Extracting Sequence Features to Predict Protein–DNA Interactions: A Comparative Study

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td><a href="http://dx.doi.org/10.1093/nar/gkn361">http://dx.doi.org/10.1093/nar/gkn361</a></td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:2757494">http://nrs.harvard.edu/urn-3:HUL.InstRepos:2757494</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#0AP">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#0AP</a></td>
</tr>
</tbody>
</table>
Extracting sequence features to predict protein–DNA interactions: a comparative study

Qing Zhou1,* and Jun S. Liu2

1Department of Statistics, University of California, Los Angeles, CA 90095 and 2Department of Statistics, Harvard University, Cambridge, MA 02138, USA

ABSTRACT

Predicting how and where proteins, especially transcription factors (TFs), interact with DNA is an important problem in biology. We present here a systematic study of predictive modeling approaches to the TF–DNA binding problem, which have been frequently shown to be more efficient than those methods only based on position-specific weight matrices (PWMs). In these approaches, a statistical relationship between genomic sequences and gene expression or ChIP-binding intensities is inferred through a regression framework; and influential sequence features are identified by variable selection. We examine a few state-of-the-art learning methods including stepwise linear regression, multivariate adaptive regression splines, neural networks, support vector machines, boosting and Bayesian additive regression trees (BART). These methods are applied to both simulated datasets and two whole-genome ChIP-chip datasets on the TFs Oct4 and Sox2, respectively, in human embryonic stem cells. We find that, with proper learning methods, predictive modeling approaches can significantly improve the predictive power and identify more biologically interesting features, such as TF–TF interactions, than the PWM approach. In particular, BART and boosting show the best and the most robust overall performance among all the methods.

INTRODUCTION

Transcription factors (TFs) regulate the expression of target genes by binding in a sequence-specific manner to various binding sites located in the promoter regions of these genes. A widely used model for characterizing the common sequence pattern of a set of TF-binding sites (TFBSs), often referred to as a motif, is the position-specific weight matrix (PWM). It assumes that each position of a binding site is generated by a multinomial probability distribution independent of other positions. Since 1980s, many computational approaches have been developed based on the PWM representation to ‘discover’ motifs and TFBSs from a set of DNA sequences (1–6). See ref. (7,8) for recent reviews. From a discriminant modeling perspective, a PWM implies a linear additive model for the TF–DNA interaction. Since non-negligible dependence among the positions of a binding site can be present (9,10), methods that simultaneously infer such dependence and predict novel binding sites have been developed (11–13). Approaches that make use of information in both positive (binding sites) and negative sequences (non-binding sites) have also been developed (14–16). In addition, since a TF often cooperates with other TFs to bind synergistically to a cis-regulatory module (CRM), CRM-based models have been proposed to enhance the accuracy in predicting TFBSs (17–23).

Although predictive accuracies of these PWM-based methods for TFBSs and CRMs are still not fully satisfactory, statistical models being employed are already quite intricate. It is extremely difficult to build more complicated generative models that are both scientifically and statistically sound. First, the data used to estimate model parameters are limited to only several to several tens of known binding sites. With this little information, it is hardly feasible to fit a complicated generative model that is useful for prediction. Second, the detailed mechanism of TF–DNA interaction, which is likely gene-dependent, has not been understood well enough so as to suggest faithful quantitative models. For example, it is well-known that nucleosome occupancy and histone modifications play important roles in gene regulation in eukaryotes, but it is not clear how to incorporate them into a TF–DNA binding model. The predictive modeling approach described in this article seems to provide a different angle to account for such complications.

Recently, the abundance of ChIP-based TF-binding data (ChIP-chip/seq) and gene-expression data has brought up the possibility of building flexible predictive models for capturing sequence features relevant to TF–DNA interactions. In particular, ChIP-based data not
only provide hundreds or even thousands of high resolution TF-binding regions, but also give quantitative measures of the binding activity (ChIP-enrichment) for such regions. Treating gene expression or ChIP-chip intensity values as response variables and a set of candidate motifs (in the form of PWMs) and/or other sequence features as potential predictors, predictive modeling approaches use regression-type statistical learning methods to train a discriminative model for prediction. In contrast to those PWM-based generative models constructed from biophysics heuristics (24–26), predictive modeling approaches aim to learn from the data a flexible model to approximate the conditional distribution of the response variable given the potential predictors. In doing so, they also pick up relevant sequence features.

An attractive feature of the predictive modeling approach is its simple conceptual framework that connects genes’ ‘behavior’ (i.e. expression) with their (promoter regions’) sequence characteristics, thus effectively using both positive and negative information. In addition, a predictive model can avoid overfitting and be self-validated via a proper cross-validation procedure instead of solely relying on experimental verifications. This is especially useful in studying biological systems, since specific model assumptions and many rate constants are often difficult to validate due to the complexity of the problem. Instead of building a conglomerate of many intricate methods to predict global transcription regulation networks, we focus here on a more humble goal: to understand the general framework of predictive modeling and to provide some insights on the use of different machine learning tools.

Although several predictive modeling approaches have been developed in the past few years (27–31), there is still a lack of formal framework for and a systematic comparison of different yet very much related approaches. In this article, we formalize the predictive modeling approach for TF–DNA binding and examine a few contemporary statistical learning methods for their power in expression/ChIP-intensity prediction and in selection of relevant sequence features. The methods we examine and compare include stepwise linear regression, neural networks, multivariate adaptive regression splines (MARS) (32), support vector machines (SVM) (33), boosting (34), and Bayesian additive regression trees (BART) (35). A special attention is paid to the Bayesian learning strategy BART, which, to the best of our knowledge, has never been used to study TF–DNA binding, but shows the best overall performance.

**MATERIALS AND METHODS**

A basic assumption of all predictive modeling approaches is that sequence features in a certain genomic region influence the target response measurement. This is in principle true for many biological measurements. For ChIP-chip data, the enrichment value can be viewed as a surrogate of the binding affinity of the TF to the corresponding DNA segment. Influential sequence features may include motifs of the target TF and its co-regulators, genomic codes for histone modifications or chromatin remodeling and so on. For expression data, TF–DNA binding affinity, which influences the expression of a target gene, is determined by the genomic sequence surrounding the binding site.

**A general framework**

The input data for fitting a predictive model are a set of n DNA sequences \{S_1, S_2, \ldots, S_n\}, each with a corresponding response measurement \(y_i\), which may be mRNA expression values, ChIP-chip fold changes (often in the logarithmic scale), or categorical (e.g. active versus inactive, in versus out of a gene cluster). We write \([y_i, S_i]\), for \(i = 1, 2, \ldots, n\).

By feature extraction (next section), which is perhaps the most important but often lightly treated step, each sequence \(S_i\) is transformed into a multi-dimensional data point \(x_i = [x_{i1}, \ldots, x_{ip}]\) including, for example, the matching strength of a certain motif, the total energy of a certain periodic signal, etc. Then, a statistical learning method is applied to ‘learn’ the underlying relationship \(y=f(x)\) between \(x\) and \(y\) and to identify influential features. In comparison to the standard statistical learning problem, a novel feature of the problem described here is that the covariates are not given a priori, but need to be ‘composed’ from the observed sequences by the researcher.

