ICCM 2007 · Vol. IV 151–167

Predictive Modeling Approaches for Studying Protein-DNA Binding

Jun S. Liu* Qing Zhou†

Abstract

Understanding how and predicting where proteins interact with DNA are important problems in biology. Currently, computational methods for predicting binding sites of the proteins are mostly based on generative models in the form of position-specific weight matrices (also called "motifs"). We present here a systematic study of predictive modeling approaches to embryonic gene regulation. In such an approach, the genomic sequence information is combined with gene expression or other information regarding the biological system through sequence feature extraction and selection. Sequence features to be extracted include matching scores to existing TF binding motifs, frequencies of short words, certain periodic signals, a measure of cross-species conservation, etc. Feature selection is achieved by a statistical learning method that relates the gene expression values (or measures of other kind of biological properties) with some of the extracted sequence features.

Keywords and Phrases: Predictive modeling, statistical learning, motif discovery, gene regulation, transcription factor.

1 Introduction: central dogma of molecular biology

The complete information that defines the characteristics of a living cell within an organism is encoded in the form of a moderately simple molecule, deoxyribonucleic acid, or DNA. The building blocks of DNA are four nucleotides, abbreviated by their attached organic bases as A, C, G, and T. A-T and C-G are complimentary bases, between which hydrogen bonds can form. A DNA molecule consists of two long chains of nucleotides that are complimentary to each other and joined by hydrogen bonds twisted into a double helix. The specific ordering of these nucleotides, the "genetic code", is the means by which information is stored that completely defines all functions within a cell. With the recent development of very fast sequencing technology, genetic sequence databases such as GenBank have

*Department of Statistics, Harvard University, Cambridge, MA 02138, USA. Email: jliu@stat.harvard.edu

†Department of Statistics, University of California, Los Angeles, CA 90095, USA. Email: zhou@stat.ucla.edu

sustained an exponential growth rate since 1982, housing more than 65 billion DNA bases now.

Although all the cells in an organism contain the same DNA sequences, they display different physiological characteristics within different tissues, developmental stages, and environmental conditions. The central dogma of molecular biology dictates that DNA is transcribed into RNA, which serves as a transient template to make proteins, the basic building blocks of the cellular life. The collection of proteins synthesized in a particular cell state determines the cell's current physiological functions. If a protein is being synthesized at a certain state, its coding DNA (called a gene) is defined as being "active" or "expressed". Thus, a cell in a particular physiological state can be roughly viewed as a mechanical system in which each protein is either turned on (active) or turned off (inactive).

In many organisms, the DNA that codes for proteins (genes) is only a small portion of the total genomic DNA. The non-coding components of DNA, which were initially considered as "junk" sequences, contain the information for activating and deactivating the genes/proteins. Most of the control sequences for a gene lie in the *upstream regulatory region*, which is the couple of thousands base pairs long region directly before the gene (also called the transcription regulatory region - TRR, or the promoter). Transcribing or activating a gene requires not only the DNA sequence in the TRR, but also many proteins called transcription factors (TFs). TFs regulate their target genes' expression by binding in a sequence-specific manner to various binding sites located in the promotor regions of these genes. Several statistical models have been developed to characterize the common sequence pattern, often referred to as a TF binding motif (TFBM), of DNA sites bound by a TF. Most widely used is the position-specific weight matrix (PWM) model, which assumes that each position of a binding site is generated by a multinomial probability distribution independent of other positions. Many computational methods have been developed based on the PWM representation to "discover" motifs from a set of DNA sequences likely to be bound by a common TF [1, 2]. See [3, 4] for recent reviews.

Characterizing the TFBMs and predicting TF binding sites (TFBS) are crucial tasks for studying how the cell regulates its genes in response to developmental and environmental changes. With the availability of the complete genome sequences and high-throughput experimental techniques such as gene expression microarrays, it has become a reality to predict genome-wide TFBMs and TFBSs efficiently with the aid of bioinformatic tools, which can then lead to a deeper understanding on gene regulatory networks.

2 Statistical generative models for transcription regulation

Suppose we have n observed DNA sequences, $\mathbf{S}_1, \ldots, \mathbf{S}_n$, of lengths $L_1, \ldots, L_n$, respectively (the L's are typically in the range of hundreds to thousands). The basic generative motif finding model assumes that in these sequences there are segments of length W (typically in the range of 10 to 20), called "binding sites",

that are iid realizations from a probability distribution, such as a hidden Markov model or a product multinomial model, the latter often being called the position-specific weight matrix (PWM) model. Since both the model parameters and the locations of these binding sites are unknown, one can utilize the missing data formulation and estimate the binding sites either with the EM algorithm or with Markov chain Monte Carlo (see [2, 3, 27]).

From the discriminant modeling perspective, the PWM implies a linear additive model for TF-DNA interaction trained from the positive sequences (i.e., those containing binding sites) only. There are also approaches developed to make use of the information in both the positive, i.e., binding sites, and the negative sequences, i.e., non-binding sites (see [10]–[12]). Recently, several lines of experimental evidence have also demonstrated the existence of non-negligible dependence among the positions of a binding site (e.g., [5, 6]). Methods that simultaneously infer such dependence and discover novel binding sites have been developed, and were shown to outperform the PWM in both *de novo* motif discovery and site scan [7, 8, 9]. In addition, a TF often cooperates with other TFs to bind synergistically to regulatory regions. Such a region contains multiple TFBS's and is called a *cis*-regulatory module (CRM). Various models have been proposed for the CRM, including logistic regression [13], hidden Markov models [14]–[19], and a hierarchical mixture model [20].

