Computational Tools for Protein-DNA Interactions

Christopher Kauffman and George Karypis

Abstract

Interactions between DNA and proteins are central to living systems, and characterizing how and when they occur would greatly enhance our understanding of working genomes. We review the different computational problems associated with protein-DNA interactions and the various methods used to solve them. A wide range of topics is covered including physics-based models for direct and indirect recognition, identification of transcription factor binding sites, and methods to predict DNA-binding proteins. Our goal is to introduce this important problem domain to data mining researchers by identifying the key issues and challenges inherent to the area as well as provide directions for fruitful future research.

Interactions between deoxyribonucleic acid (DNA) and proteins are widely recognized as central to living systems. These interactions come in a variety of forms including repair of damaged DNA and transcription of genes into RNA. More recently it has been found that, by binding to certain DNA segments, proteins can promote or repress the transcription of genes in the vicinity of the binding site. Proteins of this kind are referred to as transcription factors (TFs). The number of TFs in an organism appears to be related to the complexity of the underlying genome: as the number of of genes increases, the number of TFs increases according to a power law [1]. This many-fold increase of TFs appears to be required in order to manage transcription in higher organisms.

Characterizing how and when protein-DNA interactions occur would greatly enhance our understanding of the genome at work. A full picture of the interactions will eventually allow characterization of which genes are transcribed at any given time in order for the organism to react dynamically to a changing environment. Protein-DNA interactions are studied both in the wet lab and computationally. Here a synergy exists: lab experiments provide data and problems for computational methods to solve while computation provides hypotheses which guide additional lab experiments.

The goal of this article is to review three major areas of interest for computational studies of protein DNA interactions: (1) physics-based studies of protein-DNA interaction, (2) identification of transcription factor binding sites, and (3) identification of DNA-binding proteins.

How Many Binding Proteins Exist?

Accounts of how many DNA-binding proteins exist vary through the literature. Attention is particularly focused on transcription factors. Older sources estimated that 2-3% of a prokaryotic genome and 6-7% of a eukaryotic genome encodes DNA-binding proteins [2]. This number was taken from the automatic gene annotation tool PEDANT [3]. Though contemporary estimates of the number of transcription factors range as high as 10% of all mammalian genes [4], averaging across genomes in the DBD database [5] classifies 4.65% of Metazoan (animal) genes as transcription factors (806 genes per animal genome). [1].

According to gene ontology annotations in PEDANT, there are currently 1714 genes in the human genome identified as coding for DNA-binding proteins with 885 of them identified as

Table 1: Counts of genes with DNA-binding annotations for human in the AMIGO gene ontology browser. Count is the raw number of genes, %all is the percentage of all genes that have the given GO term, %func is the percentage of genes with a molecular function which have the given GO annotation.

GO Term	Count	%all	%func
All gene products	18269	100.0	115.6
Molecular Function Given	15801	86.5	100.0
DNA-binding	2375	13.0	15.0
Transcription factor activity	969	5.3	6.1

transcription factors[1]. This is slightly smaller than the numbers currently in the AMIGO gene ontology browser[2] which are given in Table 1.

For researchers interested in DNA-binding protein structures, the protein data bank (PDB) [6] currently holds structures for 2372 proteins with DNA-binding gene ontology terms while 1400 of these actually have DNA structural information present in them. However, many of these structure entries are redundant in that their sequences are nearly identical: the largest data set of nonredundant structures reported in the literature contains 179 proteins [7]. See the discussion on available data sets later in this article.

Most proteins are composed of several independent units called *domains*. A domain which interacts with DNA is referred to as a *DNA-binding domain* and contains a structural motif that enables binding (see section 7.4 of [8]). A DNA-binding protein has a binding domain and possibly several other domains that determine its function. Multiple copies of DNA-binding domains may be present in a DNA-binding protein. This leads to some ambiguity in the literature as "DNA-binding protein" may sometimes refer only to the binding domain or the whole protein including both binding and nonbinding domains. Here we deal solely with interactions between binding domains and DNA. An overview of DNA-binding domains can be found in [2].

Physical Models and Energetics

Insight can be gained about DNA-protein interactions by studying them using physics models. Approaches in the literature examine bound protein-DNA complexes and either apply existing software to obtain interaction energy or develop new energy functions. Both approaches make use of complexed structures from the PDB. The goals of such studies are usually to establish why binding happens, to quantify energy changes between the bound and unbound states, and to understand how mutation in either protein or DNA may affect binding affinity. Basic understanding of binding physics guides both the development of transcription factor binding site models and the generation of protein and DNA features used in machine learning.

Early Work

An early review of the structure motifs used by transcription factors provided a number of principles used by the proteins to recognize DNA [9]. Subsequent studies on protein-DNA interactions characterized binding based on the frequency of protein residue contacts with nucleic acids in crystallized complexes [10, 11]. The propensity for each type of contact to form was

[1] `http://pedant.gsf.de/pedant3htmlview/pedant3view?Method=analysis&Db=p3_p168_Hom_sapie`
[2] `http://amigo.geneontology.org/cgi-bin/amigo/browse.cgi`

calculated by comparing the expected and observed number of contacts between each type of amino acid and nucleic acid to the number expected if contacts formed solely based on the frequency of each residue/nucleic acid type. The resulting propensities seemed to agree fairly well with the limited available experimental data: data from [12] is fit in [13] and in many cases the correct residue/base-pair combinations of zinc finger variants were predicted correctly. Stormo reviews some early developments of representations of binding preferences [14]. Simple models of recognition such as the theory of a linear code were popular early on: structural motifs in proteins allow matching of specific amino acids to specific DNA base pairs. This idea is still relevant to certain families of transcription factors such as the zinc finger domains [15].

Once a sufficient number of different DNA-binding protein families became available, it became apparent that various protein structures use diverse means of binding and achieving binding specificity to targeted sequences of DNA calling for more complex modeling techniques [16]. The current understanding is that binding is a stochastic process making probabilistic models more appropriate for model-ling protein-DNA interactions [13, 17]. Additionally, contacts between the protein and DNA backbone, between the protein backbone and DNA, and the presence of different types of interactions (electrostatic, van der Waals, hydrogen bonds, water-mediated bonds) has led to more detailed consideration of the energy change involved in binding.

Physics of Recognition Mechanisms

Protein-DNA-binding is thought to occur because the bound pair has lower free energy than the unbound molecules. A variety of factors governing free energy change are considered by Jayaram and coworkers in [18] such as desolvation of DNA and protein and electrostatic and van der Waals interactions. Some of these factors affect all molecular interactions, but order-studies of protein-DNA binding have identified two categories of binding mechanisms which allow specificity to be achieved. The first category involves energetically favorable interactions between atoms in the protein and DNA, sometimes called direct recognition or base recognition. The second category concerns the energy required to deform DNA to accommodate binding to the protein, referred to as indirect recognition or shape recognition. Both categories are described in detail in a review of recognition mechanisms [16] as well as a more recent review of the subject [19]. A few studies estimate both direct and indirect energies (e.g. [18]) while other work has studied direct [20] or indirect [21] recognition mechanisms separately. The review by Rohs et al. [19] advocates the concludes that these two recognition mechanisms are used together by almost all DNA-binding proteins, that binding site specificity is achieved by combining direct and indirect effects.

