Prediction of transcription factor binding to DNA using rule induction methods

Mikael Huss[1, *] and Karin Nordström[2]

[1]School of Computer Science and Communication, Royal Institute of Technology, 100 44 Stockholm, Sweden
hussm@nada.kth.se
[2]Department of Cell and Organism Biology, Lund University, 223 62 Lund, Sweden; Current address: Discipline of Physiology, The University of Adelaide, Adelaide, SA 5005, Australia
Karin.Nordstrom@adelaide.edu.au

Abstract

In this study, we seek to develop a predictive model for finding the strength of binding between a particular transcription factor (TF) variant and a particular DNA target variant. The DNA binding paired domains of the Pax transcription factors, which are our main focus, show seemingly fuzzy and degenerate binding to various DNA targets, and paired domain-DNA binding is not a problem well suited for previously proposed algorithms. Here, we introduce a simple way to use rule induction for predicting the strength of TF-DNA binding. We have created a dataset consisting of 597 example cases for paired domain-DNA binding by collecting information about all published and quantified interactions between TF and DNA sequence variants. Application of the rule induction based method on this dataset yields a high, although far from ideal accuracy of 69.7% (based on cross-validation), but perhaps more importantly, several useful rules for predicting the binding strength have been found. Although the primary motivation for introducing the rule induction based methods is the lack of efficient algorithms for paired domain-DNA binding prediction, we also show that the method can be applied with some success to a more well-studied TF-DNA binding prediction task involving the early growth response (EGR) TF family.

Summary

The transcription of DNA into mRNA is initiated and aided by a number of transcription factors (TFs), proteins with DNA-binding regions that attach themselves to binding sites in the DNA (transcription factor binding sites, TFBSs). As it has become apparent that both TFs and TFBSs are highly variable, tools are needed to quantify the strength of the interaction resulting from a certain TF variant binding to a certain TFBS. Ideally, one would like to have a method where any combination of TF amino acids are allowed to interact with any TFBS nucleotide, and vice versa. Rule induction algorithms might be such a method. We used a simple way to predict interactions between protein and DNA: given experimental cases from the literature where the interaction strength between two sequences has been quantified, we created training vectors for rule induction by regarding each amino acid and nucleotide position as a single feature in the example vector. The resulting interaction strength was used as the target class or value. These training vectors were then used to build a rule induction model.

We applied the rule induction method to two protein families – transcription factors from the Pax and the early growth response (EGR) families – and their corresponding DNA targets.

* Corresponding author (phone: +46 (0)704 21 30 93)

The main focus of the study was that of the less well-studied problem of Pax paired domain-DNA binding. For this problem, we collected a comprehensive dataset, which we have made available as supplementary material, hoping that it will prove useful for future investigators. Also, for the Pax problem, we found sequence/binding strength correlations using measures from information theory. Prediction results were reasonably good: the rule induction approach achieved a correlation coefficient of 0.52 on unseen (and noisy) examples in the EGR case, and a classification accuracy of 69.7% for the paired domain, as evaluated by cross-validation.

We conclude that rule induction methods can be useful for predicting binding strength between protein and DNA, given training examples where individual sites and classes are varied and the resulting interaction strength is quantified. Even when no knowledge about specific interactions is included, sites that have been shown to be of importance from mutagenesis and crystallography appear in the rules.

Introduction

The transcription of DNA into mRNA is initiated and aided by a number of transcription factors (TFs), proteins with DNA-binding regions that attach themselves to transcription factor binding sites (TFBSs) in the DNA. As it has become apparent that both TFs and TFBSs are highly variable, tools are needed to quantify the strength of the interaction (and, subsequently, rate of DNA transcription) resulting from a certain TF variant binding to a certain TFBS. Many studies exploring the prediction of such interaction strengths have been performed (see for example [1-6]), most of which are tailored to the EGR (early growth response) zinc finger proteins. In such studies it has been common to assume independence between individual DNA-protein interactions; the binding of, say, amino acid 3 from the TF protein sequence to nucleotide 5 from the TFBS nucleotide sequence has been assumed to be independent from all other bindings. As some authors have pointed out, this independence (or ”additivity”) assumption is clearly unrealistic, although it often gives a good approximation [7]. Ideally, one would like to have a method where any combination of TF amino acids are allowed to interact with any TFBS nucleotide, and vice versa. We believe that rule induction algorithms are such a method.

Rule induction is an ideal method for finding multi-dimensional (“many-to-one” and “one-to-many”) relationships in TF-TFBS binding data. Rule induction methods use input data to construct models based on given example vectors, which contain the relevant example attributes (also called features) and the target class (in classification problems) or value (in regression problems). Rules are expressed in terms of the attribute values and can be examined and readily understood by a human user, i.e. it is always trivial to deduce the model’s prediction pathway. Our initial focus was to understand the seemingly fuzzy and degenerate binding of Pax transcription factor proteins to DNA. The DNA-binding domain of the Pax transcription factors is called the *paired domain*, which is the only part of the Pax TFs that we will be concerned with in the following. Previously published predictive models do not seem to be readily applicable to the paired domain-DNA binding problem. Some of them [6, 8] are explicitly constructed for analyzing EGR protein-DNA binding, while those models that could be generalized to any transcription factor family present other difficulties when trying to adopt them for studying paired domain-DNA binding. One such method is based on selection experiments [5], but as far as we are aware, suitable data of this kind are not available for the paired domain. Two other methods are based on scoring pairs of interacting nucleotide and amino acid residues [1, 4]. These methods could in principle be applied to the paired domain, but would at present, due to insufficient constraints on the “allowed” nucleotide-amino acid contacts, run into the problem of combinatorial explosion. Since the

paired domain has 128 (or 129) amino acids that could each potentially interact with one or more of ≈ 20 nucleotides, the number of possible configurations which would have to be scored is prohibitively high.