Although it has been commonly acknowledged that all the models mentioned above are at best crude approximations to the underlying TF-DNA binding mechanism, it is extremely difficult to build more complex models that are both scientifically and statistically sound. First, the data used to infer a motif model usually contain only tens of known binding sites. With this little information, a complicated generative model can easily over-fit the data, rendering it useless for making predictions. Second, the detailed mechanism of TF-DNA interaction, which is likely gene-dependent, has not been fully understood, although some qualitative descriptions exist.

Recently, the development of chromatin immunoprecipitation followed by microarray (ChIP-chip) technology has enabled the scientist to obtain genome-wide binding regions for a TF under certain cellular conditions. The advantage of ChIP-chip data is that they 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. With abundance of such type of data, it is now hopeful that one may be able to build a more flexible model than weight matrices to capture sequence features that can be predictive of TF-DNA interactions.

Several *predictive modeling* (PM) approaches have been developed in recent years to study sequence motif features and gene expression or ChIP-chip data jointly, e.g., [21]–[25]. In contrast to many previous methods that directly build generative statistical models in the sequence space (e.g., [27, 17, 20]), PM approaches treat the gene expression or ChIP-intensity values as response variables, and regard a set of candidate sequence motifs (in the form of PWM) as potential predictors. In [23], a stepwise linear regression method was used to infer motif patterns that are of value to predict the response variable, whereas in [24], the

method of multiple adaptive regression splines (MARS, [28]) was used instead. In [25], an even more ambitious goal was attempted: predicting gene expression from sequences.

A distinctive advantage of the PM approach is that it provides a coherent framework to connect "behaviors" of genes (e.g., expression levels) with their composition (i.e., genomic sequence), thus effectively using both positive and negative information. In addition, a predictive model can be self-validated and avoid overfitting via a proper cross-validation procedure instead of relying on anecdotal biological evidences, which may be biased or inaccurate. This is especially useful in studying biological systems, since specific model assumptions are often not available due to the comlexity of the problem. For example, the concept of weight matrix motif has dominated computational *cis*-regulatory analyses. But it is also well known that the short motif sites by themselves are not specific enough to direct accurate TF recognition. For eukaryotes, nucleosome occupancy and histone modifications clearly play important roles in gene regulation. There is also evidence that sequence features other than TF binding motifs are important for both nucleosome occupancy and TF binding [32, 33]. As illustrated in this paper, the PM approach provides a useful tool for the researcher to explore in this direction.

3 A framework for predictive modeling

A basic assumption of all PM approaches is that certain sequence features influence the response measurement in either a linear or a nonlinear way. This is in principle true for many biological measurements. For example, 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 other than the binding motif of the target TF may correspond to binding motifs of co-factors, genomic codes for histone remodeling, and so on. Thus, the PM approach consists of two generic steps: *Step* 1, feature extractions and *Step* 2, feature selections.

The input data for fitting a predictive model are a set of DNA sequences, $\{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_n\}$, which correspond to potential binding regions, each with a corresponding response value, e.g., expression or ChIP-chip fold change value (in the logarithmic scale), y_i. We write $\{(y_i, \mathbf{S}_i)$, for $i = 1, 2, \ldots, n\}$. In *Step* 1, we map each $\mathbf{S}_i$ to a feature space composed of generic features, background word frequencies, a set of motif scores derived from both known motifs documented in biological databases and *de novo* motif finding software, and others such as certain periodic properties and structural properties. Thus, each sequence $\mathbf{S}_i$ is transformed into a multi-dimensional data vector representing p features: $\mathbf{x}_i = [x_{i1}, \ldots, x_{ip}]$. For example, in Conlon et al. [23], we used a fast *de novo* motif search method MDscan [26] to first generate a large set of putative TFBMs (typically in the range of tens to hundreds). Then each sequence $\mathbf{S}_i$ is scored against each putative motif M_j to get matching score x_{ij} (which can be intuitively understood as the number of matching sites in $\mathbf{S}_i$).

In *Step* 2, we apply a statistical learning method to infer the relation between

the response variables, i.e., to fit the model

$$y_i = f(\mathbf{x}_i) + \epsilon_i. \tag{3.1}$$

Or, if the y_i are categorical responses (such as binary), one can model their probabilities of belonging to a particular class, i.e., fitting a model of the form

$$P(y_i = k) = g(\mathbf{x}_i). \tag{3.2}$$

The most well-known "learning" model for (3.1) is the linear regression, whereas for (3.2) the logistic regression. Many other statistical learning methods, such as multivariate adaptive regressions (MARS), neural networks (NN), support vector machines (SVM), boosting, Bayesian additive regression trees (BART), etc., have been developed over the past few decades to counter the high-dimensionality and nonlinearity problems. At a conceptual levels, all of these methods are composed from a set of simpler units (such as a sum of 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" and the ways of combining these units, these methods have different sensitivies, tolerance on nonlinearity, and ways of coping with over-fitting. In the following sections, we first report an early successful study of histone modifications using the PM approach via a novel nonlinear dimension-reduction and regression method. We then move on to a comparative study of the utility of various advanced statistical learning tools, reporting their performances in the application to a human TF (Oct4) activity data, and in a simulation study.