Specificity Tests: Mutating DNA and Protein Sequences

A common use of binding energetics models is to study DNA mutations and their effects on binding energy. Determining which DNA sequences result in low-energy binding to the protein indicates the protein's likely binding sites on the genome [22, 20]. The models of binding energetics can be verified using experimental techniques which measure binding specificity of proteins. These are reviewed by Stormo and Zhao in [23]. Aside from pure physics-based approaches, machine learning has been employed in some cases to aid in this task. An interesting example is in [24] where DNA-binding sequences for proteins were predicted by training a perceptron on deformation energy. This is in contrast to transcription factor binding site location, described in the next sections, which employ statistics and motif identification rather than physical models.

Transcription Factor Binding Site Identification

Transcription factors (TFs) are DNA-binding proteins whose primary purpose is to regulate the transcription of genes. Though there are some exceptions, many TFs accomplish regulation by binding to DNA at specific sites. The presence of the bound TF will attract or obstruct RNA polymerase thus promoting or repressing gene expression, respectively. TFs appear in greater abundance in eukaryotes and higher animals allowing more complex regulatory control of how and when genes are transcribed [1].

In order to form a picture of the working genome, it has become important to identify the genes that TFs affect by finding the genomic locations to which they bind. Computational tools comprise an important part of this discovery process.

Reviews of TF Binding Site Discovery

Transcription factor binding site identification is a well-studied area but continues to develop rapidly. Here we mention a few good reviews of the area which are useful to understanding the data and tools available for analysis.

Narlikar and Ovcharenko [25] provide a good overview of the lab science behind TF binding site identification and 184 citations to past literature. They include a brief section on computational tools to derive transcription factor properties/models like position weight matrices. Computational methods for discovering other genomic regulatory elements are also discussed.

Hannenhalli gives a review of current computational techniques for various representations of TF binding sites and how they are derived [26]. The review also briefly covers techniques for identification of other regulatory modules. Vingron and coworkers describe recent developments of computational techniques that expand on the capabilities of TF binding site identification [27]. An older but focused review from Bulyk is in [28] while Elnitski and coworkers cover methods specific to TF binding sites in mammals [29].

Charoensawan and coworkers give a current review of the resources available for study of TFs including databases of TFs with known binding sites and the types of annotations available for the TFs [1].

Finally, Das and Dai surveyed motif discovery algorithms which may be of use to determine appropriate algorithms for a particular task [30]. Supplementing this is a slightly older benchmark of motif discovery algorithms performed by Tompa and coworkers [31]. A blind evaluation of the algorithms was done on both synthetic and experimental data which may still be used as a guide for algorithm selection.

Motif Identification

Typically biologists are interested in which genes a TF regulates. This can be determined by identifying the genomic locations to which the TF binds. In motif identification, one starts with a collection of DNA sequences thought to contain TF binding sites. The computational task is to identify the TF binding site amongst these DNA sequences. Early approaches used simple models such as exact DNA motif sequences. These have largely been supplanted by position weight matrices (PWMs, alternatively referred to as position specific scoring matrices, PSSMs) as they more accurately model the probabilistic nature of binding. Though the assumption of independent contributions from each position of PWM is not entirely realistic [32], PWM methods are sufficient for the purpose of motif identification [33, 34, 35] especially when used in the context of locating entire regulatory modules [36]. More sophisticated models explore interdependence of DNA positions [37, 38] and use prior probability models based on TF class [39].

The newest models incorporate additional information specific to the experimental technique used to derive the DNA sequence collection [40].

An alternative to direct motif detection is phylogenetic footprinting. Homologous genomes are aligned to identify conserved noncoding regions which are likely to assume regulatory roles such as working as a TF binding sites. A number of such approaches are reviewed in [41, 42].

The function of a new gene can be inferred from the TFs associated with it. Using a library of transcription factor binding sites, one can detect TF binding sites in the noncoding region near a gene. Enrichment of a particular TF indicates the gene may share a function with other genes that the TF affects [43, 44].

Obtaining DNA Sequences for Motif Identification: Experimental Methods

Computational motif identification requires a collection of DNA sequences which contain a DNA-binding motif. Several wet lab techniques can provide such a collection by determining the approximate genomic location TF binding sites. Chromatin immunoprecipitation (ChIP) is a fundamental tool used in most wet lab TF binding site identification techniques. ChIP allows an *in vivo* snapshot of the proteins bound to DNA to be obtained. Traditionally ChIP was followed by microarray analysis, together called ChIP-chip [45]. More recently, the ChIP-seq approach follows chromatin immunoprecipitation with sequencing of the DNA [46]. Another wet lab technology directly measures *in vitro* binding affinities between DNA and proteins using protein binding microarrays [47].

Alternatively, co-regulated genes may be used as a source for approximate TF binding sites. Genes that are up- and down-regulated together are typically affected by the same TFs. Thus, the noncoding regions near these genes constitute a collection of DNA sequences which are likely to contain binding sites for a TF [48].

Identification of Binding Proteins and Binding Residues

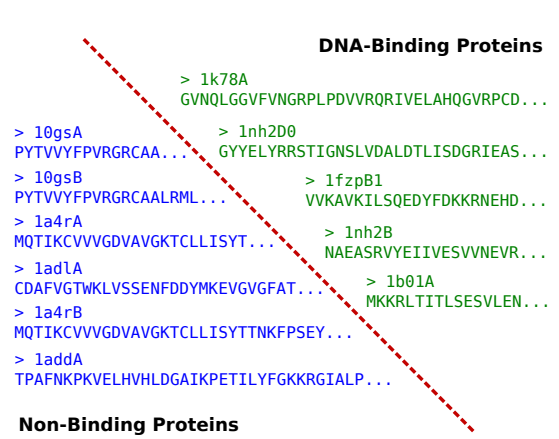
While studies of transcription factors tend to focus on DNA motifs and binding locations in the genome, attributes specific to DNA-binding proteins are also of interest. After isolating a new protein, biologists frequently want to discern its function. Data mining may be used to distinguish DNA-binding proteins from other types. Once it is established that a given protein interacts with DNA, a biologist may be interested in which of the protein's residues are involved with binding. Computational methods are of service here again to perform binding residue identification.

Both binding protein and binding residue identification may be addressed using techniques from supervised machine learning. The goal is to train a model which differentiates between the binding (positive) class and nonbinding (negative) class. The classes may represent either whole proteins or individual residues. The usual process for supervised learning is the following: establish a set of proteins as training examples, determine which features of the proteins will be given to the computational model as input, and then train the model to discriminate between binding and nonbinding classes. Predictive performance is evaluated on proteins which are excluded from the training process in order to judge the method's capabilities on future data.

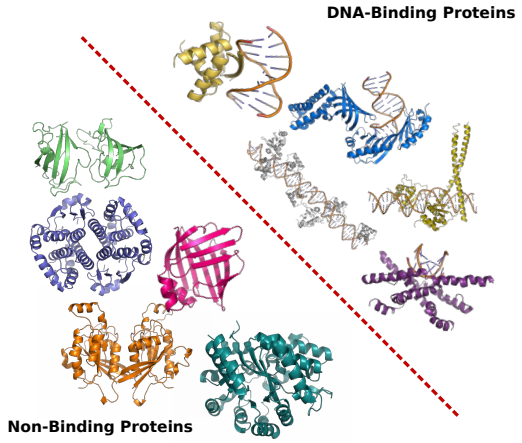
Whole Protein versus Residue-level Predictions

Most methods focus on predicting at either the whole protein level or residue level. Some methods accomplish both tasks simultaneously, but for the most part, addressing these two problems calls for different techniques.