When the response \(Y\) is the observed log-ChIP-intensity of a transcription factor \(P\) to a DNA sequence \(S\) (not restricted to the short binding site), the predictive modeling framework can be derived from a chemical physics perspective of the biochemical reaction \(P+S=PS\), where \(PS\) is the TF–DNA complex. At temperature \(T\), the equilibrium association constant is

\[
K_\text{eq}(S) = \frac{[PS]}{[P][S]} = \exp(-\Delta G(S)/RT),
\]

where \(\Delta G(S)\) is the Gibbs free energy of this reaction for the sequence \(S\). The log-enrichment of the TF–DNA complex \(PS\) can be expressed as

\[
\log \frac{[PS]}{[S]} = \log[P] - \Delta G(S)/RT \equiv f(S).
\]

Suppose that \(f(S)\) can be written as a function of the extracted sequence features \(X = [X_1, \ldots, X_p]\) and that the observed ChIP-intensity gives a noisy measure of the enrichment of the TF–DNA complex with an additive error \(\epsilon\) in the logarithmic scale. Then, Equation (1) becomes

\[
Y = f(X) + \epsilon,
\]

which serves as the basis for our statistical learning framework.

Many learning methods such as MARS, neural networks, SVM, boosting and BART, to be reviewed later, are composed of a set of simpler units (such as a set of ‘weak learners’), which make them flexible enough to approximate almost any complex relationship between responses and covariates. However, due to the nature of their basic learning units, these methods differ in their sensitivity, tolerance on nonlinearity and ways of coping with overfitting. As shown in later sections, BART and boosting are particularly attractive for our task due to their use of weak learners as the basic learning units.
Feature extraction

The goal of this step is to transform the sequence data into vectors of numerical values. This is often the most critical step that determines whether the method will be ultimately successful for a real problem. In ref. (27,28), the extracted features are k-mer occurrences, whereas in ref (29–31), are motif matching scores (which may differ depending on how one scores a motif match) for both experimentally and computationally discovered motifs. In ref. (36), features include both motif scores and histone modification data; and in ref. (37), a periodicity measure is further added to the feature list. However, a general framework and a reasonable criterion for comparing different feature extraction approaches are still lacking.

Here, we extract three categories of sequence features from a repeat-masked DNA sequence: the generic, the background and the motif features. Generic features include the GC content, the average conservation score of a sequence and the sequence length. The length of a sequence is defined as the number of nucleotides after masking out repetitive elements. It is included in our model to control for a potential confounding effect on experimental measures.

Potential confounding effects on repetitive elements. It is included in our model to control for this effect on experimental measures. As shown in our analysis, a periodicity measure is further added to the feature list. However, a general framework and a reasonable criterion for comparing different feature extraction approaches are still lacking.

Background and the motif features. Generic features include the GC content, the average conservation score of a sequence and the sequence length. The length of a sequence is defined as the number of nucleotides after masking out repetitive elements. It is included in our model to control for a potential confounding effect on experimental measures.

Note that the single nucleotide frequency (k = 1) is equivalent to the GC content.

Motif features of a DNA sequence are derived from a precompiled set of TF-binding motifs, each represented by a PWM. The compiled set includes both known motifs from TF databases and new motifs found from the positive ChIP sequences in the data set of interest using a de novo motif search tool. We fit a heterogeneous (i.e. segmented) Markov background model for a sequence to account for the heterogeneous nature of genomic sequences. Intuitively, this model assumes that the sequence in consideration can be segmented into an unknown number of pieces and, within each piece, the nucleotides follow a homogeneous first-order Markov chain (39). Using a Bayesian formulation and an MCMC algorithm, we estimate the background transition probability of each nucleotide by averaging over all possible segmentations with respect to their posterior probabilities. Suppose that the current sequence is \( S = R_1 R_2 \cdots R_w \), the PWM of a motif of width \( w \) is \( \Theta = \Theta(j) \) (\( i = 1, \ldots, w, j = A, C, G, T \)), and the background transition probability of \( R_i \) given \( R_{i-1} \) is \( \theta_0(R_i|R_{i-1}) \) (\( 1 \leq i \leq L \)) in the estimated heterogeneous background model. For each \( w \)-mer in \( S \) (both strands), say \( R_i \cdots R_{i+w-1} \), we calculate a probability ratio

\[
r_i = \frac{\Theta(R_{i+1})}{\theta_0(R_{i+1}|R_{i-1})},
\]

and define the motif score for this sequence as \( \log(\sum_{i=1}^{m} r_i/L) \), where \( r_i \) is the \( i \)-th ratio in descending order. In this article, we take \( m = 25 \). This value was chosen according to a pilot study on the Oct4 ChIP-chip data set see Results section, for which we observed optimal and almost identical discriminant power for motif scores defined by \( m \geq 20 \) (data not shown).

Statistical learning methods

Let \( Y \) be the response variable and \( X = [X_1, \ldots, X_p] \) its feature vector. The main goal of statistical learning is to ‘learn’ the conditional distribution \( [Y | X] \) from the training data \( \{(y_i, x_i) | i = 1, \ldots, n\} \). We consider the general learning model (2), where \( Y \) is a continuous variable and \( \epsilon \) is the observational error (e.g. Gaussian noise). Dependent on the posited functional form for \( f(X) \) and the method to approximate it, different learning methods have been developed.

Linear regression. This classic approach assumes that \( f(X) \) is linear in \( X \), i.e.

\[
f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j.
\]

The standard estimate of the \( \beta \)’s, which is also the maximum likelihood estimate under the Gaussian error assumption, is obtained by minimizing the sum of squared errors between the fitted and the observed \( Y \)’s. When it is suspected that not all the features are needed in Equation (5), a stepwise approach (called stepwise linear regression) is often used to select features according to a model comparison criterion, such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). With an initial linear regression model, this approach iteratively adds or removes a feature according to the model comparison criterion, until no further improvement can be obtained. Another way to select features is achieved by adding a penalty term, often in the form of the \( L^1 \) norm of the fitted coefficients, to the sum of squared errors and then minimizing this modified error function (40). This last approach has become a recent research hotspot in statistics. In this study, we employ AIC-based stepwise approaches for feature selection in linear regression methods.