In this study, we used the software package Rule Discovery System (RDS), a commercial product from Compumine (Stockholm, Sweden; http://www.compumine.com/) which is, however, freely available for academic use. The primary reason for using this package was the possibility to include additional domain knowledge encoded as PROLOG rules. We use this feature for trying different representations of amino acids and nucleotides (see below and Methods). Rule induction methods aim at generating useful classification rules. These can be displayed in the form of decision trees (using recursive partitioning) where each bifurcating fork indicates an attribute that will determine the resulting class. Since the input variables have to be split to binary variables the resulting trees often produce noisy and unstable predictors. This can be improved by fitting many resulting trees to bootstrap-resampled versions of the data and combine them by majority vote using bagging (*b*ootstrap *agg*regating). Covering algorithms take each class separately, and try to cover all examples in that class, at the same time excluding examples not in the class. These algorithms operate by adding tests to the rule that is under construction, always trying to create a rule with maximum accuracy. Whereas decision tree algorithms choose an attribute to maximize the separation between the classes (using an information gain criterion), the covering algorithm chooses an attribute-value pair to maximize the probability of the desired classification.

We used a simple way to predict interactions between TFs and TFBSs: given experimental cases from the literature where the interaction strength between two sequences has been quantified, we created training vectors for rule induction by regarding each amino acid (for TFs) and nucleotide (for TFBSs) position as a single feature in the example vector. Where the interaction strength had been numerically quantified (as in the EGR case), the logarithm of this value was used as a target variable for regression. When this was unavailable the interaction strength was described in terms of a few categories or classes (as in the paired domain case), and the class was used as the target variable for classification. These training vectors were then used to build a rule induction model.

As it is well-known that data-driven model building is highly dependent on a suitable representation of the training data (see for instance [9]), we used two alternative representation schemes in addition to the default representation. In the default mode, amino acids and nucleotides were represented simply as letters. In the first of the alternative schemes, the “enriched” representation, amino acids and nucleotides were represented in terms of their physico-chemical characteristics (see Methods). The second alternative representation, the “numerical” representation, described each amino acid or nucleotide with three numerical values ([10] and see Methods). The numerical representation scheme thus gives the finest granularity while the ”default” is the roughest.

To evaluate the performance of the rule induction approach on a known problem, we constructed models for a zinc finger data set previously analyzed in the literature [6]. Zinc finger domains are found in many eukaryotic transcription factors. Many earlier studies on prediction of protein-DNA interaction specificity have focused on zinc finger-DNA complexes, primarily for zinc finger regions from the EGR (early growth response) protein family (for example [2, 4, 5]). In these complexes, the DNA and protein regions relevant for DNA-protein binding are known with good precision, and the identities of amino acid and nucleotide residues in the relevant sequence positions have been systematically varied in experiments. These studies have given valuable insight into the possibility of ‘protein-DNA recognition codes’ (which may be probabilistic in nature; see [11]).

For paired domains, no systematic experiments on this scale exist, and the regions relevant for DNA binding have not been pinpointed to the same degree as in the EGR family. Therefore no predictive paired domain-DNA binding models have been proposed. Our main aim in initiating this study was to understand the logic of paired domain-DNA binding. Paired domain (Pax) transcription factors have fundamental roles in the development of primarily the nervous system and its associated sensory organs but also of other peripheral organ systems. Many severe human developmental defects, such as the Waardenburg syndrome [12] and Aniridia [13] are caused by paired domain mutations, and Pax defects have also been suggested to be involved in cancerogenesis (see [14, 15]). The paired domain consists of 128 amino acids organized as two helix-turn-helix motifs joined via a linker region. The crystal structure for the paired domain bound to its DNA recognition sequence has been established for two Pax proteins [16, 17]. While paired domains bind to similar consensus binding sites *in vitro*, genetically defined sites vary (e.g. [18-20]). Furthermore, the paired domain recognition site is unusually long (16-20 nucleotides) compared to other DNA-binding proteins.

After showing that rule induction can give reasonable results on a previously studied dataset, without having been tuned to incorporate any domain specific information about protein-DNA contacts, we attacked the difficult problem of paired domain-DNA binding. Many direct DNA-protein contacts are known from experiments, and the corresponding sequence positions tend to be conserved. Our aim is to understand how non-conserved sequence positions modify the strength of paired domain-DNA binding, or in other words, to identify more subtle interactions that are not immediately deduced from crystallography data. We also evaluated whether including information gain profiles of individual sites would improve the prediction quality, and found this not to be the case. Using the entire problem space, we found that paired domain-DNA interaction can be predicted with an accuracy (≈ 70%) that is far from perfect but good compared to a random guess (≈ 43%). More importantly to the molecular biologist, we found a number of highly predictive and biologically understandable rules that can be used to predict and understand paired domain-DNA binding.

Results

Applying rule induction to zinc finger-DNA binding data

We tested our method on previously analyzed data [6] in which a microarray approach was used to generate a comprehensive dataset, where the binding of wild-type zinc fingers and four mutant zinc finger variants to all their possible tri-nucleotide targets was quantified (see [6] for more details on zinc fingers and this data set in particular). Because of the systematic design of the experiments used to arrive at this dataset, we were able to use the binding measurements (after taking logarithms) as continuous target values for regression. We used RDS to construct a range of models (using recursive partitioning, covering and bagging) from this data set, and tested the resulting models' performance on independently reported zinc finger-DNA binding experiments [22]. The training examples were vectors consisting of ten amino acids followed by three nucleotides involved in direct binding, and the target variable (the natural logarithm of K_d). We tried all three representations of the nucleotides and amino acids.

We evaluated the performance using the correlation coefficient. The models were able to learn the training data set quite well, with correlation coefficients of up to 0.84 (Table 1), as assessed by leave-one-out cross-validation using the 320 examples. Upon testing on the independent cases, the best model (covering, numerical representation) gave a correlation coefficient of 0.52. We find this performance to be satisfactory, considering that (i) we have not included any problem-specific information about individual nucleotide-amino acid

contacts, and (ii) the test cases come from experiments performed in a different lab with different experimental conditions. Indeed, one of the experiments in the test data [22] was also performed in the training data set [6] and the reported binding strength differs by a considerable margin between the two. (In the training data set, the protein-DNA combination in question has a K_d of ≈ 0.38, while in the test set it has K_d ≈ 0.15.)

