about 1300 disordered segments. The experimental data is acquired mostly from the missing coordinates in X-ray crystallography derived structures or chemical shifts generated with the NMR. Some of the sequences in this database include functional and structural annotations for the disordered segments.

2.3.2 **Datasets**

To prepare the dataset that is used to design and validate our method, we first collected 4289 protein complexes (structures that include protein interacting with a ligand) from PDB on Mar 28, 2008. These complexes concern interaction between a protein and a small peptide (i.e., a short AA chain). This peptide is a putative MoRF (putative disordered binding site) whose sequence is between 5 and 25 residues. This size is consistent with the related works that developed MoRF predictors (Oldfield et al., 2005; Cheng et al., 2007; Dosztányi et al., 2009). Next, we remove complexes for which the interaction between the two AA chains is not significant enough to be considered as biologically relevant. We measure whether a biologically relevant interaction occurs by calculating the change of accessible surface area (ΔASA) between unbound (two separate chains) and bound (a complex with two chains) states. We utilize the BALL library (http://www.bioinf.uni-sb.de/OK/BALL/) to calculate ΔASA, and we considered an interaction as spurious if its $\Delta ASA < 400$ Å^2 (Jones & Thornton, 1996; Vacic et al., 2007). The cutoff is intended to be small enough to catch small interfaces between the two chains and, at the same time, large enough to remove spurious contacts. As a result, 452 complexes were removed. Of the remaining complexes, 3148 that include globular partners with > 70 AAs (Jones, 1998) were kept. The cutoff at 70 AAs was chosen to avoid discarding shorter folded domains. The remaining MoRFs were mapped to the UniProtKB/Swiss-Prot v56.8 and UniProtKB/TrEMBL v39.8 databases using FASTA algorithm (http://fasta.bioch.virginia.edu/fasta_www2/fasta_down.shtml) with e-value set as 1000 (Pearson, 1988). 1805 MoRF segments were successfully mapped to their parent sequences; in the remaining cases the MoRFs were too short to uniquely map to the UniProt or could not be found. 842 MoRFs were left after removing duplicates and MoRFs that include ambiguous AAs, such as X. We evaluated whether the 842 MoRFs are disordered when unbound. The analysis based on the protocol described in (Gunasekaran et al., 2004) shows that all MoRFs are disordered in isolation, see Figure 4. The AAs that form these MoRFs were annotated in the parent sequences, and these sequences were used to develop and assess our predictor. As a result, each of the 842 sequence in our dataset is annotated with one MoRF segment, which length varies between 5 and 25 AAs.

Each MoRF was classified into one of the four types: α (helix), β (strand), γ (coil), or complex based on the largest percentage value of their secondary structure types assigned by DSSP (Kabsch & Sander, 1983). If for a given MoRF there was no clear preponderance of any secondary structure type (at least 1% greater than the other two types), we categorized it as a complex MoRF. Only the residues in the interface were counted in the secondary structure classification. Among the

842 MoRFs, there are 181 helical, 34 strand, 595 coil and 28 complex MoRF regions and two annotations with unspecified secondary structure.

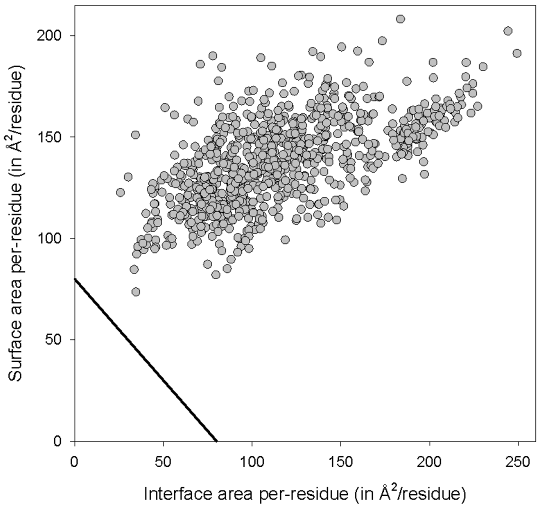

Figure 4. Gunasekaran-Tsai-Nussinov (Gunasekaran et al., 2004) graph for the 842 MoRFs. The plot provides a scale that measures confidence with which one can say whether a protein is ordered or disordered. The farther the point, which corresponds to a given chain, is from the dividing black line (boundary), the greater the confidence with which a protein can be classified into either of the classes. Points above the line correspond to disordered chains.

We also annotated the MoRF segments as those related to immune response and others. We used text mining of HEADER, TITLE and KEYWDS records in each PDB entry (each complex that was used to identify our MoRFs) to look for keywords such as histocompatibility, MHC, IgG, antigen, antibody, HLA, T cell, B cell, heavy chain, light chain, FAB fragment, and cycophilin. As a result, we identified 120 immune-related MoRFs.

The annotated MoRF regions were used to select full chains in the UniProt and the remaining AAs in these chains (all AAs except the residues that compose the MoRF) were by default assumed to be non MoRFs. We anticipate that some of the (default) non MoRF residues could in fact correspond to MoRFs, i.e., our

MoRF annotations are incomplete. We address this issue when we design and evaluate our method.

We divided the dataset into two parts: training and test sets. This division was performed to assure that chains in the training and test set share low sequence similarity. This means that a simple sequence similarity (sequence alignment) cannot accurately identify MoRF regions in the test set based on the MoRF annotations in the training set. We used CD-Hit (Huang et al. 2010) to cluster sequences in the entire dataset with identity > 30%. This resulted in 427 clusters with 274 of them that included only 1 sequence. We then assigned each cluster to either training or test set at random. This assures that training and test sets have similar number of chains and that similarity between sequences in these two datasets is below 30%. The training dataset was used to develop the method (to perform feature selection and parameterize the prediction algorithm) and test dataset was used to evaluate and compare it with other existing methods.

2.3.3 **Test and evaluation protocols**

2.3.3.1 Evaluation measures

We compare a prediction for a given sequence with its native/true annotation using two types of assessment: (1) per residue assessment which evaluates predictions for individual AAs; and (2) per sequence assessment which looks at the sequence as a whole. The prediction of MoRFs is performed for each AA in the input protein sequence. Each prediction consists of a numerical score *p* that quantifies propensity (probability) of this residue to form a MoRF segment and a binary score that categorizes the AA as MoRF or non MoRF residue.

The first per sequence measure is success rate. Use of this measure is motivated by the fact that there might be some un-annotated MoRF regions in our dataset which, if predicted, would count as false predictions. This success rate was originally used to evaluate predictions of B-cell epitopes and was designed to deal with the incompleteness of their annotations (Rubinstein, 2009). To calculate this measure, we compare the average predicted probability/propensity of residues in the native MoRF region to the average probability of the whole sequence, and we assign a score to each sequence. For i^{th} sequence in a dataset, the success rate S_i is calculated as follows:

$$Ave_{MoRF} = \frac{\sum P_{MoRF}}{n_{MoRF}} \quad , Ave_{nonMoRF} = \frac{\sum P_{nonMoRF}}{n_{nonMoRF}}$$

$$\begin{cases} Ave_{MoRF} > Ave_{nonMoRF} & S_i = 1 \\ O.w. & S_i = 0 \end{cases}$$

Total success rate S is calculated by averaging the per sequence scores over all sequences in a given dataset:

$$S = \frac{\sum_i S_i}{n_{sequences}}$$

Since probabilities of the predicted MoRFs should be higher than the non MoRFs, a correctly predicted sequence should have $S_i = 1$. After averaging over all chains, the success rate is a real value in the [0,1] range where 0 and 1 mean that all proteins were predicted incorrectly and correctly, respectively. Higher value of S indicates a better prediction quality.

For the per residue evaluations, we use three criteria that assess binary predictions:

$$\text{Accuracy} = (\text{TP}+\text{TN})/(\text{TP}+\text{FP}+\text{TN}+\text{FN})$$

$$\text{True positive rate} = \text{TPR} = TP\,/\,(TP + FN) = TP\,/\,N_{MoRF}$$

$$\text{False Negative rate} = \text{FPR} = TN\,/\,(TN + FP) = TN\,/\,N_{nonMoRF}$$

where TP is the number of true positives (correctly predicted MoRF residues), FP denotes false positives (non MoRF residues that were predicted as MoRF), TN denotes true negatives (correctly predicted non MoRF residues), FN stands for false negatives (MoRF residues that were predicted non MoRF), respectively. The accuracy values range between 0 and 1 and it is equal one when all residues are predicted correctly.

Another per residue prediction assessment method is based on calculating the area under the receiver operating characteristic (ROC) curve. The ROC curve is used to examine the predicted probabilities/propensities. The values of probabilities *p* (between 0 and 1) generated by a given prediction method are binarized such that all residues with probability equal or greater than a given threshold are set as MoRFs and all other residues are set as non MoRFs. The thresholds are varied between 0 and 1 (they are set to each of the values of *p*) and for each threshold the TPR and the FPR are calculated. We use the area under the corresponding curve (AUC), i.e., curve created by adjacent TPR vs. FPR points, to quantify the predictive quality.

We calculate the abovementioned criteria (success rate, accuracy, TPR, FPR, and AUC) using full protein chain. However, we also perform the same evaluation on a specific fragment of the sequence, which is motivated by the incompleteness of the MoRF annotations. We evaluate using the regions which are less likely to contain unannotated MoRF residues. Since MoRF regions are defined as small segments in a larger segment of disorder, we anticipate that the sequence surrounding a given MoRF region is less likely to contain unannotated MoRFs when compared to the remaining part of the chain. This means that annotations of non MoRFs are possibly more accurate in this area. Consequently, we perform "second" evaluation using a fragment of protein sequence that consists of the MoRF region with *n* AAs and *n* flanking AAs on each side of this region.

2.3.3.2 Biserial and φ correlations

Biserial correlation is used to measure correlation of two quantities where one is binary and the other is continuous. Given binary variable *X*, we divide values of the continuous variable *Y* to two groups: 0 and 1, based on their corresponding values of *X*. The biserial correlation is calculated as:

$$corr(x, y) = \frac{M_0 - M_1}{S_n}\sqrt{\frac{n_0 n_1}{n^2}}$$

where S_i is the standard deviation of *X* and M_0 and M_1 are mean values for group 0 and group 1 with sizes n_0 and n_1 respectively.

We use biserial correlation when designing our method to perform feature selection i.e. to quantify the correlation of a given input feature with the native (binary) annotation of MoRFs. We perform this by calculating an average biserial correlation over 5 training folds using the training dataset. We use this average to sort the features in the descending order.