Neural networks. We focus here on the most widely used single hidden layer feed-forward network, which can be viewed as a two-stage nonlinear regression. The hidden layer contains \( M \) intermediate variables \( Z = [Z_1, \ldots, Z_M] \) (hidden nodes), which are created from linear combinations of the input features. The response \( Y \) is modeled as a linear combination of the \( M \) hidden nodes,

\[
Z_m = s(\alpha_m + X \beta_m), \quad m = 1, \ldots, M, \quad f(X) = \beta_0 + Z \beta.
\]
where \( \mathbf{x}_m \) and \( \mathbf{\beta} \) are \( p \)-dimensional column vectors and \( s(v) = 1/(1+e^{-v}) \) is called the ‘activation function’. The unknown parameters \( \mathbf{\theta} = (\{\alpha_{m0}, \mathbf{\alpha}_m\}_1^M, \mathbf{\beta}_0, \mathbf{\beta}) \) are estimated by minimizing the sum of squared errors, \( R(\mathbf{\theta}) = \sum_i(y_i - f(x_i))^2 \), via the gradient descent method (i.e. moving along the direction with the steepest descent), also called back-propagation in this setting. In order to avoid overfitting, a penalty is often added to the loss function: \( R(\mathbf{\theta}) + \lambda \| \mathbf{\theta} \|^2 \), where \( \| \mathbf{\theta} \|^2 \) is the sum of squares of all the parameters. This modified loss function is what we use in this work for training neural networks. Here, \( \lambda \) is referred to as the weight decay.

**SVM.** Suppose each data point in consideration belongs to one of the two classes, the SVM aims to find a boundary in the \( p \)-dimensional space that not only separates the two classes of data points but also maximizes its margin of separation. The method can also be adapted to deal with regression problems, in which case the prediction function is represented by

\[
f(\mathbf{X}) = \mathbf{\alpha}_0 + \sum_{i=1}^n \mathbf{\alpha}_i K(\mathbf{X}, \mathbf{x}_i),
\]

where \( K \) is a chosen kernel function (e.g. polynomial or radial kernels). The problem of deciding the optimal boundary is equivalent to minimizing

\[
\sum_{i=1}^n V(y_i - f(\mathbf{x}_i)) + \frac{C}{2} \sum_{i,j} \mathbf{\alpha}_i \mathbf{\alpha}_j K(\mathbf{x}_i, \mathbf{x}_j),
\]

where \( V(\cdot) \) is an ‘\( \varepsilon \)-insensitive’ error measure and \( C \) is a positive constant that can be viewed as a penalty to overly complex boundaries (40). The computation of SVM is carried out by convex quadratic programming. The training data whose \( \mathbf{\alpha}_i \neq 0 \) in the solution (6) are called support vectors.

**Additive models: MARS, boosting and BART.** Another large class of learning methods, based on additive models, approximates \( f(\mathbf{X}) \) by a summation of many non-linear basis functions, i.e.

\[
f(\mathbf{X}) = \mathbf{\beta}_0 + \sum_{m=1}^M \mathbf{\beta}_m g_m(\mathbf{X}; \gamma_m),
\]

where \( g_m(\mathbf{X}; \gamma_m) \) is a basis function with parameter \( \gamma_m \). Different forms of basis functions with different ways to select and estimate them give rise to various learning methods, including MARS (32), boosting (34) and BART (35) among others. In some sense, neural networks and SVM can also be formulated as additive models. In MARS, the collection of basis functions is

\[
C = \{(X_j - t)^+ + (t - X_j)^+, \ 0 < t \leq \max(X_j), \ i = 1, \ldots, n, \ j = 1, \ldots, p\},
\]

where \( ^+ \) means positive part: \( x^+ = x \) if \( x > 0 \) and \( x^+ = 0 \) otherwise. These basis functions are called linear splines. The \( g_m(\mathbf{X}; \gamma_m) \) in MARS can be a function in \( C \) or a product of up to \( d \) such functions. The learning of MARS is performed by forward addition and background deletion of basis functions to minimize a penalized least square loss, which incurs a penalty of \( \lambda \) per additional degree of freedom of the model. Consequently, \( \lambda \) and \( d \) determine the model complexity of MARS.

Regression trees are widely used as basic functions for boosting and BART. Let \( T \) denote a regression tree with a set of interior and terminal nodes. Each interior node is associated with a binary decision rule based on one feature, typically of the form \( \{X_j \leq a\} \) or \( \{X_j > a\} \) for \( 1 \leq j \leq p \). Suppose the number of terminal nodes is \( B \). Then the tree partitions the feature space into \( B \) disjoint regions, each associated with a parameter \( \mu_b \) \((b = 1, \ldots, B)\) (Figure 1A). According to this tree and its associated parameters represents a piece-wise constant function with \( B \) distinct pieces (Figure 1B). We denote this function by \( g(\mathbf{X}; T, \mu) \), where \( \mu = [\mu_1, \ldots, \mu_B] \). Then, the additive regression tree model approximates \( f(\mathbf{X}) \) by the sum of \( M \) such regression trees:

\[
f(\mathbf{X}) = \mathbf{\beta}_0 + \sum_{m=1}^M g(\mathbf{X}; T_m, \mu_m),
\]
in which each tree \( T_m \) is associated with a parameter vector \( \mu_m (m = 1, \ldots, M) \). The number of trees \( M \) controls the model complexity, and it is usually large (100–200), which makes the model flexible enough to approximate the underlying relationship between \( Y \) and \( X \).

In the original AdaBoost algorithm (34), an additive model is produced via iteratively re-weighting the training data according to the training error of the current classifier. Later, a more general framework was developed to formulate boosting algorithms as a method of function gradient descent (41), which is what we use in this work. For a gradient boosting machine under square loss, given a current additive model \( f^{(m-1)}(X) = \beta_0 + \sum_{k=1}^{m-1} g(X; T_k, \mu_k) \), a regression tree \( g(X; T_m, \mu_m) \) is fitted to the current residuals and then the additive model is updated to

\[
f^{(m)}(X) = f^{(m-1)}(X) + \nu \cdot g(X; T_m, \mu_m),
\]

where \( 0 < \nu \leq 1 \) is the shrinkage parameter (learning rate).

Applying this recursion for \( M \) iterations gives the boosting additive tree model as in Equation (7).

BART completes Bayesian inference on model (7) with prescribed prior distributions for both the tree structure and the associated parameters, \( \beta_0, \mu_m \) and the variance \( \sigma^2 \) of Gaussian noise \( \varepsilon \) (2). The prior distribution for the tree structure is specified conservatively so that the size of each tree is kept small, which forces it to be a weak learner. The priors on \( \mu_m \) and \( \sigma^2 \) also contribute to alleviating over-fitting. A Markov chain Monte Carlo method (BART MCMC) was developed (35) for sampling from the posterior distribution \( P((T_m, \mu_m)^M_{m=1}, \beta_0, \sigma^2 | \{(y_i, x_i)\}^n_{i=1}) \), in which both the tree structures and the parameters are updated simultaneously. Given a new feature vector \( x^* \), BART predicts the response \( y^* \) by the average response of all sampled additive trees instead of using only the best model. In this sense, BART has the nature of Bayesian model averaging.