4 RSIR and histone modification prediction

4.1 SIR and RSIR methods

Here we assume that the response variable y (i.e., gene expression values) is dependent of the gene's upstream features through a smooth function $f(\cdot)$ of k linear combinations of these features, i.e.,

$$y = f(\boldsymbol{\beta}_1^T\mathbf{x}, \boldsymbol{\beta}_2^T\mathbf{x}, \ldots, \boldsymbol{\beta}_k^T\mathbf{x}, \epsilon),$$

where k, β's, and $f(\ )$ are unknown. The expression-score data y are usually high-dimensional and noisy, so a direct fit of f using non-parametric methods is impractical. It is thus desirable to estimate the linear combinations without fitting f. This task can be accomplished by SIR [34], which was originally developed for dimension reduction and data visualization. After having obtained $\boldsymbol{\beta}_1^T\mathbf{x}, \boldsymbol{\beta}_2^T\mathbf{x}, \ldots, \boldsymbol{\beta}_k^T\mathbf{x}$, we can identify the influential individual xs with nonzero contributions to the linear combinations and their corresponding sequence features. However, since many of the $\mathbf{x}$'s are highly correlated, a direct application of the SIR method, which is equivalent to sequentially solving the eigen-value problems:

$$\arg\max_{\boldsymbol{\beta}^T\Sigma\boldsymbol{\beta}} \boldsymbol{\beta}^T M\boldsymbol{\beta}$$

where $\Sigma = Cov(\mathbf{x})$ and $M = Cov[E(\mathbf{x}|y)]$ (which is estimated by slicing the y), often results in highly variable solutions. In [35], we introduced the regularized SIR method, which is equivalent to solving

$$\arg \max_{\boldsymbol{\beta}^T(\Sigma+\epsilon I)\boldsymbol{\beta}} \boldsymbol{\beta}^T M \boldsymbol{\beta},$$

and can greatly reduce the estimation variability.

4.2 Histone modification data analysis via RSIR

Now we come to the biology part. Gene activities in eukaryotic cells are concertedly regulated by TFs and chromatin structure. The basic repeating unit of chromatin is the nucleosome, an octamer containing two copies each of four core histone proteins. While nucleosome occupancy in promoter regions typically occludes TF binding, thereby repressing global gene expression, the role of histone modification is more complex [36]–[38]. Histone tails can be modified in various ways, including acetylation, methylation, phosphorylation, and ubiquitination. Even the regulatory role of histone acetylation, the best characterized modification to date, is still not fully understood [39, 40].

Each of the four core histones contains several acetylable sites at their amino terminus tails. Genome-wide histone acetylation data from Saccharomyces cerevisiae [41, 42] have offered new opportunities for us to evaluate the regulatory effects of histone acetylation at these lysine sites. In particular, both H3 and H4 acetylation levels were found to be positively correlated with gene transcription rates. However, a subtle but important issue in analyzing such data is that effects of other potentially important factors not included in the analysis, generally termed as confounding factors, cannot be revealed by simple correlation plots. It is unclear, for example, how much regulatory information associated with histone acetylation is redundant with the genomic sequence information. To gain insights into this, we conducted a predictive modeling analysis by combining acetylation [41, 42, 44], nucleosome occupancy [46, 43, 42], gene upstream sequence information [45], and gene expression data [46]–[48] to investigate the effect of histone acetylation in the context of other regulatory factors in S. cerevisiae.

We analyzed two recent genome-wide histone acetylation datasets (more details in Yuan et al. [33]). Pokholok et al. [42] measured acetylation levels at three different sites, H3K9, H3K14, and H4, with the last referring to nonspecific acetylation on any of the four acetylable lysines on H4 tails. A typical analysis, when both histone acetylation data on a single site (for example, H3K9) and transcription rate data are available, is to simply correlate the two sets of measurements and to report the apparent significant statistical correlation between the two. When data on multiple acetylation sites are available, a slightly more formal analysis is to fit a linear regression model with multiple acetylation covariates. However, gene regulation is a complex process involving many contributing factors. Probably the best characterized factor for controlling gene transcription is the upstream sequence information. Although histone acetyltransferases (HATs) and histone deacetylases (HDACs) do not have obvious sequence specificity them-

selves, they may be recruited by TFs that recognize specific sequences. Thus, sequence information is an important confounding factor.

We tested using two different sequence motif based-methods to account for the *cis*-regulatory information and observed that the two methods gave remarkably consistent results. Here we present results from using MDscan [26], which first infers sequence motif information *de novo* based on the gene transcription rate data. In particular, this algorithm searched for enriched sequence motifs of widths 5 to 15 in the promoter sequences, resulting in 580 statistically significant, possibly overlapping, candidate TFBMs (p-value < 0.05). We then used these motif patterns to scan all promoter regions for matches so as to compute a motif score for each TFBM at each promoter. We used both a linear regression procedure, Motif Regressor [23], and RSIR to select 33 motifs that are significantly influential of the transcription rate.

As an alternative approach to account for the *cis*-regulatory information, we directly used the 666 TFBMs reported by Beer and Tavazoie [25], which is a combination of computational predictions using AlignACE [15] and 51 experimentally derived ones. Out of these 666 motifs, our linear regression and RSIR procedures found 15 that are highly relevant to predicting gene transcription rates.

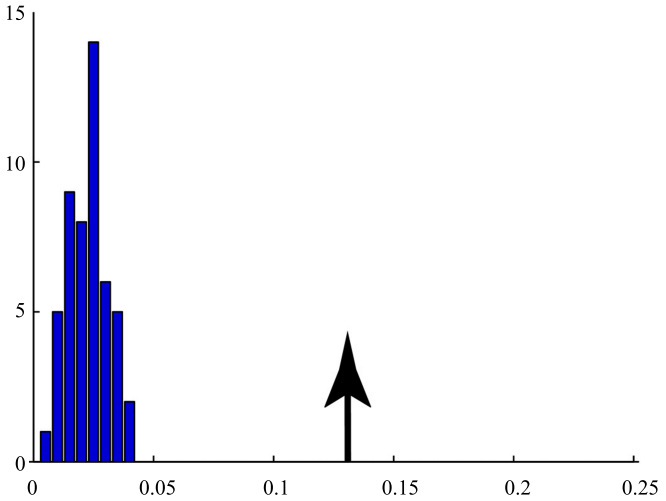

Figure 1 Model validation by comparing the R^2 for the real versus randomly permutated datasets. The R^2 obtained by applying the motif selection and fitting Equation (4.1) (with sequence motif information only) procedures to randomly permutated and real data. The histogram is obtained based on 50 randomly permutated samples. The arrow on the right marks the R^2 for the real data. Results for the coding regions are represented here.