For binary input features we use φ coefficient (Ernest, 1991), which quantifies correlation when both variables are binary. Using notation from Figure 5 we define φ coefficient as follows:

$$\varphi = \frac{P_{00}P_{11} - P_{10}P_{01}}{\sqrt{P_1 Q_1 P_2 Q_2}}$$

We scale φ to [-1,1] range as φ/φ$_{\max}$ where φ$_{\max}$ is defined as

$$\varphi_{\max} = \frac{\sqrt{Q_2 P_1}}{\sqrt{Q_1 P_2}} \qquad forP_2 \geq P_1$$

$$\begin{array}{cc|c|c|c}
 & & \multicolumn{2}{c}{\textbf{Variable 1}} & \\
 & & 0 & 1 & \\
\hline
 & 0 & P_{00} & P_{01} & Q_1 \\
\hline
\textbf{Variable 2} & 1 & P_{10} & P_{11} & P_1 \\
\hline
 & & Q_2 & P_2 &
\end{array}$$

Figure 5. Matrix that defines combinations of values of two binary variables. In case of the MoRF prediction, variable 1 corresponds to the native MoRF annotations and variable 2 could be an input feature or a binary MoRF prediction.

2.3.3.3 Test protocols

To guarantee an unbiased evaluation of our method (by the fact that we use a training dataset), we divide the original dataset of 842 chains into a training set, which is used to develop the method, and an independent set (that includes sequences that are dissimilar to sequences in the training dataset), which is used to evaluate the final design of our method and compare it with existing methods. The design, which includes feature selection, parameterization of the Support Vector Machine (SVM) and selection of the final method, uses the training set with the 5-fold cross validation protocol. This is performed to assure that our method does not overfit the training dataset, and thus it can provide equally good predictions on the test set. To perform 5-fold cross validation we divide the training set into 5 equal-sized subsets of protein chains. We use four of these subsets to form a training dataset that is utilized to compute the model and the fifth subset constitutes a test set that is used to perform the evaluation. This procedure is repeated five times, each time choosing a different fold as the test set. Finally, to estimate the performance, the results from the 5 test folds are averaged. We note that sequence in that training set are clustered based on their similarity, as explained in section 2.3.2. When selecting the five fold, the sequences in the same cluster are kept together. This assures that sequences between the folds share low similarity below 30%, which is also true when comparing training and test datasets.

2.3.3.4 Significance Test

A test of statistical significance is performed to verify whether or not a given result occurred by chance. The significance level or *p*-value that is given by a significance test represents the probability of observing the result by chance. Therefore, lower values of *p*-value correspond to the results that have higher significance.

To evaluate the statistical significance of the improvements offered by our methods (when compared with other methods), we compare 10 paired results quantified using success rate and AUC that are obtained using the bootstrapping with 50% of randomly selected test chains. We determine normality of a given measurement with Anderson-Darling test at the 0.05 significance. For normal measurements, we use paired t-test, and otherwise we use Wilcoxon rank-sum test. We use thresholds of 0.5 and 0.01 for the resulting *p*-value, i.e., results with the *p*-value lower than these thresholds are assumed to be significant.

3 Sequence-based MoRF prediction

3.1 Overall architecture

Figure 6 shows the overall architecture of our method, which is called MoRFpred. The first step is to calculate a feature set that represents each residue in the input sequence using a sliding window, i.e., the calculation of the features is based on a segment of residues centered over the predicted residue. The use of the window is a popular approach in the design of similar sequence-based predictors (Kurgan and Disfani, 2011). In the second step, the feature vector is fed into a linear SVM to calculate propensity of a given input residue to form a MoRF region. We do not describe SVM in this thesis as this is out of scope of this work; the reader is referred to (Fan et al. 2008) for the details. Finally, in the third step, these propensities are merged with the results of alignment of the input protein against a set of MoRF annotated proteins in a training dataset to produce the final propensities. Following we describe the details.

In the first step, we use the protein sequence to predict the following 1D descriptors from the sequence: (1) disorder; (2) solvent accessibility; and (3) b-factor. These descriptors are used to calculate the feature set. To predict disorder, we utilize IUPredL and IUPredS (Dosztányi et al., 2005), DISOPRED2 (Ward et al., 2004), DISOclust (McGuffin, 2008) and MFDp (Mizianty et al. 2010). These predictors represent the four major classes of disorder predictors, see section 2.2.2. The Real-SPINE3 (Farragi et al., 2009) is used for the prediction of the relative solvent accessibility and PROFbval (Schlessinger et al., 2006) is used for the B-factor prediction. The choice of the disorder predictors is based on the evaluation in CASP8 where we picked top available methods (Kurgan & Disfani, 2011). For all of the methods we acquired the standalone version. We also calculate Position Specific Scoring Matrix (PSSM) profiles generated with PSI-BLAST (Altschul et al., 1997) which summarizes information from multiple sequence alignment. Finally, we represent various biophysical properties AAs with the amino acid indices from AAindex database. All predictors were run using default parameters. The PSSM profiles were generated using the non-redundant (nr) database from NCBI, which was filtered using PFILT (Jones & Swindells, 2002) to remove low-complexity regions, transmembrane regions, and coiled-coil segments.

The acquired information is used to build a feature set. A sliding window of size 25 is used on each residue to generate features pertaining to that residue. The size of the window is determined based on the average size of the MoRF regions which is 12. We initially calculate a large number of features using a relatively wide window to later filter them out. We keep the features which improve the quality of MoRF predictions. These features are used as the input to the SVM. Choice of SVM is motivated by its successful application in prediction of disordered regions (Ishida & Kinoshita, 2008; Mizianty et al. 2010) and B-factors (Chen et al., 2007).

Due to a large number of samples (amino acids that need to be predicted), we decided to use a fast SVM implementation. Therefore we chose liblinear (Fan et al. 2008) that was previously utilized in MFDp (Mizianty et al. 2010), which is one of the top methods to predict disorder. The output of our method is a real value that quantifies the probability of a given residue to form a MoRF region. These values are binarized using a threshold of 0.5, i.e., amino acids predicted with probability > 0.5 are assumed to form MoRFs. Finally, we use PSI-BLAST to align the input protein against proteins in the training set. We transfer the annotations of MoRF regions from the aligned proteins into the input protein and merge them with the predictions from SVM to get the final results.

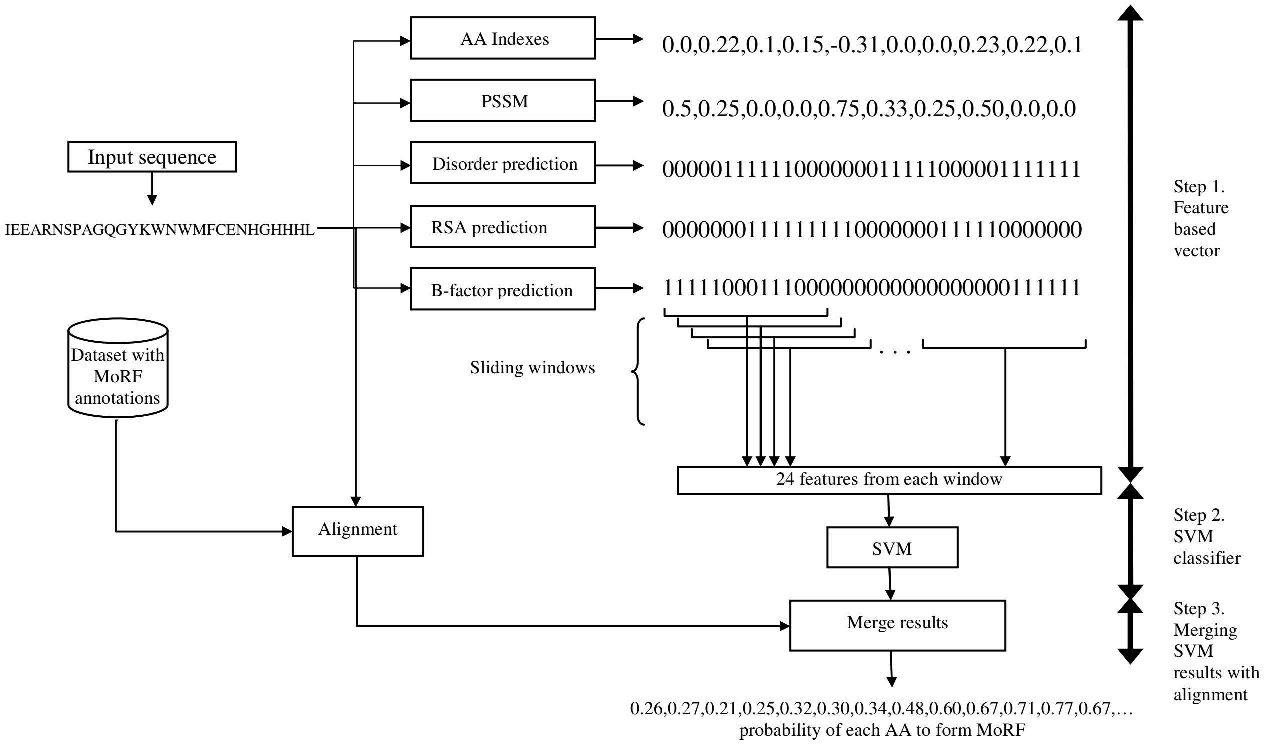

Figure 6. Architecture of the MoRFpred method.

3.2 Feature based sequence representation

Using predictions mentioned in the previous section, we build 5 types of features that are based on the alignment, amino acid indexes, and predicted disorder, solvent accessibility, and flexibility. For each type of features we calculate several per residue and aggregated features as explained below. The total number of features is 1764.

Each residue in a given input protein chain is a sample which is described by a set of features. For each residue, we include information about the residue itself and its neighbors. To do so, we create a sliding window of size 25 that is centered on the predicted residue and we extract information from this window to calculate the feature set. For the residues on both termini (ends) of the sequence where there are no neighboring residues on the right or left side, we fill these positions in the window with default values. Calculations of the features for each position in the window was motivated by the previous methods in this field, including α-MoRF-PredI and α-MoRF-PredII, which used attributes such as predicted disorder and secondary structure. When calculating the features, we used the predictions in two forms: the probabilities (propensities) and the corresponding binary values.

We also generate another, novel group of features which provide information about a segment in the sequence rather than an individual residue (a position in the window). These features are created by aggregating raw values over a window of a certain size. Simple aggregations include averaging a quantity over the window for real valued data or calculating the content for binary valued predictions. We also aggregate by calculating a difference between an average value in a smaller (inner) window and a larger (outside) window; see Figure 7. We utilize this aggregation to contrast the values calculated using amino acids that are close to the being predicted residue against the values associated with residues in a wider neighborhood in a sequence. This is motivated by the fact that MoRF segments are usually surrounded by larger disordered segments. While size of the entire/sliding window is fixed at 25, the size of the inner window is adjusted.