RESULTS

We applied the predictive modeling approach outlined above to recently published ChIP-chip data sets of the TFs Oct4 and Sox2 in human embryonic stem cells (ESC) (42), with a comparative study on various mentioned learning methods using 10-fold cross-validations (CVs). We discovered consistently a Sox–Oct composite motif from both the Oct4 and the Sox2 ChIP-chip data sets using a de novo motif search algorithm. This motif is known to be recognized by the protein complex of Oct4 and Sox2, the target TFs in the ChIP-chip experiments, and thus we included it in our precompiled motif set. In addition, we extracted 223 motifs from TRANSFAC release 9.0 (43) and literature survey to compile a final list of 224 PWMs. Please see Supplementary Text for the details. The sequence sets, ChIP-enrichment values and feature matrices are available at http://www.stat.ucla.edu/~zhou/ChIPLearn/. We also conducted a simulation study to further verify the performances of the methods.

The Oct4 ChIP-chip data in human ESCs

Boyer et al. (42) reported 603 Oct4-ChIP-enriched regions (positives) in human ESCs with DNA microarrays covering −8 kb to +2 kb of ~17000 annotated human genes. We randomly selected another 603 regions with the same length distribution from the genomic regions targeted by the DNA microarrays (negatives), i.e. [−8, +2] kb of the annotated genes. The ChIP-intensity measure defined as the average array-intensity ratio of ChIP samples over control samples was attached to each of the 1206 ChIP-regions. We treated the logarithm of the ChIP-intensity measure as the response variable \( Y \) and the features extracted from the genomic sequences as the explanatory variables \( X \). This produced a data set of 1206 observations with 269 features (224 PWMs, 3 genetic features, 10 dimers, 32 trimers: see Equation 3), called the Oct4 data set henceforth.

CVs. We compared the following statistical learning algorithms via a 10-fold CV procedure on the Oct4 data set: (i) LR-SO/LR-Full, linear regression using the Sox–Oct composite motif only or using all the 269 features; (ii) Step-SO/Step-Full, stepwise linear regression starting from LR-SO or starting from LR-Full; (iii) NN-SO/NN-Full, neural networks with the Sox–Oct composite motif feature or all the features as input; (iv) Step + NN, neural networks with the features selected by Step-SO as input; (v) MARS using all the features; (vi) SVM for regression with various kernels; (vii) Boost, boosting with regression trees as base learners and (viii) BART with different number of trees.

The observations were divided randomly into 10 subgroups of equal size. Each subgroup (called ‘the test sample’) was left out in turn and the remaining nine subgroups (called ‘the training sample’) were used to train a model using one of the above methods. Then the observed responses of the test sample were compared to those predicted by the trained model. We used the correlation coefficient between the predicted and the observed responses as a measure of the predictive power of a tested method. This measure is invariant under linear transformation, and can be understood as the fraction of variation in the response variable attributable to the explanatory features. We call this measure the CV-correlation (CV-cor) henceforth. Treating the LR-SO method as the baseline for our comparison, which only involves the single target motif, we are interested in testing whether other genomic sequence features can significantly influence the prediction of the ChIP enrichment, which can be viewed as a surrogate of the TF–DNA binding affinity. The predictive modeling approach employed here facilitates us a coherent framework to identify useful features, which may lead to testable hypotheses that enhance our understanding of the protein–DNA interaction.

As reviewed in the MATERIALS AND METHODS section and listed in Table 1, sophisticated learning methods have their respective tuning parameters. We conducted CVs for a wide range of these tuning parameters and identified the optimal value in terms of CV-cor. A proper common practice in the field is to further divide the samples to tune these parameters. But, here we chose not to do so since
we were also interested in comparing the robustness of different methods with different tuning parameters. The detailed description and results of the CVs, including the way we chose the tuning parameters, are given in the Supplementary Text.

The overall comparison results are briefly summarized in Table 1. The average CV-cor (over 10 test samples) of LR-SO was 0.446. Among all the linear regression methods, Step-SO achieved the highest CV-cor of 0.535. NN-SO and Step + NN showed a slight improvement over LR-SO after tuning weight decay and the number of hidden nodes, while the performance of NN-Full (optimal CV-cor < 0.38) was unsatisfactory. SVM showed 23% improvement over LR-SO and it was quite robust with stable CV-cor for C ≥ 1. The optimal results of MARS, boosting and BART were all substantially superior to the other methods. However, they had different degrees of robustness to their respective tuning parameters. The CV-cor of MARS could drop below 0.46 if one did not choose a good value for its penalty cost λ, while the other two methods were much more robust. For example, the CV-cor of BART with all the tested numbers of trees (M) ranging from 20 to 200 varied only between 0.592 and 0.6, which uniformly outperformed the optimal tuned results of all the other methods. This suggests that in practice the users may not need to worry too much about the tuning parameters in BART and boosting, but apply these methods with the default settings comfortably to their data sets. This is a significant advantage over some other learning methods such as NN and MARS. The SD in CV-cor across the 10 test samples for all the methods were quite comparable (around 0.04–0.05) except that the LR-Full and NN methods showed larger variability.

A common method for predicting whether a particular DNA sequence can be bound by a TF is to score the sequence by the PWM of the TF. This is equivalent to using LR-SO for the binding prediction. Thus, our study demonstrated that sophisticated statistical learning tools such as MARS, SVMs, boosting and BART can all significantly improve the basic LR-SO prediction by including more sequence features.

**Table 1. Ten-fold CVs of the Oct4 ChIP-chip data**

<table>
<thead>
<tr>
<th>Method</th>
<th>Tuning parameters</th>
<th>CV-cor</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR-SO</td>
<td></td>
<td>0.446 (0%)</td>
<td>0.044</td>
</tr>
<tr>
<td>LR-Full</td>
<td></td>
<td>0.491 (10%)</td>
<td>0.064</td>
</tr>
<tr>
<td>Step-SO</td>
<td></td>
<td>0.535 (20%)</td>
<td>0.045</td>
</tr>
<tr>
<td>Step-Full</td>
<td></td>
<td>0.513 (15%)</td>
<td>0.054</td>
</tr>
<tr>
<td>NN-SO</td>
<td># of nodes, weight decay</td>
<td>0.468 (5%)</td>
<td>0.063</td>
</tr>
<tr>
<td>Step + NN</td>
<td># of nodes, weight decay</td>
<td>0.463 (4%)</td>
<td>0.067</td>
</tr>
<tr>
<td>MARS</td>
<td>Interaction d, penalty λ</td>
<td>0.580 (30%)</td>
<td>0.043</td>
</tr>
<tr>
<td>SVM</td>
<td>Cost C</td>
<td>0.547 (23%)</td>
<td>0.054</td>
</tr>
<tr>
<td>Boost</td>
<td># of trees M</td>
<td>0.581 (30%)</td>
<td>0.048</td>
</tr>
<tr>
<td>BART</td>
<td># of trees M</td>
<td>0.600 (35%)</td>
<td>0.044</td>
</tr>
</tbody>
</table>

Reported here are the average CV-cors. The percentage in the parentheses is calculated by the percent of improvement over the CV-cor of LR-SO. SD is the standard deviation across 10 test samples. The detailed definitions of tuning parameters are reviewed in the MATERIALS AND METHODS section and further discussed in Supplementary Text.