Interestingly, we found that the models based on recursive partitioning and/or bagging resulted in overfitting on this problem. The covering models, which yield more compact hypotheses with fewer rules, were able to generalize better (Table 1). We hypothesize that the sampling of amino acid sequences was insufficient (there were only five zinc finger variants) to build a detailed model with many classification rules.

Table 1. Prediction performance on zinc fingers using a selection of methods and data representations (320 data points in training data, 14 data points in test data). Note that the recursive partitioning-based methods suffer from heavy overfitting: the training set results are much better than the test set results.

Method	Representation[1]	CC (training)[2]	CC (test)[3]
Covering	Numerical	0.68	0.52
Covering	Default	0.69	0.36
Recursive partitioning	Default	0.74	0.35
Bagging with rec. part.	Numerical	0.84	0.14

As we obtained good performance level on this microarray-based study with both amino acid and nucleotide sequences varied, and we had thus shown that rule induction can give good results on a previously studied dataset, without having been tuned to incorporate any domain specific information about protein-DNA contacts, we attacked the more difficult problem of paired domain-DNA binding.

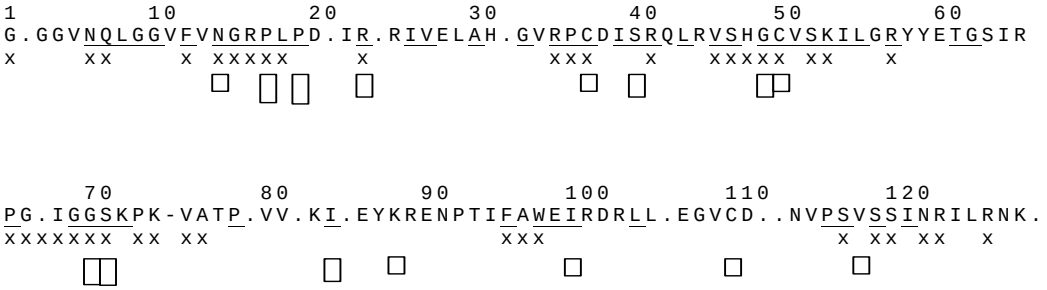
Application of an information gain measure on paired domain-DNA binding data

Since the paired domain was expected to be harder to model, we wanted to extract sites with the highest information gain for fine-tuning of the training data. We collected 597 binding experiments from the literature and assigned the interaction strength in each experiment to one of three classes (++: strong interaction, +: weak interaction, -: no interaction). While the EGR data set provided us with quantified binding information in the form of binding coefficient, the Pax data set mostly consisted of qualitative results, sometimes only in the form of gel images. Thus, conversion into classes instead of numerical values was necessary.

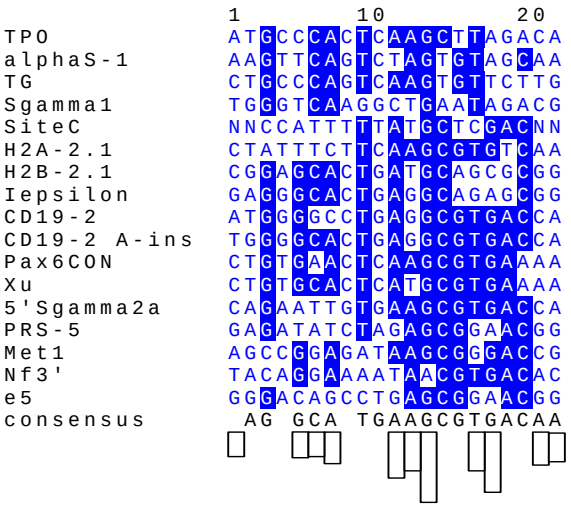
Using the full set of 597 example cases, we calculated information gain profiles quantifying the usefulness of each single sequence position for predicting binding strength (Figure 1a, b). For the purposes of these calculations, we considered only the identity of each nucleotide or amino acid, with no special encoding of their physico-chemical character or other features. Specifically, we used the information gain ratio instead of the standard information gain measure, which favors attributes that can take on a large number of values [23]. Many amino acid positions that we found to show substantial information content have previously been described to be associated with binding specificity. These include position 17 and 48 (Figure

[1] Refers to default or numerical feature representation.
[2] Correlation coefficient, estimated by leave-one-out cross-validation on the training data.
[3] Correlation coefficient, estimated by testing on 14 independent examples collected from [24].

A



B



C

Nucleotide	Amino acid	Inf. Gain ratio
13	104	0.0906
18	128	0.0884
13	128	0.0848
13	101	0.0832
13	70	0.0829
13	71	0.0829
15	128	0.0816
13	37	0.0809
13	17	0.0798
17	101	0.0797
13	117	0.0793
6	128	0.0793
18	21	0.0791
13	80	0.0781
13	55	0.0781
14	128	0.0763
17	128	0.0761
16	128	0.0760
18	79	0.0756
17	104	0.0754

Figure 1. Paired domain information gain profiles. In a) a consensus paired domain with the conserved residues (invariant in most wildtype proteins [60]) shown underlined. Sites that vary too much to reach a meaningful consensus are shown with a period ('.'). DNA contact points of Pax6 and Paired [16, 17] are shown under the sequence with 'x'. Between position 74 and 75 (where Pax3 and Pax7 have an additional residue) a gap in the alignment is shown with a hyphen ('-'). The information gain ratios for the 15 most informative sites are shown as histobars directly under the consensus sequence. The height of the bar is proportional to the information gain ratio of the corresponding site. Note that many of the positions with high information peaks are located close to residues that make DNA contacts. Panel b) shows a nucleotide alignment of a subset of the DNA sites analyzed, together with the information gain profile for the ten most informative nucleotide sites under the alignment. The sites are named as in their primary reference. The N-terminal end binds to site 10-21, the C-terminal to 1-7, and the linker to 8-15. In c) the top 20 combinations of sites are shown in the order of the information they provide. Column 1 shows nucleotide positions numbered 1-21 as in the alignment in panel b); column 2 shows the amino acid residues numbered 1-128 as in panel a) above; and 3) finally shows the information that doublet provides.

a) that have been shown to be naturally occurring Pax3 mutations in families with the Waardenburg syndrome [24], and residues in areas where many residues contact the DNA directly [16, 17], such as positions 70 and 71. These two have also been shown to be responsible for the lack of binding in *Acropora millepora* PaxD [25]. Other results were somewhat surprising and included amino acid residues 84, 88, and 109, located in regions of low or no direct DNA contact. One can hypothesize that they have a crucial role in maintaining the protein 3D backbone.

