A Taxonomy of protein binding problems

(Research Exam)

Fayyaz ul Amir Afsar Minhas
Department of Computer Science
Colorado State University
Fort Collins, Colorado, 80523-1873
fayyazafsar@gmail.com

The understanding of protein interactions with other molecules and ions through the identification of the binding interfaces involved in them, is of critical importance in biology and medicine. In this research exam, the focus is on developing a taxonomy of protein binding problems and on computational approaches for identifying different types of protein interfaces involved in them. In particular, four different types of protein interfaces are considered: protein-protein, protein-nucleic acid (DNA & RNA) and protein-metal ions. Similarities and differences in the nature and characteristics of binding interfaces in these protein interactions are discussed and a review of various computational methods for their identification is presented.

List of Initial Papers

1. S. Ahmad and K. Mizuguchi, "Partner-Aware Prediction of Interacting Residues in Protein-Protein Complexes from Sequence Data," PLoS One, vol. 6, no. 12, Dec. 2011.
2. Y. Qi, M. Oja, J. Weston, and W. S. Noble, "A Unified Multitask Architecture for Predicting Local Protein Properties," PLoS ONE, vol. 7, no. 3, p. e32235, Mar. 2012.
3. Y. Ofran, V. Mysore, and B. Rost, "Prediction of DNA-Binding Residues from Sequence," Bioinformatics, vol. 23, no. 13, p. i347–i353, Jul. 2007.
4. H. Kazan, D. Ray, E. T. Chan, T. R. Hughes, and Q. Morris, "RNAcontext: A New Method for Learning the Sequence and Structure Binding Preferences of RNA-Binding Proteins," PLoS Comput Biol, vol. 6, no. 7, Jul. 2010.
5. A. Passerini, M. Lippi, and P. Frasconi, "Predicting Metal-Binding Sites from Protein Sequence," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 1, pp. 203–213, Feb. 2012.

Committee

- Dr. Asa Ben-Hur, Advisor, Dept. of Computer Science, Colorado State University.
- Dr. Charles Anderson, Dept. of Computer Science, Colorado State University.
- Dr. Bruce Draper, Dept. of Computer Science, Colorado State University.

1 Introduction

Proteomics is the study of structure and function of proteins and it has emerged as a very important branch of biology with strong linkages to other branches of science owing to the vast variety of problems being studied in it. Proteins are of particular interest to biologists because they form the functional backbone of innumerable biologically important processes. The role proteins play in cellular functions can be appreciated by considering that approximately 50% of the dry weight of the human body is protein [1]. Another example that illustrates the importance of proteins is that of human blood which consists of red and white blood cells. 92% of the dry weight of all red blood cells is a protein called hemoglobin [2]. Hemoglobin gives the blood its distinct red color and is involved in the transfer of oxygen to cells which is absolutely critical for survival. Other functions of proteins include but are not limited to: enzymatic processing, cell signaling, transportation, immunological responses, muscular contractions, hormonal processes, structural stability, gene expression control [3]. Almost all protein functions are possible only through the interaction or binding of proteins with other molecules such as other proteins, nucleic acids, ligands and metal ions. For example, the ability of hemoglobin to bind to iron allows it to transport oxygen.

Study of protein binding is important in understanding protein function and disease mechanisms. It is also important for drug design, discovery and effectiveness studies. When proteins bind other molecules or ions, they do so at specific interfaces called binding sites on the proteins. These interfaces are the focus of this research exam. In particular, it discusses different approaches for computational prediction of binding sites on proteins in their interactions with the following:

1. Other proteins
   Proteins can bind to other proteins resulting in protein-protein interactions or protein complexes. The problem of identifying protein-protein binding interfaces is examined in relatively more detail in this research exam in comparison to other types of interfaces.
2. Nucleic Acids
   Specific methods for prediction of binding interfaces between proteins and nucleic acids (Ribonucleic acid (RNA) and Deoxyribonucleic acid (DNA)) molecules are also considered. Of particular interest in this area are the so called transcription factor proteins which bind to DNA and regulate the process of transcription. Transcription is the process through which the DNA sequence of a gene is expressed as RNA. In the context of nucleic acids, binding sites on both the binding protein and the nucleic acid sequence are of interest to biologists. In this exam, we discuss computational prediction approaches for both kinds of binding sites.
3. Ligands and Metal Ions
   Ligands are substances (usually small molecules) that bind to a biomolecule such as a protein. Examples of ligands include lipids, nucleic acids or drug-like molecules such as penicillin. This research exam targets specific methods that predict the binding interfaces of proteins to metal ions such as iron or zinc.

The rest of the research exam is organized as follows: Section-2 details the biological basis of protein interactions and interfaces. Such information is critical to understanding how proteins interact and what

kind of features can be used to predict interfaces involved in such interactions. It also renders more motivation for developing automated binding site prediction techniques for protein interactions. Section-3 presents a detailed view about the specific nature of different types of protein interactions and interfaces listed above. It explains a number of existing techniques used for predicting interfaces in different types of interactions. The methods discussed in section-3 have been selected on the basis of their accuracy and conceptual novelty. Shortcomings of different methods are also detailed in this section. Section-4 identifies general issues and open problems in the field.

2 Nature of protein binding interfaces

Proteins are composed of one or more polypeptide chains having a complex three dimensional structure. Proteins consist of an arbitrary-length sequence of 20 amino acids. The structure of a protein can be viewed at four different levels (primary, secondary, tertiary and quaternary) as shown in Figure 1. This figure also illustrates the concepts of domains and motifs in proteins.

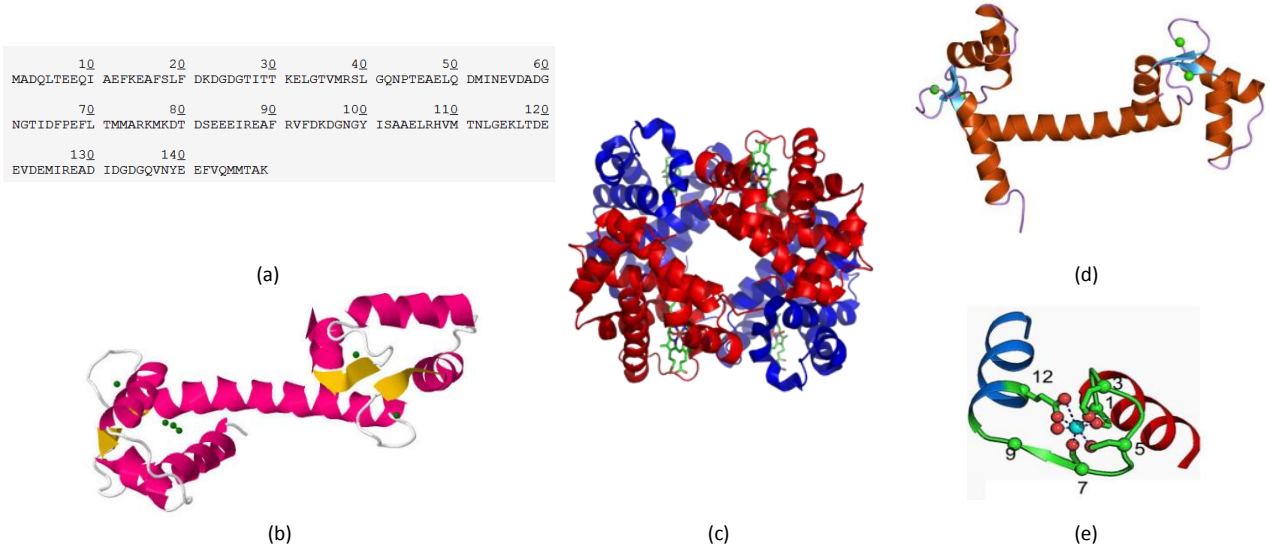

Figure 1 Different concepts related to the structure of a protein. Images are taken from the Protein Data Bank (PDB) [4]. (a) Primary structure of the protein Calmodulin. Primary structure is the sequence of amino acids (or residues) of a protein. The amino acids are held together by peptide bonds (not shown) and as a consequence a protein is sometimes called a polypeptide. (b) The secondary and tertiary structures of Calmodulin (PDB ID: 1CLL). Neighboring amino acids in the protein sequence interact with one another through hydrogen bonds (not shown) and give rise to secondary structures such as α-helices, β-sheets and loops shown in red, yellow and off-white, respectively. Secondary structures in a protein chain arrange themselves in a configuration called the tertiary structure of the protein. Folding of the protein into its tertiary structure allows residues that are not sequence neighbors to come close to each other spatially. Note that all residues in the tertiary structure are linked to one another and this allows a protein chain to be represented by its sequence. (c) Multiple protein chains can interact with one another to give rise to the quaternary structure of a protein (also called a protein complex). Shown in (c) is the quaternary structure of hemoglobin which consists of 4 polypeptide chains. (d) Proteins are composed of domains which are units in the protein that can evolve, function and exist independently of the rest of the protein chain. Shown in (d) is the EF-Hand domain which is found in Calmodulin and other calcium binding proteins such as LCP1. Proteins that contain the same domain are said to belong to the same family. (e) Structural motifs are structurally similar constructs that can occur in different proteins. Unlike domains, structural motifs cannot exist independently. Shown in (e) is the EF-Hand calcium (shown in cyan) binding motif. Calmodulin contains four of these motifs and as a consequence can bind up to four calcium ions. It should be noted that structural motifs are different from sequence motifs which are biologically significant amino acid (for proteins) or nucleotide (for DNA and RNA) sequences patterns. For example, the consensus sequence of the EF-hand calcium binding structural motif shown in (e) is ExxxxxxxxDx[DN]x[SDN]Gx[LVI]x[ESD]xxE where 'x' denotes a don't-care position and residues in square brackets can substitute one another at the same location. Such consensus sequences or sequence logos are used to represent sequence motifs (see Figure 7). For more information on protein structure, the interested reader is referred to [3]. In this figure, a cartoon representation of proteins has been shown. An alternate visualization scheme is the space filling view of Figure 2.

As stated earlier, most of the molecular functions of proteins are made possible only through their interactions with other molecules and ions. In order to fully appreciate the importance of identifying the sites of these interactions on proteins, complications involved in their computational prediction and the role that computational methods can play in this area, a thorough description of molecular recognition or binding in proteins is warranted.

In order to explain the basis of molecular recognition in proteins with respect to their structural and physiochemical properties, a number of models have been proposed. One of the first such models is the "lock-and-key" model [5] which dictates that structural shape complementarity between the protein and the binding substance is essential for binding to occur. However, this model is too restrictive as shape complementarity, although playing an important role, is not the sole cause of binding [6].

A more realistic view is provided by the "induced-fit" model which states that shape complementarity plays an important role in binding but the binding process is also driven by usually non-covalent intermolecular forces (such as van der Waals interactions, hydrophobic effects, Hydrogen bonding etc.) resulting from the interaction between the protein and the binder. It also states that the binding process can cause conformational changes in the protein leading to an induced fit of the binder in the protein [7]. Thus, along with shape complementarity, complementarity in the physiochemical properties of a binder and its target protein determines binding. For example, hydrogen bond donors in a protein occur opposite to hydrogen bond acceptors in the binder, non-polar groups occur opposite to other non-polar groups and positive charges occur opposite to negative charges and so on. The induced fit model also explains the fact that not all intermolecular interactions occur between pre-existing complementary surfaces. One or both of the molecules involved in the interaction can undergo structural conformational changes to make binding possible. Thus, the task of predicting binding sites in proteins translates to identifying areas in a protein where such complementarities between the protein and its binder either pre-exist or can result from a binding-triggered conformational change. A more detailed model for binding is the conformational selection model discussed in [8].

Figure **2** renders an example of molecular recognition in proteins. Here, the protein Calmodulin is shown to bind to a protein (Edema Factor) from the Anthrax bacteria [9]. This binding is possible because of a conformational change in both proteins that causes their hydrophobic residues to bind. Calmodulin is an important calcium signaling protein involved in a variety of functions in the human body ranging from neuronal spiking to muscular contractions and its binding to the Edema Factor disrupts its functionality. This interaction is one of the mechanisms through which Anthrax affects cells in the human body. Another similar example is the ultimate destruction of a human immune cell by the binding of its CD4 protein with the Human Immunodeficiency Virus's (HIV) gp120 protein [10]. Characterization of the interface between these proteins can help in developing treatment drugs such as inhibitors and antagonists or to introduce targeted mutations aimed at making this binding infeasible. These examples not only illustrate the importance of studying protein binding but they also underline how the identification of binding sites on proteins is critical to understanding biological processes.