**Figure 2.** The posterior inclusion probability $P_{in}$ of all the features in descending order in the BART model for the Oct4 data set.

**Sequence features selected by BART.** We chose BART with 100 trees to perform a detailed study on the full data set (with 1206 observations). The posterior mean tree size (of a base learner) measured by the number of terminal nodes was 2.31. Thus, each tree in the model was indeed quite simple. We define the posterior inclusion probability $P_{in}$ of a feature by the fraction of sampling iterations in which the feature is included in the additive trees. Figure 2 shows the $P_{in}$’s for all the 269 features in descending order, and 22 of them are >0.5 (Table 2). In addition, the average number of times a feature is used in each posterior sample of the BART model (a feature can be used a number of times by a BART model since the model is an aggregation of many regression trees), called the per-sample inclusion (PSI), is also reported in the table as another measure of the importance of the feature.

The top feature under both the overall and the PSI statistics is the Sox-Oct composite motif, consistent with the existing biological knowledge that Sox2 is one of the most important co-regulators of Oct4 and they form a protein complex to bind to the composite sites. Besides, we found another eight motifs with $P_{in} > 0.5$, among which OCT_Q6, OCT1_Q6 and OCT1_Q5_01 represent all the variants of the Oct4 motif in our compiled list. This implies the high sensitivity of our method in detecting functional motifs. The remaining five motifs, Hsf1, Uf1h3b, Nfy_Q6, E2F and E2F1, may be co-regulators of Oct4 or other functional TFs in ESCs. As reported recently (44), the TF Nfyb regulates ES proliferation in both mouse and human ESCs by binding specifically to the motif Nfy_Q6 in the promoter regions of ES-upregulated genes. Nfyb were also reported to co-activate genes with E2F via adjacent binding sites (45). The significant roles of both motifs in our model suggest that their co-activation of genes in ESCs may be recruited or enhanced by Oct4 binding. The motif Uf1h3b contains the consensus of the Klf4 motif (CCCCRCGCC) (46). The TF Klf4, together with Oct4, Sox2 and c_Myc, can reprogram highly differentiated somatic cells back to pluripotent ES-like cells (47). In a reported Oct4 enhancer (48) there is a highly conserved
site that matches the consensus of the Uf1h3b motif within 40 bp of a known Sox–Oct site. These external data confirmed the biological relevance in ESCs of at least six out of the nine motifs identified by the BART model (indicated in Table 2). It is interesting to note that the Hsf1 motif is not highly enriched in the positive ChIP regions (with small t-statistic), yet it has a very high P_in, indicating that it may interact with other factors in a nonlinear way. The sequence length was selected in the model as well, which served to balance out the potential bias in ChIP-intensity caused by the length difference of repeat elements in the original sequences.

A surprising yet interesting finding is the high inclusion probabilities of many nonmotif features, such as ‘GC’, ‘CAA’, ‘CCA’, ‘AA’ and ‘G/C’. This is also true for the learning results of other methods, such as stepwise linear regression and boosting, especially among the features with significant roles in these models (Figure 3). It is possible that some of these words are directly responsible for the interaction strength between Oct4 and its DNA-target region, such as ‘CAA’ and ‘AA’, which occur in the Oct4 motif consensus (ATG\text{CAA}AT). They may also contribute to the bending of the DNA sequences and thereby promote the assembly of elaborate protein–DNA interacting structures (49). To further verify their effect in predictive modeling, we excluded nonmotif features from the input, and applied BART (with 100 trees), MARS ($d=1$, $\lambda=6$), MARS ($d=2$, $\lambda=20$) and Step-SO to the resulting data set to perform 10-fold CVs. The parameters for MARS were chosen based on the CV results in Table 1 (also see Supplementary Text). The CV-corrs turned out to be 0.510, 0.511, 0.478 and 0.456 for the above four models, respectively, which decreased substantially (about 12–15%) compared to the CV-corrs of the corresponding methods with all the features.

Using this data set, we also compared the use of heterogeneous and homogeneous Markov background models for motif feature extraction. The heterogeneous background model is discussed in the Feature extraction section and is used for computing motif features in the above analysis. For the homogeneous background model, we used all the nucleotides in a sequence to build a first-order Markov chain. With these two different background models, we calculated motif scores for all the Oct4-family matrices in the 224 PWMs, i.e. the Sox–Oct composite motif, OCT1_Q6, OCT_Q6 and OCT1_Q5_01. We observed that, for all the four matrices, the motif scores under the heterogeneous background model showed higher correlations with the log-ChIP intensity than the scores under the homogeneous background model (Supplementary Table 1). We further computed the t-statistic for each motif score between the

---

### Table 2. The top 22 features in the BART model for the Oct4 ChIP-chip data

<table>
<thead>
<tr>
<th>Feature</th>
<th>$P_{in}$</th>
<th>PSI</th>
<th>$t$</th>
<th>Notes/Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>SoxOct\textsuperscript{a}</td>
<td>1.000</td>
<td>3.50</td>
<td>14.0</td>
<td>CWTTNWATGYAAAT</td>
</tr>
<tr>
<td>GC</td>
<td>1.000</td>
<td>1.56</td>
<td>2.7</td>
<td>Background</td>
</tr>
<tr>
<td>CAA</td>
<td>0.998</td>
<td>2.18</td>
<td>4.4</td>
<td>Background</td>
</tr>
<tr>
<td>CCA</td>
<td>0.987</td>
<td>1.45</td>
<td>–5.2</td>
<td>Sequence length</td>
</tr>
<tr>
<td>Length</td>
<td>0.964</td>
<td>1.37</td>
<td>7.8</td>
<td>Sequence length</td>
</tr>
<tr>
<td>HSF1\textsubscript{Q6}</td>
<td>0.963</td>
<td>1.32</td>
<td>1.2</td>
<td>TTCTRGAAVNTTCTYM</td>
</tr>
<tr>
<td>AA</td>
<td>0.945</td>
<td>1.41</td>
<td>2.7</td>
<td>Background</td>
</tr>
<tr>
<td>G/C</td>
<td>0.870</td>
<td>1.46</td>
<td>0.9</td>
<td>GC content</td>
</tr>
<tr>
<td>UF1H3b\textsubscript{Q6}\textsuperscript{a}</td>
<td>0.821</td>
<td>1.14</td>
<td>7.8</td>
<td>GCCCCWCCCCRCC</td>
</tr>
<tr>
<td>CA</td>
<td>0.760</td>
<td>1.19</td>
<td>–6.1</td>
<td>Background</td>
</tr>
<tr>
<td>CGC</td>
<td>0.723</td>
<td>1.04</td>
<td>4.6</td>
<td>Background</td>
</tr>
<tr>
<td>AAT</td>
<td>0.693</td>
<td>0.96</td>
<td>0.7</td>
<td>Background</td>
</tr>
<tr>
<td>cs</td>
<td>0.691</td>
<td>0.91</td>
<td>8.0</td>
<td>Conservation</td>
</tr>
<tr>
<td>OCT\textsubscript{Q6}</td>
<td>0.649</td>
<td>0.95</td>
<td>10.2</td>
<td>TNATTTCATN</td>
</tr>
<tr>
<td>NFY\textsubscript{Q6}\textsuperscript{a}</td>
<td>0.628</td>
<td>0.84</td>
<td>2.8</td>
<td>TRRCCAATSRN</td>
</tr>
<tr>
<td>CGA</td>
<td>0.622</td>
<td>0.87</td>
<td>7.7</td>
<td>Background</td>
</tr>
<tr>
<td>OCT1\textsubscript{Q6}</td>
<td>0.589</td>
<td>0.79</td>
<td>8.8</td>
<td>ATGCAAATNA</td>
</tr>
<tr>
<td>OCT1\textsubscript{Q5_01}\textsuperscript{a}</td>
<td>0.575</td>
<td>0.78</td>
<td>10.2</td>
<td>TNATTTCATW</td>
</tr>
<tr>
<td>GA</td>
<td>0.564</td>
<td>0.78</td>
<td>–0.5</td>
<td>Background</td>
</tr>
<tr>
<td>E2F\textsubscript{Q3_01}</td>
<td>0.540</td>
<td>0.75</td>
<td>8.6</td>
<td>TTTSGCCSG</td>
</tr>
<tr>
<td>GCA</td>
<td>0.528</td>
<td>0.70</td>
<td>2.5</td>
<td>Background</td>
</tr>
<tr>
<td>E2F\textsubscript{Q4_01}</td>
<td>0.512</td>
<td>0.68</td>
<td>8.8</td>
<td>TTTSGCCSG</td>
</tr>
</tbody>
</table>