We also calculated information gain ratio profiles quantifying the usefulness of all combinations of nucleotide and amino acid pairs for predicting the binding strength (Figure 1c, for the 20 most informative combinations). A few sites appear several times in this list. These include amino acid residues 101 and 104, two sites with low individual information gain, but which are fairly conserved across wildtype paired domains [26]. Nucleotides bound by the N-terminal half of the paired domain (nucleotides 10-21 [16]) provide more information for determining binding specificity (Figure 1c), which is supported by the fact that the C-terminal is dispensable for DNA-binding by some paired domains [17, 27].

Interestingly, these preliminary studies were of little value when it came to building global prediction models (see below). The single positions with high information gain values did not necessarily appear in any reliable classification rules. Apparently, combinations of several amino acid and nucleotide sequence positions are necessary to determine binding specificity; knowing one or two positions is not enough.

Applying rule induction methods on paired domain-DNA binding data

Using the same 597 case dataset as above, we applied rule induction to the paired domain. The best prediction results were obtained using the numerical representation (see Methods) and a bagging algorithm on top of recursive partitioning (Table 2). The predictive accuracy was estimated using 5-fold cross-validation, which means that results from five runs, in each of which 80% of the data was used for building the predictive model and 20% was used purely for testing the resulting model's accuracy, were averaged. The 20% of the examples used for "blind testing" in each of the 5 runs were cycled so that in the end, all examples had been used for testing. Bagging with recursive partitioning yielded an accuracy of 69.7% (Table 2), whereas random guessing would have yielded 42.7%. A more detailed look at this best performing method (Table 3) shows that the cases of strong binding give the highest precision and recall values.

Table 2. Prediction performance on the Paired domain using different methods and data representations (597 data points). The performance was evaluated by 5-fold cross-validation.

Method	Default[4]	Enriched[5]	Numerical values[6]
Recursive partitioning	54.7%	54.7%	53.0%
Covering	57.7%	59.4%	55.0%
Bagging with rec. part.	65.0%	62.9%	69.7%
Bagging with covering	60.7%	61.0%	59.2%

[4] Default feature representation.
[5] Prediction accuracy using the enriched feature representation.
[6] Prediction accuracy when amino acid and nucleotide residues were assigned numerical values.

In addition, we evaluated a feature selection strategy, where only the 15 amino acid attributes with the highest information gain ratios according to our preliminary studies were used in building the prediction model. This model performed much worse than the models built from all amino acid attributes (not shown). The explanation for this is probably that individual sequence positions are not informative enough; the correlations between sequences and binding strengths occur at the level of combinations of nucleotide and amino acid positions. This can also be seen by inspecting the classification rules from the best models (Figure 2). The first rule gives an example of a non-binding case ('-'). As noted above, site 70 has been shown to affect the lack of binding in *A. millepora* PaxD [25] and is thus consistent with our finding. Amino acid 55 is located in the linker region of the paired domain, and one can therefore assume that mutations at this site affects how the TF fits into the grooves of the TFBS. Nucleotide 1 and amino acid 70 might have a strong correlation with non-binding classification, as they also appear in rule 4, but this time combined with amino acid 82.

Table 3. Precision and recall for the best-performing method on the Pax data set (recursive partitioning with bagging, using numerical representation). The method yielded an overall accuracy of 69.7% with 527 rules.

Class	Precision[7]	Recall[8]
-	0.67188	0.65152
+	0.60345	0.47619
++	0.75172	0.86166

The second rule, which gives a strong ('++') binding profile, shows the importance of nucleotide 17 (Figure 1b) and amino acid 37, which are both highly conserved. Furthermore amino acids 37 and 75 are located in a core part of the binding region of the paired domain (Figure 1a, and [16, 17]). The presence of key amino acids and nucleotides at these sites thus make sense for strong binding between the TF and TFBS. Rule 3 shows the collaborative effect of four nucleotides from the core of the conserved TFBS sequence. These correlate with an amino acid position known to contact this region [16, 17] to generate "weak binding" classification.

The vast majority of classification rules (Figure 2) include both amino acid and nucleotide attributes, underlining the fact that neither the paired domain nor the target DNA site can in themselves guarantee good binding; it is the relation between the two that determines the binding strength. Even if there are not enough 'good' rules to cover the whole problem space, those rules that we found to have high accuracy are highly useful where applicable.

Discussion

We have shown that rule induction can be applied to TF-DNA binding specificity prediction. No prior knowledge about DNA binding has been incorporated and all sequence positions have been considered equally. The results therefore provide an unbiased model of protein-DNA interactions. Our approach is reasonably successful on the previously analyzed EGR transcription factor family, but we mainly want to introduce it as a powerful tool in more complex cases, such as paired domain-DNA interaction prediction. In order to address this

[7] True positives / (true positives + false positives). This is also called the positive predictive value, or PPV.