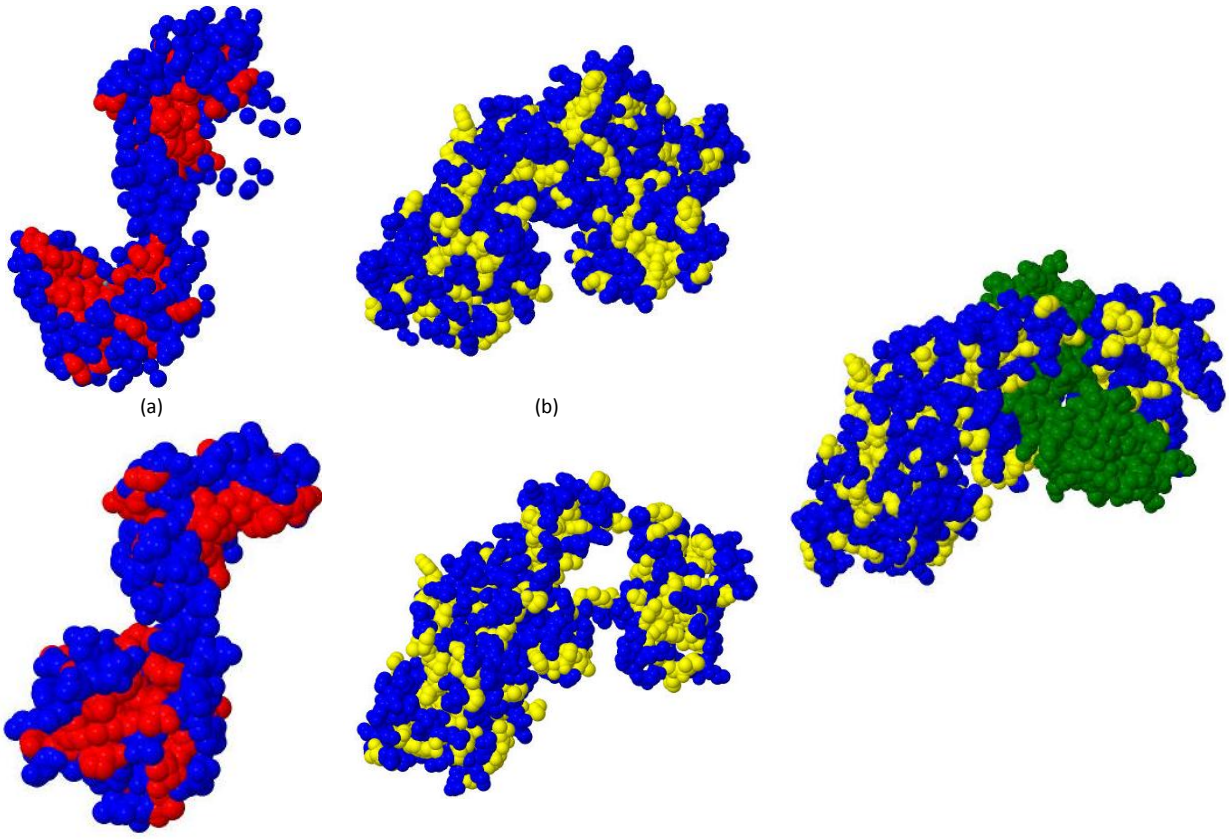

Figure 2 Binding of Calmodulin to the Edema Factor protein from Anthrax bacteria. (a) Unbound 'S' like structure of Calmodulin shown with its hydrophobic residues in red (PDB Entry 1CLL). These hydrophobic surfaces do not like to stay in contact with water. (b) The unbound structure of Edema Factor from Anthrax bacteria with its hydrophobic residues shown in yellow (PDB entry 1K8T). (c) The complex (bound) structure of Calmodulin (in green) with the Edema Factor (yellow and blue) (PDB entry 1K93) (d) The bound structure of Calmodulin from (c) with the Edema Factor removed from view. (e) The bound structure of the Edema Factor from (c) with Calmodulin removed from view. Notice the difference in conformation between (b) and (e). Also notice (from (c) and (e)) that the hydrophobic residues of Calmodulin bind to the hydrophobic residues of the Edema Factor. These figures were created using Jmol [11].

2.1 Finding binding sites

A number of experimental approaches can be used to find binding interfaces on proteins. If the three dimensional structure of a complex is known (for instance, through crystallography and Nuclear Magnetic Resonance studies), then binding interfaces can be found using distance thresholds and changes in Accessible Surface Area (ASA) as explained below.

ASA is the area of the protein that is accessible to a solvent [12]. The residues whose ASA decreases upon binding can be considered to be part of the protein interface [13]. For example, interacting residues in the protein-protein interaction database InterPare [14] are defined to be the ones whose ASA changes by more than 1 $Å^2$. However, this definition ignores the fact that during the process of binding, different conformational changes can lead to changes in ASA of residues which are not a part of the physical binding site. Another consequence of such conformational changes is that some buried residues (with small ASA) will move to the interface and, in effect, increase their ASA.

Another definition of binding interfaces derived from the 3D structure of a complex is based upon the distance between interacting components (residues, atoms or nucleotides). Amino acids on a protein (or

nucleotides on a nucleic acid) that are within 4-6 Å of the interacting substance are defined to be part of the binding interface as a bond is assumed to exist between them. This definition allows for different types of interactions [13] that operate over different distances (covalent bonds, disulfide bonds, salt bridges, hydrogen bonds, Vander-Waals forces) between the interacting molecules [3] and also for the variability in sizes of the protein residues themselves [15].

Different methods in the literature use different techniques to define what qualifies as a binding site. Most methods use a simple distance based threshold. Other approaches for determining binding sites include Voronoi Tesselations [14], setting different distance cut-offs for different amino acids etc. However, the differences between interaction sites defined by all these methods is fairly small [14]. One the other hand, the thresholds used in a method can potentially play a significant role. Gromiha et al. [16] have used inter-residue interaction energy to define binding interfaces. They have found that, for their data set of 306 proteins, only 28% of residues were common between the sets of interacting residues defined using interaction energy and those found through a distance criteria. They have pointed out that using the distance based definition of interacting residues, 5.7% of residues have strong repulsive energies and may not be truly interacting.

Different types of biological assays are employed in the experimental determination of protein binding and identification of binding sites. Since these assays are not the focus of this research exam, therefore relevant biological techniques are only referenced here without a detailed description. For protein-protein interactions, methods such as Yeast Two-Hybrid (Y2H) [17], Affinity Purification/Mass Spectrometry (APMS) [18], Protein micro-arrays [19] etc. are employed. The binding site of a Protein-RNA interaction on the RNA can be detected using Crosslinking and Immunoprecipitation (CLIP) [20] and its derivatives [21], [22], etc. Assays for identification of Protein-DNA interactions include Chromatin Immunoprecipitation (ChIp) and its derivatives [23]. The locations of the binding sites for ligands can be determined using crystallographic data, nuclear magnetic resonance or from mutagenesis experiments [24].

Most of these biological assays are time consuming and expensive to perform. Computational approaches present a fast and inexpensive tool for the prediction of putative binding sites. Such computational approaches can also help in performing these biological experiments by narrowing down possible binding sites.

2.2 Characteristics of binding interfaces

Some of the general features that distinguish binding sites from non-binding sites in proteins (for all types of interactions) include structure, physiochemical properties, evolutionary conservation etc. In this section, similarities and differences between the characteristics of binding vs. non-binding regions among different categories of protein interactions are discussed. Information contained in this section is expected to help explain the choice of features made in different computational binding interface prediction methods presented in the rest of the research exam. However, it should also be noted that different proteins can have very different binding interface characteristics.

2.2.1 Structural properties

Interfaces between proteins and smaller substrates (ligands) are typically cavities and concave clefts. The surface area of proteins buried in these types of interactions is usually smaller (up to a few hundred Å^2) in comparison to the flatter, much larger (1200-2000 Å^2) and more structurally intricate surfaces involved in protein-protein interactions [25]. It should be noted that residues forming a binding site in a protein may not be contiguous in sequence. This holds for most types of protein interactions [3], [26].

In terms of secondary structure, binding areas on proteins tend to favor β-sheets in comparison to α-helices. Moreover, loops in interfaces tend to be longer [27]. The sterically unfavorable right handed γ-helix occurs in protein binding interfaces with high-specificity .

Most of the time, the residues involved in forming an interaction lie on the surface of a protein which permits the use of structural properties (from the unbound state of the protein) such as solvent accessible surface area (ASA) [28][29], Protrusion Index [30], Depth Index [30] to be employed for prediction of binding sites. However, at times, the binding process itself can result in a conformational change in the protein leading to changes in the degree of exposure of different residues in the protein to the surface [15]. Features based on these conformational changes have been utilized in [30] for the prediction of hotspot residues (see section 2.2.3).

Post Translational Modifications[1] (PTMs) can affect the structure and binding characteristics [32],[33] of a protein. For example, phosphorylation of a target residue in or near a Calmodulin (CaM) binding domain in a protein can change the affinity of the protein to bind CaM [31]. However, most of the existing methods for binding site prediction do not consider the role PTMs play.

Structure can be employed as a good predictor of protein binding sites but there is an exponentially growing gap between the number of known structures [34] (which is much smaller) and the number of proteins with known sequence data. Therefore, using structure limits the applicability of a protein binding site prediction approach. Some of the structural descriptors used in computational methods for binding site prediction are [15]: neighbor list (residues that lie in spatial vicinity to the residue in question), ASA (can be calculated from protein structure using programs like DSSP [35]), relative ASA (ASA expressed as a fraction of the overall surface of the residues that is exposed to the solvent), residue interface propensity [36] B-factor (approximates the flexibility of a residue), secondary structure (what kind of secondary structure does a residue occur in), sequence distance (in number of residues) between residues that are structurally contiguous etc. These descriptors can either be calculated from the known unbound structures of a protein or be estimated from its predicted or approximated structure using methods that predict structure from sequence such as [37]. A theoretical problem with using unbound structures for training a binding predictor is that proteins can undergo significant structural changes upon binding. Thus, structure based methods trained on known complexes can fail to identify the binding sites in many unbound structures and this particular true if the size of training data

[1] PTMs [31] are chemical modifications of a protein after its creation through the process of translation within the cell. PTMs extend the range of possible functions that a protein can perform or change its behavior by attaching other biochemical functional groups to it such as acetates and phosphates.

is small. Prediction of protein tertiary structure may not always be feasible (e.g., if the protein does not have related proteins (homologs) with known structure).

2.2.2 Sequence properties

The composition and propensity of different residues has been observed to be different in binding areas in comparison to non-binding areas. Such differences were employed as features in [38] for protein-protein interaction site prediction. For example, the propensity of Tryptophan (W) to occur in protein-protein interfaces was found to be higher than that of Alanine (A). Protein-DNA binding sites tend to be highly enriched in positively charged Arginine (R) and Lysine (K) and lack negatively charged Aspartic acid (D) and Glutamic acid (G) [39]. The characterization of these differences using a variety of feature representations allows for sequence based prediction of protein interfaces.

2.2.3 Hotspots

All residues in a binding interface do not contribute equally to the binding energy [40]. For example, in a typical 1200-2000 $Å^2$ less than 5% of interface residues contribute more than 2kcal/mol to binding. In small interfaces, this can be related to a single residue. These residues form what are called hotspots in protein interfaces and they can be of great importance in understanding protein interactions. They can also be desired drug targets [41]. However, only a few existing approaches target the interface identification problem at the hot-spot level [10],[42],[43]. Hotspots can be identified by measuring the decrease in binding affinity between a protein and its binder upon mutating residues within the known binding site [44]. Typically, the residues are mutated to Alanine and the process is called Alanine scanning mutagenesis [45]. In the presence of well resolved three dimensional structure of a protein, in-silico Alanine scanning can be used. If the mutation of a residue to Alanine changes the binding energy of the protein to its partner substantially (typically G > 2.5 kcal/mol), then this residue is considered a hotspot residue. Reliable methods for identifying hotspot residues can be used to design small drug like molecules which can bind to the hotspots on proteins and prevent complex formation. This can be a significant step forward in rational drug design.

2.2.4 Physiochemical properties

In terms of their physiochemical properties, both protein-protein interaction sites and high affinity binding cavities for other ligands are often less polar or more hydrophobic in comparison to the rest of the protein [46]. Other physiochemical properties that have been employed in binding site prediction include residue side chain polarity and charge, amphiphilicity index etc.

2.2.5 Evolutionary conservation

Evolutionary conservation of an amino acid in a family of proteins has been used as an indicator of functionally important sites within a protein. Interface residues tend to be slightly more conserved than other surface residues in proteins. However, the degree of conservation needs to be very high (more than 90% for protein-protein interactions [47]) to infer that the two proteins would bind to similar targets. The degree of conservation between interface residues and those in the interior of the protein is not significantly different [48], [49]. This is true because many conserved residues are buried in the interior of the protein and contribute to protein folding and stability.

Evolutionary features used in the binding site prediction methods include sequence profiles, Position Specific Scoring Matrices (PSSMs), residue conservation scores, conservation of physiochemical properties etc. A good review of different methods to score residue conservation is given in [50].

Homology or sequence conservation information can be used to predict the structure of a protein based upon the structure of related proteins (using approaches such as SWISS-MODEL [51], ESyPred3D [52] and others) and the resulting predicted structure can then be used for predicting protein interfaces. The Structural similarity between two proteins is, in general, a good indicator of existence of similar binding interfaces.

Other sources of information used in binding site prediction include cellular localization, protein-protein interaction profiles and functional annotations.

3 Predicting binding sites

As discussed earlier, computational approaches for the prediction of binding interfaces in proteins are of particular interest to bioinformaticians and biologists as they present an inexpensive and fast alternative to experimental techniques. Due to this interest, a large number of computational approaches exist in the literature for predicting different types of protein binding sites. In this research exam, an analysis of the overall status of the field is presented. Some interesting methods are discussed in detail. An attempt is made to identify the general challenges, difficulties and opportunities in this research area.

In order to aid the discussion of different approaches for binding site prediction, a simple and generic mathematical formulation of the problem is described henceforth. Given a protein P and its binding partner Q (another protein, a ligand, a metal ion or a nucleic acid molecule), the objective of binding site prediction is to find a scoring function $S(.,.)$ which will score the binding propensity of different components p (residues, motifs or domains) on the protein P to those on its binding partner Q. Mathematically,

$$S(.,.) = f\big(\boldsymbol{\phi}\big((p,P),(q,Q)\big);\boldsymbol{\theta}\big) \quad (1)$$

Here, f is the prediction function with parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ denotes the feature descriptors used as input. Some instances of the above general problem formulation require that the binding components on both the protein and its binding partner be localized. Scoring functions of such problem instances can be expressed as $S(p|p \in P, q|q \in Q) = f\big(\boldsymbol{\phi}\big((p,P),(q,Q)\big);\boldsymbol{\theta}\big)$, where q is a component on the binding partner Q and $p|p \in P$ indicates that the specific occurrence of p on P is being considered while generating the prediction. This is particularly desirable for protein-protein and protein-nucleic acid binding site predictors. However, as discussed in the next sections, not all protein-protein or protein-nucleic acid binding site predictors do this. Other formulations may just predict the binding propensity of $p \in P$ with its partner Q regardless of the binding component q on Q. Such methods thus calculate $S(p|p \in P, Q) = f\big(\boldsymbol{\phi}\big((p,P),Q\big);\boldsymbol{\theta}\big)$. Yet other types of binding site predictors may choose not to consider the specific binding partner Q at all and the scoring functions of such methods can be expressed as $S(p|p \in P) = f\big(\boldsymbol{\phi}\big((p,P)\big);\boldsymbol{\theta}\big)$. Methods for predicting generic functional sites on a

protein fall into this category. Still other methods can simply calculate $S(p,.) = f(\boldsymbol{\phi}(p); \boldsymbol{\theta})$, i.e., without any regard to the protein on which p lies. Domain-Domain interaction prediction methods are examples of this specific formulation. Depending upon what the objective of scoring is, what kinds of data is used and what the exact formulation of the predictor f is, the input descriptor representation $\boldsymbol{\phi}$ can be modified accordingly. For example, kernel based methods may not even require an explicit feature representation. The parameters $\boldsymbol{\theta}$ can be chosen empirically or estimated using training data through machine learning techniques. The output of the scoring function f can be class labels (interacting or not), numerical scores indicating the binding propensity or probabilistic values. Henceforth, we discuss computational methods for prediction of protein binding interfaces in interactions of proteins with different types of interaction partners.