To assess the significance of our model for controlling the confounding effects due to sequence information, we randomly permuted the transcription rates data 50 times and repeated the same statistical procedures: identifying motif candidates using MDscan, selecting the most significant motifs using RSIR, and fitting

the linear regression model. The distribution of R^2 obtained for these randomized data, as well as the R^2 value for the original data, was shown in Figure 1. The largest motif-based R^2 observed in randomized data was 0.038, which is significantly below the motif-based R^2 for the real data.

The combined transcriptional control by TFBMs, nucleosome occupancy, and histone acetylation is modeled as:

$$y_i = \alpha + \sum_j \beta_j x_{ij} + \sum_j \eta_j z_{ij} + \delta w_i + \epsilon_i, \tag{4.1}$$

where the x_{ij} values are the three histone acetylation levels (corresponding to H3K9, H3K14, and H4, respectively), the z_{ij} values are the corresponding scores to the 33 selected motifs, and w_i is the nucleosome occupancy level. The results are shown in Table 1. One can see that a simple regression of transcription rates against histone acetylation without considering any other factors gave an R^2 of 0.1841, implying that about 18% of the variation of the transcription rates is attributable to histone acetylation. In contrast, the regression of transcription rates against motif scores and nucleosome density levels (no histone acetylation) gave an R^2 of 0.1997. The comprehensive model with all the variables we considered bumped up the R^2 to 0.3262, indicating that the histone acetylation does have a significant effect on the transcription rate, although not as high as that in the naive model.

Table 1: Model performance (adjusted R^2) with different covariates

Ace. sites	-	Seq	Nuc	S/N	-	Seq	Nuc	S/N
-	0	0.14	0.11	0.20	0	0.13	0.14	0.22
H3K9,14	0.18	0.27	0.26	0.32	0.10	0.21	0.25	0.31
H4	0.08	0.21	0.25	0.31	0.02	0.15	0.21	0.28
H3K9,14,H4	0.18	0.27	0.27	0.33	0.20	0.26	0.26	0.31

The adjusted R^2 for the linear regression model (Equation (4.1)) containing different regulatory factors (Nuc (N), nucleosome occupancy; Seq (S), sequence information).

5 A case study: Oct4 ChIP-chip data in human ESCs

5.1 The data and learning methods

The DNA microarray used in [49] covers −8 kb to +2 kb of ∼16,000 annotated human genes. We identified consistently a Sox-Oct composite motif from both the Oct4 and the Sox2 ChIP-chip data sets using the *de novo* motif search algorithm CisModule [20] with heterogeneous Markov background [50]. This motif is known to be recognized by the protein complex of Oct4 and Sox2, the target TFs in the ChIP-chip experiments. Noting that this motif is identical to the Sox-Oct

composite motif detected from an independent Oct4 ChIP-PET data set in mouse ESCs [51], we included this motif in our pre-compiled motif set. In addition, we included all the 219 known high-quality PWMs from TRANSFAC release 9.0 [52] and the PWMs of four TFs with known functions in ES cells from the literature [53], to compile a final list of 224 motif PWMs.

Boyer *et al.* [49] reported 603 Oct4-ChIP enriched regions (positives) in human ESCs. We randomly selected another 603 regions with the same length distribution from the genomic regions targeted by the DNA microarray (negatives), i.e. $[-8, +2]$ kb of the 16,000 human genes. A ChIP-intensity measure, which is defined as the average array-intensity ratio of ChIP samples over control samples, is attached to each of the 1206 ChIP-regions. We treat the logarithm of the ChIP-intensity measure as the response variable, and the features extracted from the genomic sequences as explanatory variables. This produced a data set of 1206 observations with 269 features (explanatory variables).

We compare the following methods for statistical learning on this data set: (1) LR-SO, linear regression using the Sox-Oct composite motif only; (2) LR-Full, linear regression using all the 269 features; (3) Step-SO, stepwise linear regression starting from LR-SO; (4) Step-Full, stepwise linear regression starting from LR-Full; (5) NN-SO, neural networks with the Sox-Oct composite motif feature as input; (6) NN-Full, neural networks with all the features as input; (7) MARS, multivariate adaptive regression splines using all the features; (8) Boost, boosting with regression tree as base learner; (9) SVM, support vector machine for regression with various kernels; (10) BART, Bayesian additive regression trees with different number of trees.

5.2 Comparison results

A ten-fold cross-validation procedure was conducted as follows. We first divided the observations into ten subgroups of equal size at random. Each time, we left one subgroup (called "the test sample") out and used the remaining nine subgroups (called "the training sample") to train a model using one of the above methods. Then, we predicted the responses for the test sample based on the trained model and compared them with the observed responses. This process was continued until every subgroup had served as the test sample once. In this section, we use the correlation coefficient between the predicted and observed responses as a measure of the goodness of model performance. This measure is invariant under linear transformation, and can be intuitively understood as the fraction of variation in the response variable that can be explained by the features (covariates). We call this measure the CV-correlation, or CV-cor, henceforth.