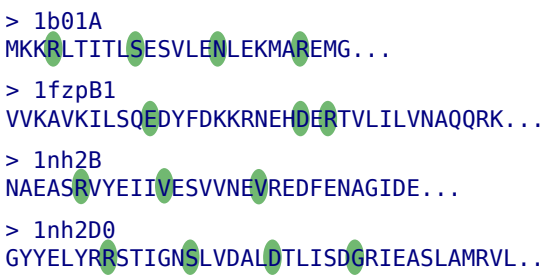
Figure 1: Identification Tasks Suitable for Machine Learning



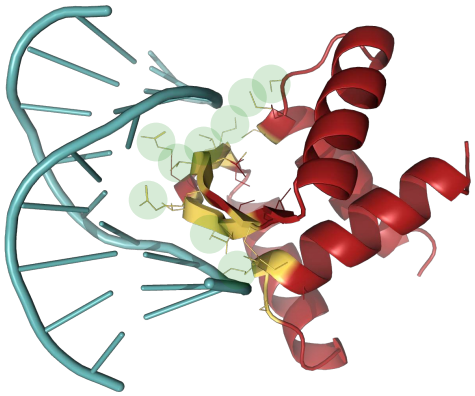
(a) DNA binding-protein identification using sequence



(b) DNA binding-protein identification using structure



(c) DNA binding-residue identification using sequence



(d) DNA binding-residue identification using structure

In the first case, the task is to identify DNA-binding proteins amongst proteins with other functions. This has increasing relevance as both sequencing and structural genomics projects have dramatically increased the number of proteins with unknown function. A variety of methods have been developed to accomplish this task [49, 50, 51, 52, 53, 54, 7, 55].

Prediction of DNA-binding residues assumes that that the protein under scrutiny binds DNA and predicts which residues are involved at the interface. Again, a wide array of approaches utilizing both sequence and structure features have been developed for residue-level prediction [56, 57, 50, 58, 59, 60, 61, 62, 63, 64, 56, 57, 65, 66, 67, 68].

While DNA-binding protein predictions are used primarily to elucidate the function of a new protein, there are several uses for DNA-binding residue predictions. They may be used to guide wet lab mutation experiments that affect binding affinity between protein and DNA. Rather than trying every residue in the protein, attention may be focused on mutating only residues which are predicted to play a role in binding. When structure is available but unbound, it may be possible to use predicted binding residues used to help identify the geometric binding site on protein as has been done for small ligands [69], though no work has yet explored this approach for DNA-protein interactions.

Prediction of DNA-binding Function

In the current literature, most methods approach binding protein and binding residue identification assuming either (1) the protein of interest has known structure, or (2) only the the protein's sequence is available. A third class of methods, known as homology modeling or threading, make predictions by assessing the compatibility of a target protein with DNA-binding structures.

Prediction from Structure

Knowledge of the protein's structure can be very helpful in determining its DNA-binding status. The structure may come from several sources. Traditionally, a protein's structure has been determined experimentally due to specific interest in how it fulfills its role in a biological system. Thus X-ray diffraction is used to determine the structure of protein-DNA complexes and this information is deposited in structural databases, primarily the PDB. These database entries provide examples for learning predictive models as the protein's function is typically well characterized. In some cases, two structures of the protein are available: the bound complex which has DNA present (*holo* protein conformation) and the unbound protein with no DNA present (*apo* conformation).

Though studies of single proteins have traditionally been the source for structure information, structural genomics projects are producing the structures of many new proteins for which no function information is available [70]. DNA-binding proteins produced by structural genomics efforts are usually determined in their apo (unbound) conformation. Annotating the protein as DNA-binding would greatly illuminate its biological role.

A very simple method of determining whether a protein is DNA-binding is to identify similar structures of known function using any of a number of structure alignment methods. However, the presence of a good structural match does not definitively establish the function of protein as similar structure/different function proteins exist. DNA-binding residues can be inferred as those structurally aligned to known binding residues.

Rather than rely directly on structural similarity to known DNA binding proteins to classify the function of new proteins, there are several lines of research which exploit structure features for identification of DNA-binding proteins. Examples of these include direct use of structural motifs and electrostatics to predict function, or the encoding of structural information into features amenable to machine learning methods [51, 53, 71, 61, 72, 54]. Such structural features and techniques are also heavily employed in methods for binding residue prediction [50, 58, 61, 57, 68].

Prediction from Sequence

Difficulty determining a protein's structure has motivated the development of binding predictors which utilize only sequence information. Such methods predict whether a protein binds DNA and which residues are involved in the process without relying on the geometry of the protein.

Aside from using standard sequence database searches such as BLAST and PSI-BLAST, few purely sequence-based methods are available for binding protein prediction [49]. This is likely due to the difficulty of encoding an information about an entire protein in the type of fixed-length feature vector required by most machine learners. Due to the simplicity of representing single-residue sequence information to a machine learner, more work has focused on methods for binding residue prediction from sequence [59, 61, 62, 73, 74, 75, 67, 56, 67, 66].

There have been some claims that these "template-free" models, which do not consider structural aspects of the protein, give inferior performance to their structure-based counterparts [55]. This may simply be due to sequence-based methods relinquishing potentially useful

structure information in order to make predictions when it is not available. When a good structure model is available, binding predictions will likely be improved by employing it. However, no true head-to-head benchmark between structure and sequence methods has yet been executed to illustrate the superiority of one over the other.

Homology Model-ling and Threading

A technique that has proved effective for DNA-binding prediction but does not constitute traditional machine learning is *homology modeling* and its relative *threading*. In both techniques, a *target* protein with unknown structure is modeled by identifying a *template* protein of known structure. The target sequence is then mapped onto the known template structure and refined (e.g. [76]). The key to this process is identifying a good template with known structure, which is usually done via a combination of sequence similarity and energy calculations on the sequence-structure mapping. In that sense, homology modeling may be likened to a nearest neighbor computation with sequence-structure compatibility acting as a specialized similarity measure. However, mapping a target sequence onto the template to produce a model structure is above and beyond the typical nearest neighbor method.

Threading methods can handle both whole protein and residue-wise binding prediction [7]. Merely finding a good template match which is a DNA-binding protein is an indication that the target may also bind DNA, but it is not sufficient evidence to declare the target is DNA-binding. Other nonbinding templates will likely score well, necessitating additional information such as interaction energy analysis between the target's model structure and DNA [55]. After structurally mapping a target onto its template, binding residues may be identified based on the corresponding binding residues in the template.

When no structure is available for a target protein, in some cases it may be possible to generate a full three-dimensional model using homology modeling or threading. In most cases, homology models are not entirely accurate, but for the purpose of determining whether the protein binds DNA, recent work has demonstrated that the use of homology models has promise [22, 77]. The second study also showed good prediction performance when training on both bound and unbound structures, making it viable for characterization of structural genomics targets.

Producing a homology model of the protein's structure may fail for several reasons, most commonly because no suitable template is available. Dependence on a good structural template is the primary disadvantage of template-based methods [55]. According to the literature, this happens with some frequency: over 40% of DNA-binding proteins have no suitable template for homology modeling ([56], section 1.2).

Machine Learning Features

Numerous features have been employed in prediction schemes for binding proteins and binding residues. These are divided into structure and sequence features. There is mild overlap in some cases: for instance, secondary structure is available from the protein's structure or it may be predicted from sequence.

Structural Features

- **Electrostatic Potentials** Molecular dynamics software is used to compute the charges for each atom which is usually averaged to assign an electrostatic score to each residue [50, 51, 52, 53]. Software is also available for this specific task [78].