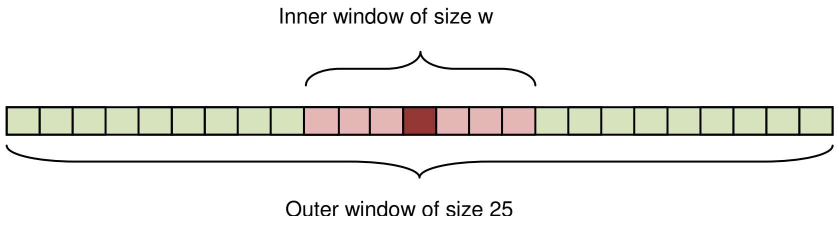

Figure 7. Sliding window used to create feature set. A sliding window of size 25 centered on the target residue (dark red) is used to create per residue features. The inner window of size *w* is shown in red. The flanking area, which corresponds to the outside window, is shown in green.

Table 2 describes details about *per residue* (calculated for each position in the sliding window) and *aggregated* features for each feature type. Note that we calculate the disorder-based features for each of the 5 disorder predictors.

Our dataset is heavily unbalanced, i.e., the numbers of MoRF and nonMoRF residues are very different. To be more exact, there is only 1 MoRF residue for every 46 non MoRF residues. This imbalance is likely to bias a prediction method to under-predict or completely ignore the MoRF regions. To avoid this, we undersample the non MoRF residues. We test three ways to undersample. As motivated in the last paragraph in section 2.3.3.1, in the first sampling strategy we use the non MoRF residues that are the flanking residues of the MoRF residues (*local sampling*); this results in 2:1 ratio between non MoRF and MoRF residues. We also use *random sampling* with the same 2:1 ratio (two nonMoRFs for each MoRF) and higher 3:1 ratio.

Feature type	Input type		Description	Window size	Number of features
Per residue	Disorder, RSA, B-factor		For each prediction method, we include binary values and probabilities in a window. 7 (methods: 5 disorder + RSA + B-factor) * 25 (window size) * 2 (binary and probability) = 350 features.	$w = 25$	350
	PSSM generated with PSI-BLAST		For each residue a matrix of size 7*20 = 140 is included in the features where each row is a window of size 7 centered on the main residue and each column contains values corresponding to different amino acids.	$w = 7$	140
Aggregated	Disorder	Average probability	Average of probability over the window of size *w*.	$w = \{2*n+1 \mid n=2,..,12\}$	170
		Content	Content of binary prediction over the window of size *w*.	$w = \{2*n+1 \mid n=2,..,12\}$	
		Average difference	Difference of probability averages in an inside window of size *w* and an outside window of size 25.	$w = \{2*n+1 \mid n=2,..,7\}$	
		MinMax average	Difference of minimum average in an inside window of size *w* from maximum average in an outside window of size 25.	$w = \{2*n+1 \mid n=2,..,7\}$	
	Relative solvent accessibility (RSA)	Average RSA	Average of RSA values over the window of size *w*.	$w = \{2*n+1 \mid n=2,..,7\}$	24
		Standard deviation (stdv)	Standard deviation of RSA values over the window of size *w*.	$w = \{2*n+1 \mid n=2,..,7\}$	
		Content	Content of binary prediction over the window of size *w*.	$w = \{2*n+1 \mid n=2,..,7\}$	
		Stdv difference	Difference of standard deviation in an inside window of size *w* and an outside window of size 25.	$w = \{2*n+1 \mid n=2,..,7\}$	
	B-values	Minimal B-factor	Minimum of normalized B-factor over the window of size *w*.	$w = \{2*n+1 \mid n=2,..,7\}$	18
		Content	Content of binary prediction over the window of size *w*.	$w = \{2*n+1 \mid n=2,..,7\}$	
		Content difference	Difference of content in an inside window of size *w* and an outside window of size 25.	$w = \{2*n+1 \mid n=2,..,7\}$	
	AA Indices	Average	Average of amino acid index over a window of size *w*.	$w = 15$	1062
		Average difference	Difference of averages in an inside window of size *w* and outside window of size 25.	$w = 15$	

Table 2. Description of features considered in building the proposed MoRFpred. We describe per residue and aggregated features and categorize them based on the type of information they utilize. We briefly describe each feature type and specify window sizes that used to calculate them. For features which calculate the difference between the outside and inner windows, the size of the inner window is specified by parameter *w* and size of the outside window = 25–*w*. The difference is calculated by subtracting the value for the inner window from the value for the outside window.

3.3 Feature selection and parameterization of SVM

Feature selection methods are used to select a subset of relevant features to improve the performance of machine learning-based classification methods. We perform feature selection in 3 steps. First, a scoring function is used to rank the features based on their relevance/relation to MoRF annotations in training dataset. Second, features with lower ranks (below a certain threshold) are removed. Third, a best first search is implemented to pick features that improve predictive results based on cross validation on the training dataset.

We repeat feature selection 9 times, considering three ranking functions executed for the three sampling strategies. We rank the features based on:

- Their average (over 5 training folds) biserial correlations with annotation of MoRFs using the complete training set, i.e., using all residues in the training set (referred to as *complete correlation ranking*)
- Their average (over 5 training folds) biserial correlations with annotation of MoRFs for the MoRF residues and the flanking residues, i.e., using the same residues as in the local sampling (*local correlation ranking*)
- Their average success rate calculated when using a single feature on training set to predict the annotation of MoRFs in 5 fold cross validation (*success rate ranking*). The predictions are performed using a linear kernel SVM classifier with the default complexity parameter $C = 5$ (Fan et al. 2008).

We sort the features in the descending order for each of the three rankings and we remove features with correlation < 0.05 for the complete and local correlation rankings and with success rate < 0.5 in case of the success rate ranking. We selected the thresholds to remove only the irrelevant/poorly performing features.

		Whole Sequence					Flanking Region					Average (whole and flanking)	
Sampling	Feature selection	ACC	TPR	FPR	Success rate	AUC	ACC	TPR	FPR	Success rate	AUC	AUC	Success rate
local	Complete ranking	0.948	0.183	0.034	0.665	0.642	0.682	0.183	0.063	0.637	0.616	0.629	0.651
	Local ranking	0.788	0.391	0.203	0.748	0.632	0.650	0.391	0.218	0.696	0.632	0.632	0.722
	Success rate ranking	0.503	0.596	0.499	0.720	0.564	0.566	0.596	0.450	0.705	0.598	0.581	0.713
	Combined	0.920	0.245	0.064	0.703	0.654	0.686	0.245	0.088	0.703	0.665	0.660	0.703
random 3:1	Complete ranking	0.929	0.205	0.055	0.696	0.664	0.660	0.205	0.106	0.632	0.584	0.624	0.664
	Local ranking	0.503	0.637	0.500	0.722	0.599	0.559	0.637	0.481	0.694	0.620	0.609	0.708
	Success rate ranking	0.740	0.428	0.253	0.751	0.630	0.614	0.428	0.291	0.663	0.579	0.604	0.707
	Combined	0.931	0.225	0.053	0.696	0.674	0.679	0.225	0.088	0.691	0.611	0.643	0.694
random 2:1	Complete ranking	0.456	0.767	0.551	0.774	0.672	0.447	0.767	0.716	0.679	0.570	0.621	0.727
	Local ranking	0.504	0.599	0.498	0.698	0.572	0.577	0.599	0.434	0.698	0.614	0.593	0.698
	Success rate ranking	0.178	0.947	0.839	0.765	0.636	0.378	0.947	0.914	0.615	0.548	0.592	0.690
	Combined	0.454	0.768	0.553	0.762	0.653	0.442	0.768	0.725	0.601	0.539	0.596	0.682

Table 3. Comparison of results of MoRF prediction using different feature selection methods and different sampling strategies. The results are based on the cross validation on the training dataset. Rows list individual setups, which consider three sampling strategies and 3 feature selection approaches. We also use a combined feature set which implements a union of the features selected by the three selection approaches. The columns list results when evaluation is performed using the whole chain, using only the flanking region (see Section 2.2 in the main text), and the average of the two.

The 0.05 threshold removes features that have virtually no correlation with the outcomes. Similarly, the 0.5 value for the success rate ranking removes features that provide predictions equivalent to a random predictor. We then execute the best first search algorithm on the sorted list of the remaining features. In this algorithm we start with the top ranked feature and we continue by adding one (next-ranked) feature at a time. A given feature is added into the current feature set if it results in improved prediction quality when compared with the methods that uses the current feature set. The predictions are based on a linear kernel SVM classifier with the default parameter $C = 5$ and a modified version of the 5 fold cross validation. The modification of the cross validation is meant to prevent overfitting (due to the large number of feature sets that are considered) and simulate predictions on the independent dataset. We use 4 of the 5 folds to implement the 4 fold cross validation and we keep the 5th fold as an independent test set; we refer to this as 4+1 cross validation. We calculate the success rates for both the cross validation and the independent test set and compare these with the currently best success rates. If the newly added feature improves the success rate by at least 0.01 on both tests then we add the feature. If the success rate improves in only one or none of the tests we discard the feature and move on to the next ranked feature.

Table 3 shows the results of the cross validation on the training set for the 9 feature selection setups; 3 (sampling strategies) * 3 (ranking methods) = 9 setups. For each sampling, the last row of the table presents results of a model that uses a feature set that combines the features selected by all three feature selection methods. For each setup, we present evaluations on the whole sequence and on the flanking regions. We select the best performing setup by considering predictive performance on both the flanking region and the whole sequence. Considering the average (over the flanking region and the whole sequence) AUCs and success rates (the last two columns in Table 3) we observe that the model based on the local sampling and combined features have the highest average AUC and a reasonably high success rate. We select this setup to implement our prediction method.

Next, the selected feature set is used to parameterize the SVM model, i.e., to optimize value of parameter C, utilizing the 4+1 cross validation on the training dataset. We consider $C = 2^x$, where x = -13, -12, …, 8, 9, and select $C = 2^{-6}$ which has the highest success rate on the independent test fold, see Figure 8. We also observe that the SVM generates similarly good results for a relatively wide range of values of *C*, between 2^{-8} and 2^{8}.

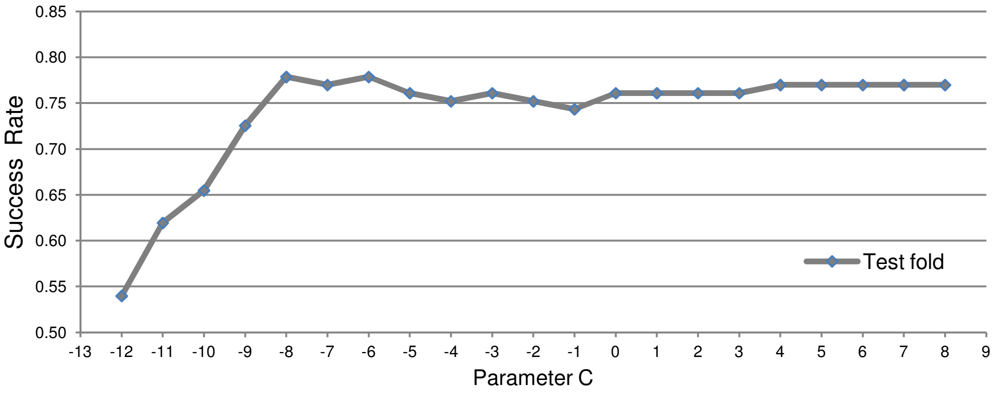

Figure 8. Results of parameterization of parameter *C* for the SVM classifier that uses the combined feature set selected based on the local sampling. The vertical axis represent success rate and horizontal axis shows $\log_2^{C}$.