The cross validation results are summarized in Table 2. The average CV-cor (over 10 cross validations) of LR-SO is 0.446, which is the lowest among all the linear regression methods. All the other methods used more features and predicted better, demonstrating that sequence features other than the target motif contribute to the prediction of ChIP-intensity. In Step-SO, we started from the LR-SO model and used the stepwise method (with both forward and backward steps) to add or delete features in the linear regression model based the AIC criterion

(see R function "step"). The Step-Full was performed similarly, but starting from the LR-Full model. Among all the linear regression methods, Step-SO achieved the highest CV-cor of 0.535.

Table 2: Ten-fold cross validations of the Oct4 ChIP-chip data

Method	Tuning parameters	Optimal Cor
LR-SO	–	0.446 (0%)
LR-Full	–	0.491 (10%)
Step-SO	–	0.535 (20%)
Step-Full	–	0.513 (15%)
NN-SO	# of nodes, weight decay	0.468 (5%)
Step+NN	# of nodes, weight decay	0.463 (4%)
MARS	interaction d, penalty λ	0.580 (30%)
SVM	cost C	0.547 (23%)
Boost	# of iterations	0.586 (32%)
BART	# of trees N	0.600 (35%)

Note: Reported here are the average CV-correlations (Cor). The percentage in the parentheses is calculated by Cor/Cor(LR-SO)−1.

For neural networks (implemented in R package "nnet"), we tested its performance with all combinations of different number of hidden nodes (2, 5, 10, 20, 30) and weight decay (0, 0.5, 1.0, 2.0). However, even the optimal results were not satisfactory. The NN-SO showed a slight improvement in CV-cor over that of LR-SO, while the neural network with all the features as input encountered a severe overfitting problem, resulting in a CV-cor < 0.38. The NN reached an optimal CV-cor of 0.463 with 2 hidden nodes. MARs (R package "mda") is sensitive to the choice of the penalty parameter λ, and the optimal CV-cor of 0.580 was reached when $\lambda = 6$. Support vector regression (ϵ-SVR) as defined in [29] was applied to this dataset, with the implementation of LIBSVM [54] in R-package "e1071". We tested the linear, radial basis, polynomial (3rd-order), and sigmoid kernels and found that the radial basis kernel performed the best. The optimal CV-cor of 0.548 was reached when the cost parameter $C = 1$. The boosting method (the R package "mboost" [55]) using regression trees with a maximum depth of 2 as the base learner performed quite robustly for this dataset, with CV-cor ranging from 0.532 to 0.586 with various stopping rules. For BART, we ran 20,000 iterations after a burn-in period of 2,000 iterations, as implemented in the R package "BayesTree". We tested the method with the number of trees ranging from 20 to 200. Notably, BARTs with different number of trees reached CV-cor's between 0.592 and 0.6, which outperformed all the other methods in terms of both CV-cor and robustness.

5.3 An analysis of selected sequence features

The top feature chosen by BART is the Sox-Oct composite motif, which is consistent with the existing biological knowledge that Sox2 is one of the most important cooperative TFs of Oct4 and they form a complex to bind to the composite sites. The next three important variables are all background features, "GC", "CAA",

and "CCA" with $P_{in} > 0.98$. Other two background variables, "AA" and "G/C", also have high posterior inclusion probabilities. It is interesting to note that the t-values for "GC", "AA", and "G/C" are not that significant, implying that they may be cooperating with other variables to affect the TF-DNA binding. The frequency of "CAA" is significantly higher and the frequency of "CCA" is significantly lower in the positive ChIP-regions than in the negative ones (see their t-values). It is possible that these words are responsible for the interaction strength between the TF Oct4 and its DNA target region, given that "CAA" occurs in the Sox-Oct motif consensus.

In addition to the Sox-Oct motif, we found eight motifs with $P_{in} > 0.5$, among which OCT_Q6, OCT1_Q6 and OCT1_Q5_01 are variants of the Oct4 motif, implying that Oct4 may work with different cooperative factors to control the transcription of different target genes. The remaining five motifs, Hsf1, Uf1h3b, Nfy_Q6, E2F, and E2F1, may be cooperative factors of Oct4 or other functional TFs in ESCs. We note that the sequence length is also selected in the model, which serves 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 inclusion of many non-motif features in the optimal BART model. This is also true for the learning results of other methods, such as stepwise linear regressions (data not shown). To further verify their effect in predictive modeling, we excluded non-motif 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 reduced data set. The CV-correlations were 0.510, 0.511, 0.478, and 0.456 for the above four models, respectively, which decreased substantially (about 12∼15%) compared to those of the corresponding methods with all the features. One almost obtains no improvement (2%) in predictive power by taking more motif features in the linear regression. However, if the background and other generic features are incorporated, the stepwise regression improved dramatically (20%).

Using this data set, we also compared the use of heterogeneous and homogeneous Markov background models for motif feature extraction. For the homogeneous background model, we used all the nucleotides in a sequence to build a first-order Markov chain. Intuitively, the heterogeneous background model [50] assumes that the sequence in consideration can be segmented into pieces and within each piece the nucleotides follow a homogeneous first-order Markov chain. Using a Bayesian formulation and an MCMC algorithm, we estimate the background transition probability of each nucleotide. 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 observe that, for all the Oct4-family matrices, the motif scores under the heterogeneous background model show higher correlations with the log-ChIP intensity than the scores under the homogeneous background model. We further computed the t-statistic for each motif score between the positive and negative ChIP-regions. Similarly, using the heterogeneous background model enhances the separation between the positive and negative regions by resulting in larger t-statistics (Table 3).