- **Dipole and quadrupole moments** Charge moments measure how widely distributed electric charge is across the protein. Fairly simple methods can calculate the electric dipole and quadrupole from structure and according to the cited study, dipoles in combination with overall charge make a fairly discriminatory feature between binding and nonbinding proteins [79].
- **Structural Motifs** Certain structural motifs (patterns) are known for interaction with DNA. Identifying such a motif in a novel protein can lend support to its classification as a DNA-binding protein [51].
- **Structural Neighborhood** A simple representation of residue environment is to count the other amino acids inside a ball centered on the residue of interest [61].
- **Surface Curvature** In order to accommodate bound DNA, proteins may exhibit certain curvature, at least locally at the binding site [52].
- **Secondary Structure** Proteins assume local, repeated geometric patterns called secondary structure which may be calculated from its coordinates [80] or may be predicted from sequence [81, 82]. Several studies have shown that secondary structure is not a particularly informative feature for DNA-binding identification [58, 83].
- **Solvent Accessible Surface Area (SASA)** Binding residues are almost always well exposed to solvent to enable them to form contacts with DNA making SASA a useful predictive feature. Like secondary structure, SASA can be calculated from the protein structure or predicted from sequence. Some studies limit their focus to only surface residues from the outset [57].

Sequence Features

- **Amino Acid Sequence** The most common feature to any sequence-based predictor, the protein's amino acid sequence provides baseline information to the predictor. Raw sequence is usually encoded as a 20-dimensional binary vector. Positively charged residues such as arginine are more likely to interact with the negatively charged backbone of DNA according to both physical and statistical studies [83].
- **Residue Class/Type** The twenty amino acids may be grouped according to physical properties such as charge and hydrophobicity which is then used as an additional sequence feature such as the six classes in [66].
- **Sequence Profiles** The majority of machine learning approaches to bioinformatics problems now employ sequence profiles rather than raw sequence as profiles are generally acknowledged to provide better information. Profiles are usually generated using PSI-BLAST [84] and represent the probability of substituting a different amino acid for the one observed at a specific position. This is encoded as a twenty-dimensional vector at each sequence position, positive numbers indicating favorable substitutions and negative numbers unfavorable substitutions.
- **Global Composition of AAs** When attempting to identify DNA-binding proteins counts or frequencies of each type of amino acid are often used, typically as a 20-dimensional vector. Pairs of adjacent residues have also been used as a compositional feature [85].
- **Hydrophobicity** Measures of residue hydrophobicity, the degree to which the residue is repelled by water, are a commonly used feature. A typical example is the hydrophobicity scale in [86] which assigns a fixed numerical value to each of the twenty amino acid types.

- **Evolutionarily Conserved Residues** Residues that mitigate interactions between proteins and DNA are usually conserved through evolution. Thus identifying conserved residues can yield a powerful feature. This may be done using only sequence or combined with structural information to yield collections of conserved residues which are proximal in space [87, 54]. However, both approaches assume a sufficient number of close homologs to the target are available.

Additionally, sequence features are commonly augmented via *sliding windows* to capture the local sequence environment of a residue. Features of residues immediately to the left and right are concatenated onto those of a central residue before being presented to the machine learner. Window sizes between one (only the central residue) and eleven (five residues on either side) are commonly used. Many of the features described above are used in sliding windows in the approaches that describe them.

Machine Learning Tools

Most standard machine learning tools have been applied to DNA-binding protein and DNA-binding residue prediction. The short list includes support vector machines (SVMs) [56, 71], neural networks [53, 57], decision trees [67], Bayesian inference [63], logistic regression [77], and random forests [54, 66].

Comparisons between methods to determine an optimal approach are hindered by the different data used for evaluation and the variation of basic assumptions amongst studies. For example both [56] and [57] do binding residue prediction, but the former uses only sequence-based features while the latter uses structural information and evaluates performance only on surface residues. Direct comparison of their reported performance is not particularly informative.

Data Sets

If possible, new studies of DNA-protein interactions should employ a data set that has already been used in the literature. This facilitates direct comparison to previous efforts. Some common data sets in use are listed in Table 2 with relevant properties. In these cases we have checked the sequence independence of the data sets to verify whether they correspond to the levels reported in the literature. Of particular interest are the the data sets used for DBD-Hunter. These are the largest and most sequence-independent data sets in the literature making them a good place to start for new work. The data sets in Table 2 are available as supplemental information to this paper. Additionally, there are several new databases devoted solely to protein-DNA interactions which aggregate and augment information available from the PDB [88, 89].

For new data sets, authors should report the maximum level of sequence similarity amongst proteins in the set. The similarity level should be kept at or below 30-35% to be comparable to current methods. This can be accomplished using a sequence clustering program such as `blastclust` (available from NCBI) to group similar sequences and then select a single representative from each cluster. It is also important to eliminate proteins that are subsequence of other proteins in a dataset which can also be done with `blastclust`. For example, the following use of `blastclust` will cluster sequences at 35% identity and detect subsequences that are as little as 10% of the length of other sequences.

```
blastclust -i seq.fa -o seq.bc -S 35 -L 0.1
```

This is the mechanism that was used to analyze sequence redundancy of the datsets in Table 2. Another popular method of clustering is the PISCES web server for sequence culling [91].

Table 2: Commonly used data sets for DNA-binding protein and DNA-binding residue identification.

ID	Study	Notes
DB179	[7]	179 DNA-binding proteins, almost entirely nonredundant at 40% sequence identity
NB3797	[7]	3797 nonbinding proteins, significant redundancy at 35% sequence identity level (only 3482 independent clusters)
APO/HOLO104	[7]	104 unbound/bound pairs of DNA-binding proteins, maximum 30% identity, 10 apo/holo pairs have less than 90% sequence identity.
PD138	[77]	138 DNA-binding proteins, almost entirely nonredundant at 35% sequence identity, divided into seven structural classes
APO/HOLO54	[77]	54 unbound/bound pairs of DNA-binding proteins, maximum BLAST e-value for pairs of 0.1, 100% identity between APO/HOLO pairs. A few homologous sequences are present.
DISIS	[56]	78 DNA-binding proteins, close to nonredundant at 20% sequence identity
PDNA62	[58]	62 DNA-binding proteins, 78 chains, 57 nonredundant sequences at 30% identity.
NB110	[58]	110 nonbinding proteins, nonredundant at 30% sequence identity level, derived from the RS126 secondary structure data set [90] by removing entries related to DNA.
BIND54	[53]	Reported as 54 binding proteins, actually 58 chains, nonredundant at 30% sequence identity, original list of proteins was reported in [2].
NB250	[53]	250 nonbinding proteins, mostly nonredundant at 35% sequence identity
DBP374	[66]	374 DNA-binding proteins, significant redundancy at 25% sequence identity level
TS75	[66]	75 DNA-binding proteins, designed to be independent from DBP374 and PDNA62 but has some redundant entries in both at 35% sequence identity level

When dividing the data set for cross-validation, ensure divisions are done at the protein level even for binding residue prediction: residues from the same protein should not appear in both training and testing sets. When reporting performance, a variety of measures should be included, particularly an ROC analysis [92] and the Matthews correlation coefficient.

Current State of the Art

The current crop of DNA-binding protein predictors provide good results when sequences homologous to the target protein are available. Table 3 and Table 4 compile results for DNA-binding protein and DNA-binding residue prediction respectively. These tables should be interpreted carefully keeping the following points in mind. The methods are grouped by row based on the dataset which is used in evaluation, many of which appear in Table 2. Refer to it for details on the level of sequence redundancy of the dataset: moderate levels of sequence redundancy artificially make it easy to achieve good predictions rates. Within each dataset, evaluation strategies vary between studies as some use leave-one-out cross validation (also referred to as "jackknife" evaluation), while others employ 5-fold or 10-fold cross-validation. Where possible, we have included footnotes on the strategy as these variation in splits of training/testing sets make also affect the inferred performance. Finally, background information is used in various ways by the different methods. For example, many methods use a sequence database to generate PSI-BLAST profiles [56, 57, 75] while the threading methods [77, 7, 55] rely on large numbers of known structures for their template libraries. Changing background information can affect performance as is noted for DBD-HUNTER between [7] and [55].