The ‘t’ is the t-statistics between the positive and negative ChIP regions. \textsuperscript{a}Motifs with reported functions in ESCs.

---

![Figure 3.](image_url)
positive and negative ChIP-regions. Similarly, using the heterogeneous background model enhanced the separation between the positive and negative regions by resulting in larger t-statistics (Supplementary Table 1).

**Prediction and validation in mouse data.** The trained predictive models are useful computational tools to predict whether a piece of DNA sequence can be bound by the TF. They can be utilized to identify novel binding regions outside of the microarray coverage or not detected (false negatives) by the ChIP-chip experiment, and also to predict binding regions in other closely related species. As a proof of the concept, we applied the trained Oct4 BART and boosting models in the above human data to discriminate ~1000 Oct4-bound regions in mouse ESCs (50) from 2000 random upstream sequences with the identical length distribution. After extracting the same 269 features, we predicted ChIP-binding intensities of the 3000 sequences by the BART and boosting models, respectively. As a comparison, we also scanned each sequence to compute its average matching score to the Sox–Oct composite motif found by de novo search (called scanning method) for measuring the likelihood to be bound by the TF as suggested by many other studies (29). By gradually decreasing the threshold value, we obtained both higher sensitivity and more false positive counts for each method. We focused on the part where all the three methods predicted <500 random sequences as being bound by Oct4. As reported in Figure 4, both the BART and boosting models significantly reduced the number of false positives (random sequences) for almost all the sensitivity levels compared to the scanning method, which corresponded to an average of ~30% decrease in the false positive rate. Note that the Oct4 target genes identified by the ChIP-based assays are substantially different (with <10% of overlapping targets) between human and mouse (50). Thus, our result here represents an unbiased validation of the computational predictions. It also suggests that the binding pattern of Oct4 as characterized by our predictive models is very similar between the two species, even though its target genes may be quite different.

**The Sox2 ChIP-chip data in human ESCs**

We next applied our method to the 1165 Sox2 positive-ChIP regions identified by (42), accompanied by the same number of randomly selected regions from [−8, +2] kb of a gene with the same length distribution. Although known to co-regulate genes in the undifferentiated state, the target genes of Sox2 and Oct4 identified by ChIP-based experiments are substantially different (42). Besides testing our method, we are also interested in comparing the sequence features of the binding regions of these two regulators.

We extracted the same 269 features as in the previous subsection for each sequence in this Sox2 data set, and conducted 10-fold CVs to study the performances of the aforementioned statistical learning methods (more details in Supplementary Text). As shown in Table 3, BART with different tree numbers (CV-cors between 0.561 and 0.572) again outperformed all the other methods with optimal parameters, while boosting and MARS showed slightly worse but comparable performances. The improvement of these three methods over LR-SO, the baseline performance, was >54%. It is important to note that BART again performed very robustly while many other methods, such as NN, MARS and SVM, showed more variable performances with different choices of their tuning parameters (Supplementary Text). In addition, the SD in CV-corr for BART was the smallest among all the methods. We further applied Step-SO, MARS (d = 1, λ = 10) and BART (M = 100), with only motif features as input, and obtained CV-cors of 0.465, 0.469 and 0.500, respectively. One sees that all of them performed significantly worse than the corresponding methods using all the features. This confirms our hypothesis that nonmotif features are essential components of the underlying model for TF–DNA binding.

We re-applied BART with 100 trees to the full Sox2 data set with 2330 observations. There are 29 features with $P_m > 0.5$, including three generic features, 10 background frequencies and 16 motifs (Supplementary Table 2). As indicated in the Table, 7 out of the top 9 motifs were reported to be recognized by functional TFs in ESCs or differentiation. In agreement with the fact that the target TF is Sox2, the top three features include the Sox–Oct composite and the Sox2 motifs. Other potential motifs identified by BART that co-occur with Sox2 binding include Nanog ($P_m = 0.71$), Nkx2.5 ($P = 0.70$), U1Hb3b ($P = 0.67$), P53 ($P = 0.64$) and Gata-binding proteins ($P = 0.59$). Among them, Nanog is known as a crucial TF that co-regulates with Sox2 and Oct4, and P53 is a marker gene in ESCs (51,52). Interestingly, the Sox2 motif is overwhelmingly overrepresented ($p = 2 \times 10^{-12}$) in the Nanog-bound regions of up-regulated genes in mouse ESCs (53). These analyses strongly support the tight regulatory interaction
Table 1. The same notations and definitions of tuning parameters are used as in Table 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>CV-cor (SD)</th>
<th>Method</th>
<th>CV-cor (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR-SO</td>
<td>0.358 (0%)</td>
<td>LR-Full</td>
<td>0.494 (38%)</td>
</tr>
<tr>
<td>Step-SO</td>
<td>0.513 (43%)</td>
<td>Step-Full</td>
<td>0.509 (42%)</td>
</tr>
<tr>
<td>NN-SO</td>
<td>0.364 (2%)</td>
<td>Step+NN</td>
<td>0.465 (30%)</td>
</tr>
<tr>
<td>MARS</td>
<td>0.553 (54%)</td>
<td>SVM</td>
<td>0.526 (47%)</td>
</tr>
<tr>
<td>Boost</td>
<td>0.560 (56%)</td>
<td>BART</td>
<td>0.572 (60%)</td>
</tr>
</tbody>
</table>

The same notations and definitions of tuning parameters are used as in Table 1.

between Sox2 and Nanog in both human and mouse embryonic development. Nkx2.5 and Gata4 are known cooperative TFs in endoderm differentiation (54). Both of them are suppressed in ESCs but highly expressed once the cells start to differentiate into early endoderm. These results suggest a hypothesis for further investigation that genes repressed by Sox2 in ESCs may be activated later for endoderm development once Gata4 and Nkx2.5 are expressed and bind to their motifs in the Sox2 bound regions (Figure 5). Such competitive binding of Gata/Nkx2-families could be an efficient mechanism to accelerate the termination of Sox2-bound repression when differentiation is initiated.