Table 3: Comparison between the heterogeneous (ht) and the homogeneous (hm) background models for motif feature extraction

Motif	Sox-Oct	OCT1_Q6	OCT_Q6	OCT1_Q5_01
Cor(ht)	0.442	0.261	0.303	0.295
Cor(hm)	0.421	0.221	0.267	0.256
t-stat(ht)	14.04	8.77	10.19	10.16
t-stat(hm)	13.06	7.38	8.87	8.68

6 A simulation study

We performed a simulation study as a final test on the effectiveness of the PM approaches. We generated 1,000 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 a probability-ratio score for each inserted site. For each motif, we obtained the sum of the site scores for a sequence, denoted by $Z_1, \ldots, Z_4$ for Oct4, Sox2, Nanog, and Nkx2.5, respectively. Then we defined the motif score for a sequence by $X_j = \log(\max(Z_j, 1))$ for $j = 1, \ldots, 4$. Denote by X_5 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 each sequence was simulated as:

$$Y = X_1(1 + 0.5X_2 + 0.3X_3 + 0.4X_4) + \sqrt{X_1X_3X_4} + 2X_5 + \epsilon, \qquad (6.1)$$

where $\epsilon \sim N(0, \sigma^2)$. This model states that X_1 is the target TF with three interactive factors (X_2, X_3, X_4), and the GC content (X_5) has a positive effect on the level of ChIP-intensities. The signal-to-noise-ratio (SNR) of a simulated data set is defined as $Var(Y)/\sigma^2 - 1$. We simulated 10 independent sequence sets, and then generated observed ChIP-intensities with $SNR = 1/0.6$, $1/1$, and $1/2$, respectively.

We applied exactly the same sequence feature extraction procedure as in the previous sections to the simulated data sets. Since stepwise linear regression, MARS, SVM, boosting and BART showed more promising learning ability in the Oct4 and Sox2 data sets, we tested only these five methods in this simulation study. To quantify their performance, we calculated the average correlation coefficient between predicted and the true ChIP-intensities, and compared the motifs selected by each method to the true ones. 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. We used MARS with $d = 1, \lambda = 6$, boosting with 100 iterations, the radial basis SVM with $C = 1$ and BART with 100 trees here given that these were their optimal tuning parameters in the Oct4 data set, which is roughly of the same size as the simulated data. The comparison of the results of these methods are given in Table 4.

Table 4: Performance comparison on the simulated data sets

Method	SNR	1/0.6	1/1	1/2
Step-LR	Cor	0.732(0.012)	0.703(0.017)	0.637(0.018)
	N_T	3.8(0.42)	3.8(0.42)	3.3(0.67)
	N_F	4.9(2.64)	8.6(3.63)	7.5(3.54)
MARS	Cor	0.732(0.014)	0.704(0.019)	0.651(0.031)
	N_T	3.9(0.32)	3.5(0.53)	3.3(0.48)
	N_F	4.4(1.78)	4.2(1.81)	4.6(2.80)
SVM	Cor	0.798(0.010)	0.737(0.020)	0.626(0.021)
Boost	Cor	0.810(0.009)	0.787(0.013)	0.725(0.016)
BART	Cor	0.805(0.011)	0.779(0.011)	0.720(0.011)
	N_T	4.0(0.00)	3.6(0.70)	3.5(0.71)
	N_F	1.8(1.40)	2.5(1.18)	2.5(1.43)

Note: Reported are the averages results (standard errors in the parentheses) of 10 independent data sets. "Cor" is the correlation between predicted and true ChIP-intensities. N_T and N_F are the numbers of true and false motifs identified.

As expected, with the increase of the noise level, the average correlation and the accuracy of motif identification decreased for all the tested methods. However, even when the SNR is as low as 1/2, the correlation is still above 0.7 for the two additive-tree-based methods, boosting and BART, which again demonstrates their strong capability of approximating complicated non-linear functions from training data. We further tested the accuracy of these approaches in detecting true motifs that determine the ChIP-intensity. When we set the threshold of P_{in} to be 0.7, BART identified on average more than 85% of the true motifs with at most 2.5 falsely included motifs. At comparable sensitivity levels (N_T), BART reported significantly fewer false positives (N_F) for all the SNR levels than stepwise linear regression and MARS (Table 4). For either SVM or boosting, we were not able to check the features (covariates) utilized by these methods given that the methods (and the implemented programs) work more like a black-box with no explicit feature selection criteria.

7 Discussion

We have demonstrated in this article how the predictive modeling approach, with the help of sophisticated statistical learning tools, can reveal subtle sequence signals that may influence TF-DNA binding and generate testable hypotheses. Compared with some other more systems-based approaches to gene regulation, such as building a large system of differential equations or inferring a comprehensive Bayesian network, PM approaches are more direct, intuitive, theoretically solid (as many in-depth statistical learning theories have been developed), and easily validated (i.e., using cross-validations). It can generate straightforward testable hypotheses.

In the Oct4 ChIP-chip data, PM approaches not only unambiguously identified the binding motifs for the target TF, but also discovered several known or potential cooperative TFs, such as E2F and NFY. As a principled way of utiliz-

ing both positive (i.e., binding sequences) and negative (nonbinding) information, this approach provides another way to detect *cis*-regulatory motifs besides the well-established generative model-based motif discovery methods (e.g., [1, 2, 56, 57]). The PM approaches appear to be more sensitive than both *de novo* motif discovery approaches and motif scan approaches (based on experimentally validated TF motifs) since they effectively utilize the ChIP-intensity or expression information in the model. For instance, with only the positive-ChIP regions, we were not able to detect E2F or NFY motifs by conducting only *de novo* motif searches.