3.4 Alignment-based MoRF prediction

We align proteins in the test set against chains from training set which are annotated with MoRF regions using PSI-BLAST with default parameters. For each sequence in the test set then we get a number of matching/similar sequences in the training set; this number depends on the e-value that quantifies similarity. These matches indicate sequences in the training set that are (partly) aligned with our target sequence, i.e., the sequence from the test set. If the amino acids that are aligned between the target and the training sequence contain a MoRF region, then we annotate these amino acids in the target sequence as MoRFs. We tested different thresholds for the e-value using the training dataset, by merging the results of the SVM with the annotations transferred through the alignment. We picked e-value = 0.5 which provides the best AUC and success rate. We use the sequences with e-value ≤ 0.5 and discard the remaining matches.

We add the annotations acquired from the alignment to the SVM predictions by updating the probabilities generated by the SVM. For the residues that are predicted as MoRFs by alignment and as non MoRFs by SVM (SVM generates probability < 0.5), we add 1 to the probabilities generated by SVM and divide the result by 2; as a result these residues will be predicted as MoRF residues. We use the probability generated by the SVM for the remaining residues.

3.5 Prediction of MoRF regions by merging SVM and alignment

We compute the SVM model on the locally sampled training set using the combined feature set with $C=2^{-6}$ and test it on the independent test set. Table 4 presents results of prediction before and after merging alignment-based predictions. The results are slightly improved after merging the alignment–based annotations; the AUC is improved by 1% and TPR by 3%. We also evaluate the alignment only-based results in the last row. We observe that although alignment helps to improve the predictive performance of the SVM, it cannot be used alone as an accurate predictor of the MoRF regions.

	Whole Sequence					**Flanking Regions**				
Predictor	**ACC**	**TPR**	**FPR**	**Success rate**	**AUC**	**ACC**	**TPR**	**FPR**	**Success rate**	**AUC**
SVM	0.937	0.226	0.048	0.714	0.663	0.706	0.226	0.059	0.752	0.678
SVM + Alignment	0.937	0.254	0.049	0.718	0.673	0.711	0.254	0.065	0.754	0.684
Alignment	0.980	0.039	0.001	0.043	NA	0.679	0.039	0.008	0.038	NA

Table 4. Comparison of performance of MoRFpred before and after the addition of the alignment-based predictions. We use the best selected (using training dataset) SVM model and we train it on the training dataset. The alignment is performed against the training dataset. The results are based on the independent test dataset. Alignment generates only binary predictions and thus its AUC cannot be calculated. The two main columns list results when evaluation is performed using the whole chain and using only the flanking region (see Section 2.2 in the main text).

4 Results

4.1 Comparison with existing methods

We empirically compare our MoRFpred method with the three available MoRF predictors, namely α-MoRF-PredI, α-MoRF-PredII, and ANCHOR. We evaluate results on the independent test set on both whole sequences and the area containing MoRF and its flanking region, see Table 5. The α-MoRF-PredI and α-MoRF-PredII predictors provide only the binary values, which mean that AUC cannot be calculated for these methods.

We observe a relatively large gap between the success rates of α-MoRF-PredI and α-MoRF-PredII predictors and the results generated by ANCHOR and MoRFpred. This is due to the fact that the former two predictors were developed to identify MoRF regions that form only the alpha helixes upon binding. In contrast, ANCHOR and MoRFpred are designed to identify all types of MoRF regions. Thus, we focus on comparing results of MoRFpred and ANCHOR.

Table 5 shows that MoRFpred outperforms ANCHOR in terms of both AUC and success rate by 7 and 10%, respectively. We note that the improvements are consistent for the evaluations with the whole sequences and the flanking regions. The differences in accuracies between the whole sequences and flaking region are due to different ratios of non-MoRF to MoRF residues. The binary predictions generated by our method are characterized by low FPR and relatively high TPR. To compare the binary predictions side by side with the other methods, we added two last rows in Table 5 where we match MoRFpred's TPR and FPR to the highest TPR and lowest FPR of the other methods, respectively We match these by adjusting the thresholds on the predicted probabilities and we perform that separately for the evaluations on the whole sequence and the flanking regions. These results demonstrate that MoRFpred outperforms the competing solutions by providing substantially higher TPRs given similar FPRs and lower FPRs for comparable TPRs.

	Whole Sequence					**Flanking Region**				
Predictor	**ACC**	**TPR**	**FPR**	**Success rate**	**AUC**	**ACC**	**TPR**	**FPR**	**Success rate**	**AUC**
α-MoRF-PredI	0.946	0.123	0.037	0.158 ++	NA	0.668	0.123	0.065	0.129 ++	NA
α-MoRF-PredII	0.889	0.258	0.098	0.303 ++	NA	0.673	0.258	0.124	0.263 ++	NA
ANCHOR	0.740	0.389	0.253	0.611 ++	0.600 ++	0.640	0.389	0.237	0.659 ++	0.590 ++
MoRFpred (SVM + alignment)	0.937	0.254	0.049	0.718	0.673	0.711	0.254	0.065	0.754	0.684
MoRFpred (to match the highest TPR)	0.854	0.389	0.137	0.718	0.673	0.696	0.389	0.153	0.754	0.684
MoRFpred (to match the lowest FPR)	0.948	0.222	0.037	0.718	0.673	0.711	0.254	0.065	0.754	0.684

Table 5. Comparison of prediction results on the test dataset. The last two rows show results for MoRFpred where the binary predictions were calculated (by adjusting the threshold on the probabilities) to match the highest TRP and FPR generated by the existing methods. The two main columns list results when evaluation is performed using the whole chain and using only the flanking region (see Section 2.2 for details). α-MorfPredI and α-MorfPredII generate only binary predictions and thus their AUC cannot be calculated. Statistical significance of the differences in the success rates and AUC between the MoRFpred and the other three methods is shown next to the success rate and AUC values, where ++ and + denote that the improvement is significant at the *p*-value < 0.01 and < 0.05, respectively.

Figure 9 presents the ROC curves for MoRFpred and ANCHOR. The ROC curves zoom on the low FPR values below 0.1, which is motivated by the imbalanced nature of our dataset. Higher FPRs would lead to a significant over-prediction of the MoRF residues. We observe a large separation between MoRFpred and ANCHOR across the entire range of the FPR values. We also note that addition of the alignment into MoRFpred also results in improvements for all values of FPRs.

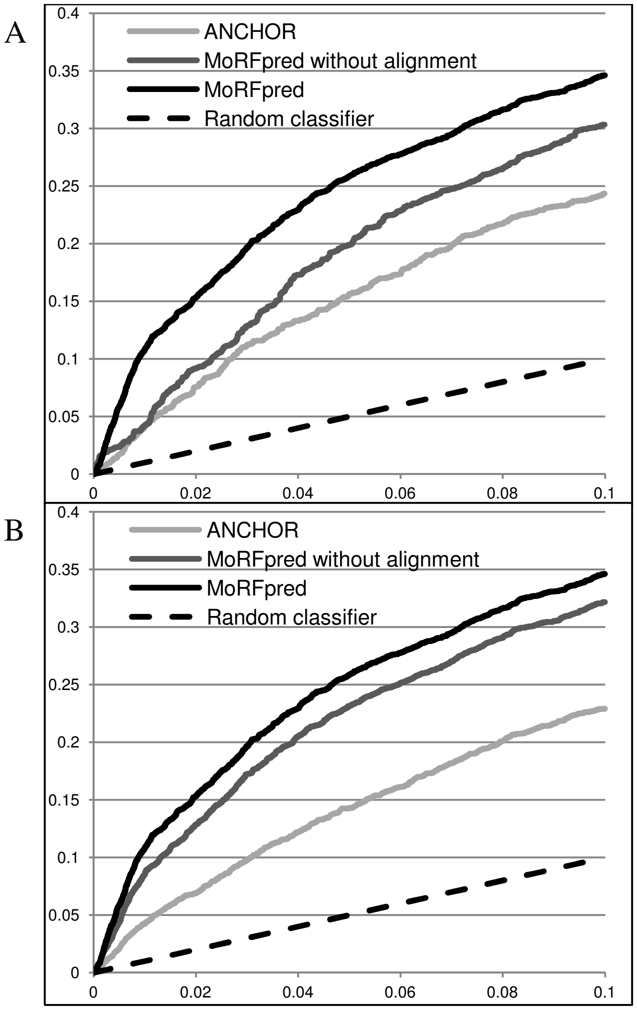

Figure 9. Comparison of ROCs for MoRFpred and ANCHOR on the test dataset. Panel A compares ROCs for when evaluations is performed using the whole sequences and panel B when using the flanking region. The ROC curves are provided for the FPR < 0.1.

4.1.1 Evaluation for different MoRF types

In section 2.2.1, we discuss the fact that MoRFs often fold into a specific secondary structure upon binding and therefore they are grouped as helix, sheet, and coil types. MoRF regions that include two types of the secondary structures are referred to as complex regions. We evaluate the considered methods separately for each MoRF type, see Table 6.

Table 6 shows that MoRFpred outperforms the other three methods with respect to the success rates for each MoRF type. Using the AUC measure, MoRFpred again improves over ANCHOR in all cases. The evaluation on the α-MoRF type shows a visible improvement of the α-MoRF-PredI and α-MoRF-PredII when compared to their predictions on the other MoRF types. This improvement is expected since these methods were designed for the prediction of the helix-type MoRFs. However, ANCHOR and MoRFpred are still better than α-MoRF-Pred methods for all MoRF type. The success rates of MoRFpred are higher by 4%, 12%, and over 5% than ANCHOR for the prediction of α-MoRFs, coil-MoRFs, and complex-MoRFs, respectively. Results also show that all methods perform relatively poorly for the predictions of the β-MoRFs, although MoRFpred still outperforms the other solutions. However, we note relatively low numbers of the β- and complex-MoRFs which could affect validity of our conclusions.

The alignment only-based predictions have low TPRs coupled with very low (close to zero) FPRs for all MoRF types. This shows that alignment predicts only a few MoRFs but with high quality. The alignment contributes 3 to 6% to the TPR of the MoRFpred for the helix, sheet, and coil MoRF types.