Furthermore, we compared the BART models inferred from the Oct4 and the Sox2 data sets, and found that BART identified 22 and 29 features with \( P_{in} > \) 0.5, respectively. Among them, 10 features are in common, which is much higher than that expected by chance with a \( P \)-value of \( 1 \times 10^{-5} \) and a 4.2-fold enrichment. This is consistent with the known co-regulation of Oct4 and Sox2 in ES cells. A notable common motif feature with high \( P_{in} \) in both models is the Uf1h3b motif, which contains the consensus of Klf4 whose motif has not been included in TRANSFAC yet. This result predicted Klf4 as a common co-regulator of Oct4 and Sox2 in ESCs. The ChIP-chip data of Plath and colleagues (Sridharan et al., manuscript in preparation) provided an experimental validation of this prediction, in which the respective target genes of Oct4, Sox2, Klf4 and other important ESC regulators were identified. Among the \( \sim 1400 \) Klf4 target genes, 500 and 335 of them were also bound by Oct4 and Sox2, respectively. Furthermore, the actual binding regions shared at least 650 bp for \( >75\% \) of the common targets. These experimental data confirmed the co-regulation between Klf4 and the other two master TFs. On the other hand, sequence features specific to each data set provide a basis to distinguish the binding patterns of Oct4 and Sox2. For example, Nfy was only identified with high \( P_{in} \) in the Oct4 data set, whereas Nanog only co-occurred in the Sox2 bound regions. The missing of Nanog in the Oct4 BART model is consistent with an independent observation that the Nanog motif is not enriched in the Oct4-bound enhancers in mouse ESCs (53), although significant overlaps in target genes of these two TFs were reported (42,50). One possible explanation, which awaits future experimental investigations, is that the direct DNA binding of Nanog may depend on its interactive TFs. In the presence of Oct4, Nanog may not bind to DNA directly but co-regulate genes with Oct4 via protein–protein interactions (55).

Figure 5. A hypothesis of competitive binding between Sox2 and Gata4/Nkx2.5. In undifferentiated ES cells, Sox2 binds to a regulatory sequence (bracket region) to repress a target gene, while Gata4 and Nkx2.5 are not expressed. Later upon differentiation, Gata4 and Nkx2.5, both highly expressed, out-compete Sox2 to bind to the same region, thus terminating the repression of the downstream gene.

A simulation study

We performed a simulation study as a final test on the effectiveness of the predictive modeling approach, which allows us to evaluate its accuracy in identifying true sequence features, especially motifs that determine the observed measurements. We generated 1000 sequences, each from a first-order Markov chain, of length uniformly distributed between 800 and 1200. For each of the first 500 sequences, we inserted one, two or three Oct4 motif sites with probability 0.25, 0.5 or 0.25, respectively. Furthermore, we inserted one site for each of the three motifs, Sox2, Nanog and Nkx2.5, independently with probability 0.5. We calculated the score of an inserted site by Equation (4). For each motif we obtained the sum of the site scores for a sequence, denoted by \( Y_j \) over the noise variance \( \sigma_j^2 \).

Denote by \( X_j \) the GC content of a sequence. We normalized these five features by their respective standard deviations so that the rescaled features have a unit variance. Then, the observed ChIP-intensity \( Y \) for a sequence was simulated as:

\[
Y = f(X) + \varepsilon \\
= X_1(1 + 0.5X_2 + 0.3X_3 + 0.4X_4) + \sqrt{X_1X_3X_4 + 2X_3 + \varepsilon,}
\]

where \( \varepsilon \sim N(0, \sigma^2) \). This model states that \( X_1 \) is the primary TF with three interactive factors \( X_2, X_3, X_4 \) and the GC content \( X_3 \) has a positive effect on the ChIP-intensity. The signal-to-noise-ratio (SNR) of a simulated data set is defined as \( Var(f(X))/\sigma^2 \equiv Var(Y)/\sigma^2 - 1 \), which is the ratio of the variance of the true ChIP-intensity \( f(X) \) over the noise variance \( \sigma^2 \). We simulated 10 independent sequence sets, and then generated observed ChIP-intensities according to (8) with \( \text{SNR} = 1/0.6, 1/1 \) and 1/2, respectively.

We applied the same sequence feature extraction as in the previous sections to the simulated data sets and...
Table 4. Performance comparison on the simulated data sets

<table>
<thead>
<tr>
<th>Method</th>
<th>SNR</th>
<th>1/0.6</th>
<th>1/1</th>
<th>1/2</th>
</tr>
</thead>
<tbody>
<tr>
<td>EP-cor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step-LR</td>
<td></td>
<td>0.579 (0.009)</td>
<td>0.497 (0.012)</td>
<td>0.368 (0.010)</td>
</tr>
<tr>
<td>N_T</td>
<td>3.8 (0.42)</td>
<td>3.8 (0.42)</td>
<td>3.3 (0.67)</td>
<td></td>
</tr>
<tr>
<td>N_F</td>
<td>4.9 (2.64)</td>
<td>8.6 (3.63)</td>
<td>7.5 (3.54)</td>
<td></td>
</tr>
<tr>
<td>EP-cor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MARS</td>
<td></td>
<td>0.579 (0.011)</td>
<td>0.498 (0.013)</td>
<td>0.376 (0.018)</td>
</tr>
<tr>
<td>N_T</td>
<td>3.9 (0.32)</td>
<td>3.5 (0.53)</td>
<td>3.3 (0.48)</td>
<td></td>
</tr>
<tr>
<td>N_F</td>
<td>4.4 (1.78)</td>
<td>4.2 (1.81)</td>
<td>4.6 (2.80)</td>
<td></td>
</tr>
<tr>
<td>EP-cor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boost</td>
<td></td>
<td>0.603 (0.007)</td>
<td>0.528 (0.010)</td>
<td>0.405 (0.011)</td>
</tr>
<tr>
<td>N_T</td>
<td>3.9 (0.32)</td>
<td>3.7 (0.48)</td>
<td>3.5 (0.53)</td>
<td></td>
</tr>
<tr>
<td>N_F</td>
<td>0.6 (0.70)</td>
<td>2.4 (1.58)</td>
<td>5.9 (1.85)</td>
<td></td>
</tr>
<tr>
<td>EP-cor</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BART</td>
<td></td>
<td>0.636 (0.009)</td>
<td>0.551 (0.007)</td>
<td>0.416 (0.006)</td>
</tr>
<tr>
<td>N_T</td>
<td>4.0 (0.00)</td>
<td>3.6 (0.70)</td>
<td>3.5 (0.71)</td>
<td></td>
</tr>
<tr>
<td>N_F</td>
<td>1.8 (1.40)</td>
<td>2.5 (1.18)</td>
<td>2.5 (1.43)</td>
<td></td>
</tr>
</tbody>
</table>