DISIS provides a model of how machine learning methods can be applied to DNA-binding residue prediction [56]. For researchers implementing new sequence-based methods, it serves as a good example of how to describe the features derived for proteins, machine learning tools employed, and the evaluation framework used to gauge performance. The general methodology is equally applicable to set up DNA-binding protein prediction experiments. The only exception is that future studies should report a variety of performance measures, particularly a Matthews correlation coefficient (MCC) and receiver operator characteristic (ROC). The work of Langlois and Lu [49] is an excellent example of how to compare new work to older studies.

DBD-Threader provides a state-of-the-art threading approach which is likely amongst the best predictors when good templates are available for a target [55]. New structure-based methods should compare against its performance again with the addition of reporting performance in terms of ROC.

Most current DNA-binding classification methods rely upon the availability of similar proteins, either explicitly in the case of threading methods, or implicitly through the similarity measures used in machine learning methods and sequence comparison. When a homologs to the target protein are not available, the task of identifying DNA-binding proteins and residues is significantly more difficult. The work in [22] and [77] finds that homology modeling will usually fail when no good template is found. For sequence-based methods, this situation can be simulated by leaving out an entire structural classes while training. Testing on the left out structural fold led to only a modest drop in prediction performance for a sequence-based machine learner according to a small scale study in [95]. Thus, sequence-based methods may be the best approach when predictions for truly new proteins are required.

The number of experimentally verified DNA-binding structures is likely to continue increasing which will extend the capabilities of similarity-based methods. However, until homologs are available for all protein families, predicting DNA-binding attributes of new proteins is likely to remain a challenge.

Table 3: Summary of DNA-binding protein prediction results

Method	Type	Dataset	ACC	SEN	SPE	MCC	ROC
BLAST in [49]	Q		79.3	27.8	90.4	21.5	66.0
Langlois and Lu [49][a]	Q		89.1	48.1	98.0	66.2	90.3
Langlois and Lu, LOO [49][b]	Q	BIND54 and	-	-	-	-	91.1
Nimrod et al. [54]	T	NB250 [53]	-	87.0	94.0	-	-
Stawiski et al. [53]	T		92.0	81.0	94.4	74.0	-
Szilagyi and Skolnick [77]	T		-	89.0	85.0	73.0	93.0
BLAST in [49]	Q		81.4	80.8	81.8	70.4	90.5
Langlois and Lu [49]	Q	PDNA62[c],	89.9	84.6	93.6	84.9	97.1
Ahmad and Sarai [79]	T	NB110 [58]	83.9	80.8	87.0	68.0	-
Szilagyi and Skolnick [77]	T		-	92.0	85.0	79.0	95.0
BLAST in [49]	Q	Bhardwaj [71]	82.4	75.2	86.1	70.2	90.3
Langlois and Lu [49][a]	Q		94.7	88.4	97.9	88.8	96.7
Bhardwaj et al. CV5[71][d]	T		89.1	82.1	93.9	-	-
Bhardwaj et al. [71][d]	T	Bhardwaj [71]	90.3	67.4	94.9	-	-
Nimrod et al. [54][e d]	T	Filtered[f]	-	73.6	94.9	-	-
BLAST in [49]	Q	Langlois. [72]	72.7	42.7	83.2	70.4	90.5
Langlois and Lu [49][a]	Q		89.6	69.3	96.7	74.3	91.3
AdaC4.5 in Langlois et al. [72][d]	T		88.5	66.7	96.3	-	88.7
BLAST in [49]	Q	PD138,	71.8	79.7	61.8	45.1	80.1
Langlois and Lu [49]	Q	NB110 [77]	85.9	89.9	80.9	74.8	93.4
Szilagyi and Skolnick [77]	T		-	-	-	74.0	93.0
Nimrod et al. [54]	T		-	90.0	90.0	80.0	96.0
BLAST in [49]	Q	LEAC35 [60]	72.9	59.4	80.4	46.3	74.9
Langlois and Lu [49]	Q		84.0	68.8	92.4	69.5	92.3
BLAST in [49]	Q	LEAC25 [60]	69.4	42.6	82.4	28.6	67.8
Langlois and Lu [49]	Q		84.7	64.8	94.4	66.2	91.5
Szilagyi and Skolnick [77]	T	APO54 [77]	-	85.0	85.0	72.0	93.0
DBD-Hunter [7]	T		84.0	66.0	93.0	-	-
Szilagyi and Skolnick [77]	T	HOLO54 [77]	-	80.0	85.0	68.0	91.0
DBD-Hunter [7]	T		89.0	68.0	93.0	-	-
PSI-BLAST in [7]	Q		-	44.0	99.3	56.0	-
PSI-BLAST (Uniprot DB) in [55]	Q		-	43.0	99.3	55.3	-
DBD-Hunter [7][g]	T	DB179 and	-	58.0	99.5	69.0	-
DBD-Hunter [55][g]	T	NB3797 [7]	-	56.0	99.6	68.1	-
DBD-Threader [55]	Q		-	61.0	99.2	68.0	-
PROSPECTOR [93]	Q		-	53.0	99.1	60.9	-
Ahmad et al. [58]	Q	NRTF-915 [58]	64.5	68.6	63.4	-	-
Method	**Type**	**Dataset**	**ACC**	**SEN**	**SPE**	**MCC**	**ROC**

Columns are: **Method** with citation; **Type** of 'T' for structure-based and 'Q' for sequenced-based; **Dataset** which was used in evaluation; **ACC** for accuracy; **SEN** for sensitivity; **SPE** for specificity; **MCC** for Matthews Correlation Coefficient, scaled to -100 to 100; **ROC** for area under receiver operating curve, scaled to 0 to 100.

[a] 10-fold cross-validation

[b] Leave-one-out cross-validation

[c] PDNA62 is referred to as PD78 in [77].

[d] 5-fold cross-validation

[e] Reported in Supplementary Text S1 of [54].

[f] Filtered the dataset in [71] to be nonredundant at 20% sequence identity.

[g] The differing performance of DBD-Hunter between [7, 55] is due to an updated template library.

