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 TFs 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.

<table>
<thead>
<tr>
<th>GO Term</th>
<th>Count</th>
<th>%all</th>
<th>%func</th>
</tr>
</thead>
<tbody>
<tr>
<td>All gene products</td>
<td>18269</td>
<td>100.0</td>
<td>115.6</td>
</tr>
<tr>
<td>Molecular Function Given</td>
<td>15801</td>
<td>86.5</td>
<td>100.0</td>
</tr>
<tr>
<td>DNA-binding</td>
<td>2375</td>
<td>13.0</td>
<td>15.0</td>
</tr>
<tr>
<td>Transcription factor activity</td>
<td>969</td>
<td>5.3</td>
<td>6.1</td>
</tr>
</tbody>
</table>

transcription factors

For researchers interested in DNA-binding protein structures, the protein data bank (PDB) 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. 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. Subsequent studies on protein-DNA interactions characterized binding based on the frequency of protein residue contacts with nucleic acids in crystallized complexes. The propensity for each type of contact to form was

2. [http://amigo.geneontology.org/cgi-bin/amigo/browse.cgi](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] fits 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 modeling 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 conclusion 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]. Supplemented 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 [10].

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 [11, 12].

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 [13, 14].

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 \textit{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 \textit{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 [18].

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.
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 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 [50, 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 ([50], 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. 

• **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.

• **Structural Neighborhood** A simple representation of residue environment is to count the other amino acids inside a ball centered on the residue of interest.

• **Surface Curvature** In order to accommodate bound DNA, proteins may exhibit certain curvature, at least locally at the binding site.

• **Secondary Structure** Proteins assume local, repeated geometric patterns called secondary structure which may be calculated from its coordinates or may be predicted from sequence. Several studies have shown that secondary structure is not a particularly informative feature for DNA-binding identification.

• **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.

**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.

• **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.

• **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 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.