MoRF type		Whole Sequence							Flanking Region						
# (%) of MoRF segments	Predictor	ACC	TPR	FPR	Success rate		AUC		ACC	TPR	FPR	Success Rate		AUC	
Helix **97 (23%)**	α-MorfPredI	0.930	0.176	0.056	0.320	++	NA		0.648	0.176	0.115	0.258	++	NA	
	α-MorfPredII	0.847	0.403	0.144	0.598	++	NA		0.677	0.403	0.186	0.546	++	NA	
	ANCHOR	0.623	0.545	0.376	0.866	+	0.635	++	0.657	0.545	0.286	0.876	=	0.662	++
	MoRFpred	0.937	0.357	0.052	0.907		0.747		0.741	0.357	0.066	0.907		0.763	
	Alignment only	0.982	0.063	0	0.093		NA		0.68	0.063	0.010	0.093		NA	
Sheet **15 (4%)**	α-MorfPredI	0.961	0.099	0.018	0.067	++	NA		0.697	0.099	0.009	0.067	++	NA	
	α-MorfPredII	0.936	0.224	0.046	0.200	++	NA		0.706	0.224	0.058	0.200	++	NA	
	ANCHOR	0.866	0.168	0.117	0.333	++	0.506	+	0.681	0.168	0.067	0.600	++	0.554	++
	MoRFpred	0.934	0.149	0.047	0.600		0.654		0.685	0.149	0.052	0.733		0.698	
	Alignment only	0.974	0.043	0.004	0.067		NA		0.681	0.043	0.006	0.067		NA	
Coil **288 (69%)**	α-MorfPredI	0.954	0.084	0.027	0.094	++	NA		0.677	0.084	0.039	0.08	++	NA	
	α-MorfPredII	0.912	0.175	0.073	0.198	++	NA		0.667	0.175	0.096	0.156	++	NA	
	ANCHOR	0.811	0.308	0.178	0.528	++	0.595	++	0.630	0.308	0.216	0.583	++	0.555	++
	MoRFpred	0.937	0.206	0.048	0.653		0.634		0.697	0.206	0.067	0.701		0.638	
	Alignment only	0.978	0.029	0.002	0.028		NA		0.68	0.029	0.008	0.021		NA	
Complex **19 (4%)**	α-MorfPredI	0.946	0.332	0.043	0.389	++	NA		0.663	0.332	0.157	0.278	++	NA	
	α-MorfPredII	0.860	0.467	0.133	0.500	++	NA		0.708	0.467	0.162	0.500	++	NA	
	ANCHOR	0.590	0.640	0.411	0.833	++	0.658	+	0.645	0.640	0.352	0.722	++	0.692	++
	MoRFpred	0.940	0.369	0.050	0.889		0.760		0.736	0.369	0.066	0.833		0.767	
	Alignment only	0.982	0	0.001	0		NA		0.649	0	0	0		NA	

Table 6. Comparison of prediction results for different MoRF types on the test dataset. Comparison of prediction results for different MoRF types on the test dataset. The two main columns list results when evaluation is performed using the whole chain and using only the flanking region (see Section 2.2 in the main text). α-MorfPredI and α-MorfPredII generate only binary predictions and thus their AUC cannot be calculated. Statistical significance of the differences in the success rates and AUC between the MoRFpred and the other three methods is shown next to the success rate and AUC values, where ++, +, and = denote that the improvement is significant at the *p*-value < 0.01, at *p*-value < 0.05, and that the difference is not significant, respectively.

MoRFpred produces the most accurate results for the α-MoRFs, as evidenced by high AUC and success rate values. This could originate from the fact that helixes are local (in the sequence) structures and thus hey are easier to capture using a window-based approach that is implemented by our method. In contrast, β-sheets can span over large stretches of the sequence, and thus the window may not be sufficient to find them.

MoRF type # (%) of MoRF segments	Predictor	Whole Sequence							Flanking Region						
		ACC	TPR	FPR	Success rate		AUC		ACC	TPR	FPR	Success rate		AUC	
Immune response-related **74 (18%)**	α-MoRF-PredI	0.958	0	0.019	0	++	NA		0.691	0	0	0	++	NA	
	α-MoRF-PredII	0.921	0.016	0.057	0.027	++	NA		0.681	0.016	0.021	0.014	++	NA	
	ANCHOR	0.824	0.214	0.161	0.5	=	0.573	++	0.654	0.214	0.149	0.635	=	0.569	+
	MoRFpred	0.932	0.156	0.049	0.581		0.568		0.716	0.156	0.033	0.662		0.583	
	Alignment Only	0.976	0	0	0		NA		0.691	0	0	0		NA	
Other **345 (82%)**	α-MoRF-PredI	0.945	0.143	0.039	0.191	++	NA		0.664	0.143	0.077	0.157	++	NA	
	α-MoRF-PredII	0.885	0.298	0.104	0.362	++	NA		0.672	0.298	0.143	0.316	++	NA	
	ANCHOR	0.729	0.419	0.265	0.635	++	0.608	++	0.638	0.419	0.253	0.664	++	0.595	++
	MoRFpred	0.937	0.273	0.049	0.748		0.692		0.711	0.273	0.072	0.774		0.701	
	Alignment Only	0.98	0.045	0.001	0.052		NA		0.677	0.045	0.009	0.046		NA	

Table 7. Comparison of prediction results for immune function-related and other proteins on the test dataset. The two main columns list results when evaluation is performed using the whole chain and using only the flanking region (see Section 2.2 in the main text). α-MorfPredI and α-MorfPredII generate only binary predictions and thus their AUC cannot be calculated. Statistical significance of the differences in the success rates and AUC between the MoRFpred and the other three methods is shown next to the success rate and AUC values, where ++, +, and = denote that the improvement is significant at the *p*-value < 0.01, at *p*-value < 0.05, and that the difference is not significant, respectively.

Table 7 shows evaluations for the immune function-related MoRFs vs. the remaining MoRFs. Our method outperforms the other approaches for the non-immune MoRFs. However, for the immune function-related MoRFs, the improvements offered by our MoRFpred are smaller, i.e., about 3-8% in success rate and about 1% in AUC when compared with the runner-up ANCHOR. We note that all considered method perform relatively poorly for these MoRFs, which motivates further research in this area. We hypothesize that the immune function-related MoRFs are distinct from other MoRF types and thus their prediction may need a dedicated method.

4.2 Similarity analysis

In this section we investigate the hypothesis that MoRF regions have non-random similarity to each other. If true, this could be used to validate our claim that some of our false positive MoRF predictions might correspond to true MoRF regions. To this end, we create 4 different sets of protein segments which are used to investigate the similarity:

- Set of the native MoRFs in the test set.
- Set of random segments generated from test set that have the same length distribution and number when compared to the set of native MoRFs.
- Set of predicted MoRF segments that have at least 50% overlap with the native MoRFs in the test set.
- The predicted MoRF segments that have no overlap with the native MoRFs (predicted "false positive" MoRFs).

We use the native MoRFs in the training set as our reference population against which we align the four abovementioned sets. We measure the similarity using EMBOSS needle (Rice et al. 2000) (http://www.ebi.ac.uk/Tools/psa/emboss_needle/) with default parameters.

Each random segment, native/true, and predicted MoRF is aligned against the 421 native MoRFs in the reference population and we use the maximum score from the 421 similarities. We obtain four sets of scores for the native test set, random set, predicted overlapping MoRFs, and predicted non-overlapping MoRFs. Using these scores, we generate distributions which are fitted into the data using the EasyFit software (http://www.mathwave.com/products/easyfit.html).

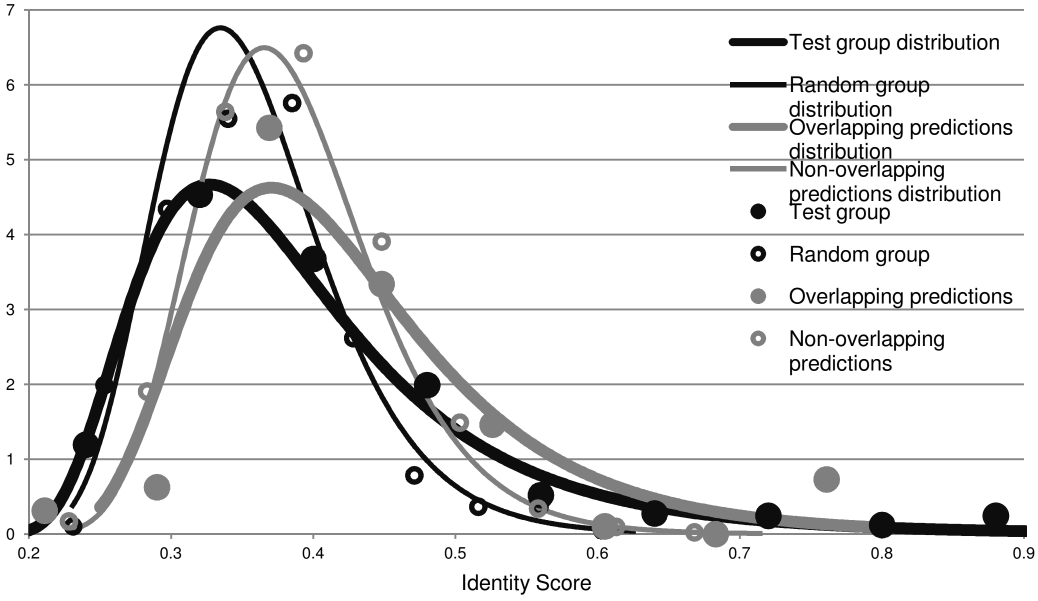

Figure 10. Similarity comparison of native MoRF and predicted MoRF. Similarity between the native MoRFs in the test set (test group), the random segments in the test set (random group), the MoRFs predicted in the test set by MoRFpred which overlap with the native MoRFs (overlapping predictions), and the MoRFs predicted in the test set by MoRFpred which do not overlap with the native MoRFs (non-overlapping predictions that correspond to false positive predictions) and the native MoRFs in the training dataset. The distributions, which are based on the Pearson 5 function, were fitted using EasyFit. The *x*-axis shows the similarity between the segments measured with EMBOSS needle and *y*-axis shows the relative number of segments.

We tried 6 commonly used types of distributions including normal, log-normal, gamma, beta, Pearson 5, and Pearson 6 distributions. Their fit into the data was evaluated using the Kolmogorov-Smirnov goodness of fit test. We use the Pearson 5 distribution which provided the best rank when considering the four sets of similarities.