Table 4: Summary of DNA-binding residue prediction results

Method	Type	Dataset	ACC	SEN	SPE	MCC	ROC
DBD-Hunter [7]	T	DB179[a] [7]	90.0	72.0	93.0	-	-
DBD-Threader [55][b]	Q		87.5	60.0	92.0	52.0	-
DISPLAR [57][c]	T	DISPLAR [57]	76.4	60.1	-	-	-
Ahmad et al. [58]	T	PDNA62 [58]	40.3	81.8	79.1	-	-
DP-BIND Structure [61]	T		78.1	79.2	77.2	49.0	84.0
Ahmad and Sarai [59]	Q		66.4	68.2	66.0	-	-
DP-BIND Sequence [61]	Q		76.0	76.9	74.7	45.0	83.6
BindN [62]	Q		70.3	69.4	70.5	-	75.2
Ho et al. [73]	Q	PDNA62 [58]	80.2	80.1	80.2	-	-
BindN-RF [74]	Q		78.2	78.1	78.2	-	86.1
BindN+ [75]	Q		79.0	77.3	79.3	44.0	85.9
Carson et al. [67]	Q		78.5	79.7	77.2	57.0	85.7
DISIS [56]	Q	DISIS [56]	89.0	-	-	-	-
Carson et al. [67]	Q		86.4	84.6	87.0	72.5	93.1
DNABINDPROT [68]	T		78.6	9.3	90.5	-	-
DP-BIND [61, 94][d]	T	Ozbek [68]	74.0	63.7	75.0	-	-
DISPLAR [57][d]	T		80.1	45.1	86.2	-	-
DBD-HUNTER [7][d]	T		95.2	43.6	44.5	-	-
Random Forests, Wu et al. [66]	T	DBP374 [66]	91.4	76.6	73.2	70.0	91.3
BindN [62]	Q	TS75 [66]	-	-	-	-	78.2
SVM, Wu et al. [66]	Q	3.5Å cutoff[e]	-	-	-	-	84.3
Random Forests, Wu et al. [66]	Q		-	-	-	-	85.5
Random Forests, Wu et al. [66]	Q	TS75 [66]	80.5	67.2	81.8	34.1	-
DP-BIND [61]	Q	4.5Å cutoff[e]	78.0	67.8	79.0	31.6	-
Random Forests, Wu et al. [66]	Q	TS75 [66]	78.2	51.4	84.6	34.1	-
DISIS [56]	Q	6.0Å cutoff[e]	81.6	7.7	99.2	19.0	-
Method	**Type**	**Dataset**	**ACC**	**SEN**	**SPE**	**MCC**	**ROC**

Columns are: **Method** with citation; **Type** of 'T' for structure-based and 'Q' for sequenced-based; **Dataset** which was used in evaluation; **ACC** for accuracy; **SEN** for sensitivity; **SPE** for specificity; **MCC** for Matthews Correlation Coefficient, scaled to -100 to 100; **ROC** for area under receiver operating curve, scaled to 0 to 100.

[a] Only did binding residue prediction on 103 proteins predicted as DNA-binding proteins by DBD-Hunter and DBD-Threader respectively. Average per-protein statistics reported.

[b] Estimated from Figure 3 of [55].

[c] Evaluation was done only on surface residues only.

[d] Results reported in Supplementary Table S2 of [68].

[e] Refers to the distance cutoff determining DNA-binding residues.

Future Directions

Machine learning has already impacted the study of protein-DNA interactions, particularly the identification of DNA-binding proteins. These innovations are set to continue down a number of avenues. The capability of machine learning to identify binding residues in a protein may be used to guide physical simulations of protein-DNA interactions. This capability has been utilized in some studies of protein interactions with small molecules to guide ligand docking [69] and improve models of the binding site [96]. The same methodology may also be employed for DNA-binding proteins.

Another avenue of pursuit is applying machine learning to identify the genomic binding sites for transcription factors. There has already been some work done to develop models for various structural classes of TFs [39]. Features of both the genome binding site (DNA sequence) and the protein are used to train classifier for each TF family. An analogous problem in cheminformatics is to classify small molecules according to whether they activate a particular protein. Recent work which employs multitask learning [97] to characterize active compounds for different proteins [98] may carry over directly to the case of TF binding site identification on multiple TFs.

Finally, a true head-to-head comparison of the various methods for DNA-binding protein identification and DNA-binding residue prediction would guide further development in this area. Dividing a benchmark into sequence-based and structure-based predictions would elucidate how much inference capability is gained when a protein's structure is available.

Acknowledgments

This work was supported in part by NSF (IIS-0905220, OCI-1048018, IOS-0820730), NIH (RLM008713A), and the Digital Technology Center at the University of Minnesota. Access to research and computing facilities was provided by the Digital Technology Center and the Minnesota Supercomputing Institute.

References

[1] V. Charoensawan, D. Wilson, and S. A. Teichmann, "Genomic repertoires of DNA-binding transcription factors across the tree of life.," *Nucleic Acids Res*, Jul 2010.

[2] N. M. Luscombe, S. E. Austin, H. M. Berman, and J. M. Thornton, "An overview of the structures of protein-DNA complexes.," *Genome Biol*, vol. 1, no. 1, p. REVIEWS001, 2000.

[3] M. C. Walter, T. Rattei, R. Arnold, U. Güldener, M. Mnsterkötter, K. Nenova, G. Kastenmller, P. Tischler, A. Wülling, A. Volz, N. Pongratz, R. Jost, H.-W. Mewes, and D. Frishman, "Pedant covers all complete refseq genomes.," *Nucleic Acids Res*, vol. 37, pp. D408–D411, Jan 2009.

[4] V. A. Kuznetsov, O. Singh, and P. Jenjaroenpun, "Statistics of protein-DNA binding and the total number of binding sites for a transcription factor in the mammalian genome.," *BMC Genomics*, vol. 11 Suppl 1, p. S12, 2010.

[5] D. Wilson, V. Charoensawan, S. K. Kummerfeld, and S. A. Teichmann, "Dbd–taxonomically broad transcription factor predictions: new content and functionality.," *Nucleic Acids Res*, vol. 36, pp. D88–D92, Jan 2008.

[6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucl. Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000.

[7] M. Gao and J. Skolnick, "Dbd-hunter: a knowledge-based method for the prediction of DNA-protein interactions.," *Nucleic Acids Res*, vol. 36, pp. 3978–3992, Jul 2008.

[8] T. A. Brown, *Genomes*. Oxford: Wiley-Liss, 2nd ed., 2002.

[9] C. O. Pabo and R. T. Sauer, "Transcription factors: Structural families and principles of DNA recognition," *Annual Review of Biochemistry*, vol. 61, no. 1, pp. 1053–1095, 1992. PMID: 1497306.

[10] Y. Mandel-Gutfreund, O. Schueler, and H. Margalit, "Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles.," *J Mol Biol*, vol. 253, pp. 370–382, Oct 1995.

[11] N. M. Luscombe, R. A. Laskowski, and J. M. Thornton, "Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level.," *Nucleic Acids Res*, vol. 29, pp. 2860–2874, Jul 2001.

[12] J. R. Desjarlais and J. M. Berg, "Length-encoded multiplex binding site determination: application to zinc finger proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 91, no. 23, pp. 11099–11103, 1994.

[13] Y. Mandel-Gutfreund and H. Margalit, "Quantitative parameters for amino acid-base interaction: Implications for prediction of protein-DNA binding sites," *Nucleic Acids Research*, vol. 26, no. 10, pp. 2306–2312, 1998.

[14] G. D. Stormo, "DNA binding sites: representation and discovery.," *Bioinformatics*, vol. 16, pp. 16–23, Jan 2000.

[15] A. Klug and J. W. Schwabe, "Protein motifs 5. zinc fingers.," *FASEB J*, vol. 9, pp. 597–604, May 1995.

[16] A. Sarai and H. Kono, "Protein-DNA recognition patterns and predictions.," *Annu Rev Biophys Biomol Struct*, vol. 34, pp. 379–398, 2005.

[17] C. O. Pabo and L. Nekludova, "Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition?," *J Mol Biol*, vol. 301, pp. 597–624, Aug 2000.

[18] B. Jayaram, K. McConnell, S. B. Dixit, A. Das, and D. L. Beveridge, "Free-energy component analysis of 40 protein-DNA complexes: a consensus view on the thermodynamics of binding at the molecular level.," *J Comput Chem*, vol. 23, pp. 1–14, Jan 2002.

[19] R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann, "Origins of specificity in protein-dna recognition.," *Annu Rev Biochem*, vol. 79, pp. 233–269, 2010.