The comparative study on several learning methods suggests BART as a good candidate for analyzing high-dimensional genomic data because of both its predictive power and its interpretability. First, like boosting, BART is an ensemble learning method, which approximates an unknown relationship by the aggregation of a large number of simple models (small trees). Second, the Bayesian formulation of BART leads to a posterior sample of the predictive models, which helps predict the response of a new observation by using the average of all sampled trees. This is effectively a model averaging approach, which is known to improve the model's predictive power [58]. Finally, BART MCMC updates features in additive trees according to the joint posterior distribution. Thus, it has a coherent variable selection component based on the posterior probabilities. This provides an sensible way to identify important sequence features that contribute to TF-DNA interaction. Note that Step-SO is equivalent to MotifRegressor [23] and MARS is equivalent to MARSMotif [24], with all the known and discovered (Sox-Oct) motifs as input. Thus, our study demonstrates that BART with all three categories of features outperformed MotifRegressor and MARSMotif significantly.

With the rapid accumulation of such large-scale data, we believe that, as illustrated by this work, flexible statistical learning methods on well-designed sequence features will be very useful for the understanding of TF-DNA interaction and the development of predictive approaches in *cis*-regulatory analysis.

References

[1] Stormo, G.D. and Hartzell, G.W. (1989) Identifying protein-binding sites from unaligned DNA fragments, *Proc. Natl. Acad. Sci. USA*, 86, 1183-1187.

[2] Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wooton, J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, *Science*, 262, 208–214.

[3] Jensen, S.T., Liu, X.S., Zhou, Q., and Liu, J.S. (2004) Computational discovery of gene regulation binding motifs: a Bayesian perspective, *Statist. Sci.*, 19, 188–204.

[4] Elnitski, L., Jin, V.X., Farnham, P.J., and Jones, S.J.M. (2006) Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques. *Genome Res.*, 16, 1455–1464.

[5] Benos, P.V., Lapedes, A.S., and Stormo, G.D. (2002) Probabilistic code for DNA recognition by proteins of the EGR family, *J. Mol. Biol.*, 323, 701–727.

[6] Bulyk, M.L., Johnson, P.L.F., and Church, G.M. (2002) Nucleotides of tran-

scription factor binding sites exert interdependent effects on the binding affinities of transcription factors, *Nucleic Acids Res.*, 30, 1255–1261.
[7] Barash, Y., Elidan G., Friedman, N., and Kaplan, T. (2003) Modeling dependence in protein-DNA binding sites, *Proc. Int. Conf. Res. Comp. Mol. Biol.*, 7, 28–37.
[8] Zhou, Q. and Liu, J.S. (2004) Modeling within-motif dependence for transcription factor binding site predictions, *Bioinformatics*, 20, 909–916.
[9] Zhao, Y., Huang, X.H., and Speed, T.P. (2005) Finding short DNA motifs using permuted Markov models, *J. Comput. Biol.*, 12, 894–906.
[10] Workman, C.T. and Stormo, G.D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity, *Pac. Symp. Biocomput.*, 5, 467–478.
[11] Smith, A.D., Sumazin, P., and Zhang, M.Q. (2005) Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc. Natl. Acad. Sci. USA*, 102, 1560–1565.
[12] Hong, P., Liu, X.S., Zhou, Q., Lu, X., Liu, J.S., and Wong, W.H. (2005) A Boosting approach for motif modeling using ChIP-chip data, *Bioinformatics*, 21, 2636–2643.
[13] Wasserman, W.W. and Fickett, J.W. (1998) Identification of regulatory regions which confer muscle-specific gene expression, *J. Mol. Biol.*, 278, 167–181.
[14] Frith, M.C., Hansen, U., and Weng, Z. (2001) Detection of cis-element clusters in higher eukaryotic DNA, *Bioinformatics*, 17, 878–889.
[15] Roth, F.P., Hughes,J.D., Estep, P.W., Church, G.M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotech*, 16, 939–945
[16] Xing, E.P., Wu, W., Jordan, M.I., and Karp, R.M (2003) LOGOS: A modular Bayesian model for de novo motif detection, *IEEE CSB2003*.
[17] Thompson, W., Palumbo, M.J., Wasserman, W.W., Liu, J.S., and Lawrence, C.E. (2004) Decoding human regulatory circuits, *Genome Res.*, 14, 1967–1974.
[18] Gupta, M. and Liu, J.S. (2005) *De novo* cis-regulatory module elicitation for eukaryotic genomes, *Proc. Natl. Acad. Sci. USA*, 102, 7079–7084.
[19] Zhou, Q. and Wong, W.H. (2007) Coupling hidden Markov models for the discovery of cis-regulatory modules in multiple species, *Ann. Appl. Statist.*, 1, 36–65.
[20] Zhou, Q. and Wong, W.H. (2004) CisModule: *De novo* discovery of cis-regulatory modules by hierarchical mixture modeling, *Proc. Natl. Acad. Sci. USA*, 101, 12114–12119.
[21] Bussemaker, H.J., Li, H., and Siggia, E.D. (2001) Regulatory element detection using correlation with expression, *Nat. Genet.*, 27, 167–171.
[22] Keles, S., van der Laan, M., and Eisen, M.B. (2002) Identification of regulatory elements using a feature selection method, *Bioinformatics*, 18, 1167–1175.
[23] Conlon, E.M., Liu, X.S., Lieb, J.D., and Liu, J.S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis, *Proc. Natl. Acad.*