[8] True positives / (true positives + false negatives)

Example rule 1.
Generated by the default representation.

nt 1 **is** T
aa 55 **is not** G
aa 70 **is not** G

Prediction: - (no binding)
Reliability: 5/5 (training set), 2/2 (test set)

Example rule 2.
Generated by the default representation.

nt 17 **is** G
aa 37 **is not** C
aa 75 **is** Q
aa 81 **is not** I

Prediction: ++ (strong binding)
Reliability: 10/10 (training set), 2/2 (test set)

Example rule 3.
Generated by the enriched representation.

nt 13 **is** G
nt 14 **is** C
nt 15 **is** G
nt 19 **is** C
aa 55 **is** polar

Prediction: + (weak or ambiguous binding)
Reliability: 11/11 (training set), 4/4 (test set)

Example rule 4.
Generated by the numerical representation.

nt 1 **has steric bulk descriptor value less than or equal to** -0.8203
aa 70 **has hydrophobicity descriptor value less than or equal to** –1.115
aa 82 **has hydrophobicity descriptor value larger than** 2.17

Prediction: - (no binding)
Reliability: 11/11 (training set), 3/3 (test set)

Figure 2. Paired domain decision rule examples. The decision rules were extracted with high prediction accuracy, and were generated using three different data representations in RDS. Note that all rules use both nucleotides (nt) and amino acid (aa) residues to predict binding strength. The rules consist of logical statements involving the nucleotide and amino acid positions. In the default representation, only the identity of each amino acid or nucleotide is considered. In the enriched representation, each amino acid is also described with attributes relating to hydrophobicity, charge, aromaticity, and special characteristics. Thus, rules generated using this representation can contain logical relations of the form "amino acid 2 is hydrophobic", in addition to rules involving residue identities. In the numerical representation, nucleotides and amino acid residues are encoded by numerical biophysical descriptor values, which are then used to generate rules. The three nucleotide descriptors roughly represent size, electronic properties and hydrophobic properties, and the three amino acid descriptors roughly correspond to a hydrophobicity measure, a steric bulk measure, and a polarity measure [10, 57]. In this case, rules are not of the form "X is Y" or "X is not Y", but of the form "X is larger than Y" (or "X is larger than or equal to Y") or "X is smaller than Y" (or "X is smaller than or equal to Y"). The Predictions line shows the predicted binding class for examples matching the rule shown. The Reliability line lists the performance of the rule on example cases in the training set (which consists of a randomly selected 80% of the examples) and the test set (which consists of the remaining 20%).

problem, we have constructed a data set based a large number of experiments culled from the literature. This data set, which is provided as supplementary material and could prove useful for future investigators, is in itself one of the main contributions of this paper. Our rule induction method, applied on this data set, shows a fair if unspectacular prediction performance, but more importantly, it identifies a number of classification rules of high predictive accuracy. These rules are expressed in terms of both amino acids and nucleotides, and may help in identifying true physical interactions. Our study also pointed out two interesting details. Firstly, in the paired domain case, data pre-processing by feature selection using information gain (which measured correlations between sequence positions or pairs of positions and the binding strength) was useless for building the prediction models. This suggests that paired domain-DNA binding specificity is determined by higher-level, complex interaction patterns, where the identities of individual residues are relatively unimportant in themselves: the right combination is what matters. Secondly, we found that data representation was fairly unimportant in this problem domain. The relatively sophisticated numerical representation did, on the whole, perform best, but only by a small margin compared to the rough default method.

Methods

Binding data

EGR binding data. We used a previously studied comprehensive dataset (see supplementary material in [6]), where the binding of wild-type zinc fingers and four mutant zinc finger variants to all their possible trinucleotide targets was quantified. Our training vectors consisted of the amino acids of zinc finger 2 involved in binding (10 in number), followed by the interacting nucleotides (3 in number) and the target variable (the natural logarithm of the K_d value of the interaction; as there were several replicates of each experiment, we used the average K_d value in each case). Finally we tested the resulting models' performance on independently reported zinc finger-DNA binding experiments from the literature [22]. The examples found in that paper were manually converted into the same format with 10 amino acids, 3 nucleotides and (the natural logarithm of) one K_d value.

Paired domain binding data. For paired domains we had to retrieve data from a large range of publications and treat them in a consistent way to build our example vectors.

Paired domains can be either 128 or 129 amino acids long, as paired domains of the Pax3/7 class have an extra amino acid inserted at position 75. We thus used 129 attributes for the amino acid positions, and set position 75 to a sequence gap in all the paired domains with 128 amino acids. We elected to use 21 attributes (corresponding to 21 nucleotide positions) in the DNA sequence targeted by the paired domain. The specific nucleotide positions used in each case were obtained by making alignments of the DNA sequences using a few early publications as primary guidelines (mainly [19, 21, 28, 29]). We thereafter assumed that the corresponding nucleotide positions were comparable as far as amino acid - nucleotide interactions were concerned. Due to this way of presenting and analyzing the data set, DNA sequences that were difficult to align convincingly had to be left out from the remaining analysis. Since the binding strengths were reported in a wide variety of ways in the articles, from numerical values to pictures of spots on a gel, the assignment of classes to examples was a difficult task. The dataset is therefore, in machine learning jargon, noisy. As an additional difficulty, the data set is not evenly sampled: the available binding data in many cases uses different DNA sequences for each paired domain.

Binding data was retrieved from the following references: [16, 18, 19, 21, 24, 25, 28-55]. We used binding data for paired domains only when these could be analyzed without interference

from homeodomains or other parts of the Pax sequence. The complete set consisting of 597 binding experiments used to build and evaluate the models. 158 DNA-sites were aligned with ClustalW [56], adjusted by eye, and trimmed down to 21 nucleotides. The sites were adjusted using a range of sites tested [21] as a guide, using their nucleotides 4-24, as well as [19, 28, 29]). No gaps were allowed in the nucleotide alignment. Nucleotide sequences that could not be convincingly aligned with the other sites, such as 5aCON [36], were excluded from the remaining analyses. 103 complete 128-129 amino acid paired domains (wildtype and those altered by mutagenesis) were aligned using ClustalW. Gaps were allowed between positions 74 and 75, where wildtype Pax3 and Pax7 have an additional residue, but at no other positions.

Binding strengths were extracted from the publications and given three possible denominations: '-' for no binding, '+' for weak or ambiguous binding, and '++' for definite binding. Since much published binding data is unquantified, we used comparisons between known sites in different experiments to determine the level of binding, using [21] as a primary guide. For several paired domain–DNA binding site combinations, multiple references were found, confirming our initial judgments. Ambiguities between publications led to the exclusion of the data point(s) in question.