Figure 10 depicts the distributions of the four sets of similarities. We observe that the distribution of the similarities for the native test group (using native MoRFs) has a higher and longer right tail when compared with the distribution of the random group (for random segments). This means that the native MoRF have higher similarity to each other when compared to their similarity with randomly selected segments. This motivated the use of the alignment in our predictor. Moreover, both distributions for the predicted MoRFs are also characterized by higher than random similarity. The shift to the right of the distribution for the overlapping (with native MoRFs) predicted MoRFs when compared with the distribution for the native test group means that our overlapping predictions tend to focus on the MoRF segments that are similar to the MoRFs in the training set. However, this bias is relatively minor, considering that the two distributions are shifted by only 0.05. Most importantly, the distribution for the predicted non-overlapping MoRFs (predicted false positives) is shifted toward higher similarity when compared with the random group, which suggests that some of our false positives (putative MoRFs) may correspond to native MoRFs.

4.3 Probability scores identify high quality predictions

We demonstrate that probabilities that are generated by MoRFpred can be used to select predictions that have higher quality. Figure 11 plots positive predictive value (PPV) for MoRF predictions (probability > 0.5) and negative predictive value (NPV) for non MoRF predictions (probability < 0.5) against the binned prediction probabilities generated by MoRFpred on the test dataset. The PPV and NPV values quantify the predictive performance of MoRFpred when it predicts MoRF and non MoRF residues, respectively. The non MoRF (negative) predictions for the low probabilities between 0 and 0.25, which account for 20% of all predictions, have substantially higher NPV when compared with the predictions with higher probabilities, e.g. in 0.4 to 0.5 range. The same is true for the MoRF (positive) predictions. We observe that for high probabilities between

0.7 and 1, our method provides a much higher PPV when compared with the predictions for probabilities closer to 0.5 (between 0.5 and 0.6). To sum up, Figure 11 demonstrates that predictions with probabilities farther away from the 0.5, which is the threshold to differentiate between MoRF and non MoRF residues, are characterized by higher predictive quality. This means that a user should be more confident with the predictions associated with either low or high probabilities.

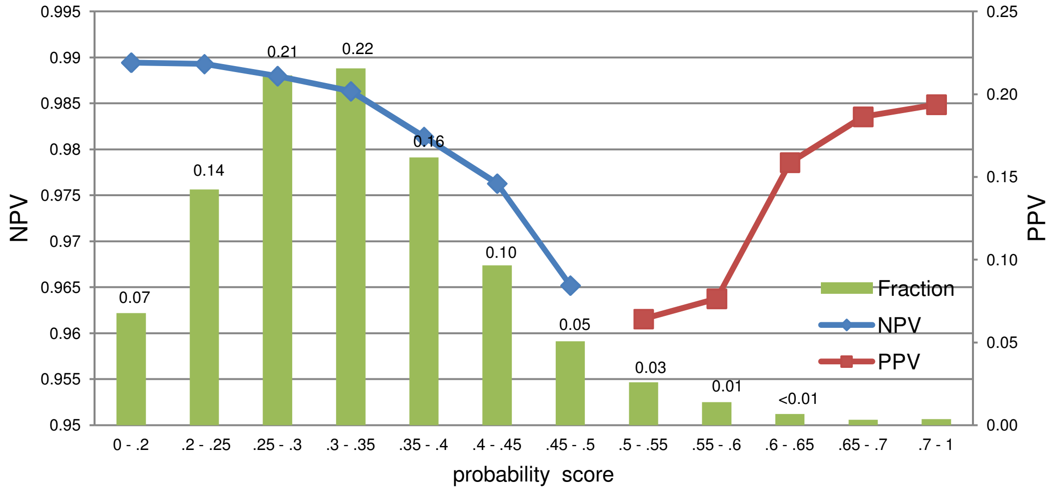

Figure 11 . Relation between predictive quality and the magnitude of the probabilities generated by MoRFpred on the test dataset. Values of probabilities are binned and shown on the *x*-axis. The left *y*-axis shows the percentage of correctly predicted non MoRF residues (NPV), which quantifies predictive quality when probabilities are below 0.5. The right *y*-axis corresponds to the percentage of correctly predicted MoRF residues (PPV), which evaluates predictive quality when probabilities are above 0.5. The bars indicate the fraction of residues for a given range of the probability.

4.4 Analysis of selected features

We describe a few potential sequence-derived markers of MoRF residues based on the features that were selected to implement the MoRFpred. The MoRFpred uses 24 features which were selected using three different feature selection methods. To analyze the selected features, we sort them in the ascending order based on their average ranking for the rankings generated by the three selection protocols; see section 3.3 for details. We calculate the average values of the top ranked features for the native MoRF and non MoRF residues (in flanking region), respectively. These averages for the five top-ranked features are compared in Figure 12. The values of these features have opposite signs for the native MoRF and non MoRF residues. However, the large and overlapping standard deviations (denoted by the error bars) show that they could not be used individually to accurately identifies MoRFs. This is why we employ multiple features in our prediction model.

The three left-most sets of bars in Figure 12 represent the same type of features, which is based on the average difference of disorder probabilities (refer to Table 2 for definition) calculated using predictions from IUPred with w = 15 and DISOPRED2 with w = 5 and w=15, respectively. These features were designed to contrast the value of the predicted disorder propensities in a MoRF region (inner window) and the sequence segments that flank this region (outside window). According to the study by Oldfield *et al.* (2005), MoRF regions are short ordered segments inside a larger disordered segment. Therefore, we expect a higher average of predicted disorder propensities in the outside window when compared to inner window, which should result in a positive value for our features for the native MoRF residues. This is confirmed in Figure 12 where the three features have, on average, positive values for the MoRF residues and negative (near zero) values for the non MoRF residues.

The right-most two sets of bars in Figure 12 show average difference-based features (features based on the differences in values between the inner and outside windows) calculated using two AAindexes, which quantify stability (Zhou & Zhou 2004) and hydrophobicity (Nozaki & Tanford, 1971), respectively. The stability scale is a quantity used to characterize the contributions of individual residues to stability of a protein fold where higher values mean higher stability. We observe that the average difference for residues in the native MoRF region for the stability-based feature is negative. That means that residues in the MoRF region have higher stability when compared to the surrounding residues. This agrees with the underlying biology, since MoRF residue should be more stable to transition into the structured state when compared to the flanking residues that are likely to be (more) disordered. The last feature is based on hydrophobicity. The negative value of this feature for the native MoRF residues indicates that these residues are, on average, more hydrophobic than the surrounding residues.

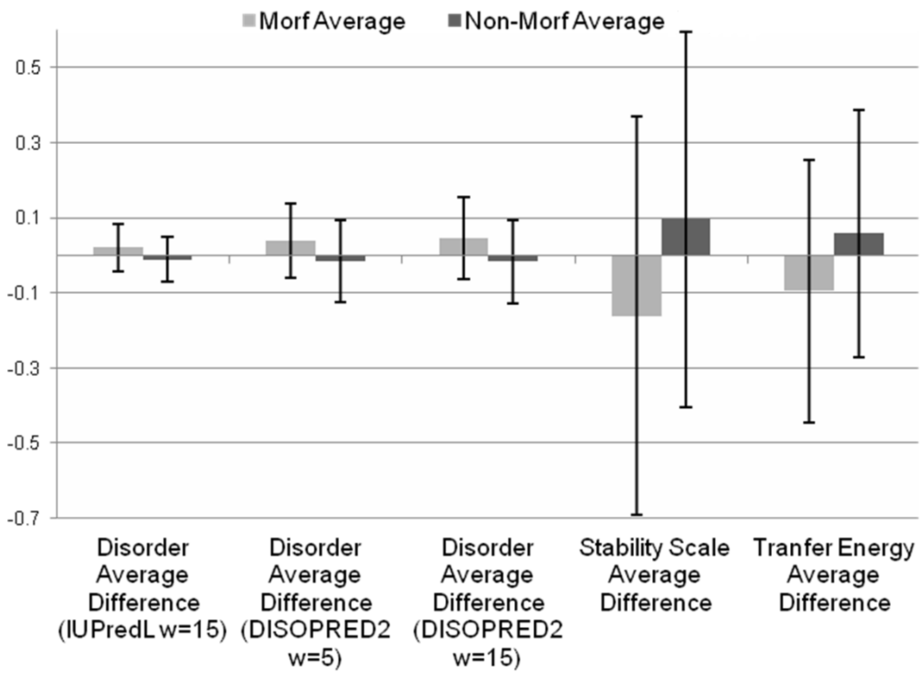

Figure 12. Analysis of the top-ranked features that can be used to characterize MoRFs. The average values of the top 5 ranked features used by MoRFpred, which are shown on the *x*-axis, for the native MoRF residues (light gray bars) and native non MoRF residues (dark gray bars) are compared. The corresponding standard deviations are shown using the error bars. The selected five features represent an average difference of a given quantity, which is described in Table 2. Negative values mean that average in the inner window of size w was higher than the average of the flanking areas.

Our features reveal a few interesting sequence-derived markers of MoRF residues. These residues are less disordered, more stable and more hydrophobic when compared to the disordered residues that surround them in a protein chain. This is in line with the observations in (Meszaros et al., 2007) that the local increase in the hydrophobicity in a disordered segment is a characteristic of binding sites in IDPs. Bastolla *et al.* (2005) show a strong positive correlation between a hydrophobicity profile and a contact matrix (i.e., the residue-residue contacts that quantify stability), which describes stability of the protein structure. This supports our result that also shows that increase in both stability and hydrophobicity are indicative of MoRF residues. Importantly, our model shows that these markers can be derived directly from the sequence, based on the predicted (with IUPred and DISOPRED2) disorder and the two AA scales (Zhou & Zhou 2004; Nozaki & Tanford, 1971)

4.5 Case studies

4.5.1 Case studies for true positive predictions

Two case studies are used to demonstrate the MoRFpred predictions. They were selected to represent two situations, when the MoRF region are overpredicted and where they are underpredicted. Moreover, the first case concerns a long MoRF segment, while the second concerns a short segment.

The first case study is the *transcriptional intermediary factor-2 isoform 2* protein which was collected from UniProt, and for which the MoRF was extracted using the PDB complex *1m2z_B*. This protein has 1394 residues and contains a coil MoRF region which is 21 residues long.

Figure 13 visualizes predictions for this protein from all considered predictors: α-MoRF-PredI, α-MoRF-PredII, ANCHOR, and MoRFpred. We observe that all methods were able to (partially) identify the native MoRF region. However this is not the only predicted MoRF and the ammount of MoRFs is overpredicted by all methods. MoRFpred has the least predicted MoRFs by predicting 89 residues as MoRFs when compared to α-MoRF-PredI with 171 MoRF predictions, α-MoRF-PredII with 306 MoRF predictions, and ANCHOR with 876 MoRF predictions.

The second case study is a 89 residues long *H2A class histone* protein for which MoRF region was extracted from the *1ydp_P* complex from PDB. The native MoRF region in this protein folds into a coil, which is located near the C-termini and is 9 residues long. Figure 14, shows that α-Morf-PredI and α-Morf-PredII did not predict any MoRF residues in this sequence, which is correct since these methods are designed to predict α-MoRFs.