3.1 Protein-Protein Binding Interfaces

Protein-Protein interactions are absolutely critical to a huge number of biologically important processes and the identification of the protein binding sites involved in these predictions can help provide insights into the mechanisms by which different proteins fulfill their roles. A large number of computational approaches exist for predicting the binding sites involved in protein-protein interactions (see reviews [39], [53–57]). For protein-protein binding interfaces, the binding partner Q in equation (1) is another protein.

3.1.1 Classification of existing approaches

A classification of different approaches for protein-protein binding site prediction is given in Table 1. The table classifies a selection of existing methods on the basis of the following criterion:

a. Use of information about binding proteins

Some protein-protein binding site prediction methods [58][59][60], simply predict binding affinities between pairs of motifs or domains independent of the proteins on which these motifs or domains lie. In terms of equation (1), these scoring functions of these approaches can be written as $S(p,q)$. These methods typically try to *explain* a protein-protein interaction network through bindings between domains or motifs. For example, [58] tries to find the minimum number of domain-domain interactions using linear programming such that interaction constraints in a known protein-protein interaction network are satisfied. The underlying assumption of these methods is that the binding affinity between motifs is solely dependent upon the motifs themselves irrespective of the proteins on which they lie. However, the nature of binding sites is actually dependent upon the overall structural and chemical composition of the proteins involved in the binding [61]. Other methods predict the putative binding sites on any given protein irrespective of its specific binding partners, i.e., they identify structural or sequence constructs where a protein can bind *any* other protein. Such methods are called partner independent predictors. With equation (1) in mind, their scoring functions can be represented as $S(p|p \in P) = f\big(\boldsymbol{\phi}\big((p,P)\big); \boldsymbol{\theta}\big)$. A good review of such approaches is given in [56]. These methods ignore the fact that the binding propensity of a residue (or any other construct such as a motif or domain) is also dependent upon the nature of residues on its target protein. Moreover, partner-specific binding site predictions allow a more detailed and precise

understanding of the nature of protein-protein binding. In terms of equation (1), the scores generated by such methods can be expressed as $S(p|p \in P, q|q \in Q) = f\big(\boldsymbol{\phi}\big((p,P),(q,Q)\big);\boldsymbol{\theta}\big)$. However, only a few existing methods generate partner specific predictions.

b. Descriptors used

Binding site prediction methods can be classified on the basis of the type of information that they use to make the predictions, i.e., on the basis of what makes up $\boldsymbol{\phi}$ in equation (1). Most of the methods rely on structural features of proteins. Some methods also use homology or conservation information. However, there do exist some approaches that use sequence information alone.

c. Interface level

Protein interactions can be seen to occur at three different levels of a concept hierarchy: whole-proteins, domains/motifs or residues. Binding site prediction approaches can operate at either the domain/motif level or at the residue level, i.e., depending upon how the components of a protein ($p \in P$) are defined in equation (1). An example prediction from a method that operates at the motif level might look like *"Motif M on protein A interacts with protein B".* On the other hand, binding site prediction approaches working at the residue level attempt to identify individual residues that are a part of the binding interface on a protein. Some of the early methods for binding site prediction represent binding sites as surface patches on proteins where a patch is defined to be a group of neighboring residues in the three dimensional structure of a protein.

d. Prediction of hotspots

Ofran and Rost [10] indicate that generic approaches (like ISIS [62] or ProMate [27]) for the prediction of binding interfaces between proteins actually identify specific residues that are crucial for stabilizing the interactions, i.e. hotspots. Recently, some methods that explicitly predict whether an interface residue is a hotspot residue or not have emerged.

e. Types of interactions

The nature of interaction sites and the level of difficulty involved in predicting them vary depending upon the types of protein-protein interactions being considered [63]. This relates to the grouping of proteins P and Q in equation (1). For example, interfaces in antigen-antibody interactions are structurally very different from other types of interactions [64]. Detection of binding sites on antibodies is considerably easier in comparison to antigenic proteins [15]. A similar difference exists in the prediction of homo- versus hetero- oligomeric interaction sites with the former being much easier than the latter. Majority of the existing methods do not consider these variations during training and prediction. Although this ensures larger applicability of the resulting algorithm but it comes at the cost of loss in accuracy. Some of the existing methods that focus on peculiar types of interactions have shown that considering the interaction type can improve accuracy. However, for many types of interactions there are not enough training examples available and this hinders interaction-type-specific training.

f. Prediction scheme

Protein-Protein binding site prediction techniques can be classified on the type of the function f used in equation (1) for producing the scores. Almost all prediction schemes use parametric

functions. Some methods use empirical scoring functions with weighted terms for contribution from different input data and the parameters $\boldsymbol{\theta}$ are chosen manually. However, the design of such functions requires significant physical insight. Methods that perform automated selection of the parameters and the choice of the functional form are more desirable. These methods fall under the realm of machine learning. Researchers have applied both supervised (e.g., linear regression [65], neural networks [66], support vector machines [67], conditional random fields [68], random forests [69]) and unsupervised schemes (e.g., Bayesian probabilistic models for modeling protein-protein interaction networks to predict protein specific domain-domain interactions as in InSite [61]). The choice of a particular machine learning method is dependent upon the type of data being used and the specific output requirements of the approach.

3.1.2 Qualitative analysis of existing methods

As is evident from Table 1, most of the methods require structural information to make predictions. However, as described earlier, the use of structure information can limit the applicability of a method as the number of proteins with known structure is small. In this research exam, we focus on techniques that use sequence or sequence-derived information only. Another general observation from Table 1 is that most of the existing approaches perform partner-independent prediction of binding sites. In this research exam, we focus on existing approaches that perform partner-aware predictions. A naive approach to generate binding propensities between pairs of constructs (amino acids, domains or motifs) on two proteins using a partner-independent predictor is to simply add the prediction scores assigned to each construct by the predictor, i.e., in terms of equation (1), $S(p|p \in P, q|q \in Q) = S(p|p \in P) + S(q|q \in Q)$. However, methods that consider pair-wise interactions between proteins during training perform significantly better [70].

One of the first approaches to perform partner-specific binding site predictions is InSite [61]. InSite does not require binding site information during training and uses non-structural information about a protein to make its predictions. InSite learns possible binding pairs at the motif level given a protein-protein interaction network using a probabilistic model. InSite is able to integrate a wide variety of data sources such as protein-protein interaction information from a number of different biological assays along with any indirect evidence of protein-protein or domain-domain interaction like expression correlation, Gene Ontology (GO) annotations [71] and domain fusion. InSite uses a library of conserved sequence motifs (such as ProSite [72] or Pfam [73]) and generates a probabilistic score as the binding propensity between two motifs. However, two motifs can obtain different scores depending upon the proteins they are occur on. InSite has been used to understand the mechanics of different diseases in light of its predicted sites of interactions between proteins. However, InSite can only predict those motifs that are part of the input motif library, i.e., it cannot find sequence motifs automatically. Additionally, It is more interesting to obtain binding site information between two interacting proteins at the residue level as it allows for a more detailed understanding of the interaction.

A non-parametric approach for identifying binding site residues in protein sequences using sequence information alone, called PIPE-Sites, is given by Amos-Binks et al. [74]. This method is based on the Protein-Protein Interaction Prediction Engine (PIPE) [75] and is able to perform partner-aware predictions. According to an independent assessment by Park et al. [76], PIPE compares well with the

best existing binding predictors. For any two proteins *A* and *B* and a given data set of known protein-protein interactions (called interaction list), PIPE selects a 20 residue long sliding window w_a from protein *A.* It then creates a set *R* of the interaction partners of all proteins in the interaction list that contain a window matching w_a (within a pre-specified similarity threshold). PIPE then finds all the proteins from *R* that match a sliding window w_b from protein *B*. The number of proteins in *R* that contain w_b is defined as the score $S(w_a,w_b)$. This score is computed for all possible windows in the two proteins and yields a 3D surface called PIPE landscape. Peaks in this landscape are potential binding sites between the two proteins. PIPE-Site uses a heuristic approach to find peaks in the landscape generated by PIPE. It is important to note that PIPE-Site does not use information about known binding sites during its operation. Amos-Binks et al. have shown that this simple approach is able generate accurate results when tested over human and yeast proteomes. Although an independent validation of this approach using a non-redundant protein interaction list needs to be performed, the PIPE landscape can prove to be a helpful feature for binding site prediction in conjunction with other features and a more sophisticated classifier. PIPE-Sites can perform poorly in cases where the specificity of binding interfaces of a protein is low, e.g., Calmodulin binding proteins.

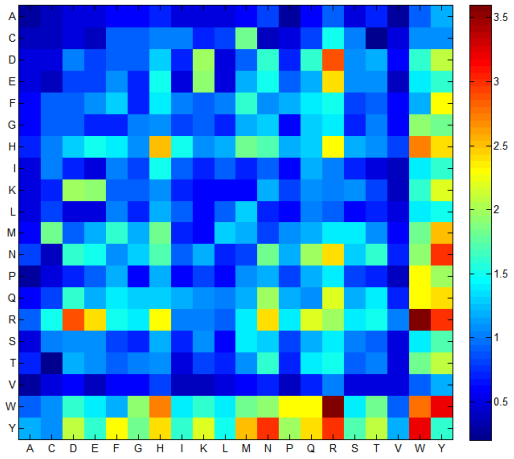

Figure 3 Binding propensity of different amino acid pairs plotted from the data in supplementary table-1 of [70]. A single residue shows different binding propensities with different resdiues. For example, the pairing of Proline (P) with Tryptophan (W) is highly preferred in comparison to its pairing with Alanine (A) in protein-protein interfaces. Only some of the amino acids, such as Tyrosine (Y), Valine (V) and Alanine (A), have relatively uniform binding propensities for different amino acids. They have also discussed the physiochemical basis of the observed differences in propensities.

Ahmad and Mizuguchi [70] discuss the effects of performing partner-aware versus partner-independent binding site prediction. Their analysis of the binding propensities of residue pairs in protein-protein interfaces clearly shows that the binding propensity of a residue is strongly dependent upon its partner in other proteins (see Figure 3). On the basis of these differences in propensities, they hypothesized that considering residue pairs on interacting proteins in binding site prediction can improve performance and found out that this is in fact the case. Their proposed approach uses features based on residue pairs in proteins. First, for any given protein *P*, its position specific scoring matrix (PSSM) [77] and position dependent 1-spectrum feature representations are obtained. The PSSM of protein P with length n_P is given by the $20 \times n_P$ matrix $PSSM(P)_{i,j} = c_i(P_j),\ i = 1 \dots 20, j = 1 \dots n_P$ in which each element $c_i(P_j)$ is the (relative) frequency of occurrence of amino acid i at location j in a multiple sequence alignment

with protein P. Thus, PSSM quantifies the degree of evolutionary conservation of different residues at different locations in the protein. The position specific 1-spectrum representation of protein P is another $20 \times n_P$ matrix given by $PD_1(P)_{i,j} = \begin{cases} 1 & if\ P_j = i \\ 0 & otherwise \end{cases}, i = 1 \dots 20, j = 1 \dots n_P$. Figure 4 details the construction of residue pair features using $PSSM$ and PD_1. It also explains the training and evaluation processes used in the method. It is important to note that the feature representations used in this approach are position dependent. Such feature representations offer better sensitivity in prediction as they allow for the learning of position specific differences in interface and non-interface residues [78]. The findings from their performance evaluation can be summarized as follows:

a. *Predictors trained on residue pairs outperform partner-independent predictors*
b. *Combination of multiple window sizes for different feature types and the averaging of outputs from neural networks improves performance*
c. *Complexes with large conformational changes are difficult to predict*

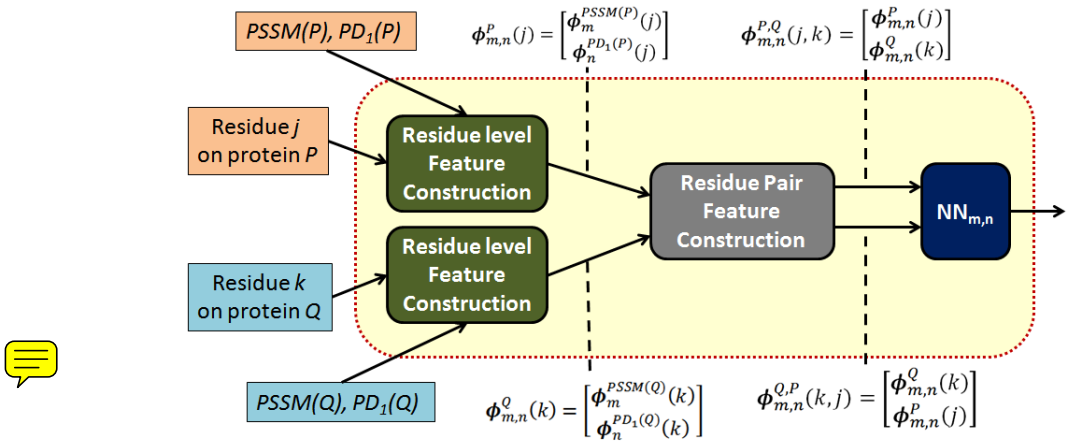

Figure 4 Feature extraction, training and evaluation in [70]. The feature representation of single residue j in a n_P residue long protein P is constructed by considering m and n columns from $PSSM(P)$ and $PD_1(P)$ centered at j. The PSSM feature vector representation of the residue using a length m window is a $(20 \times m)$ dimensional column vector obtained by the concatenation of m columns of $PSSM(P)$ centered at position j and is denoted by $\boldsymbol{\phi}_m^{PSSM(P)}(j)$. Similarly, the $(20 \times n)$ dimensional column vector $\boldsymbol{\phi}_n^{PD_1(P)}(j)$ denotes its PD-1 representation with window length n. The complete feature representation for residue j in protein P is given by $\boldsymbol{\phi}_{m,n}^{P}(j)$. Valid indices for j are $j \in [l(m,n) + 1, n_P - l(m,n)]$, where $l(m,n) = max(0,(max(m,n)-1)/2)$. Next, residue-pair features $\boldsymbol{\phi}_{m,n}^{P,Q}(j,k)$ are extracted. In particular, for a residue j in protein P and a residue k in protein Q, $\boldsymbol{\phi}_{m,n}^{P,Q}(j,k)$ is constructed by concatenating the individual feature representations of the two residues. A neural network denoted by $NN_{m,n}$ is then trained over these residue pairs using their corresponding labels (interacting or not-interacting). Since $\boldsymbol{\phi}_{m,n}^{P,Q}(j,k)$ and $\boldsymbol{\phi}_{m,n}^{Q,P}(k,j)$ represent the same physical interaction pair, therefore both of these instances are used to train the neural network with the same label. A total number of 5 window sizes ($m,n \in \{0,1,3,5,7\}$) for each type of feature representation are employed. This leads to 24 possible feature representations since $m = n = 0$ is an empty representation. A total of 24 different neural networks (each operating at different window sizes) are trained and, during evaluation, their outputs are summed to produce the final score for every pair of residues, i.e., in terms of equation (1), $S(j|j \in P, k|k \in Q) = \sum_{m,n}\left(NN_{m,n}\left(\boldsymbol{\phi}_{m,n}^{P,Q}(j,k)\right) + NN_{m,n}\left(\boldsymbol{\phi}_{m,n}^{Q,P}(k,j)\right)\right)$.