Sequence descriptors

We initially modeled each sequence with one attribute for each of its nucleotides, allowing as a value either a letter corresponding to each of the four naturally occurring nucleotides, or 'N' for unknown, followed by the attributes corresponding to the amino acid residues, allowing as a value a letter corresponding to either one of the 20 naturally occurring amino acids or a hyphen ('-') for a gap. This initial coding was identical for each of the three representation schemes. In the default scheme the initial representation was not modified. In the numerical representation scheme, each letter was expanded into three numerical values corresponding to three biophysical descriptor scales of amino acids and nucleotides, respectively [10, 57]. These descriptors were generated by compressing a large number of previously used descriptors using principal component analysis and the partial least squares algorithm to yield more compact, compound descriptors. The nucleotide descriptors are derived from 24 different physico-chemical properties and can be said to represent size, electronic properties and hydrophobic properties. The amino acid descriptors are derived from 26 physico-chemical properties, and the three descriptors we use roughly correspond to, respectively, a hydrophobicity measure, a steric bulk measure, and a polarity measure [10]. In the enriched representation scheme, only the letters corresponding to amino acid positions were expanded. It described amino acids in terms of polarity (polar or nonpolar), hydrophobicity (hydrophobic or non-hydrophobic), charge (charged or not) and "special characteristics" (whether the amino acid has "special properties" like the –SH group in cysteine and the ring structure in proline; apart from these, glycine is also included among the "special" amino acids).

Information gain

The information gain (*IG*) for an attribute or a *feature* (here denoted as *f*) with respect to the set of classes (here denoted as *X*) was calculated as:

$$IG(f,X) = H(X) - H(X|f)$$

where $H(X)$ is the entropy of X :

$$H(X) = -\sum_{x \in X} P(x) \ln P(x)$$

and $H(X|f)$ is the entropy of X when the value v of feature f is known:

$$H(X|f) = -\sum_{v} P(v) \sum_{x \in X} P(x|v) \ln P(x|v).$$

$P(x)$ is the probability that an example belongs to class x, $P(v)$ is the probability that feature f has value v, and $P(x|v)$ is the probability that an example belongs to class x given that feature f has value v.

However, the information gain tends to favor attributes with many possible values. Thus for example, an attribute corresponding to a highly variable amino acid position might get a high information gain value even if the variations are not strongly correlated with the class. Therefore, we used the information gain ratio (*IG-ratio*), which is the information gain divided by a correction term that describes information needed to determine the value of the attribute:

$$IG\text{-}ratio\ (f,X) = IG(f,X) / (-\sum_{v} P(v) \ln P(v)).$$

Above, the observed nucleotides and amino acids at specific sequence positions were considered as the outcomes of random variables with alphabets Ω_n and Ω_a comprising, respectively, letters corresponding to the four nucleotides (A, G, T, C) and 'N' for unknown, and letters corresponding to 20 natural amino acids and '-' for a gap. Similarly, the binding strength was interpreted as the outcome of a random variable X with possible outcomes in Ω_c = (-, +, ++).

Rule induction

We used the rule induction package Rule Discovery System (RDS [58]) to build models from the example cases. RDS is freely available for academic use and allows the user to create classification rules either in the form of decision trees (using *recursive partitioning*) or lists of potentially overlapping rules (using *covering*). Recursive partitioning gives rise to decision trees, which are appropriate for the task at hand, as they handle discrete data, evaluate information context-specifically, and represent extracted knowledge in a comprehensible manner. Covering algorithms have the additional advantage of often yielding more compact hypotheses than decision trees [59].

In addition to these methods, RDS also supports ensemble methods such as *bagging,* which works by generating a set of hypotheses by repeated subsampling of the data. At prediction time, the predictions of different hypotheses are combined to yield a single prediction. RDS furthermore allows the user to specify background knowledge about the problem domain in PROLOG format. This feature was used for encoding the enriched and numerical representations described above.

Supplementary material

The paired domain data set can be downloaded as a tab-separated text file from http://www.csc.kth.se/~hussm/pax/pax-data.txt and as an ARFF (Attribute-Relation File Format) file from http://www.csc.kth.se/~hussm/pax/pax-data.arff.

Acknowledgements

We thank Per Lidén at Compumine for valuable discussions and Joel Westerberg for computer assistance.

References

1. Suzuki M, Yagi N: **DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families**. *Proc Natl Acad Sci U S A*, **91**(26):12357-12361, 1994.

2. Choo Y, Klug A: **Physical basis of a protein-DNA recognition code**. *Curr Opin Struct Biol*, **7**:117-125, 1997.

3. Greisman HA, Pabo CO: **A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites**. *Science*, **275**(5300):657-661, 1997.

4. Mandel-Gutfreund Y, Margalit H: **Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites**. *Nucleic Acids Res*, **26**(10):2306-2312, 1998.

5. Benos PV, Lapedes AS, Stormo GD: **Probabilistic code for DNA recognition by proteins of the EGR family**. *J Mol Biol*, **323**:701-727, 2002.

6. Bulyk ML, Johnson PLF, Church GM: **Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors**. *Nucleic Acids Res*, **30**(5):1255-1261, 2002.

7. Benos PV, Bulyk ML, Stormo GD: **Additivity in protein-DNA interactions: how good an approximation is it?** *Nucleic Acids Res*, **30**(20):4442-4451, 2002.

8. Desjarlais JR, Berg JM: **Toward rules relating zinc finger protein sequences and DNA binding site preferences**. *Proc Natl Acad Sci U S A*, **89**:7345-7349, 1992.

9. Langley P, Simon H: **Applications of machine learning and rule induction**. *Communications of the ACM*, **38**:54-64, 1995.

10. Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S: **New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids**. *J. Med. Chem.,* **41** (14), 2481 - 2491, 1998.

11. Benos PV, Lapedes AS, Stormo GD: **Is there a code for protein-DNA recognition? Probab(ilistical)ly.** *BioEssays*, **24**(5):466-475, 2002.