[20] J. E. Donald, W. W. Chen, and E. I. Shakhnovich, "Energetics of protein-DNA interactions.," *Nucleic Acids Res*, vol. 35, no. 4, pp. 1039–1047, 2007.

[21] K. A. Aeling, M. L. Opel, N. R. Steffen, V. Tretyachenko-Ladokhina, G. W. Hatfield, R. H. Lathrop, and D. F. Senear, "Indirect recognition in sequence-specific DNA binding by escherichia coli integration host factor: the role of DNA deformation energy.," *J Biol Chem*, vol. 281, pp. 39236–39248, Dec 2006.

[22] A. V. Morozov, J. J. Havranek, D. Baker, and E. D. Siggia, "Protein-DNA binding specificity predictions with structural models.," *Nucleic Acids Res*, vol. 33, no. 18, pp. 5781–5798, 2005.

[23] G. D. Stormo and Y. Zhao, "Determining the specificity of protein-dna interactions.," *Nat Rev Genet*, vol. 11, pp. 751–760, Nov 2010.

[24] K. A. Aeling, N. R. Steffen, M. Johnson, G. W. Hatfield, R. H. Lathrop, and D. F. Senear, "DNA deformation energy as an indirect recognition mechanism in protein-DNA interactions.," *IEEE/ACM Trans Comput Biol Bioinform*, vol. 4, no. 1, pp. 117–125, 2007.

[25] L. Narlikar and I. Ovcharenko, "Identifying regulatory elements in eukaryotic genomes.," *Brief Funct Genomic Proteomic*, vol. 8, pp. 215–230, Jul 2009.

[26] S. Hannenhalli, "Eukaryotic transcription factor binding sites–modeling and integrative search methods.," *Bioinformatics*, vol. 24, pp. 1325–1331, Jun 2008.

[27] M. Vingron, A. Brazma, R. Coulson, J. van Helden, T. Manke, K. Palin, O. Sand, and E. Ukkonen, "Integrating sequence, evolution and functional genomics in regulatory genomics.," *Genome Biol*, vol. 10, no. 1, p. 202, 2009.

[28] M. L. Bulyk, "Computational prediction of transcription-factor binding site locations.," *Genome Biol*, vol. 5, no. 1, p. 201, 2003.

[29] L. Elnitski, V. X. Jin, P. J. Farnham, and S. J. M. Jones, "Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques.," *Genome Res*, vol. 16, pp. 1455–1464, Dec 2006.

[30] M. K. Das and H.-K. Dai, "A survey of DNA motif finding algorithms.," *BMC Bioinformatics*, vol. 8 Suppl 7, p. S21, 2007.

[31] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. D. Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu, "Assessing computational tools for the discovery of transcription factor binding sites.," *Nat Biotechnol*, vol. 23, pp. 137–144, Jan 2005.

[32] M. L. Bulyk, P. L. F. Johnson, and G. M. Church, "Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors.," *Nucleic Acids Res*, vol. 30, pp. 1255–1261, Mar 2002.

[33] P. V. Benos, M. L. Bulyk, and G. D. Stormo, "Additivity in protein-DNA interactions: how good an approximation is it?," *Nucleic Acids Res*, vol. 30, pp. 4442–4451, Oct 2002.

[34] M. F. Berger and M. L. Bulyk, "Protein binding microarrays (pbms) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins.," *Methods Mol Biol*, vol. 338, pp. 245–260, 2006.

[35] T. Kaplan, N. Friedman, and H. Margalit, "Ab initio prediction of transcription factor targets using structural knowledge," *PLoS Comput Biol*, vol. 1, p. e1, 06 2005.

[36] P. V. Loo and P. Marynen, "Computational methods for the detection of cis-regulatory modules.," *Brief Bioinform*, vol. 10, pp. 509–524, Sep 2009.

[37] Y. Barash, G. Elidan, N. Friedman, and T. Kaplan, "Modeling dependencies in protein-DNA binding sites," in *Proceedings of the seventh annual international conference on Research in computational molecular biology*, RECOMB '03, (New York, NY, USA), pp. 28–37, ACM, 2003.

[38] Q. Zhou and J. S. Liu, "Modeling within-motif dependence for transcription factor binding site predictions.," *Bioinformatics*, vol. 20, pp. 909–916, Apr 2004.

[39] L. Narlikar, R. Gordan, U. Ohler, and A. J. Hartemink, "Informative priors based on transcription factor structural class improve de novo motif discovery.," *Bioinformatics*, vol. 22, pp. e384–e392, Jul 2006.

[40] M. Hu, J. Yu, J. M. G. Taylor, A. M. Chinnaiyan, and Z. S. Qin, "On the detection and refinement of transcription factor binding sites using chip-seq data.," *Nucleic Acids Res*, vol. 38, pp. 2154–2167, Apr 2010.

[41] Z. Zhu, Y. Pilpel, and G. M. Church, "Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (tfcc) algorithm.," *J Mol Biol*, vol. 318, pp. 71–81, Apr 2002.

[42] H. J. Bussemaker, H. Li, and E. D. Siggia, "Regulatory element detection using correlation with expression," *Nat Genet*, vol. 27, pp. 167–174, Feb. 2001.

[43] M. C. Frith, Y. Fu, L. Yu, J.-F. Chen, U. Hansen, and Z. Weng, "Detection of functional DNA motifs via statistical over-representation.," *Nucleic Acids Res*, vol. 32, no. 4, pp. 1372–1381, 2004.

[44] H. G. Roider, T. Manke, S. O'Keeffe, M. Vingron, and S. A. Haas, "Pastaa: identifying transcription factors associated with sets of co-regulated genes.," *Bioinformatics*, vol. 25, pp. 435–442, Feb 2009.

[45] T. H. Kim and B. Ren, "Genome-wide analysis of protein-DNA interactions.," *Annu Rev Genomics Hum Genet*, vol. 7, pp. 81–102, 2006.

[46] P. J. Park, "Chip-seq: advantages and challenges of a maturing technology.," *Nat Rev Genet*, vol. 10, pp. 669–680, Oct 2009.

[47] M. L. Bulyk, "Analysis of sequence specificities of DNA-binding proteins with protein binding microarrays.," *Methods Enzymol*, vol. 410, pp. 279–299, 2006.

[48] J. van Helden, A. F. Rios, and J. Collado-Vides, "Discovering regulatory elements in non-coding sequences by analysis of spaced dyads.," *Nucleic Acids Res*, vol. 28, pp. 1808–1818, Apr 2000.

[49] R. E. Langlois and H. Lu, "Boosting the prediction and understanding of DNA-binding domains from sequence.," *Nucleic Acids Res*, vol. 38, pp. 3149–3158, Jun 2010.

[50] S. Jones, H. P. Shanahan, H. M. Berman, and J. M. Thornton, "Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins," *Nucleic Acids Research*, vol. 31, no. 24, pp. 7189–7198, 2003.

[51] H. P. Shanahan, M. A. Garcia, S. Jones, and J. M. Thornton, "Identifying DNA-binding proteins using structural motifs and the electrostatic potential.," *Nucleic Acids Res*, vol. 32, no. 16, pp. 4732–4741, 2004.

[52] Y. Tsuchiya, K. Kinoshita, and H. Nakamura, "Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces.," *Proteins*, vol. 55, pp. 885–894, Jun 2004.

[53] E. W. Stawiski, L. M. Gregoret, and Y. Mandel-Gutfreund, "Annotating nucleic acid-binding function based on protein structure.," *J Mol Biol*, vol. 326, pp. 1065–1079, Feb 2003.