Ahmad and Mizuguchi [70] use an ensemble of neural networks to produce the predictions. Use of large margin kernel methods such as the one proposed in [78] and incorporating predicted ASA, predicted secondary structure and PIPE landscapes has the potential of improving performance even further. The use of kernel methods would make the extraction of residue-pair features much easier, e.g., tensor product kernels. For a detailed description of tensor product kernels and their applicability to a very similar problem, the interested reader is referred to [80].

Most of the existing methods use a classifier (such as a neural network or SVM) that assigns labels (interacting or non-interacting) to either residues or residue pairs. As a consequence, these approaches do not take into account the inter-relations between neighboring residues when generating predictions, i.e., each residue is treated independent of any other. The only inter-relations that can exist between neighboring residues in such approaches arise from overlap between adjacent spatial patches or sequence windows. Liu et al. [81] have pointed out that considering inter-relations between spatially neighboring residues by using a Hidden Markov (HM) SVM [82] can improve prediction performance. HM-SVMs generate sequence labels for the whole protein sequence instead of producing independent predictions for each residue. In doing so, they are able to consider long distance inter-relations between labels. Other approaches that use similar ideas include: Bayesian networks [83], Conditional random fields [84] and Hidden Markov Models [85]. However, the effectiveness of using this property in sequence-only and partner aware predictions still needs to be investigated.

In terms of machine learning concepts, an interesting approach is given by Qi et al. [86]. It presents the prediction of multiple local properties (such as solvent accessibility, secondary structure etc.) and the identification of protein or DNA binding residues in a protein as a single multi-task learning problem. This allows the learning method to leverage the commonalities and inter-relations in all these tasks to produce better predictions than learning to solve each problem separately. This improvement is particularly noticeable in tasks for which the amount of available training data is small. The multi-task learning problem is solved using a deep neural network architecture. However, their approach does not perform partner-specific prediction of DNA or protein binding residues and can benefit from it. Based on this idea, It would be interesting to implement and study a partner-aware protein binding site predictor that uses its own predictions of quantities that are relevant to binding site prediction (such as ASA, secondary structure, conformational change upon complex formation, PTM sites, DNA, ligand and metal ion binding sites, etc.).

In terms of interface level of prediction, the authors of [87] have found out that simultaneous prediction of protein, domain and residue level interactions with information flows between the predictors at different levels can improve the prediction performance at all levels.

As is evident from Table 1, most of the methods for predicting hot-spots in protein interactions perform a two class classification to discriminate between hot-spot and non-hot-spot residues in protein interfaces for which the complex structure is known. However, it will be more useful to do so directly from the sequence. Nguyen et al. [88] have shown that sequence based predictors of hotspots can offer comparable performance to state of the art methods that use bound structure. However, their approach does not produce partner-specific predictions.

1 **Table 1 Qualitative comparison of different protein-protein binding site prediction methods.**

Method	Year	Partner Specific	Interface Level	Hot spots	Types of Interactions	Descriptors	Prediction Scheme
Cons-PPISP [89]	2005	No	Residue	No	General	PSI-BLAST profile, Solvent accessibility	Neural Networks
Promate [27]	2004	No	Patch	No	Heteromeric Transient (No antibodies)	Conservation, Secondary structure Amino acid pairing	Naive Bayesian
PINUP [36]	2006	No	Residue	No	Transient (No antibodies or antigens)	ASA, Residue interface propensity, Side chain energy, PSI-BLAST profile	Empirical Scoring Function
PPI-PRED [90]	2005	No	Patch	No	Results reported separately for different types	Surface shape, Electrostatic potential	SVM
SPPIDER [79]	2007	No	Residue	No	Both hetero & homo complexes	Sequence, Physiochemical properties, Evolutionary profiles Tertiary structure, Predicted relative ASA, Difference between predicted & observed relative ASA	Neural Network
Meta-PPISP [91]	2007	No	Residue	No	Enzyme/Inhibitor proteins	Uses raw scores from cons-PPISP, Promate and PINUP	Linear regression
Multilevel [87]	2009	No	Protein, domain &residue	No	General	*Protein level*: Phylogenetic profiles, localization, gene expression, Interaction. *Domain level*: Phylogenetic tree correlations of Pfam alignments, number of proteins in an organism from a particular domain. *Residue Level:* PSI-BLAST profiles, predicted secondary structure, predicted ASA	A Support Vector Regression (SVR) based multi-level learning scheme that allows sharing of training data from multiple levels
InSite [61]	2007	Yes	Motif	No	General	Conserved sequence motifs, Protein-Protein Interactions, Expression correlation, Gene Ontology annotations, Domain Fusion	Unsupervised learning of binding sites from interaction data using a probabilistic model
Ahmad et al. [70]	2011	Yes	Residue	No	General	PSSM encoded sequence windows and position dependent 1-spectrum	Neural network ensemble
PIPE-Sites [74]	2011	Yes	Residue	No	General	Sequence windows and similarity of sequence windows in two proteins using PAM120 score	A Heuristic method for identifying peaks in PIPE [75] landscape Uses protein-protein interactions to deduce binding sites
Liu et al. [81]	2009	No	Residue	No	Separate evaluations for Homo- & Hetero- complexes	PSSM profiles and ASA of spatial neighbors in a window (obtained from the experimental structure of a protein)	Hidden Markov SVM
ISIS [62][10]	2007	No	Residue	Yes	Transient heterooligomer	Residue sequence windows, Evolutionary profile of all residues in window, Predicted secondary structure, Predicted ASA, Conservation score of residues	Neural Networks
KFC2 [92]	2011	Yes	Residue	Yes	General	Use complex (bound) structure information, Hydrophobicity and solvent accessibility , structural neighborhood features, biochemical contact features etc.	SVM
HSPred [93]	2011	Yes	Residue	Yes	General	Uses complex structure. Van der Waals, hydrogen bond and solvation side-chain inter-molecular energies; van der Waals, hydrogen bond and solvation environment inter-molecular energies; van der Waals side-chain intra-molecular energy.	SVM
APIS [30]	2010	Yes	Residue	Yes	General	Uses complex structure and sequence features.	SVM
Nguyen et al. [88]	2011	No	Residue	Yes	General	Frequency domain representation of amino acid sequences (Using 3D structure features improves accuracy)	Random Forests

3.1.3 Performance Analysis

A fair comparison between existing approaches is not easy. This is because different approaches use different datasets containing different types of interactions, different feature representations, different definitions of what qualifies as an interface residue, different performance metrics, different sequence similarity thresholds in the data and so on. In the knowledge of this author, the most extensive compilation of performance metrics of existing partner-independent approaches is given in [55]. However, it only lists the original results reported in various papers. Zhou and Qin [39] have compared a small number of partner-independent predictors over the same test data set and have found that the performance ranking of methods being considered is PPI-Pred < SPIDDER < cons-PPISP, Promate < PINUP < meta-PPISP. Some later methods such as [70] claim better results than some of these approaches. However, the differences are not huge and, in the knowledge of this author, no method for binding site prediction (partner-independent or partner-dependent) reports an area under the Receiver Operating Characteristics (ROC) curve (AUC) of more than 75%. This shows that significant room for improvement is present in this problem domain.

Another point about the reporting of results by different methods is that most methods report AUC scores through a cross-validation procedure. However, in the opinion of this author, evaluating a single AUC score for the whole data set might not reveal the true performance characteristics of a method. An AUC score *for each protein* is more insightful, especially for methods that operate at the residue level. A single performance metric for a given method over a data set can then be calculated by determining the mean and dispersion of these protein-level AUC scores. These proposed performance metrics are in accordance with the nature of the problem and the way these prediction approaches are used in practice: Users are interested in identifying correct binding interfaces within a protein and If a residue in a protein receives a correct high score within that protein, it does not matter how its prediction score compares with other residues in other proteins.

3.2 Protein-Nucleic Acid Binding Interfaces

Proteins interact with DNA and RNA molecules to perform a large variety of functions of critical biological importance. A protein that binds DNA (or RNA) is called a DNA (or RNA)-binding protein (DBP or RBP). Interactions of a protein with DNA are crucial in processes such as transcription, transcriptional regulation, DNA replication, recombination and repair, viral infection etc. An important category of DBPs that activate or repress gene expression by binding to DNA motifs and histones is called Transcription Factors (TFs). Most of the RBPs bind DNA on the major groove of the DNA which is wider. DBPs are composed of DNA-binding domains such as the zinc finger, helix-turn-helix, leucine zipper etc. RBPs are involved in translation of mRNA, post-transcriptional events such as RNA splicing and editing, RNA transport, regulation of RNA levels etc. [94]. The complexes resulting from the binding of RNAs to proteins are called ribonucleoprotein complexes (RNPs) and defects in these complexes or their formation can be a basis for a number of diseases [95]. RBPs contain RNA binding domains, e.g., RNA Recognition Motif, K-Homology, RGG box etc. Prediction of binding interfaces between protein and nucleic acid can be invaluable to the understanding of precise binding mechanisms, function of genomes and diseases.

In contrast to the regular double-helical structure of DNA, RNA molecules form complex secondary and tertiary structures and this makes the prediction of interactions and interaction sites for RNA more difficult than DNA. In comparison to protein-protein interactions, protein-nucleic acid interfaces are, in general, less diverse and, as a consequence, easier to predict. Figure 5 illustrates this point for a specific data set [96]. Gromiha et al. [16] have found that protein-protein and protein-nucleic acid interactions differ significantly in their amino acid, dipeptide and tripeptide binding propensities. However, protein-nucleic acid binding is mediated through the same interactions (electrostatic, van der Waals, hydrogen bonds, water-mediated bonds etc.) as protein-protein interactions. Rawat and Biswas [97] have found that the number of hydrogen bonds in protein-protein interactions is, on average, smaller than in protein-nucleic acid complexes. Like protein-protein interactions, protein-nucleic acid binding can also introduce conformational changes in both the binding protein and the DNA.

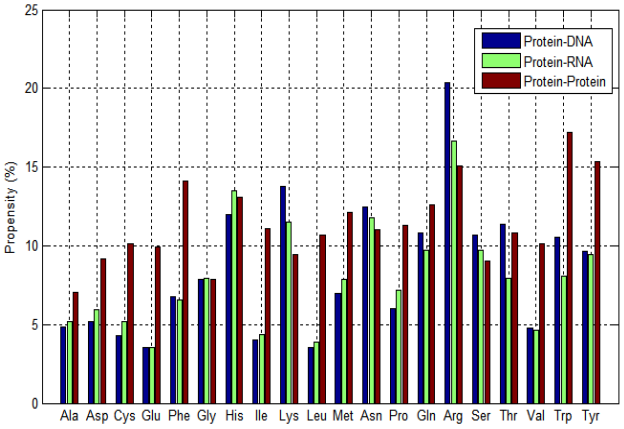

Figure 5 Single amino acid level propensities for protein-protein, protein-DNA and protein-RNA interactions. The propensity data in this figure is taken from [96]. It shows that positively charged residues (Arginine, Histidine and Lysine) show stronger binding propensities for the negatively charged DNA molecules. The Gini coefficient [98] values for protein-protein, protein-RNA and protein-DNA interactions for this data are 0.1302, 0.2856 and 0.2421, respectively. This shows that the diversity of the propensities of amino acids in protein-protein interactions is the smallest and this indicates that, in comparison to protein-nucleic acid interactions, a more intricate recognition mechanism is at work in protein-protein interactions.

While looking at interactions between nucleic acids and proteins in terms of equation (1), the binding partner Q of a protein P is a nucleic acid (DNA or RNA) molecule. It is important to note that there are two types of interfaces in a protein-nucleic acid interaction: one on the protein and the other on the nucleic acid. Identification of both of these interfaces is helpful as it can provide a fuller understanding of the binding mechanism and its biological function. In this research exam we focus on both the interfaces. However, interfaces on the DNA are discussed specifically in the context of transcription factor proteins. In the knowledge of this author, all existing computational methods predict either one of the two binding interfaces but not both. The similarity in the basic natures of protein-nucleic acid and protein-protein interface prediction problems allow for the use of a similar kind of feature representations, especially when considering binding interfaces on proteins. These features include both structural and sequence based representations [99]. Structure based methods (such as [100–102]) for identifying interaction sites on DBPs offer good performance. However, the unavailability of experimentally determined protein structure or unreliability of predicted structure (due to a lack of homologues) prevents wide-scale use of these methods. In this research exam, we focus on sequence

based prediction approaches. Henceforth, we present an overview of existing computational methods for predicting each type of binding interface for both DNA- and RNA- protein interactions.