12. Morell R, Friedman TB, Moeljopawiro S, Hartono, Soewito, Asher JH, Jr.: **A frameshift mutation in the HuP2 paired domain of the probable human homolog of murine Pax-3 is responsible for Waardenburg syndrome type 1 in an Indonesian family**. *Hum Mol Genet*, **1**(4):243-247, 1992.

13. Jordan T, Hanson I, Zaletayev D, Hodgson S, Prosser J, Seawright A, Hastie N, van Heyningen V: **The human PAX6 gene is mutated in two patients with Aniridia**. *Nat Genet*, **1**(5):328-332, 1992.

14. Chi N, Epstein JA: **Getting your Pax straight: Pax proteins in development and disease**. *Trends Genet*, **18**(1):41-47, 2002.

15. Schaner ME, Ross DT, Ciaravino G, Sorlie T, Troyanskaya O, Diehn M, Wang YC, Duran GE, Sikic TL, Caldeira S *et al*: **Gene expression patterns in ovarian carcinomas**. *Mol Biol Cell*, **14**(11):4376-4386, 2003.

16. Xu HE, Rould MA, Xu W, Epstein JA, Maas RL, Pabo CO: **Crystal structure of the human Pax6 paired domain-DNA complex reveals specific roles for the linker region and carboxy-terminal subdomain in DNA binding**. *Genes Dev*, **13**(10):1263-1275, 1999.

17. Xu W, Rould MA, Jun S, Desplan C, Pabo CO: **Crystal structure of a paired domain-DNA complex at 2.5 A resolution reveals structural basis for Pax developmental mutations**. *Cell*, **80**(4):639-650, 1995.

18. Czerny T, Busslinger M: **DNA-binding and transactivation properties of Pax-6: three amino acids in the paired domain are responsible for the different sequence recognition of Pax-6 and BSAP (Pax-5)**. *Mol Cell Biol*, **15**(5):2858-2871, 1995.

19. Epstein J, Cai J, Glaser T, Jepeal L, Maas R: **Identification of a Pax paired domain recognition sequence and evidence for DNA-dependent conformational changes**. *J Biol Chem*, **269**(11):8355-8361, 1994.

20. Punzo C, Seimiya M, Flister S, Gehring WJ, Plaza S: **Differential interactions of *eyeless* and *twin of eyeless* with the *sine oculis* enhancer**. *Development*, **129**(3):625-634, 2002.

21. Czerny T, Schaffner G, Busslinger M: **DNA sequence recognition by Pax proteins: bipartite structure of the paired domain and its binding site**. *Genes Dev*, **7**:2048-2061, 1993.

22. Choo Y, Klug A: **Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions**. *Proc Natl Acad Sci USA*, **91**:11168-11172, 1994.

23. Quinlan JR: **C4.5: Programs for machine learning**: Morgan Kaufman Publishers; 1993.

24. Fortin AS, Underhill DA, Gros P: **Reciprocal effect of Waardenburg syndrome mutations on DNA binding by the Pax-3 paired domain and homeodomain**. *Hum Mol Genet*, **6**(11):1781-1790, 1997.

25. Nordström K, Scholten I, Nordström J, Larhammar D, Miller D: **Mutational analysis of the *Acropora millepora* PaxD paired domain highlights the importance of the linker region for DNA binding**. *Gene*, **320**:81-87, 2003.

26. Gröger H, Callaerts P, Gehring WJ, Schmid V: **Characterization and expression analysis of an ancestor-type *Pax* gene in the hydrozoan jellyfish *Podocoryne carnea***. *Mech Dev*, **94**(1-2):157-169, 2000.

27. Jun S, Wallen RV, Goriely A, Kalionis B: **Lune/eye gone, a Pax-like protein, uses a partial paired domain and a homeodomain for DNA recognition**. *Proc Natl Acad Sci USA*, **95**:13720-13725, 1998.

28. Chalepakis G, Fritsch R, Fickenscher H, Deutsch U, Goulding M, Gruss P: **The molecular basis of the undulated/Pax-1 mutation**. *Cell*, **66**(5):873-884.

29. Jun S, Desplan C: **Cooperative interactions between the paired domain and homeodomain**. *Development* 1996, **122**:2639-2650, 1991.

30. Adams B, Dorfler P, Aguzzi A, Kozmik Z, Urbanek P, Maurer-Fogy I, Busslinger M: **Pax-5 encodes the transcription factor BSAP and is expressed in B lymphocytes, the developing CNS, and adult testis**. *Genes Dev*, **6**(9):1589-1607, 1992.

31. Apuzzo S, Gros P: **Site-specific modification of single cysteine Pax 3 mutants reveals reciprocal regulation of DNA binding activity of the paired and homeo domain**. *Biochem*, **41**:12076-12085, 2002.

32. Bäumer N, Marquardt T, Stoykova A, Spieler D, Treichel D, Ashery-Padan R, Gruss P: **Retinal pigmented epithelium determination requires the redundant activities of Pax2 and Pax6**. *Development*, **130**(13):2903-2915, 2003.

33. Chalepakis G, Goulding M, Read A, Strachan T, Gruss P: **Molecular basis of splotch and Waardenburg Pax-3 mutations**. *Proc Natl Acad Sci USA*, **91**(9):3685-3689, 1994.

34. Chalepakis G, Jones FS, Edelman GM, Gruss P: **Pax-3 contains domains for transcription activation and transcription inhibition**. *Proc Natl Acad Sci USA*, **91**(26):12745-12749, 1994.

35. Czerny T, Halder G, Kloter U, Souabni A, Gehring WJ, Busslinger M: ***twin of eyeless*, a second *Pax-6* gene of *Drosophila*, acts upstream of *eyeless* in the control of eye development**. *Mol Cell*, **3**:297-307, 1999.

36. Epstein JA, Glaser T, Cai J, Jepeal L, Walton DS, Maas RL: **Two independent and interactive DNA-binding subdomains of the Pax6 paired domain are regulated by alternative splicing**. *Genes Dev*, **8**(17):2022-2034, 1994.

37. Epstein JA, Shapiro DN, Cheng J, Lam PY, Maas RL: **Pax3 modulates expression of the c-Met receptor during limb muscle development**. *Proc Natl Acad Sci USA*, **93**(9):4213-4218, 1996.