[54] G. Nimrod, A. Szilagyi, C. Leslie, and N. Ben-Tal, "Identification of DNA-binding proteins using structural, electrostatic and evolutionary features.," *J Mol Biol*, vol. 387, pp. 1040–1053, Apr 2009.

[55] M. Gao and J. Skolnick, "A threading-based method for the prediction of DNA-binding proteins with application to the human genome.," *PLoS Comput Biol*, vol. 5, p. e1000567, Nov 2009.

[56] Y. Ofran, V. Mysore, and B. Rost, "Prediction of DNA-binding residues from sequence.," *Bioinformatics*, vol. 23, pp. i347–i353, Jul 2007.

[57] H. Tjong and H.-X. Zhou, "Displar: an accurate method for predicting DNA-binding sites on protein surfaces.," *Nucleic Acids Res*, vol. 35, no. 5, pp. 1465–1477, 2007.

[58] 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, pp. 477–486, Mar 2004.

[59] S. Ahmad and A. Sarai, "Pssm-based prediction of DNA binding sites in proteins.," *BMC Bioinformatics*, vol. 6, p. 33, 2005.

[60] N. Bhardwaj, R. Langlois, G. Zhao, and H. Lu, "Structure based prediction of binding residues on DNA-binding proteins.," *Conf Proc IEEE Eng Med Biol Soc*, vol. 3, pp. 2611–2614, 2005.

[61] I. B. Kuznetsov, Z. Gou, R. Li, and S. Hwang, "Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins," *Proteins*, vol. 64, pp. 19–27, 2006.

[62] 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 Res*, vol. 34, pp. W243–W248, Jul 2006.

[63] 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, 2006.

[64] N. Bhardwaj and H. Lu, "Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions.," *FEBS Lett*, vol. 581, pp. 1058–1066, Mar 2007.

[65] M. Andrabi, K. Mizuguchi, A. Sarai, and S. Ahmad, "Prediction of mono- and di-nucleotide-specific DNA-binding sites in proteins using neural networks.," *BMC Struct Biol*, vol. 9, p. 30, 2009.

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

[67] M. B. Carson, R. Langlois, and H. Lu, "Naps: a residue-level nucleic acid-binding prediction server.," *Nucleic Acids Res*, vol. 38 Suppl, pp. W431–W435, Jul 2010.

[68] 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 Res*, vol. 38 Suppl, pp. W417–W423, Jul 2010.

[69] D. Ghersi and R. Sanchez, "Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites.," *Proteins*, vol. 74, pp. 417–424, Feb 2009.

[70] I. Friedberg, "Automated protein function prediction–the genomic challenge.," *Brief Bioinform*, vol. 7, pp. 225–242, Sep 2006.

[71] N. Bhardwaj, R. E. Langlois, G. Zhao, and H. Lu, "Kernel-based machine learning protocol for predicting DNA-binding proteins.," *Nucleic Acids Res*, vol. 33, no. 20, pp. 6486–6493, 2005.

[72] R. E. Langlois, M. B. Carson, N. Bhardwaj, and H. Lu, "Learning to translate sequence and structure to function: identifying DNA binding and membrane binding proteins.," *Ann Biomed Eng*, vol. 35, pp. 1043–1052, Jun 2007.

[73] 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, 2007.

[74] 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 Suppl 1, p. S1, 2009.

[75] L. Wang, C. Huang, M. Q. Yang, and J. Y. Yang, "BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features.," *BMC Syst Biol*, vol. 4 Suppl 1, p. S3, 2010.

[76] A. Sali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints.," *J Mol Biol*, vol. 234, pp. 779–815, Dec 1993.

[77] A. Szilagyi and J. Skolnick, "Efficient prediction of nucleic acid binding function from low-resolution protein structures.," *J Mol Biol*, vol. 358, pp. 922–933, May 2006.

[78] S. Shazman, G. Celniker, O. Haber, F. Glaser, and Y. Mandel-Gutfreund, "Patch finder plus (PFplus): a web server for extracting and displaying positive electrostatic patches on protein surfaces.," *Nucleic Acids Res*, vol. 35, pp. W526–W530, Jul 2007.

[79] S. Ahmad and A. Sarai, "Moment-based prediction of DNA-binding proteins.," *J Mol Biol*, vol. 341, pp. 65–71, Jul 2004.

[80] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, pp. 2577–637, 1983.

[81] D. T. Jones, "Improving the accuracy of transmembrane protein topology prediction using evolutionary information.," *Bioinformatics*, vol. 23, pp. 538–544, Mar 2007.

[82] G. Karypis, "YASSPP: Better kernels and coding schemes lead to improvements in svm-based secondary structure prediction," *Proteins: Structure, Function and Bioinformatics*, vol. 64, no. 3, pp. 575–586, 2006.

[83] C. Kauffman and G. Karypis, "An analysis of information content present in protein-DNA interactions," in *Pacific Symposium on Biocomputing*, (Fairmont Orchid, Big Island of Hawaii), pp. 477–488, 2008.

[84] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman, "Gapped blast and psi-blast: A new generation of protein database search programs," *Nucl. Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

[85] M. Kumar, M. M. Gromiha, and G. P. S. Raghava, "Identification of DNA-binding proteins using support vector machines and evolutionary profiles.," *BMC Bioinformatics*, vol. 8, p. 463, 2007.

[86] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein.," *J Mol Biol*, vol. 157, pp. 105–132, May 1982.

[87] S. Ahmad, O. Keskin, A. Sarai, and R. Nussinov, "Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins.," *Nucleic Acids Res*, vol. 36, pp. 5922–5932, Oct 2008.

[88] T. Norambuena and F. Melo, "The protein-DNA interface database.," *BMC Bioinformatics*, vol. 11, p. 262, 2010.

[89] B. Contreras-Moreira, "3d-footprint: a database for the structural analysis of protein-DNA complexes.," *Nucleic Acids Res*, vol. 38, pp. D91–D97, Jan 2010.

[90] B. Rost and C. Sander, "Prediction of protein secondary structure at better than 70

[91] G. Wang and J. Dunbrack, Roland L., "Pisces: recent improvements to a pdb sequence culling server," *Nucl. Acids Res.*, vol. 33, pp. W94–98, 2005.

[92] T. Fawcett, "An introduction to roc analysis," *Pattern Recogn. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[93] J. Skolnick, D. Kihara, and Y. Zhang, "Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm.," *Proteins*, vol. 56, pp. 502–518, Aug 2004.

[94] S. Hwang, Z. Gou, and I. B. Kuznetsov, "Dp-bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins," *Bioinformatics*, vol. 23, no. 5, pp. 634–636, 2007.

[95] Z. Gou and I. B. Kuznetsov, "On the accuracy of sequence-based computational inference of protein residues involved in interactions with DNA.," *Trends Appl Sci Res*, vol. 3, pp. 285–291, Dec 2008.

[96] C. Kauffman, H. Rangwala, and G. Karypis, "Improving homology models for protein-ligand binding sites," in *LSS Comput Syst Bioinformatics Conference*, (San Francisco, CA), 2008. Available at http://www.cs.umn.edu/ karypis, last access 10/12/2009.

[97] R. Caruana, "Multitask learning: A knowledge-based source of inductive bias," in *ICML*, pp. 41–48, 1993.

[98] X. Ning, H. Rangwala, and G. Karypis, "Multi-assay-based structure-activity relationship models: Improving structure-activity relationship models by incorporating activity information from related targets," *Journal of Chemical Information and Modeling*, vol. 49, no. 11, pp. 2444–2456, 2009. PMID: 19842624.