3.2.1 Prediction of binding interfaces on DBPs

With the increase in amount of available data for DBPs, it is now known that binding specificity or recognition for protein-DNA interactions is possible through a diverse set of mechanisms [103]. This has led to a number of sophisticated approaches for the computational prediction of DNA binding interfaces on DBPs. Early research on binding site identification in DBPs ([100], [104–106]) has demonstrated that some important features that are able to distinguish between DNA-binding and non-DNA-binding residues on a protein can be extracted from sequence information alone. In the knowledge of this author, the first method that used sequence or sequence-derived information alone was proposed by Ahmad et al. [105]. This method uses a feature representation consisting of PSSM and sequence identity features of residues within a sliding window. These features are then given as input to a neural network classifier to produce a prediction. Ahmad et al. [105] found that addition of PSSM information improves the prediction performance of the classifier in comparison to using sequence information alone. However, the accuracy of this approach was very low. Later approaches such as DISIS [107], NAPS [108], SVM-PSSM [109], BindN-rf [110], BindN+ [111] etc. offer significantly better performance with some approaches performing roughly as good as the best structure based methods. A very recent review [99] tabulates the performance of a number of these approaches. In this research exam, DISIS [62] and NAPS [108] are presented as representative examples in the area.

DISIS [62] (DNA Interaction Sites Identified from Sequence) gets its name from a similar method for predicting protein-protein interfaces called ISIS proposed by the same group (see Table 1). Figure 6 shows the internal workings of DISIS. DISIS shows that using residue-level evolutionary profiles and conservation scores along with predicted secondary structure and ASA offer better discrimination.

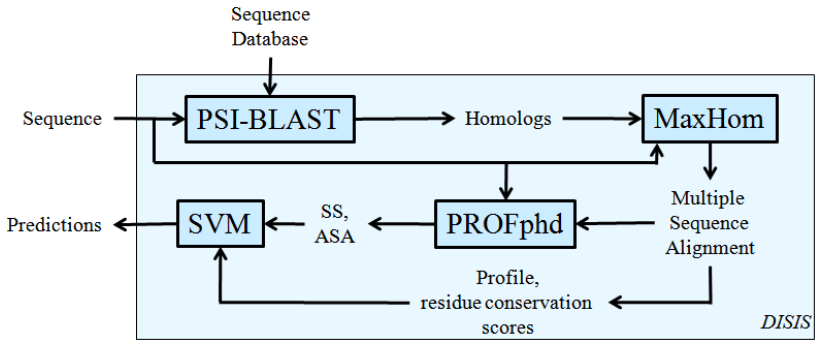

Figure 6 Internal structure of DISIS [62]: DISIS uses PSI-BLAST (3 iterations with a cut-off at 10^{-3}) to find all proteins from a sequence database related to query sequence. Using the query protein as the template, a multiple sequence alignment of the homologous proteins is obtained through MaxHom [112]. MaxHom is based on the Smith-Waterman dynamic programming algorithm. The multiple sequence alignment is used by PROFphd [113], a neural network based predictor of secondary structure (SS) and ASA. The multiple sequence alignment is also used to calculate the profile and residue conservation scores for the query sequence. For a given residue in the query sequence, the following features are given to a SVM: evolutionary profile of the residue and its 4 neighbors on each side, level of conservation along with the predicted solvent accessibility of the residue and its neighbor on each side, predicted secondary structure of the residue.

NAPS (Nucleic Acid Prediction Server) [108] is one of the most accurate sequence-based predictors currently available [99]. In comparison to DISIS, NAPS offers higher sensitivity. This is particularly true for high specificity values. For a given residue in a query protein, NAPS constructs a 301 dimensional feature

vector as follows: the amino acid for the residue is represented using a 20 dimensional binary vector (1 entry for each amino acid), a single charge attribute for each residue, the PSSM representation of a 7 element window centered at the residue of interest (140 PSSM features) and the 20 row elements from the columns of the BLOSUM62 matrix corresponding to each of the 7 residues in the window (140 BLOSUM62 features in total). PSSMs and BLOSUM62 features capture position dependent and position independent conservation information for a residue, respectively. NAPS uses residue charge information as well by assigning charges of +1, +0.5 and 0 to Arginine and Lysine, Histidines and all other residues respectively. This can help because DNA molecules are negatively charged and positively charged basic amino acid residues on proteins exhibit relatively higher binding affinities than other amino acids (see Figure 5). NAPS uses C4.5 decision tree classifiers with bagging for classification. NAPS uses the same features for predicting interface residues on RBPs as well. The authors of NAPS have compared different classification approaches. However, in terms of sensitivity at high specificity values, SVM appears to perform marginally better than the C4.5 based classifiers recommended in their paper. Physiochemical properties of amino acids are also used in BindN-rf [111] along with PSSM and conservation scores. It should be noted that BindN-rf computes, for a given residue in a protein, the average and standard deviation of the physiochemical properties of amino acids from the PSSM to which the input residue aligns. NAPS compares well with structure based prediction approaches [99]. The recently proposed sequence based multi-task prediction method by Qi et al. [86] offers roughly the same performance as NAPS.

Recently, a meta-classification approach called MetaDBSite [114] has been proposed. MetaDBSite combines the predictions of multiple existing predictors including DISIS and BindN-rf. In comparison to the individual methods used in MetaDBSite, the combination achieves marginally better average of sensitivity and specificity values. However, in the absence of a complete ROC curve, it is difficult to draw specific conclusions about the efficacy of this approach.

3.2.2 Prediction of binding sites on DNA

Identifying the locations on DNA where proteins bind can help understand the mechanisms of how DNA works. There are three major types of binding sites of proteins on DNA:

- Regulatory sites are specific cis-acting locations on the DNA where trans-acting proteins bind to regulate the expression of genes. An important and well studied category of regulatory proteins is transcription factors (TFs). TFs are responsible for regulating the process of transcription. TFs function alone or with other proteins (forming a complex) to promote (as an activator) or block (as a repressor) the recruitment of RNA polymerase[2] for specific genes. However, some TFs act both as activators and repressors for the same gene [115]. TFs are usually 4-30 base pairs long.
- Restriction sites are binding sites for proteins called restriction enzymes. Restriction enzymes, found in bacteria and archaea, cut foreign DNA as they bind to that DNA molecule. This works as a defense mechanism against invading viruses. Host DNA is methylated to prevent damage from the enzyme's activity.

[2] RNA polymerase is an enzyme that creates RNA from DNA.

- Recombination sites: Site specific recombination is an operation in which DNA-strand exchange takes place between segments possessing only a limited degree of sequence homology. Proteins involved in this process are called recombinases and the locations on the DNA where these recombinases bind the DNA are called recombination sites.

The binding specificity of different types of binding sites is different [116]. For example restriction sites are generally identical short (4-8 base pairs) sequences and binding propensity greatly decreases for less similar sequences. For example, the common restriction enzyme EcoRI recognizes the palindromic sequence GAATTC and cuts between the G and A on both stands. Thus, these binding sites can be expressed as consensus sequences or a sequence that allows for some ambiguity (for example, GTYRAC for the enzyme HincII where locations with Y and R are pyrmidines (T and C) or purines (A and G) respectively). On the other hand, regulatory and recombination sites are usually different from one another (even for the same protein) and usually sequence logos or PSSMs (also called Position-specific Weighted Matrices (PWMs)) are employed to represent these binding sites. For example, the sequence logo representation of the transcription repressor lexA is shown in Figure 7.

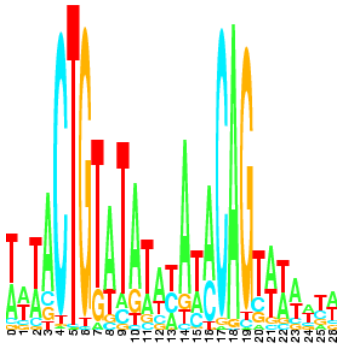

Figure 7 The sequence logo for the transcription factor lexA. The x-axis corresponds to the relative positions on the DNA whereas the height of a nucleotide is proportional to its prevalence at that location. The overall height at a location is proportional to overall conservation (across all binding targets) measured using information content at that location. Figure adapted from:

(http://www.ch.embnet.org/CoursEMBnet/Exercises06/regulatory_motifs.html)

In this research exam, approaches for finding TF binding sites (TFBS) on DNA are considered. Methods for finding these sites are considerably different from techniques for locating binding sites on proteins discussed earlier. This is due to two primary reasons:

- Over and above the prediction of TFBS, biologists are also interested in finding the PSSMs or motifs of TFBS
- The experimental techniques used in finding these binding sites are considerably different in their nature.

Most of the existing methods for finding binding sites on proteins obtain their training data from 3D structures based on a distance criteria. Whereas, methods for locating binding sites on DNA mostly use data from techniques such as DNA expression microarrays, Chromatin Immunoprecipitation (ChIP) array/seq. Microarrays can measure the expression level of each gene in an organism under various conditions. Genes showing similar expression levels under various conditions may be co-regulated by the same TF or sets of TFs. TFBS on these co-regulated genes can be found by giving sequences of regions upstream from these genes as input to motif finding algorithms. Another way to find TFBS with

expression data is to use phylogenetic profiling. This approach works on the principle that TFBS across species are more conserved than random background. Upstream regions from co-expressed genes from multiple species are given as input to a motif finder which produces the PSSMs or consensus sequences for putative TFBS. In recent years, TF binding assays such as ChIP followed by microarray (ChIP-array or ChIP-chip) or high-throughput next generation sequencing (ChIP-seq) have become more common. ChIP-seq data offers significantly better resolution (25-200 base pairs) than array based technologies. ChIP based methods not only give high-throughput but also provide quantitative measures of binding activity (ChIP enrichment). Another emerging technology is protein binding microarrays (PBMs).

In terms of prediction approaches, TFBS finding methods can be categorized as generative or predictive. Generative models of binding motifs like Gibbs motif sampler [117], MEME [118], Consensus [119] etc. generate PSSMs from a set of input sequences. PSSMs model the probability distribution of different nucleotides at a given location independent of its inter-relationships with other locations. However, methods (e.g., [120]) that consider such inter-relations have shown better performance. Some motif finders (such as SVMotif [121], DEME [122], DME [123] etc.) work in a discriminatory fashion and, as a consequence, are able to use information from both positive and negative examples. Often, TFBS occur in clusters called cis-regulatory modules (CRMs) and there are interactions amongst different TFs. Some approaches try to infer the location of these CRMs (examples include [124], [125]). A recent review of these generative approaches is given in [126] .

Predictive methods, on the other hand, treat gene expression or ChIP intensity values as output variables and usually employ regression-style discriminatory machine learning approaches for estimation of these values from DNA sequence features. In the context of discriminative approaches, PSSMs can be viewed as linear additive models. Predictive methods offer a more general and flexible framework for TFBS identification as they can easily incorporate information from both positive and negative examples. In this research exam we discuss some of these predictive approaches in detail. In terms of equation (1), these approaches can be expressed as $S(q) = f(\boldsymbol{\phi}(q, Q); \boldsymbol{\theta})$, where q is a component of DNA sequence Q and $\boldsymbol{\theta}$ are the parameters of the model. Different feature representations $\boldsymbol{\phi}(q, Q)$ used in the literature include [127]: matching strength of q with a certain motif, k-mer occurrences, histone modification data, conservation scores, GC content etc. Zhou and Liu have compared a number of different regression approaches such as Linear regression with feature selection, Neural networks with regularization, Support Vector Regression (SVR) with different kernels add additive models such as Bayesian Additive Regression Trees (BART) [128] using ChIP-chip data. They found that BART gave the best correlation between true and predicted values. However, the performance of SVMs was only marginally lower.

Recently, methods for TFBS prediction have started to incorporate the roles of epigenetic features [126]. For example, Narlikar et al. [129] have found that using an informative prior over DNA sequence positions based on a discriminative view of nucleosome occupancy can improve the chances of find the correct TFBS motif significantly (by ~52%) in comparison to a uniform prior. Yuan et al. [130] have also proposed that incorporating nucleosome occupancy along with histone acetylation level can improve prediction performance. Similar observations have also been made by other recent approaches such as [131] and [132]. Information about other types of histone modifications such as phosphorylation, methylation etc. can also impact prediction accuracy.

Most of the existing predictive approaches do not attempt to find TFBS for individual TFs. Rather, they try to locate genomic regions of regulatory elements. Recently, an approach called Chromia [133] has been proposed that predicts TFBS for individual TFs using hidden markov models. It has shown to significantly improve prediction performance.

3.2.3 Prediction of binding interfaces on RBPs

Methods for identifying binding sites on RBPs are, in character, very similar to those used for DBPs. Most of the existing sequence based methods (such as BindN+ [111], NAPS [108], PiRaNha [134], PPRInt, RNABindR, PRBR) for prediction of binding site residues on RBPs use features such as solvent accessibility, secondary structure, sequence conservation, hydrophobicity, etc. A number of structure based prediction approaches also exist [135]. Puton et al. [135] have recently published a comparison of ten prediction methods over a data set of 44 sequences from PDB. They have found that PiRaNha [134] outperforms all other approaches including the three structure based techniques considered in the comparison. They also present a novel meta-predictor by combining the predictions of three top scoring sequence based predictors (PiRaNha, BindN+ and PPRInt) from their comparison and found that their meta-predictor outperforms all existing methods (improvement of 0.013 in AUC score from PiRaNha). It is interesting to analyze the difference in prediction performance of the methods compared by Puton et al. in the light of the features used in them. Both BindN [136] and BindN+ [111] use SVMs and produce AUC scores of 0.733 and 0.821 respectively. BindN uses side chain disassociation coefficient values, hydrophobicity and residue molecular mass as features. BindN+ also incorporates evolutionary information (through PSSMs). This supports the general observation that evolutionary information can be very helpful in improving prediction accuracy. PiRaNha and NAPS (AUC: 0.679) also use PSSM information along with other features. It is interesting to notice the big difference in the performance of NAPS and BindN+ given that these two methods perform roughly the same on the task of identifying binding sites on DBPs (on the same data set) [99] and both use PSSMs. The exact reason for this difference is unknown. PiRaNha represents a residue using features such as PSSM scores, residue interface propensity (likelihood of a residue being found in an RNA-binding site taken from [137]), predicted solvent accessibility (from SABLE [138]) and hydrophobicity in a length 23 window centered at the residue of interest. Predictions are made using SVM. This shows that predicted structure features can improve classification performance.