38. Fitzsimmons D, Lutz R, Wheat W, Chamberlin HM, Hagman J: **Highly conserved amino acids in Pax and Ets proteins are required for DNA binding and ternary complex assembly**. *Nucleic Acids Res*, **29**(20):4154-4165, 2001.

39. Kalousova A, Benes V, Paces J, Paces V, Kozmik Z: **DNA binding and transactivating properties of the paired and homeobox protein Pax4**. *Biochem Biophys Res Commun*, **259**(3):510-518, 1999.

40. Kozmik Z, Daube M, Frei E, Norman B, Kos L, Dishaw LJ, Noll M, Piatigorsky J: **Role of pax genes in eye evolution: a cnidarian *PaxB* gene uniting Pax2 and Pax6 functions**. *Dev Cell*, **5**:773-785, 2003.

41. Maier H, Ostraat R, Parenti S, Fitzsimmons D, Abraham LJ, Garvie CW, Hagman J: **Requirements for selective recruitment of Ets proteins and activation of *mb-1*/Ig-a gene transcription by Pax-5 (BSAP)**. *Nucleic Acids Res*, **31**(19):5483-5489, 2003.

42. Meech R, Kallunki P, Edelman GM, Jones FS: **A binding site for homeodomain and Pax proteins is necessary for L1 cell adhesion molecule gene expression by Pax-6 and bone morphogenetic proteins**. *Proc Natl Acad Sci USA*, **96**(5):2420-2425, 1999.

43. Miller DJ, Hayward DC, Reece-Hoyes JS, Scholten I, Catmull J, Gehring WJ, Callaerts P, Larsen JE, Ball EE: ***Pax* gene diversity in the basal cnidarian *Acropora***

millepora **(Cnidaria, Anthozoa): Implications for the evolution of the *Pax* gene family**. *Proc Natl Acad Sci USA*, **97**(9):4475-4480, 2000.

44. Pellizzari L, Fabbro D, Lonigro R, Di Lauro R, Damante G: **A network of specific minor-groove contacts is a common characteristic of paired-domain-DNA interactions**. *Biochem J*, **315**(Pt 2):363-367, 1996.

45. Plaza S, de Jong D, Gehring WJ, Miller DJ: **DNA-binding characteristics of cnidarian Pax-C and Pax-B proteins *in vivo* and *in vitro*: no simple relationship with the Pax-6 and Pax-2/5/8 classes**. *J Exp Zool B Mol Dev Evol*, **299**(1):26-35, 2003.

46. Rodrigo I, Hill RE, Balling R, Münsterberger A, Imai K: **Pax1 and Pax9 activate *Bapx1* to induce chondrogenic differentiation in the sclerotome**. *Development*, **130**:473-482, 2003.

47. Sheng G, Harris E, Bertuccioli C, Desplan C: **Modular organization of Pax/homeodomain proteins in transcriptional regulation**. *Biol Chem*, **378**(8):863-872, 1997.

48. Sheng G, Thouvenot E, Schmucker D, Wilson DS, Desplan C: **Direct regulation of *rhodopsin 1* by *Pax-6/eyeless* in *Drosophila:* evidence for a conserved function in photoreceptors**. *Genes Dev*, **11**(9):1122-1131, 1997.

49. Skala-Rubinson H, Vinh J, Labas V, Kahn A, Phan Dinh Tuy F: **Novel target sequences for Pax-6 in the brain-specific activating regions of the rat aldolase c gene**. *J Biol Chem*, **277**(49):47190-47196, 2002.

50. Sun H, Dickinson DP, Costello J, Li WH: **Isolation of *Cladonema Pax-B* genes and studies of the DNA-binding properties of cnidarian pax paired domains**. *Mol Biol Evol*, **18**(10):1905-1918, 2001.

51. Sun H-M, Rodin A, Zhou Y, Dickinson DP, Harper DE, Hewett-Emmet D, Li W-H: **Evolution of paired domains: Isolation and sequencing of jellyfish and hydra *Pax* genes related to *Pax-5* and *Pax-6***. *Proc Natl Acad Sci USA*, **94**(10):5156-5161, 1997.

52. Tell G, Scaloni A, Pellizzari L, Formisano S, Pucillo C, Damante G: **Redox potential controls the structure and DNA binding activity of the paired domain**. *J Biol Chem*, **273**(39):25062-25072, 1998.

53. Vogan KJ, Gros P: **The C-terminal subdomain makes an important contribution to the DNA binding activity of the Pax-3 paired domain**. *J Biol Chem*, **272**(45):28289-28295, 1997.

54. Vogan KJ, Underhill DA, Gros P: **An alternative splicing event in the Pax-3 paired domain identifies the linker region as a key determinant of paired domain DNA-binding activity**. *Mol Cell Biol*, **16**(12):6677-6686, 1996.

55. Wheat W, Fitzsimmons D, Lennox H, Krautkramer SR, Gentile LN, McIntosh LP, Hagman J: **The highly conserved beta-hairpin of the paired DNA-binding domain is required for assembly of Pax-Ets ternary complexes**. *Mol Cell Biol*, **19**(3):2231-2241, 1999.

56. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res*, **22**(22):4673-4680, 1994.

57. Jonsson J, Eriksson L, Hellberg S, Lindgren F, Sjostrom M, Wold S: **A Multivariate Representation and Analysis of DNA-Sequence Data.** *Acta Chemica Scandinavica*, 45: 186-192, 1991.

58. Compumine: **Rule Discovery System (RDS) 1.0**. Stockholm: 2004.

59. Boström H: **Covering vs. Divide-and Conquer for Top-Down Induction of Logic Programs**. In: *Fourteenth International Joint Conference on Artificial Intelligence.* Morgan Kaufman, 1194-1200, 1995.

60. Gröger H, Callaerts P, Gehring WJ, Schmid V: **Gene duplication and recruitment of a specific tropomyosin into striated muscle cells in the jellyfish *Podocoryne carnea***. *J Exp Zool*, 285(4):378-386, 1999.