Sci. USA, 100: 3339–3344.
[24] Das, D., Banerjee, N., and Zhang, M.Q. (2004) Interacting models of cooperative gene regulation, *Proc. Natl. Acad. Sci. USA*, 101, 16234–16239.
[25] Beer, M.A. and Tavazoie, S. (2004) Predicting Gene Expression from Sequence, *Cell*, 117, 185–198.
[26] Liu, X., Brutlag, D.L., and Liu, J.S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin immunoprecipitation microarry experiments. *Nat. Biotech.*, 20, 835–39.
[27] Liu, J.S., Neuwald, A.F., and Lawrence, C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies, *J. Am. Stat. Assoc.*, 90, 1156–1170.
[28] Friedman, J.H. (1991) Multivariate adaptive regression splines, *Ann. Statist.*, 19, 1–67.
[29] Vapnik, V. (1998) *The nature of statistical learning theory* (2nd edition), Springer-Verlag, New York.
[30] Freund, Y. and Schapire, R. (1997) A decision-theoretical generalization of online learning and an application to boosting, *J. Comp. Syst. Sci.*, 55, 119–139.
[31] Chipman, H.A., George, E.I., and McCulloch, R.E. (2006) BART: Bayesian additive regression trees, *Technical Report*, Univ. of Chicago.
[32] Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thaström, A. Field, Y., Moore, I.K., Wang, J.Z., and Widom, J. (2006) A genomic code for nucleosome positioning. *Nature*, 442, 772–778.
[33] Yuan, G.C., Ma, P., Zhong, W. and Liu, J.S. (2006) Statistical assessment of the global regulatory role of histone acetylation in Saccharomyces cerevisiae. *Genome Biology*, 7, R70.
[34] Li, K.C. (1991) Sliced inverse regression for dimension reduction, *J. Am. Stat. Assoc.*, 86, 316–327.
[35] Zhong, W., Zeng, P., Ma, P., Liu, J.S., and Zhu, Y. (2005). RSIR: regularized sliced inverse regression for motif discovery. *Bioinformatics*, 21, 4169–4175
[36] Strahl, B.D. and Allis, C.D. (2000) The language of covalent histone modifications, *Nature*, 403, 41–45.
[37] Turner, B.M. (2002) Cellular memory and the histone code, *Cell*, 111, 285–291.
[38] Schreiber, S.L. and Bernstein, B.E. (2002) Signaling network model of chromatin. *Cell*, 111, 771–778.
[39] Roth, S.Y., Denu, J.M. and Allis, C.D. (2001) Histone acetyltransferases. *Annu. Rev. Biochem.*, 70, 81–120.
[40] Kurdistani, S.K. (2003) Grunstein M: Histone acetylation and deacetylation in yeast. *Nat. Rev. Mol. Cell. Biol.*, 4, 276–284.
[41] Kurdistani, S.K., Tavazoie, S., and Grunstein, M. (2004) Mapping global histone acetylation patterns to gene expression, *Cell*, 117, 721–733.
[42] Pokholok, D.K., Harbison, C.T., Levine, S., Cole, M., Hannett, N. M., Lee, T.I., Bell, G.W., Walker, K., Rolfe, P. A., Herbolsheimer, E., et al. (2005) Genome-wide map of nucleosome acetylation and methylation in yeast, *Cell*, 122, 517–527.

[43] Lee, C.K., Shibata, Y., Rao, B., Strahl, B.D., and Lieb, J.D. (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide, *Nat. Genet.*, 36, 900–905.
[44] Robert, F., Pokholok, D.K., Hannett, N.M., Rinaldi, N.J., Chandy, M., Rolfe, A., Workman, J.L., Gifford, D.K., and Young, R.A. (2004) Global position and recruitment of HATs and HDACs in the yeast genome, *Mol. Cell*, 16, 199–209.
[45] Saccharomyces Genome Database, http://www.yeastgenome.org/
[46] Bernstein, B.E., Liu, C.L., Humphrey, E.L., Perlstein, E.O., and Schreiber, S.L. (2004) Global nucleosome occupancy in yeast, *Genome Biol.*, 5, R62.
[47] Holstege, F.C., Jennings, E.G., Wyrick, J.J., Lee, T.I., Hengartner, C.J., Green, M.R., Golub, T.R., Lander, E.S., and Young, R.A. (1998) Dissecting the regulatory circuitry of a eukaryotic genome, *Cell*, 95, 717–728.
[48] Wang, Y., Liu, C.L., Storey, J.D., Tibshirani, R.J., Herschlag, D. and Brown, P.O. (2002) Precision and functional specificity in mRNA decay, *Proc Natl Acad Sci USA*, 99, 5860–5865.
[49] Boyer, L.A. *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells, *Cell*, 122, 947–956.
[50] Liu, J.S. and Lawrence, C.E. (1999) Bayesian inference on biopolymer models, *Bioinformatics*, 15, 38–52.
[51] Loh, Y.H. *et al.* (2006) The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells, *Nature Genet.*, 38, 431–440.
[52] Matys, V., Fricke, E., Geffers, R. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles, *Nucleic Acids Res.*, 31, 374–378.
[53] Zhou, Q., Chipperfield, H., Melton, D.A., and Wong, W.H. (2007) A gene regulatory network in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. USA*, 104, 16438–16443.
[54] Chang, C.C and Lin, C.J. (2001) LIBSVM : a library for support vector machines. http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
[55] Bühlmann, P. and Hothorn, T. (2007) Boosting algorithms: Regularization, Prediction and model fitting, *Statist. Sci.*, in press.
[56] Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 28–36.
[57] Liu, X., Brutlag, D.L., and Liu, J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Smp. Biocomput.* 6, 127–138.
[58] Kass, R.E. and Raftery, A.E. (1995) Bayes factors. *J. Am. Stat. Assoc.*, 90, 773–795.