3.2.4 Prediction of binding interfaces on RNA

In order to understand the mechanics of post-transcriptional regulation (PTR), the binding preferences of RBPs need to be characterized. Assays for identifying the binding sites of RBPs on RNA include Cross Linking and Immunoprecipitation (CLIP), CLIP-Array (CLIP followed by microarray analysis), CLIP-Seq (CLIP followed by high throughput sequencing), PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced CLIP), iCLIP (individual nucleotide-resolution ultraviolet CLIP) [21] etc. A number of computational approaches also exist in the literature for predicting putative RBP binding sites on RNA.

In this research exam, we first consider the recently proposed approach called RNAcontext [139]. RNAcontext, for substrings of length K on the RNA, learns a PSSM and a structural context vector that describes the relative preferences of different RNA structural constructs in putative RBP binding sites. Despite its use of RNA secondary structural information, RNAcontext does not require that the structure

of the RNA be known in advance. Rather, RNAcontext uses SFOLD [140] to estimate marginal distributions at each nucleotide over different secondary structural constructs (such as hairpin loops, unstructured components, paired components and others). RNAcontext, given a RNA sequence, makes real valued predictions of its binding affinity using the PSSM and structural context learnt by the minimization of an regularized prediction error function using a training data set for a given protein. The PSSM and structural context produced from learning allow general conclusions to be drawn for the binding affinity of a protein, e.g., the RNA binding site for protein Vts1p is twice as likely to contain a hairpin structure comparison to the binding sites of HuR. RNAcontext offers excellent performance in comparison to other existing approaches such as MEMERIS [141] and MatrixREDUCE [142]. However, one short-coming of this approach is that, during training, the learning of parameters is done separately for each protein. It can potentially benefit from a multi-task style learner. Moreover, for certain proteins (such as YB1, FUSIP1 and U1A), the prediction performance is still very low.

A more recent approach called PARalyzer uses binding evidence from PAR-CLIP data (log of the number of the number of T to C mutations in a region) and a motif finder called cERMIT to find putative binding sites using sequence information alone. PARalyzer finds a motif for the putative RNA binding site of a protein such that the average binding evidence value for sequence regions matching that motif is significantly higher than the average binding evidence. The optimization is done using a greedy search strategy that relies on local motif updates.

Recently, some researchers have taken a feature space approach using SVM type classifiers to predict RNA binding sites in proteins. Livi et al. [143] have used features based on motif scores and *k*-spectrum[3] representations to predict regulatory binding sites of the protein CELF1 with an SVM classifier. However, using this approach, it is difficult to identify any PSSMs that would make the predictions more comprehensible to a biologist. A similar approach to binding using weighted degree kernels for identifying splice sites on the RNA is proposed by [145] and followed in [146]. Splice sites are regions on the pre-mRNA where certain small nucleic RNPs (snRNPs) bind and performing splicing to produce mRNA. Mersch et al. [147] have used a 1-norm SVM to identify the binding sites of SR proteins on the RNA. They employ locality improved and combined position dependent oligomer kernels and generate highly accurate predictions. One advantage of using oligomer kernels is that they allow for a better visualization of the nature of binding sites. Moreover, they are less stringent than the position dependent k-spectrum kernels discussed in [78] for protein-protein interactions. Locality improved kernels have also been used in [148] for the recognition of translation initiation sites.

3.3 Binding of proteins to small drug-like ligands and metal ions

Most of the existing methods for identifying ligand binding sites on proteins do not consider the nature of the ligand [149]. A large variety of methods uses 3D structure of the protein to find pockets or clefts on the surface of the proteins [150] where a ligand can bind. Some methods use machine learning approaches (such as SVM) with protein sequence characteristics and homology to find these binding sites [149].

[3] The *k*-spectrum representation of a given sequence is a vector containing the frequencies of different *k*-mers (substrings of length *k* found in it) [144].

In this research exam, we focus specifically on binding sites of metal ions on proteins. The binding of metal ions to proteins is of critical importance for protein function and structure. Finding binding sites of metal ions on proteins is a difficult problem as the number of possible residue choices for an ion to bind to is exponential in the length of the protein. A recent method by Passerini et al. [26] predicts the specific residues forming a binding site for different types of metal ions using a large margin structured output learning approach. The method, given some training data, uses a greedy approach to construct a function that scores correct linkages of protein residues to the metal ions they bind than incorrect linkages. This method offers a reasonable accuracy in terms of precision / recall curves. However, in the absence of AUC or similar scores, it is difficult to gauge the overall performance of the system. A generalization of this approach can be used for prediction of binding sites in proteins as well, especially when the binding residues are not contiguous.

4 Issues and opportunities

Looking back at methods for prediction of different types of interfaces for protein interactions, it becomes apparent that certain aspects of these interactions are not captured by existing methods. In this section, we summarize these issues.

Partner aware predictions

Research has shown that binding regions on a protein can change depending upon its partner. For protein-protein and protein-ligand interactions, recent research has led to realization of the importance of making partner aware predictions and a few methods do consider the interaction partners in making binding site predictions. Not only are such predictions more accurate, they are also more meaningful and informative in comparison to partner independent predictions. However, this area still requires more sophisticated machine learning methods (e.g. large-margin methods) and features (such as predicted structure, predicted ASA etc.) as there is significant room for improvement in accuracy.

In the knowledge of this author, no methods for prediction of binding interfaces in protein-nucleic acid interactions explicitly utilize information from both the nucleic acid and protein sequence or make simultaneous predictions of binding interfaces on both the protein and the nucleic acid. One possible reason for this can be the lack of large data sets that allow for training a partner-aware predictor for protein-nucleic acid binding sites. Some protein-nucleic acid complex structures exist in PDB. These structures can be used to build a training data set in which we know the binding sites on both the protein and its binding nucleic acid partner. On the other hand, data from biological assays such as CLIP-Seq or ChIP-Seq, however, gives only the binding site on the nucleic acid and not on the protein. Using such data would require more sophisticated machine learning techniques such as Multiple Instance Learning (MIL).

Consideration to PTMs and epigenetic features

PTMs regulate binding and the site of binding can change depending upon PTMs the protein has had. In the knowledge of this author, no methods for protein-protein interactions considers such information. Some methods for identifying binding sites on DNA have obtained better performance by using

epigenetic features (such as histone modifications by phosphorylation). Incorporating similar information in protein-protein interaction site prediction can also help improve performance. One way of doing this can be to use the scores obtained from a PTM site predictor (such as [151]) for a given region on the protein.

Modeling dependencies across different types of interactions

Similar to the notion of PTMs affecting binding, dependencies exist amongst different types of interactions that a protein is involved in. For example, some drug like ligands bind to a location on a protein that is involved in the interaction of that protein with another protein. Another example is the difference in protein-protein binding characteristics of Calmodulin in its Calcium-bound and Calcium-free states. Modeling of such dependencies in the prediction process can potentially result in an improvement in performance. A related idea was used in [86] which employs multi-task learning to simultaneously predict binding sites in protein-protein and protein-DNA interactions along with other related characteristics (such as ASA and secondary structure).

Prediction of hotspots

Most of existing sequence based method do not explicitly model the fact that not all interacting components (residues or nucleotides) in an interaction contribute equally to binding stability can potentially produce more informative and accurate predictions. This is particularly true for methods dealing with protein-nucleic acid interactions as pointed out in [152] for protein-DNA interactions.

Considering the nature of data in prediction

It is important to consider the specific nature of data in making binding site predictions. For example, the binding sites obtained by mutagenesis experiments are, almost always, more accurate than the ones from Y2H techniques. A method that uses binding sites obtained from Y2H as training data should explicitly model this imprecision.

Prediction of binding-related information

Binding is not a binary (bound/unbound) phenomenon: it has a number of characteristics such as binding energy, conformational change etc associated with it. Despite the difficulty involved, it can be useful to predict these characteristics for residues in a binding interface along with the binding site. For instance, predicted binding energy for a residue can be used to identify whether a residue is a hotspot residue or not. Gromiha et al. [16] have calculated interaction energies and degrees of conformational change of all individual protein residues in different protein-protein interfaces from the 3D structure of a protein complex. However, in the knowledge of this author, no methods for predicting these binding characteristics from sequence information alone exists in the literature.

Interpretability of results

It is important that the binding predictions made by a method be interpretable. For example, it is interesting to see what specific motifs are involved in binding. Most feature based machine learning

approaches such as nonlinear SVMs or neural networks almost always work as black boxes. Some recent machine learning techniques such as Positional Oligomer Importance Matrices (POIMs) [153] and Feature Importance Ranking Measure (FIRM) [154] offer good interpretability and high accuracy for kernel based methods.

5 Acknowledgements

I am grateful to Mark Rogers and Michael Hamilton for their reviews which helped me in improving the quality of this research exam write-up.

Funding: I would like thank the US Department of State and Higher Education Commission of the government of Pakistan for their funding through Fulbright scholarship program.

6 References

[1] R. A. Freitas Jr., "Human Body Chemical Composition (section 3.1)," in *Nanomedicine: Basic Capabilities*, vol. 1, Landes Bioscience, 1999.
[2] A. M. Stadler, I. Digel, G. M. Artmann, J. P. Embs, G. Zaccai, and G. Büldt, "Hemoglobin Dynamics in Red Blood Cells: Correlation to Body Temperature," *Biophysical Journal*, vol. 95, no. 11, pp. 5449–5461, Dec. 2008.
[3] G. A. Petsko and D. Ringe, *Protein structure and function*. New Science Press, 2004.
[4] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, and C. Zardecki, "The Protein Data Bank," *Acta Crystallogr. D Biol. Crystallogr.*, vol. 58, no. Pt 6 No 1, pp. 899–907, Jun. 2002.
[5] E. Fischer, "Einfluss der Configuration auf die Wirkung der Enzyme," *Berichte der deutschen chemischen Gesellschaft*, vol. 27, no. 3, pp. 2985–2993, Oct. 1894.
[6] A. Kahraman, R. J. Morris, R. A. Laskowski, and J. M. Thornton, "Shape Variation in Protein Binding Pockets and their Ligands," *Journal of Molecular Biology*, vol. 368, no. 1, pp. 283–301, Apr. 2007.
[7] D. E. Koshland, "Application of a Theory of Enzyme Specificity to Protein Synthesis," *PNAS*, vol. 44, no. 2, pp. 98–104, Feb. 1958.
[8] H. R. Bosshard, "Molecular Recognition by Induced Fit: How Fit Is the Concept?," *Physiology*, vol. 16, no. 4, pp. 171–173, Aug. 2001.
[9] C. L. Drum, S.-Z. Yan, J. Bard, Y.-Q. Shen, D. Lu, S. Soelaiman, Z. Grabarek, A. Bohm, and W.-J. Tang, "Structural basis for the activation of anthrax adenylyl cyclase exotoxin by calmodulin," *Nature*, vol. 415, no. 6870, pp. 396–402, Jan. 2002.
[10] Y. Ofran and B. Rost, "Protein–Protein Interaction Hotspots Carved into Sequences," *PLoS Comput Biol*, vol. 3, no. 7, p. e119, Jul. 2007.
[11] *Jmol: an open-source Java viewer for chemical structures in 3D*. .
[12] M. L. Connolly, "Solvent-Accessible Surfaces of Proteins and Nucleic Acids," *Science*, vol. 221, no. 4612, pp. 709–713, Aug. 1983.
[13] C. Chothia and J. Janin, "Principles of protein-protein recognition," *, Published online: 28 August 1975; | doi:10.1038/256705a0*, vol. 256, no. 5520, pp. 705–708, Aug. 1975.
[14] S. Gong, C. Park, H. Choi, J. Ko, I. Jang, J. Lee, D. M. Bolser, D. Oh, D.-S. Kim, and J. Bhak, "A protein domain interaction interface database: InterPare," *BMC Bioinformatics*, vol. 6, p. 207, Aug. 2005.
[15] Y. Ofran, "Prediction of Protein Interaction Sites," in *Computational Protein-Protein Interaction*, CRC Press, pp. 167–184.
[16] M. M. Gromiha, N. Saranya, S. Selvaraj, B. Jayaram, and K. Fukui, "Sequence and structural features of binding site residues in protein-protein complexes: comparison with protein-nucleic acid complexes," *Proteome Science*, vol. 9, no. Suppl 1, p. S13, Oct. 2011.
[17] K. H. Young, "Yeast Two-Hybrid: So Many Interactions, (in) so Little Time...," *Biology of Reproduction*, vol. 58, no. 2, pp. 302–311, Feb. 1998.
[18] A. Bauer and B. Kuster, "Affinity purification-mass spectrometry. Powerful tools for the characterization of protein complexes," *Eur. J. Biochem.*, vol. 270, no. 4, pp. 570–578, Feb. 2003.
[19] H. Zhu, M. Bilgin, R. Bangham, D. Hall, A. Casamayor, P. Bertone, N. Lan, R. Jansen, S. Bidlingmaier, T. Houfek, T. Mitchell, P. Miller, R. A. Dean, M. Gerstein, and M. Snyder, "Global Analysis of Protein Activities Using Proteome Chips," *Science*, vol. 293, no. 5537, pp. 2101–2105, Sep. 2001.
[20] J. Ule, K. Jensen, A. Mele, and R. B. Darnell, "CLIP: A method for identifying protein–RNA interaction sites in living cells," *Methods*, vol. 37, no. 4, pp. 376–386, Dec. 2005.
[21] J. Konig, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D. J. Turner, N. M. Luscombe, and J. Ule, "iCLIP - Transcriptome-wide Mapping of Protein-RNA Interactions with Individual Nucleotide Resolution," *J Vis Exp*, no. 50, Apr. 2011.
[22] S. Kishore, L. Jaskiewicz, L. Burger, J. Hausser, M. Khorshid, and M. Zavolan, "A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins," *Nature Methods*, vol. 8, no. 7, pp. 559–564, May 2011.
[23] P. Collas, "The Current State of Chromatin Immunoprecipitation," *Molecular Biotechnology*, vol. 45, no. 1, pp. 87–100, 2010.
[24] A. A. Bliznyuk and J. E. Gready, "Simple method for locating possible ligand binding sites on protein surfaces," *Journal of Computational Chemistry*, vol. 20, no. 9, pp. 983–988, Jul. 1999.
[25] J. Liang, H. Edelsbrunner, and C. Woodward, "Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design," *Protein Sci.*, vol. 7, no. 9, pp. 1884–1897, Sep. 1998.