• **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 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.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Dataset</th>
<th>ACC</th>
<th>SEN</th>
<th>SPE</th>
<th>MCC</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST in [49]</td>
<td>Q</td>
<td>BIND54 and LOO</td>
<td>79.3</td>
<td>27.8</td>
<td>90.4</td>
<td>21.5</td>
<td>66.0</td>
</tr>
<tr>
<td>Langlois and Lu [49]</td>
<td>Q</td>
<td>NB250</td>
<td>89.1</td>
<td>48.1</td>
<td>98.0</td>
<td>66.2</td>
<td>90.3</td>
</tr>
<tr>
<td>Nimrod et al. [54]</td>
<td>T</td>
<td>-</td>
<td>-</td>
<td>87.0</td>
<td>94.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Stawiski et al. [53]</td>
<td>T</td>
<td>-</td>
<td>92.0</td>
<td>81.0</td>
<td>94.4</td>
<td>74.0</td>
<td>-</td>
</tr>
<tr>
<td>Szilagyi and Skolnick [77]</td>
<td>T</td>
<td>-</td>
<td>-</td>
<td>92.0</td>
<td>85.0</td>
<td>73.0</td>
<td>93.0</td>
</tr>
<tr>
<td>BLAST in [49]</td>
<td>Q</td>
<td>Bhardwaj</td>
<td>81.4</td>
<td>80.8</td>
<td>81.8</td>
<td>70.4</td>
<td>95.0</td>
</tr>
<tr>
<td>Langlois and Lu [49]</td>
<td>Q</td>
<td>PDNA62</td>
<td>89.9</td>
<td>84.6</td>
<td>93.6</td>
<td>84.9</td>
<td>97.1</td>
</tr>
<tr>
<td>Ahmad and Sarai [79]</td>
<td>T</td>
<td>NB110</td>
<td>83.9</td>
<td>80.8</td>
<td>87.0</td>
<td>68.0</td>
<td>-</td>
</tr>
<tr>
<td>Szilagyi and Skolnick [77]</td>
<td>T</td>
<td>-</td>
<td>-</td>
<td>90.0</td>
<td>85.0</td>
<td>79.0</td>
<td>95.0</td>
</tr>
<tr>
<td>BLAST in [49]</td>
<td>Q</td>
<td>Bhardwaj</td>
<td>82.4</td>
<td>75.2</td>
<td>86.1</td>
<td>70.2</td>
<td>90.3</td>
</tr>
<tr>
<td>Langlois and Lu [49]</td>
<td>Q</td>
<td>PD138</td>
<td>94.7</td>
<td>88.4</td>
<td>97.9</td>
<td>88.8</td>
<td>96.7</td>
</tr>
<tr>
<td>Bhardwaj et al. CV5 [71]</td>
<td>T</td>
<td>-</td>
<td>89.1</td>
<td>82.1</td>
<td>93.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bhardwaj et al. [71]</td>
<td>T</td>
<td>Filtered</td>
<td>-</td>
<td>73.6</td>
<td>94.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BLAST in [49]</td>
<td>Q</td>
<td>Langlois</td>
<td>72.7</td>
<td>42.7</td>
<td>83.2</td>
<td>70.4</td>
<td>90.5</td>
</tr>
<tr>
<td>Langlois and Lu [49]</td>
<td>Q</td>
<td>86.9</td>
<td>69.3</td>
<td>96.7</td>
<td>74.3</td>
<td>91.3</td>
<td></td>
</tr>
<tr>
<td>AdaC4.5 in Langlois et al. [72]</td>
<td>T</td>
<td>88.5</td>
<td>66.7</td>
<td>96.3</td>
<td>-</td>
<td>88.7</td>
<td></td>
</tr>
<tr>
<td>BLAST in [49]</td>
<td>Q</td>
<td>PD138</td>
<td>71.8</td>
<td>79.7</td>
<td>61.8</td>
<td>45.1</td>
<td>80.1</td>
</tr>
<tr>
<td>Langlois and Lu [49]</td>
<td>Q</td>
<td>NB110</td>
<td>85.9</td>
<td>89.9</td>
<td>80.9</td>
<td>74.8</td>
<td>93.4</td>
</tr>
<tr>
<td>Szilagyi and Skolnick [77]</td>
<td>T</td>
<td>-</td>
<td>-</td>
<td>90.0</td>
<td>90.0</td>
<td>80.0</td>
<td>96.0</td>
</tr>
<tr>
<td>BLAST in [49]</td>
<td>Q</td>
<td>LEAC35</td>
<td>72.9</td>
<td>59.4</td>
<td>80.4</td>
<td>46.3</td>
<td>74.9</td>
</tr>
<tr>
<td>Langlois and Lu [49]</td>
<td>Q</td>
<td>84.0</td>
<td>68.8</td>
<td>92.4</td>
<td>69.5</td>
<td>92.3</td>
<td></td>
</tr>
<tr>
<td>BLAST in [49]</td>
<td>Q</td>
<td>LEAC25</td>
<td>69.4</td>
<td>42.6</td>
<td>82.4</td>
<td>28.6</td>
<td>67.8</td>
</tr>
<tr>
<td>Langlois and Lu [49]</td>
<td>Q</td>
<td>84.7</td>
<td>64.8</td>
<td>94.4</td>
<td>66.2</td>
<td>91.5</td>
<td></td>
</tr>
<tr>
<td>Szilagyi and Skolnick [77]</td>
<td>T</td>
<td>AP054</td>
<td>-</td>
<td>85.0</td>
<td>85.0</td>
<td>72.0</td>
<td>93.0</td>
</tr>
<tr>
<td>DBD-Hunter [7]</td>
<td>T</td>
<td>84.0</td>
<td>66.0</td>
<td>93.0</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Szilagyi and Skolnick [77]</td>
<td>T</td>
<td>HOLO54</td>
<td>-</td>
<td>80.0</td>
<td>85.0</td>
<td>68.0</td>
<td>91.0</td>
</tr>
<tr>
<td>DBD-Hunter [7]</td>
<td>T</td>
<td>89.0</td>
<td>68.0</td>
<td>93.0</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PSI-BLAST in [7]</td>
<td>Q</td>
<td>-</td>
<td>44.0</td>
<td>99.3</td>
<td>56.0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PSI-BLAST (Uniprot DB) in [55]</td>
<td>Q</td>
<td>-</td>
<td>43.0</td>
<td>99.3</td>
<td>55.3</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>DBD-Hunter [7]</td>
<td>T</td>
<td>DB179 and LEAC35</td>
<td>-</td>
<td>58.0</td>
<td>99.5</td>
<td>69.0</td>
<td>-</td>
</tr>
<tr>
<td>DBD-Hunter [55]</td>
<td>T</td>
<td>NB3797</td>
<td>-</td>
<td>56.0</td>
<td>99.6</td>
<td>68.1</td>
<td>-</td>
</tr>
<tr>
<td>DBD-Threader [53]</td>
<td>Q</td>
<td>-</td>
<td>61.0</td>
<td>99.2</td>
<td>68.0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PROSPECTOR [93]</td>
<td>Q</td>
<td>-</td>
<td>53.0</td>
<td>99.1</td>
<td>60.9</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Ahmad et al. [58]</td>
<td>Q</td>
<td>NRTF-915</td>
<td>64.5</td>
<td>68.6</td>
<td>63.4</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