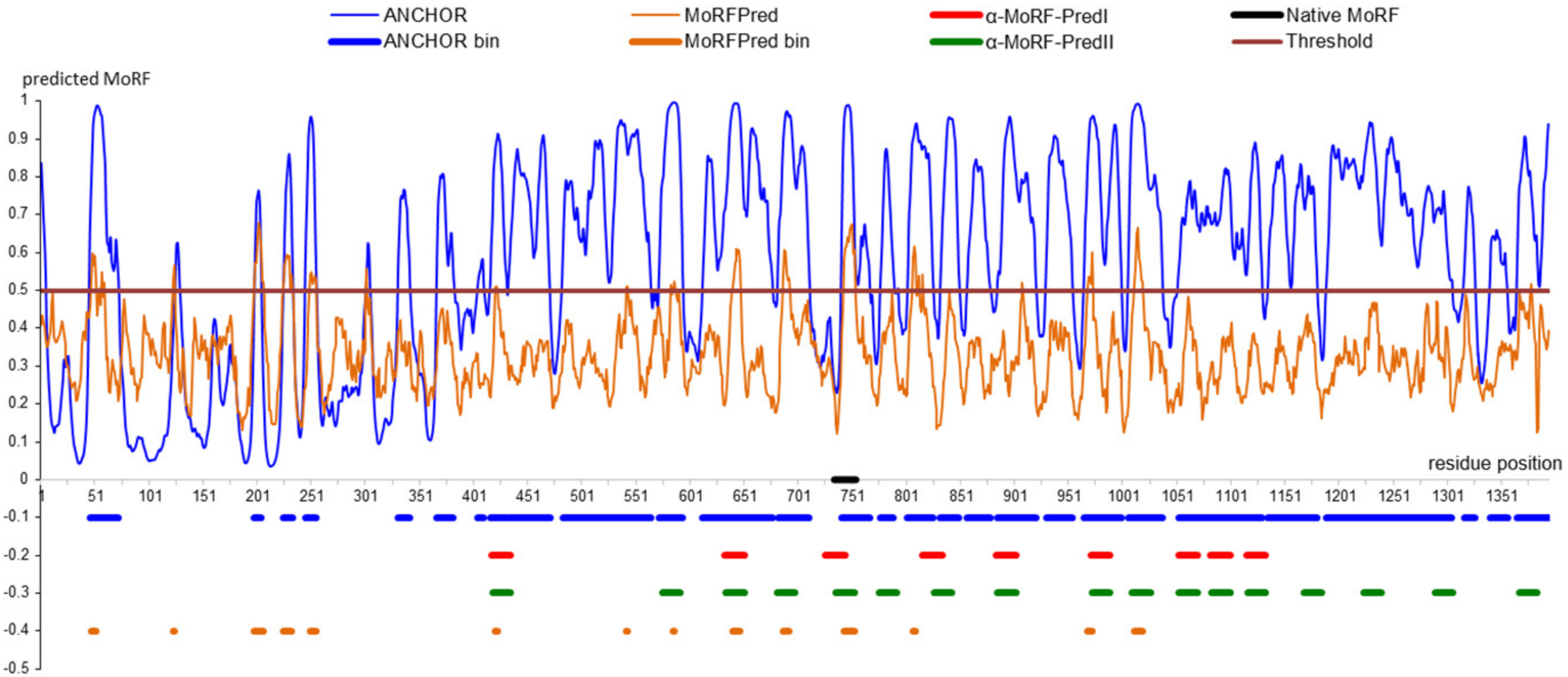

Figure 13. Prediction of MoRF residues for the transcriptional intermediary factor-2 isoform 2 protein. ANCHOR (blue lines), MoRFpred (orange lines), α-MoRF-PredI (thick red line), and α-MoRF-PredII (thick green line) predictors. Probability values are only available for ANCHOR and MoRFpred and are shown by thin blue and orange lines, respectively, at the top of the figure. The original cut-off of 0.5 for both ANCHOR and MoRFpred are shown using a brown line. The native MoRF regions are annotated using black horizontal line. The binary predictions from ANCHOR, α-MoRF-PredI, α-MoRF-PredII and MoRFpred are denoted using horizontal lines at the bottom of the figure in blue (at the -0.1 point on the *y*-axis), red (at the -0.2), green (at the -0.3), and orange (at the -0.4), respectively.

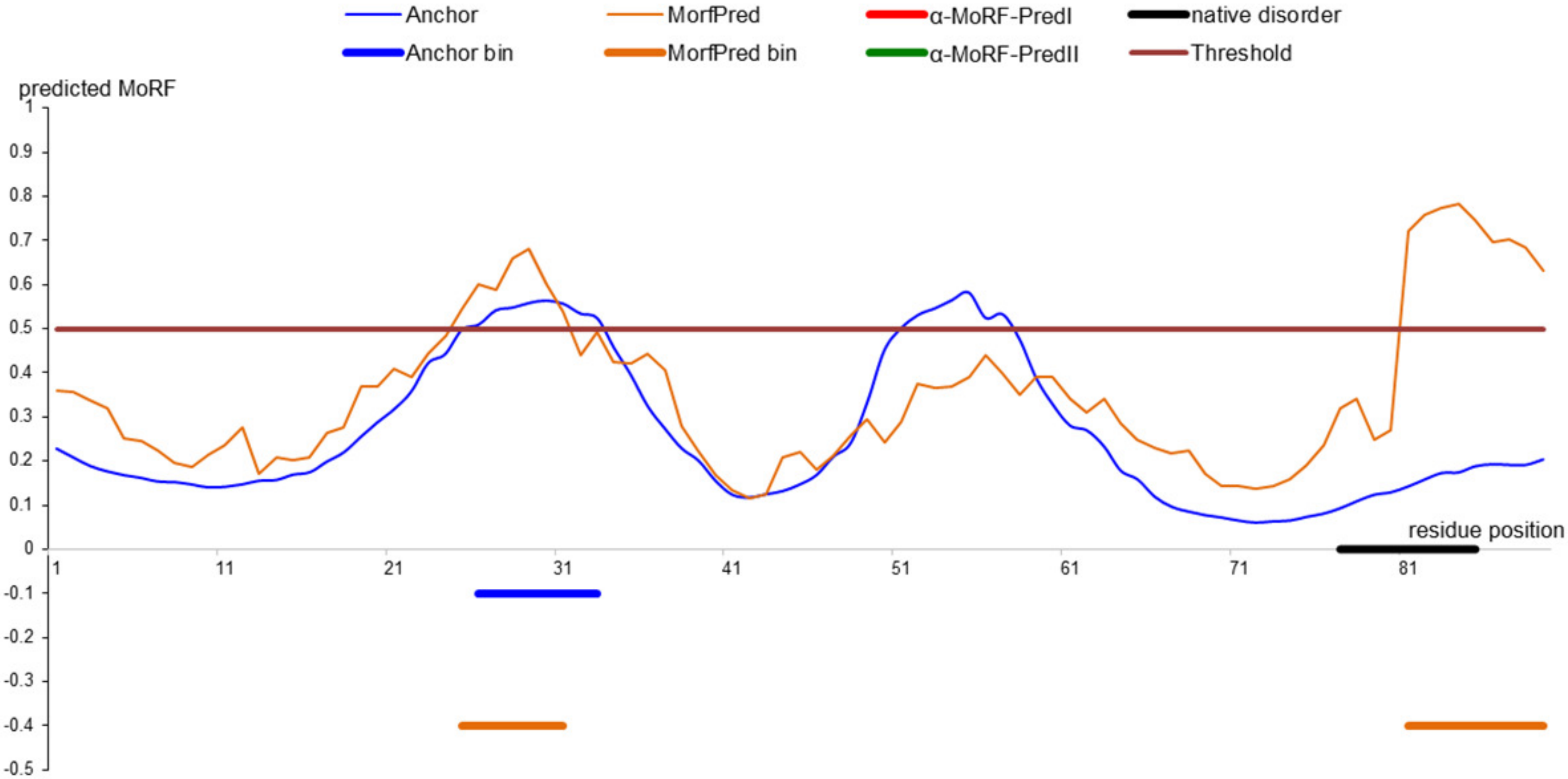

Figure 14. Prediction of MoRF residues in the Histone H2A protein. ANCHOR (blue lines), MoRFpred (orange lines), α-MoRF-PredI (thick red line), and α-MoRF-PredII (thick green line) predictors. Probability values are only available for ANCHOR and MoRFpred and are shown by thin blue and orange lines respectively. The original cut-off of 0.5 for both ANCHOR and MoRFpred are shown using a brown line. The native MoRF regions are annotated using black horizontal line. The binary predictions from ANCHOR, α-MoRF-PredI, α-MoRF-PredII and MoRFpred are denoted using blue (at the -0.1 point on the y-axis), red (at the -0.2), green (at the -0.3), and orange (at the -0.4) horizontal lines. No red and green lines means that , α-MoRF-PredI and α-MoRF-PredII predictors did not predict any of the residues as MoRF.

MoRFpred predicted the native MoRF region and another false positive MoRF region, which was also predicted by ANCHOR. We note that the probability profiles of ANCHOR and MoRFpred are very similar except for the C-terminus where the native MoRF is located. The ANCHOR outputs probabilities that are higher than the 0.5 threshold in a vicinity of 50th position, but these predictions were removed through post-processing applied by this method. MoRFpred generated probabilities < 0.5 in that region.

We note that the MoRFpred predictions in these two case studies were generated by SVM (i.e., alignment did not find any MoRF regions), which confirms that the machine learning classifier can contribute beyond what can be found based on sequence similarity.

4.5.2 **Case studies for false positive predictions**

In section 2.3.2 we argue that our dataset may contain unannotated MoRFs and thus some of the false positive MoRF predictions generated by MoRFpred might correspond to true/native MoRF regions. In section 4.2 we demonstrate that the predicted MoRF regions that have no overlap with the native MoRFs (false positive predictions) have above-random similarity to the native MoRFs. This lead us to investigate the strongest false positive MoRF predictions, i.e., predictions with the highest probability (see section 4.3). We use average (over all residues in the predicted segment) probability generated by MoRFpred to rank the false positive MoRFs. We used UniProt to find annotations of binding sites in these predicted regions. The following two case studies (among others) have binding regions in the predicted MoRF segments.

The first case is *P-selectin glycoprotein ligand 1* (PSGL-1) protein (UniProt ID Q14242) for which the predicted false positive MoRF has average probability of 0.85. The AA sequence of this region (residues 393 to 402) is DDLTLHSFLP. This region is predicted by the SVM, and was not found by alignment. It implements a few interaction sites:

- A part of the MAPK docking motif (REDREGDDLTL, residues 387-397) that helps to regulate a specific interaction in the MAPK cascade overlaps with the predicted MoRF region.
- The predicted MoRF region includes a site phosphorylated by Polo-like kinase (DDLTLHS, residues 393-399).
- Our prediction is also close to the TRAF6 binding site (PEPREDREG, residues 384-392), that acts as intracellular adaptor recruited to different receptors through its C-terminal TRAF domain.