[26] P. Frasconi and A. Passerini, “Predicting the Geometry of Metal Binding Sites from Protein Sequence.”
[27] H. Neuvirth, R. Raz, and G. Schreiber, “ProMate: a structure based prediction program to identify the location of protein-protein binding sites,” *J. Mol. Biol.*, vol. 338, no. 1, pp. 181–199, Apr. 2004.
[28] J.-L. Chung, W. Wang, and P. E. Bourne, “High-throughput identification of interacting protein-protein binding sites,” *BMC Bioinformatics*, vol. 8, no. 1, p. 223, Jun. 2007.
[29] J.-L. Chung, W. Wang, and P. E. Bourne, “Exploiting sequence and structure homologs to identify protein-protein binding sites,” *Proteins*, vol. 62, no. 3, pp. 630–640, Mar. 2006.
[30] J.-F. Xia, X.-M. Zhao, J. Song, and D.-S. Huang, “APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility,” *BMC Bioinformatics*, vol. 11, p. 174, Apr. 2010.
[31] A. Crivici and M. Ikura, “Molecular and Structural Basis of Target Recognition by Calmodulin,” *Annual Review of Biophysics and Biomolecular Structure*, vol. 24, no. 1, pp. 85–116, 1995.
[32] M. R. Wilkins and S. K. Kummerfeld, “Sticking together? Falling apart? Exploring the dynamics of the interactome,” *Trends in Biochemical Sciences*, vol. 33, no. 5, pp. 195–200, May 2008.
[33] J. van Dieck, D. P. Teufel, A. M. Jaulent, M. R. Fernandez-Fernandez, T. J. Rutherford, A. Wyslouch-Cieszynska, and A. R. Fersht, “Posttranslational modifications affect the interaction of S100 proteins with tumor suppressor p53,” *J. Mol. Biol.*, vol. 394, no. 5, pp. 922–930, Dec. 2009.
[34] M. Levitt, “Growth of Novel Protein Structural Data,” *PNAS*, vol. 104, no. 9, pp. 3183–3188, Feb. 2007.
[35] W. Kabsch and C. Sander, “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features,” *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, Dec. 1983.
[36] S. Liang, C. Zhang, S. Liu, and Y. Zhou, “Protein Binding Site Prediction Using an Empirical Scoring Function,” *Nucl. Acids Res.*, vol. 34, no. 13, pp. 3698–3707, Jan. 2006.
[37] E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, and Y. Zhou, “SPINE X: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles,” *Journal of Computational Chemistry*, vol. 33, no. 3, pp. 259–267, Jan. 2012.
[38] S. Jones and J. M. Thornton, “Prediction of protein-protein interaction sites using patch analysis,” *J. Mol. Biol.*, vol. 272, no. 1, pp. 133–143, Sep. 1997.
[39] H.-X. Zhou and S. Qin, “Interaction-site prediction for protein complexes: a critical assessment,” *Bioinformatics*, vol. 23, no. 17, pp. 2203–2209, Sep. 2007.
[40] A. A. Bogan and K. S. Thorn, “Anatomy of hot spots in protein interfaces,” *Journal of Molecular Biology*, vol. 280, no. 1, pp. 1–9, Jul. 1998.
[41] A. Ciulli, G. Williams, A. G. Smith, T. L. Blundell, and C. Abell, “Probing Hot Spots at Protein−Ligand Binding Sites: A Fragment-Based Approach Using Biophysical Methods,” *J. Med. Chem.*, vol. 49, no. 16, pp. 4992–5000, 2006.
[42] T. Pupko, R. E. Bell, I. Mayrose, F. Glaser, and N. Ben-Tal, “Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues,” *Bioinformatics*, vol. 18, no. Suppl 1, p. S71–S77, Jul. 2002.
[43] N. Tuncbag, A. Gursoy, and O. Keskin, “Identification of Computational Hot Spots in Protein Interfaces: Combining Solvent Accessibility and Inter-Residue Potentials Improves the Accuracy,” *Bioinformatics*, vol. 25, no. 12, pp. 1513–1520, Jun. 2009.
[44] J. A. Wells, “Systematic mutational analyses of protein-protein interfaces,” *Meth. Enzymol.*, vol. 202, pp. 390–411, 1991.
[45] K. Morrison and G. Weiss, “Combinatorial alanine-scanning,” *Current Opinion in Chemical Biology*, vol. 5, no. 3, pp. 302–307, 2001.
[46] S. Vajda and F. Guarnieri, “Characterization of protein-ligand interaction sites using experimental and computational methods,” *Curr Opin Drug Discov Devel*, vol. 9, no. 3, pp. 363–369, 2006.
[47] S. Mika and B. Rost, “Protein–Protein Interactions More Conserved within Species than across Species,” *PLoS Comput Biol*, vol. 2, no. 7, p. e79, Jul. 2006.
[48] D. R. Caffrey, S. Somaroo, J. D. Hughes, J. Mintseris, and E. S. Huang, “Are protein-protein interfaces more conserved in sequence than the rest of the protein surface?,” *Protein Sci.*, vol. 13, no. 1, pp. 190–202, Jan. 2004.
[49] N. V. Grishin and M. A. Phillips, “The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences,” *Protein Science*, vol. 3, no. 12, pp. 2455–2458, Dec. 1994.
[50] W. S. J. Valdar, “Scoring residue conservation,” *Proteins: Structure, Function, and Bioinformatics*, vol. 48, no. 2, pp. 227–241, Aug. 2002.
[51] K. Arnold, L. Bordoli, J. Kopp, and T. Schwede, “The SWISS-MODEL Workspace: A Web-Based Environment for Protein Structure Homology Modelling,” *Bioinformatics*, vol. 22, no. 2, pp. 195–201, Jan. 2006.
[52] C. Lambert, N. Léonard, X. De Bolle, and E. Depiereux, “ESyPred3D: Prediction of Proteins 3D Structures,” *Bioinformatics*, vol. 18, no. 9, pp. 1250–1256, Aug. 2002.
[53] N. Tuncbag, G. Kar, O. Keskin, A. Gursoy, and R. Nussinov, “A Survey of Available Tools and Web Servers for Analysis of Protein–Protein Interactions and Interfaces,” *Brief Bioinform*, vol. 10, no. 3, pp. 217–232, May 2009.
[54] I. Ezkurdia, L. Bartoli, P. Fariselli, R. Casadio, A. Valencia, and M. L. Tress, “Progress and Challenges in Predicting Protein–Protein Interaction Sites,” *Brief Bioinform*, vol. 10, no. 3, pp. 233–246, May 2009.
[55] S. J. de Vries and A. M. J. J. Bonvin, “How proteins get in touch: interface prediction in the study of biomolecular complexes,” *Curr. Protein Pept. Sci.*, vol. 9, no. 4, pp. 394–406, Aug. 2008.
[56] S. Leis, S. Schneider, and M. Zacharia, “In Silico Prediction of Binding Sites on Proteins,” *Current Medicinal Chemistry*, vol. 2010, no. 17, pp. 1550–1562, 2010.
[57] J. Fernández-Recio, “Prediction of protein binding sites and hot spots,” *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 1, no. 5, pp. 680–698, Sep. 2011.
[58] K. S. Guimarães, R. Jothi, E. Zotenko, and T. M. Przytycka, “Predicting domain-domain interactions using a parsimony approach,” *Genome Biol*, vol. 7, no. 11, p. R104, 2006.
[59] A. J. González and L. Liao, “Predicting domain-domain interaction based on domain profiles with feature selection and support vector machines,” *BMC Bioinformatics*, vol. 11, no. 1, p. 537, 2010.

[60] R. Riley, C. Lee, C. Sabatti, and D. Eisenberg, “Inferring protein domain interactions from databases of interacting proteins,” *Genome Biol.*, vol. 6, no. 10, p. R89, 2005.
[61] H. Wang, E. Segal, A. Ben-Hur, Q.-R. Li, M. Vidal, and D. Koller, “InSite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale,” *Genome Biol*, vol. 8, no. 9, p. R192, 2007.
[62] Y. Ofran and B. Rost, “ISIS: Interaction Sites Identified from Sequence,” *Bioinformatics*, vol. 23, no. 2, p. e13–e16, Jan. 2007.
[63] I. M. A. Nooren and J. M. Thornton, “Diversity of protein|[ndash]|protein interactions,” *The EMBO Journal*, vol. 22, no. 14, pp. 3486–3492, Jul. 2003.
[64] Y. Ofran, A. Schlessinger, and B. Rost, “Automated identification of complementarity determining regions (CDRs) reveals peculiar characteristics of CDRs and B cell epitopes,” *J. Immunol.*, vol. 181, no. 9, pp. 6230–6235, Nov. 2008.
[65] R. Tibshirani, “Regression Shrinkage and Selection Via the Lasso,” *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.
[66] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1999.
[67] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, “Support Vector Machines and Kernels for Computational Biology,” *PLoS Comput Biol*, vol. 4, no. 10, p. e1000173, Oct. 2008.
[68] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” in *Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, CA, USA, 2001, pp. 282–289.
[69] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[70] S. Ahmad and K. Mizuguchi, “Partner-Aware Prediction of Interacting Residues in Protein-Protein Complexes from Sequence Data,” *PLoS One*, vol. 6, no. 12, Dec. 2011.
[71] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium,” *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000.
[72] L. Falquet, M. Pagni, P. Bucher, N. Hulo, C. J. A. Sigrist, K. Hofmann, and A. Bairoch, “The PROSITE database, its status in 2002,” *Nucleic Acids Res.*, vol. 30, no. 1, pp. 235–238, Jan. 2002.
[73] A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy, “The Pfam protein families database,” *Nucleic Acids Res.*, vol. 32, no. Database issue, pp. D138–141, Jan. 2004.
[74] A. Amos-Binks, C. Patulea, S. Pitre, A. Schoenrock, Y. Gui, J. R. Green, A. Golshani, and F. Dehne, “Binding Site Prediction for Protein-Protein Interactions and Novel Motif Discovery using Re-occurring Polypeptide Sequences,” *BMC Bioinformatics*, vol. 12, no. 1, p. 225, Jun. 2011.
[75] S. Pitre, F. Dehne, A. Chan, J. Cheetham, A. Duong, A. Emili, M. Gebbia, J. Greenblatt, M. Jessulat, N. Krogan, X. Luo, and A. Golshani, “PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs,” *BMC Bioinformatics*, vol. 7, no. 1, p. 365, Jul. 2006.
[76] Y. Park, “Critical assessment of sequence-based protein-protein interaction prediction methods that do not require homologous protein sequences,” *BMC Bioinformatics*, vol. 10, p. 419, Dec. 2009.
[77] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.,” *Nucleic Acids Res*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.
[78] F. ul A. A. Minhas and A. Ben-Hur, “Multiple instance learning of Calmodulin binding sites,” presented at the Submitted for review in 11th European conference on Computational Biology (ECCB), Basel, Switzerland, 2012.
[79] A. Porollo and J. Meller, “Prediction-based fingerprints of protein–protein interactions,” *Proteins: Structure, Function, and Bioinformatics*, vol. 66, no. 3, pp. 630–645, Feb. 2007.
[80] L. Jacob and J.-P. Vert, “Protein-Ligand Interaction Prediction: An Improved Chemogenomics Approach,” *Bioinformatics*, vol. 24, no. 19, pp. 2149–2156, Oct. 2008.
[81] B. Liu, X. Wang, L. Lin, B. Tang, Q. Dong, and X. Wang, “Prediction of protein binding sites in protein structures using hidden Markov support vector machine,” *BMC Bioinformatics*, vol. 10, no. 1, p. 381, Nov. 2009.
[82] Y. Altun, I. Tsochantaridis, and T. Hofmann, *Hidden Markov Support Vector Machines*. 2003.
[83] J. R. Bradford, C. J. Needham, A. J. Bulpitt, and D. R. Westhead, “Insights into Protein–Protein Interfaces using a Bayesian Network Prediction Method,” *Journal of Molecular Biology*, vol. 362, no. 2, pp. 365–386, Sep. 2006.
[84] M.-H. Li, L. Lin, X.-L. Wang, and T. Liu, “Protein–Protein Interaction Site Prediction Based on Conditional Random Fields,” *Bioinformatics*, vol. 23, no. 5, pp. 597–604, Mar. 2007.
[85] T. Friedrich, B. Pils, T. Dandekar, J. Schultz, and T. Müller, “Modelling Interaction Sites in Protein Domains with Interaction Profile Hidden Markov Models,” *Bioinformatics*, vol. 22, no. 23, pp. 2851–2857, Dec. 2006.
[86] Y. Qi, M. Oja, J. Weston, and W. S. Noble, “A Unified Multitask Architecture for Predicting Local Protein Properties,” *PLoS ONE*, vol. 7, no. 3, p. e32235, Mar. 2012.
[87] K. Y. Yip, P. M. Kim, D. McDermott, and M. Gerstein, “Multi-level learning: improving the prediction of protein, domain and residue interactions by allowing information flow between levels,” *BMC Bioinformatics*, vol. 10, p. 241, 2009.
[88] Q.-T. Nguyen, R. Fablet, and D. Pastor, “Protein Interaction Hotspot Identification Using Sequence-based Frequency-derived Features,” *Biomedical Engineering, IEEE Transactions on*, vol. PP, no. 99, p. 1, 2011.
[89] H. Chen and H. Zhou, “Prediction of interface residues in protein–protein complexes by a consensus neural network method: Test against NMR data,” *Proteins: Structure, Function, and Bioinformatics*, vol. 61, no. 1, pp. 21–35, Oct. 2005.
[90] J. R. Bradford and D. R. Westhead, “Improved prediction of protein-protein binding sites using a support vector machines approach,” *Bioinformatics*, vol. 21, no. 8, pp. 1487–1494, Apr. 2005.
[91] S. Qin and H.-X. Zhou, “Meta-PPISP: A Meta Web Server for Protein-Protein Interaction Site Prediction,” *Bioinformatics*, vol. 23, no. 24, pp. 3386–3387, Dec. 2007.
[92] X. Zhu and J. C. Mitchell, “KFC2: A knowledge-based hot spot prediction method based on interface solvation, atomic density, and plasticity features,” *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. 9, pp. 2671–2683, Sep. 2011.