*a10-fold cross-validation
bLeave-one-out cross-validation
CPDNA62 is referred to as PD78 in [77].
d5-fold cross-validation
Reported in Supplementary Text S1 of [54].
fFiltered the dataset in [7] to be nonredundant at 20% sequence identity.
gThe differing performance of DBD-Hunter between [7, 55] is due to an updated template library.
<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Dataset</th>
<th>ACC</th>
<th>SEN</th>
<th>SPE</th>
<th>MCC</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBD-Hunter</td>
<td>T</td>
<td>DB179</td>
<td>90.0</td>
<td>72.0</td>
<td>93.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DBD-Threader</td>
<td>Q</td>
<td></td>
<td>87.5</td>
<td>60.0</td>
<td>92.0</td>
<td>52.0</td>
<td>-</td>
</tr>
<tr>
<td>DISPLAR</td>
<td>T</td>
<td>DISPLAR</td>
<td>76.4</td>
<td>60.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ahmad et al.</td>
<td>T</td>
<td>PDNA62</td>
<td>40.3</td>
<td>81.8</td>
<td>79.1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DP-BIND Structure</td>
<td>T</td>
<td></td>
<td>78.1</td>
<td>79.2</td>
<td>77.2</td>
<td>49.0</td>
<td>84.0</td>
</tr>
<tr>
<td>Ahmad and Sarai</td>
<td>Q</td>
<td></td>
<td>66.4</td>
<td>68.2</td>
<td>66.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DP-BIND Sequence</td>
<td>Q</td>
<td></td>
<td>76.0</td>
<td>76.9</td>
<td>74.7</td>
<td>45.0</td>
<td>83.6</td>
</tr>
<tr>
<td>BindN</td>
<td>Q</td>
<td></td>
<td>70.3</td>
<td>69.4</td>
<td>70.5</td>
<td>75.2</td>
<td>-</td>
</tr>
<tr>
<td>Ho et al.</td>
<td>Q</td>
<td>PDNA62</td>
<td>80.2</td>
<td>80.1</td>
<td>80.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BindN-RF</td>
<td>Q</td>
<td></td>
<td>78.2</td>
<td>78.1</td>
<td>78.2</td>
<td>86.1</td>
<td>-</td>
</tr>
<tr>
<td>BindN+</td>
<td>Q</td>
<td></td>
<td>79.0</td>
<td>77.3</td>
<td>79.3</td>
<td>44.0</td>
<td>85.9</td>
</tr>
<tr>
<td>Carson et al.</td>
<td>Q</td>
<td>DISIS</td>
<td>78.5</td>
<td>79.7</td>
<td>77.2</td>
<td>57.0</td>
<td>85.7</td>
</tr>
<tr>
<td>DISIS</td>
<td>Q</td>
<td>DISIS</td>
<td>89.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Carson et al.</td>
<td>Q</td>
<td></td>
<td>86.4</td>
<td>84.6</td>
<td>87.0</td>
<td>72.5</td>
<td>93.1</td>
</tr>
<tr>
<td>DNABINDPROT</td>
<td>T</td>
<td></td>
<td>78.6</td>
<td>9.3</td>
<td>90.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DP-BIND</td>
<td>T</td>
<td>Ozbek</td>
<td>74.0</td>
<td>63.7</td>
<td>75.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DISPLAR</td>
<td>T</td>
<td></td>
<td>80.1</td>
<td>45.1</td>
<td>86.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DBD-HUNTER</td>
<td>T</td>
<td></td>
<td>95.2</td>
<td>43.6</td>
<td>75.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Random Forests, Wu et al.</td>
<td>T</td>
<td>DBP374</td>
<td>91.4</td>
<td>76.6</td>
<td>73.2</td>
<td>70.0</td>
<td>91.3</td>
</tr>
<tr>
<td>BindN</td>
<td>Q</td>
<td>TS75</td>
<td>64.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>78.2</td>
</tr>
<tr>
<td>SVM, Wu et al.</td>
<td>Q</td>
<td>3.5Å cutoff</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>84.3</td>
</tr>
<tr>
<td>Random Forests, Wu et al.</td>
<td>Q</td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>85.5</td>
</tr>
<tr>
<td>Random Forests, Wu et al.</td>
<td>Q</td>
<td>TS75</td>
<td>80.5</td>
<td>67.2</td>
<td>81.8</td>
<td>34.1</td>
<td>-</td>
</tr>
<tr>
<td>DP-BIND</td>
<td>Q</td>
<td>4.5Å cutoff</td>
<td>78.0</td>
<td>67.8</td>
<td>79.0</td>
<td>31.6</td>
<td>-</td>
</tr>
<tr>
<td>Random Forests, Wu et al.</td>
<td>Q</td>
<td>TS75</td>
<td>78.2</td>
<td>51.4</td>
<td>84.6</td>
<td>34.1</td>
<td>-</td>
</tr>
<tr>
<td>DISIS</td>
<td>Q</td>
<td>6.0Å cutoff</td>
<td>81.6</td>
<td>7.7</td>
<td>99.2</td>
<td>19.0</td>
<td>-</td>
</tr>
</tbody>
</table>

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.

\*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.

\*Estimated from Figure 3 of [55].

\*Evaluation was done only on surface residues only.

\*Results reported in Supplementary Table S2 of [68].

\*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 and improve models of the binding site. 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. 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 to characterize active compounds for different proteins 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


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