Reported are the averages (standard deviations in the parentheses) over 10 independent data sets. N_T and N_F are the numbers of true and false motifs identified, respectively.

We calculated the correlation between a predicted response $Y$ and the true ChIP-intensity $f(X)$, denoted by $R(Y, f(X))$. Then the expected correlation between the predicted $\hat{Y}$ and a future observed response $Y = f(X) + \epsilon^*$, called the expected predictive correlation (EP-cor), was computed by

$$\frac{\text{Cov}(\hat{Y}, f(X) + \epsilon^*)}{\sqrt{\text{Var}(\hat{Y})[\text{Var}(f(X))]} = \frac{R(\hat{Y}, f(X)) \sqrt{\text{SNR}/(\text{SNR} + 1)}}{\sqrt{\text{SNR} + 1}},$$

where $\epsilon^* \sim \mathcal{N}(0, \sigma^2)$ is independent of $\hat{Y}$ and $f(X)$. When we set the threshold of $P_{\text{in}}$ to 0.7, BART identified on average >85% of the true motifs accompanied by at most 2.5 false positives. For stepwise linear regression (Step-LR), only features with a regression $P$-value < 0.01 were used for computing error rates in motif identification since including all the covariates selected by the method resulted in an overly large number of false positives. At comparable sensitivity levels ($N_T$), BART reported significantly fewer false positives ($N_F$) for all the SNR levels than Step-LR and MARS (Table 4). The relative influence of a feature selected by boosting was measured by the reduction of squared error attributable to the feature. We ranked all the selected features by their relative influence and set a reasonable threshold to obtain similar sensitivity levels as those of BART. We noticed that these two methods showed comparable error rates in motif identification for SNR ≥ 1, but the boosting method seemed to result in a higher false positive rate at the lowest SNR level (=1/2).

DISCUSSION

We have demonstrated in this article how predictive modeling approaches can reveal subtle sequence signals that may influence TF–DNA binding and generate testable hypotheses. Compared with some other more system-based approaches to gene regulation, such as building a large system of differential equations or inferring a complete Bayesian network, predictive modeling is more intuitive, more theoretically solid (as many in-depth statistical learning theories have been developed), more easily validated (by CV), and can generate more straightforward testable hypotheses.

Our main goal here is to conduct a comparative study on the effectiveness of several statistical learning tools for combining ChIP-chip/expression and genomic sequence data to tackle the protein–DNA binding problem. Because of the generality of these tools, we are able to include sequence features besides TF motifs, such as background words, GC content, and a measure of cross-species conservation. The finding that these nonmotif features can significantly improve the predictive power of all the tested methods indicates a potentially important yet less understood role these sequence features play in TF–DNA interactions. They may help the localization of a TF on the DNA for precise recognition of its binding sites, or may have a function with chromatin associated factors and histone modification activities. It is generally believed that many other factors in addition to the sequence specificity of short TFBSs contribute to TF–DNA binding. Along this direction, we have proposed a general framework to explore and characterize potentially influential factors.

In both ChIP-chip data sets, we not only unambiguously identified all the binding motifs for the target TFs, Oct4 and Sox2, but also discovered a number of verified cooperative or functional regulators in ESCs, such as Nanog, Klf4, Nfyb and P53. As a principled way to utilize both positive (i.e. binding sequences) and negative (non-binding) information, the predictive modeling approach provides a powerful alternative for detecting TF-binding motifs to those more popular generative model-based tools (3–6). Noting that the stepwise linear regression methods (Step-Full, Step-So, and Step-LR) are equivalent to MotifRegressor (29) and MARS is equivalent to MARSMotif (30) with all the known and discovered (Sox–Oct) motifs as input, we have shown that BART and boosting using all three categories of sequence features outperformed MotifRegressor and MARSMotif significantly in all of our examples.

For a generative modeling approach, separate statistical models are fitted to TF-bound (positive) and background (negative) sequences, and then discriminant analysis based on posterior odds ratio or likelihood ratio is applied to construct prediction rules. In contrast, a predictive modeling approach targets at prediction by modeling directly the condition distribution of TF-binding given extensively extracted sequence features. As shown in this article by both real and simulated data sets, modern statistical
learning tools such as boosting and BART have made it possible to estimate this conditional distribution quite accurately for the TF-binding problem. These two approaches have their own respective advantages. If the underlying data generation process is unclear or difficult to model, predictive approaches have the advantage to construct a nonparametric conditional distribution from the training data. On the other hand, generative models are usually built with more explicit assumptions that help us understand the underlying science and can capture key characteristics of a biological or physical system. Typical examples of generative models in gene-regulation problems include, for example, the mixture modeling of DNA sequence motifs (2) and the graphical model for protein–DNA interaction measured with ChIP-chip data (56).

Finally, our study suggests that the Bayesian learning method BART is a good tool for analyzing high-dimensional genomic data because of its high predictive power, its explicit quantification of uncertainty, and its interpretability. First, like boosting, BART is an ensemble learning method, which approximates an unknown relationship by an aggregation of a large number of simple models (small trees). Second, the Bayesian formulation of BART leads to not only the ‘optimal’ model, but also a posterior distribution on the space of all possible models, which can be used to predict the response of a new observation by weighted averaging predictions from all models. This model averaging approach tends to improve the model’s predictive power in general. Third, the variable selection procedure is a coherently built-in feature of BART and performs quite well in identifying important and relevant sequence features that contribute to TF–DNA interactions in all the examples. With the rapid accumulation of large-scale genomic data, we believe that flexible statistical learning methods such as BART and boosting will be very useful for studying a large class of biological problems including cis-regulatory analysis.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We appreciate the editor and two referees for helpful comments and suggestions on the manuscript. We thank the Plath lab at UCLA for sharing unpublished data to support some of the predictions in this work. This work was partially supported by NIH Grants R01-GM07899 and R01-GM080625-01A1, NSF and UCLA initiative grants. Funding to pay the Open Access publication charges for this article was provided by NIH and NSF grants.

Conflict of interest statement. None declared.

REFERENCES