[93] S. Lise, D. Buchan, M. Pontil, and D. T. Jones, “Predictions of Hot Spot Residues at Protein-Protein Interfaces Using Support Vector Machines,” *PLoS ONE*, vol. 6, no. 2, p. e16774, Feb. 2011.
[94] Y. Chen and G. Varani, “Protein families and RNA recognition,” *FEBS Journal*, vol. 272, no. 9, pp. 2088–2097, Apr. 2005.
[95] T. A. Cooper, L. Wan, and G. Dreyfuss, “RNA and Disease,” *Cell*, vol. 136, no. 4, pp. 777–793, Feb. 2009.
[96] M. M. Gromiha and K. Fukui, “Scoring Function Based Approach for Locating Binding Sites and Understanding Recognition Mechanism of Protein−DNA Complexes,” *J. Chem. Inf. Model.*, vol. 51, no. 3, pp. 721–729, 2011.
[97] N. Rawat and P. Biswas, “Shape, flexibility and packing of proteins and nucleic acids in complexes,” *Physical Chemistry Chemical Physics*, vol. 13, no. 20, p. 9632, 2011.
[98] C. Gini, “Concentration and Dependency Ratios,” *English translation in Rivista di Politica Economica*, vol. 87, pp. 769–789, original published in Italian in 1909 1997.
[99] C. Kauffman and G. Karypis, “Computational tools for protein-DNA interactions,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 14–28, Jan. 2012.
[100] S. Ahmad, M. M. Gromiha, and A. Sarai, “Analysis and Prediction of DNA-Binding Proteins and Their Binding Residues Based on Composition, Sequence and Structural Information,” *Bioinformatics*, vol. 20, no. 4, pp. 477–486, Mar. 2004.
[101] P. Ozbek, S. Soner, B. Erman, and T. Haliloglu, “DNABINDPROT: fluctuation-based predictor of DNA-binding residues within a network of interacting residues,” *Nucleic Acids Research*, vol. 38, no. Web Server, p. W417–W423, May 2010.
[102] M. Gao and J. Skolnick, “DBD-Hunter: A Knowledge-Based Method for the Prediction of DNA–Protein Interactions,” *Nucl. Acids Res.*, vol. 36, no. 12, pp. 3978–3992, Jul. 2008.
[103] R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann, “Origins of Specificity in Protein-DNA Recognition,” *Annual Review of Biochemistry*, vol. 79, no. 1, pp. 233–269, 2010.
[104] A. Sarai and H. Kono, “Protein-Dna Recognition Patterns and Predictions,” *Annual Review of Biophysics and Biomolecular Structure*, vol. 34, no. 1, pp. 379–398, 2005.
[105] S. Ahmad and A. Sarai, “PSSM-based prediction of DNA binding sites in proteins,” *BMC Bioinformatics*, vol. 6, p. 33, 2005.
[106] C. Yan, M. Terribilini, F. Wu, R. L. Jernigan, D. Dobbs, and V. Honavar, “Predicting DNA-binding sites of proteins from amino acid sequence,” *BMC Bioinformatics*, vol. 7, p. 262, May 2006.
[107] Y. Ofran, V. Mysore, and B. Rost, “Prediction of DNA-Binding Residues from Sequence,” *Bioinformatics*, vol. 23, no. 13, p. i347–i353, Jul. 2007.
[108] M. B. Carson, R. Langlois, and H. Lu, “NAPS: A Residue-Level Nucleic Acid-Binding Prediction Server,” *Nucl. Acids Res.*, May 2010.
[109] S.-Y. Ho, F.-C. Yu, C.-Y. Chang, and H.-L. Huang, “Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM-PSSM method,” *BioSystems*, vol. 90, no. 1, pp. 234–241, Aug. 2007.
[110] L. Wang, M. Q. Yang, and J. Y. Yang, “Prediction of DNA-binding residues from protein sequence information using random forests,” *BMC Genomics*, vol. 10, no. Suppl 1, p. S1, Jul. 2009.
[111] L. Wang, C. Huang, M. Yang, and J. Y. Yang, “BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features,” *BMC Systems Biology*, vol. 4, no. Suppl 1, p. S3, 2010.
[112] C. Sander and R. Schneider, “Database of homology-derived protein structures and the structural meaning of sequence alignment,” *Proteins*, vol. 9, no. 1, pp. 56–68, 1991.
[113] B. Rost, C. Sander, B. Rost, and C. Sander, “Combining evolutionary information and neural networks to predict protein secondary structure, Combining evolutionary information and neural networks to predict protein secondary structure,” *Proteins: Structure, Function, and Bioinformatics, Proteins: Structure, Function, and Bioinformatics*, vol. 19, 19, no. 1, 1, pp. 55, 55–72, 72, May 1994.
[114] J. Si, Z. Zhang, B. Lin, M. Schroeder, and B. Huang, “MetaDBSite: a meta approach to improve protein DNA-binding sites prediction,” *BMC Systems Biology*, vol. 5, no. Suppl 1, p. S7, 2011.
[115] I. Ladunga, Ed., “Chapter 1: An overview of the computational analyses and discovery of transcription factor binding sites,” in *Computational Biology of Transcription Factor Binding*, .
[116] G. D. Stormo, “DNA Binding Sites: Representation and Discovery,” *Bioinformatics*, vol. 16, no. 1, pp. 16–23, Jan. 2000.
[117] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, “Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment,” *Science*, vol. 262, no. 5131, pp. 208–214, Oct. 1993.
[118] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li, “MEME: discovering and analyzing DNA and protein sequence motifs,” *Nucleic Acids Res*, vol. 34, no. Web Server issue, p. W369–W373, Jul. 2006.
[119] G. Z. Hertz, G. W. Hartzell, and G. D. Stormo, “Identification of Consensus Patterns in Unaligned DNA Sequences Known to Be Functionally Related,” *Comput Appl Biosci*, vol. 6, no. 2, pp. 81–92, Apr. 1990.
[120] Q. Zhou and J. S. Liu, “Modeling Within-Motif Dependence for Transcription Factor Binding Site Predictions,” *Bioinformatics*, vol. 20, no. 6, pp. 909–916, Apr. 2004.
[121] M. A. Kon, Y. Fan, D. Holloway, and C. DeLisi, “SVMotif: A Machine Learning Motif Algorithm,” in *Proceedings of the Sixth International Conference on Machine Learning and Applications*, Washington, DC, USA, 2007, pp. 573–580.
[122] E. Redhead and T. L. Bailey, “Discriminative motif discovery in DNA and protein sequences using the DEME algorithm,” *BMC Bioinformatics*, vol. 8, no. 1, p. 385, 2007.
[123] A. D. Smith, P. Sumazin, and M. Q. Zhang, “Identifying Tissue-Selective Transcription Factor Binding Sites in Vertebrate Promoters,” *PNAS*, vol. 102, no. 5, pp. 1560–1565, Feb. 2005.
[124] M. Gupta and J. S. Liu, “De novo cis-regulatory module elicitation for eukaryotic genomes,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, no. 20, pp. 7079–7084, May 2005.
[125] E. Segal and R. Sharan, “A discriminative model for identifying spatial cis-regulatory modules,” *J. Comput. Biol.*, vol. 12, no. 6, pp. 822–834, Aug. 2005.
[126] B. Jiang and J. S. Liu, “Statistical Learning and Modeling of TF-DNA Binding,” in *Handbook of Statistical Bioinformatics*, H. H.-S. Lu, B. Schölkopf, and H. Zhao, Eds. Springer Berlin Heidelberg, 2011, pp. 55–72.
[127] Q. Zhou and J. S. Liu, “Extracting Sequence Features to Predict Protein–DNA Interactions: A Comparative Study,” *Nucl. Acids Res.*, vol. 36, no. 12, pp. 4137–4148, Jul. 2008.

[128] H. A. Chipman, E. I. George, and R. E. McCulloch, “BART: Bayesian additive regression trees,” *arXiv:0806.3286*, Jun. 2008.
[129] L. Narlikar, R. Gordân, and A. J. Hartemink, “A Nucleosome-Guided Map of Transcription Factor Binding Sites in Yeast,” *PLoS Comput Biol*, vol. 3, no. 11, p. e215, Nov. 2007.
[130] G.-C. Yuan, P. Ma, W. Zhong, and J. S. Liu, “Statistical assessment of the global regulatory role of histone acetylation in Saccharomyces cerevisiae,” *Genome Biol.*, vol. 7, no. 8, p. R70, 2006.
[131] S. A. Ramsey, T. A. Knijnenburg, K. A. Kennedy, D. E. Zak, M. Gilchrist, E. S. Gold, C. D. Johnson, A. E. Lampano, V. Litvak, G. Navarro, T. Stolyar, A. Aderem, and I. Shmulevich, “Genome-Wide Histone Acetylation Data Improve Prediction of Mammalian Transcription Factor Binding Sites,” *Bioinformatics*, vol. 26, no. 17, pp. 2071–2075, Sep. 2010.
[132] G. Cuellar-Partida, F. A. Buske, R. C. McLeay, T. Whitington, W. S. Noble, and T. L. Bailey, “Epigenetic Priors for Identifying Active Transcription Factor Binding Sites,” *Bioinformatics*, vol. 28, no. 1, pp. 56–62, Jan. 2012.
[133] K.-J. Won, B. Ren, and W. Wang, “Genome-wide prediction of transcription factor binding sites using an integrated model,” *Genome Biology*, vol. 11, no. 1, p. R7, 2010.
[134] Y. Murakami, R. V. Spriggs, H. Nakamura, and S. Jones, “PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences,” *Nucleic Acids Res*, vol. 38, no. Web Server issue, p. W412–W416, Jul. 2010.
[135] T. Puton, L. Kozlowski, I. Tuszynska, K. Rother, and J. M. Bujnicki, “Computational methods for prediction of protein–RNA interactions,” *Journal of Structural Biology*, no. 0.
[136] L. Wang and S. J. Brown, “BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences,” *Nucleic Acids Research*, vol. 34, no. Web Server, p. W243–W248, Jul. 2006.
[137] J. J. Ellis, M. Broom, and S. Jones, “Protein-RNA interactions: structural analysis and functional classes,” *Proteins*, vol. 66, no. 4, pp. 903–911, Mar. 2007.
[138] M. Wagner, R. Adamczak, A. Porollo, and J. Meller, “Linear Regression Models for Solvent Accessibility Prediction in Proteins,” *Journal of Computational Biology*, vol. 12, no. 3, pp. 355–369, Apr. 2005.
[139] H. Kazan, D. Ray, E. T. Chan, T. R. Hughes, and Q. Morris, “RNAcontext: A New Method for Learning the Sequence and Structure Binding Preferences of RNA-Binding Proteins,” *PLoS Comput Biol*, vol. 6, no. 7, p. e1000832, Jul. 2010.
[140] Y. Ding and C. E. Lawrence, “A Statistical Sampling Algorithm for RNA Secondary Structure Prediction,” *Nucl. Acids Res.*, vol. 31, no. 24, pp. 7280–7301, Dec. 2003.
[141] M. Hiller, R. Pudimat, A. Busch, and R. Backofen, “Using RNA secondary structures to guide sequence motif finding towards single-stranded regions,” *Nucleic Acids Res.*, vol. 34, no. 17, p. e117, 2006.
[142] B. C. Foat, A. V. Morozov, and H. J. Bussemaker, “Statistical Mechanical Modeling of Genome-Wide Transcription Factor Occupancy Data by MatrixREDUCE,” *Bioinformatics*, vol. 22, no. 14, p. e141–e149, Jul. 2006.
[143] C. Livi, L. Paillard, E. Blanzieri, and Y. Audic, “Identification of Regulatory Binding Sites on mRNA Using in Vivo Derived Informations and SVMs,” in *6th International Conference on Practical Applications of Computational Biology & Bioinformatics*, vol. 154, M. P. Rocha, N. Luscombe, F. Fdez-Riverola, and J. M. C. Rodríguez, Eds. Springer Berlin / Heidelberg, 2012, pp. 33–41.
[144] C. Leslie, E. Eskin, and W. S. Noble, “The spectrum kernel: a string kernel for SVM protein classification,” presented at the Pacific Symposium on Biocomputing, 2002, pp. 566–575.
[145] S. Sonnenburg, G. Schweikert, P. Philips, J. Behr, and G. Rätsch, “Accurate splice site prediction using support vector machines,” *BMC Bioinformatics*, vol. 8, no. Suppl 10, p. S7, 2007.
[146] M. F. Rogers, J. Thomas, A. S. Reddy, and A. Ben-Hur, “SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data,” *Genome Biology*, vol. 13, no. 1, p. R4, 2012.
[147] B. Mersch, A. Gepperth, S. Suhai, and A. Hotz-Wagenblatt, “Automatic detection of exonic splicing enhancers (ESEs) using SVMs,” *BMC Bioinformatics*, vol. 9, no. 1, p. 369, 2008.
[148] A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller, “Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites,” *Bioinformatics*, vol. 16, no. 9, pp. 799–807, Sep. 2000.
[149] C. Kauffman, G. Karypis, C. Kauffman, and G. Karypis, “Ligand-Binding Residue Prediction,” in *Introduction to Protein Structure Prediction*, John Wiley & Sons, Inc., pp. 343–368.
[150] D. Ghersi and R. Sanchez, “Beyond structural genomics: computational approaches for the identification of ligand binding sites in protein structures,” *Journal of Structural and Functional Genomics*, vol. 12, no. 2, pp. 109–117, Jul. 2011.
[151] D. Plewczynski, A. Tkacz, L. S. Wyrwicz, and L. Rychlewski, “AutoMotif server: prediction of single residue post-translational modifications in proteins,” *Bioinformatics*, vol. 21, no. 10, pp. 2525 –2527.
[152] S. Ahmad, O. Keskin, A. Sarai, and R. Nussinov, “Protein–DNA Interactions: Structural, Thermodynamic and Clustering Patterns of Conserved Residues in DNA-Binding Proteins,” *Nucl. Acids Res.*, vol. 36, no. 18, pp. 5922–5932, Oct. 2008.
[153] S. Sonnenburg, A. Zien, P. Philips, and G. Rätsch, “POIMs: Positional Oligomer Importance Matrices—Understanding Support Vector Machine-Based Signal Detectors,” *Bioinformatics*, vol. 24, no. 13, p. i6–i14, Jul. 2008.
[154] A. Zien, N. Kraemer, S. Sonnenburg, and G. Raetsch, “The Feature Importance Ranking Measure,” *arXiv:0906.4258*, Jun. 2009.