We note that ANCHOR overpredicts half of this protein as MoRF, which includes this region as well. The α-MoRF-PredII predicts this region as MoRF without overpredicting the remainder of the chain. This region was also predicted to be disordered by MFDp and IUPredS, and parts of this region were predicted by the other disorder predictors.

The second case is *putative uncharacterized protein DKFZp459P0162* (UniProt ID Q5RDR1) for which the false positive MoRF was predicted with average probability of 0.65. The AA sequence for this region (residues 212 to 222) is SPAVPNKEVTP, and it is associated with the following binding sites:

- This region covers the subtilisin/kexin isozyme-1 (SKI1) cleavage site (KEVTP, residues 218-222)

- The PAVPNK sub-segment (residues 213-218) is potentially recognized by class II SH3 domains and is involved in protein-protein interaction mediated by SH3 domains.

In contrast to the first case, this region is predicted by alignment and none of the existing predictors were able to (fully) predict this region. ANCHOR, which predicts about 1/3 of this protein as MoRF, predicts only parts of this region. The considered disorder predictors predict this region as being disordered, which provides further support for our claim that this is a strong putative MoRF.

These two case studies demonstrate that some of the false positives generated by MoRFpred may implement important binding events that require a structured conformation.

5 Summary, discussion and contributions

We introduce a new sequence-based method to predict MoRF segments, including all of their types. Our solution is based on several novel aspects. First, we utilize an updated, larger and more comprehensive dataset to build and validate MoRFpred. Second, we combine SVM-based predictions with alignment, which leads to improved predictive quality. Third, MoRFpred predictions are accompanied by probability scores which can be used to indicate more accurate predictions. Last, we use a more comprehensive set of predictive inputs when compared to existing methods: Specifically, we utilize multiple disorder predictions, predicted B-factors and RSA, evolutionary profiles based on the PSSM, and amino acid indexes to encode our inputs. We also designed a new and successful class of features to contrast the properties in the immediate neighborhood of the predicted residues with its flanking regions. Analysis of our input features shows that MoRFs are characterized by dips in disorder predictions and certain hydrophobicity- and stability-based profiles, i.e., MoRF residues have higher hydrophobicity and stability when compared to the adjacent residues.

We also hypothesize that our method can be used to identify putative MoRFs. We investigate our false positives and we show that some of them could potentially be native MoRFs. Similarity analysis shows that false positive MoRF regions predicted by MoRFpred are characterized by above-random similarity with the native MoRF regions. We used this observation, along with the fact that higher probability generated by MoRFpred corresponds to more accurate predictions, to identify and discuss a couple of interesting case studies. Finally, the similarity analysis also led us to incorporate alignment into the proposed predictor, which improved the AUC by 1%.

The following is a list of significant contributions in this work:

- We designed and developed a new sequence-based predictor that outperforms existing MoRF prediction methods.

- We adopted a new measure, success rate, which has not previously been previously used in this field, to evaluate the MoRF predictions.

- We devised a new evaluation and test method to avoid overfitting during feature selection and parameterization. This method is referred to as 4+1 cross validation.

- We provided an empirical comparison with the existing MoRF predictors.

- We identified and explained several sequence-derived markers of MoRF regions. These markers are based on hydrophobicity, stability and disorder profiles.

- We show that MoRF segments have an above-random similarity to each other, and thus it is possible to use alignment to identify some (a limited number) of the MoRF regions.

References

1. Altschul, S., Madden, T., Schäffer, A. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, **25**, 3389-3402.
2. Bastolla, U., Porto, M., Roman, H.E.,Vendruscolo, M. (2005) Principal eigenvector of contact matrices and hydrophobicity profiles in proteins., *Proteins*, **58**, 22-30.
3. Berman, H., Henrick, K., Nakamura, H., Markley, J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.,* **35**, D301-3.
4. Carolina, N. (1971) The Solubility of Amino in Aqueous Ethanol Acids and Two Glycine Dioxane Solutions Peptides, *Amino Acids*, **246**.
5. Chen, P., Wang, B., Wong, H.-S., Huang, D.-S. (2007) Prediction of protein B-factors using multi-class bounded SVM, *Protein Pept. Lett.*, **14**, 185-190.
6. Cheng, Y., Oldfield, C.J., Meng, J., Romero, P., Uversky, V.N., Dunker, A.K. (2007) Articles Mining R-Helix-Forming Molecular Recognition Features with Cross Species Sequence Alignments †, *Neural Networks*, 13468-13477.
7. Doszta, Z. (2009) Prediction of Protein Binding Regions in Disordered Proteins, *PLoS Computational Biology*, **5**.
8. Dosztányi, Z., Csizmok, V., Tompa, P., Simon, I. (2005) IUPred: web server for the pre-diction of intrinsically unstructured regions of proteins based on estimated en-ergy content, *Bioinformatics*, **21**, 3433-3434.
9. Dosztányi, Z., Mészáros, B., Simon, I. (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins, *Bioinformatics (Oxford, England)*, **25**, 2745-2746.
10. Dunker, A.K. and Kriwacki, R.W. (2011) The orderly chaos of proteins, *Sci Am*, **304**, 68-73.

11. Ernest, C., Davenport, J., El-Sanhurry, N.A. (1991) Phi/Phimax: Review and Synthesis. 1991 51:821, *Educational and Psychological Measurement* **51**, 821-828.
12. Fan, R.E., Chang, K.W., Hsieh, C.J. (2008) LIBLINEAR: A library for large linear classification, *J Mach Learn Res*, **9**, 1871-1874.
13. Faraggi, E., Xue, B., Zhou, Y. (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by fast guided-learning through a two-layer neural network, *Proteins*, **74**.
14. Gunasekaran, K., Tsai, C.J., Nussinov, R. (2004) Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J Mol Biol*., **341**,1327-41
15. He, B., Wang, K., Liu, Y., Xue, B., Uversky, V., Dunker, A. (2009) Predicting intrinsic disorder in proteins: an overview, *Cell Res.*, **19**, 929-949.
16. Huang, Y., Niu, B., Gao, Y., Fu, L., Li, W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences, *Bioinformatics*, **26**, 680-682.
17. Jain E., Bairoch A., Duvaud S., Phan I., Redaschi N., Suzek B.E., Martin M.J., McGarvey P., Gasteiger E. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website, *BMC Bioinformatics*, 10(136
18. Jones, D. and Swindells, M. (2002) Getting the most from PSI-BLAST, *Trends Biochem Sci*, **27**, 161-164.
19. Jones, S., Stewart, M., Michie, A., Swindells, M., Orengo, C., Thornton, J. (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis, *Protein Sci*, **7**, 233-242.
20. Jones, S. and Thornton, J. (1996) Principles of protein-protein interactions, *Proc Natl Acad Sci U S A*, **93**, 13-20.

21. Kabsch, W. and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers*, **22**, 2577-2637.

22. Kurgan, L. and Disfani, F.M. (2011) Structural protein descriptors in 1-dimension and their sequence-based predictions, *Current Protein and Peptide Science*, **12**, 470-489.

23. McGuffin, L. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models, *Bioinformatics*, **24**, 1798-1804.

24. Mészáros, B., Simon, I., Dosztányi, Z. (2011) The expanding view of protein-protein interactions: complexes involving intrinsically disordered proteins, *Physical biology*, **8**, 035003.

25. Meszaros, B., Tompa, P., Simon, I., Dosztanyi, Z. (2007) Molecular principles of the interactions of disordered protiens, *J. Mol. Biol.*, **372**.

26. Midic, U., Oldfield, C.J., Dunker, a.K., Obradovic, Z., Uversky, V.N. (2009) Protein disorder in the human diseasome: unfoldomics of human genetic diseases, *BMC genomics*, **10 Suppl 1**, S12.

27. Mizianty, M.J., Stach, W., Chen, K., Kedarisetti, K.D., Disfani, F.M., Kurgan, L. (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources, *Bioinformatics (Oxford, England)*, **26**, i489-496.

28. Mohan, A., Oldfield, C.J., Radivojac, P., Vacic, V., Cortese, M.S., Dunker, a.K., Uversky, V.N. (2006) Analysis of molecular recognition features (MoRFs). *Journal of molecular biology*, **362**, 1043-1059.

29. Nozaki, Y. and Tanford, C. (1971) The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale, *Journal of Biological Chemistry*, 2211-2217.

30. Oldfield, C.J., Cheng, Y., Cortese, M.S., Romero, P., Uversky, V.N., Dunker, a.K. (2005) Coupled folding and binding with alpha-helix-forming molecular recognition elements., *Biochemistry*, **44**, 12454-12470.

31. Pearson, W. and Lipman, D. (1988) Improved tools for biological sequence comparison, *Proc Natl Acad Sci U S A*, **85**, 2444-2448.

32. Rice, P., Longden, L., Bleasby, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite, Trends *in Genetics*, **16**, 276-277.

33. Rubinstein, N.D., Mayrose, I., Martz, E., Pupko, T. (2009) Epitopia: a web-server for predicting B-cell epitopes, *BMC bioinformatics*, **10**, 287.

34. Schlessinger, A., Yachdav, G., Rost, B. (2006) PROFbval: predict flexible and rigid residues in proteins, *Bioinformatics*, **22**.

35. Sickmeier, M., Hamilton, J., LeGall, T., Vacic, V., Cortese, M., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V., Obradovic, Z., Dunker, A. (2007) DisProt: the Database of Disordered Proteins, *Nucleic Acids Res*.

36. Uversky, V.N. (2011) Intrinsically disordered proteins from A to Z., *The international journal of biochemistry & cell biology*, **43**, 1090-1103.

37. Uversky, V.N. and Dunker, A.K. (2010) Understanding protein non-folding, *Biochimica et Biophysica Acta - Proteins and Proteomics*, **1804**, 1231-1264.

38. Uversky, V.N., Oldfield, C.J., Midic, U., Xie, H., Xue, B., Vucetic, S., Iakoucheva, L.M., Obradovic, Z., Dunker, A.K. (2009) Unfoldomics of human diseases: linking protein intrinsic disorder with diseases, *BMC genomics*, **10** S7.

39. Vacic V, O.C., Mohan A, Radivojac P, Cortese MS, Uversky VN, et al. (2007) Characterization of molecular recognition features, MoRFs, and their binding partners, *J Proteome Res*, **6**, 2351-2366.

40. Ward, J., McGuffin, L., Bryson, K. (2004) The DISOPRED server for the predic-tion of protein disorder, *Bioinformatics*, **20**.

41. Zhou, H. and Zhou, Y. (2004) Quantifying the effect of burial of amino acid residues on protein stability, *Proteins*, **54**, 315-322.