Improved Prediction of Transcription Binding Sites from Chromatin Modification Data

Kengo Sato*, Tom Whitington*, Timothy L. Bailey and Paul Horton

Abstract— **In this paper we apply machine learning to the task of predicting transcription factor binding sites by combining information on multiple forms of chromatin modification with the binding strength DNA site predicted by a position weight matrix. We additionally explore the effect of incorporating auxiliary features such as the distance of the site to the nearest gene's transcription start site and the degree to which the site is conserved among related species. We approach the task as a classification problem, and show that both Naïve Bayes and Random Forests can provide substantial increases in the accuracy of predicted binding sites. Our results extend previous work which simply filtered candidate sites based on H3K4Me3 chromatin modification scores. In addition we apply feature selection to explore which forms of chromatin modification and which auxiliary features have predictive value for which transcription factors.**

I. Introduction

The prediction of transcription factor binding sites (TFBS) based on a position weight matrix (PWM) to model binding sequence preference is a well studied problem [1], [2], [3]. PWMs for transcriptions factors are typically short (5-15 bases) and allow considerable degeneracy [4]. Thus a simple application of PWMs can result in numerous false positives due to the presence of non-binding site sequence regions which match the PWM well, unless the search can be limited to a small region, for example very near known promoters. This is largely possible for bacteria, and PWM based scanning has proved to be relatively effective for prediction of binding sites in bacteria [5]. However, prediction from PWM alone is increasingly difficult for higher order organisms, as binding sites distal from promoters are known to affect transcription [6]. In mammals this effect is so extreme that PWM scanning has been described as "futile" [7]. Several strategies have been employed to improve the prediction of TFBSs. Phylogenetic footprinting [8], [9] assumes that evolutionarily conserved PWM matches are more likely to be genuine sites. Other techniques include exploiting the observation that TFBSs often cluster together [10].

Chromatin modification is known to affect the binding of transcription factors (TBFS) [11], [12]. Increased accessibility of DNA to transcription factor binding can occur due to the decreased electrostatic attraction between the negatively charged DNA and the positively charged histone, when the negatively-charged acetyl group is added to the histone. Additionally, certain histone modifications such as H3K4me3 can recruit histone remodelling complexes, that increase nucleosome mobility and thereby promote DNA accessibility [13]. Whilst disparate mechanisms for chromatin modulation of transcription factor binding have been identified, the combinatorial effect of these mechanisms, driven by histone modification, has not yet been studied. Such an investigation is now possible given the availability of genome wide histone modification datasets for a large number of modifications.

II. Previous Work

Recently, Whitington *et al.* [14] showed H3K4Me3 data can improve the prediction of binding sites for many transcription factors, by using only a simple threshold filter on the H3K4Me score. Barski *et al.* [15] used Pol II, H3K4me3, and H2A.Z data with kernel density estimation functions to predict miRNA promoters. Narlikar *et al.* [16]. showed that nucleosome occupancy information can significantly improve *ab initio* TFBS motif discovery performance in yeast. Wang et al. [17] demonstrated that integrating histone modification data in the task of promoter prediction yields a significant improvement in promoter prediction performance. Won *et al.* [18] trained an HMM to predict promoter and enhancer regions using histone modification data. In another work, Won *et al.* [19] used the predicted promoter and enhancer regions to enhance prediction of cis-regulatory modules.

III. TFBS Prediction as a Classification Problem

We formulate the TFBS prediction problem as a standard binary classification task: in which one attempts to learn a function mapping a feature set onto 0 (not start of a binding site) or 1 (start of a binding site), where the start of a binding site corresponds to the first column of the transcription factor's PWM. Each genome position is an example to be classified. The starting positions of known TFBSs from the gold standard of Whitington et al. [14] are positive examples. Although binary classification is an extremely well studied problem, this formulation does present some challenges because the number of negative examples is very large, both in absolute numbers and relative to the number of positive examples.

IV. Dataset

A. TFBSs

We prepared five TFBSs datasets, one each for the transcription factors: GABP, SRF, sp1, cMyc, and NRSF. The

* Contributed Equally. Kengo Sato and Paul Horton are with the Computational Biology Research Center (AIST), Japan; Kengo Sato is also with the Graduate School of Frontier Sciences, University of Tokyo, Japan (current affiliation). Tom Whitington and Timothy Bailey are with the Institute for Molecular Bioscience, The University of Queensland, Brisbane, Australia. Our emails are {sato-kengo, horton-p}@aist.go.jp; {t.whitington, t.bailey}@imb.uq.edu.au

Whitington *et al.* [14] datasets were used for sp1 and cMyc datasets and, following the same protocol, GABP, SRF and NRSF datasets were derived from the experiments of Valouev *et al.* [20].

B. Features

- **Histone Methylation**
  We used data for various histone methylations from the study of Barski *et al.* [21]. The specific variations used were: H2BK5me1, H3K27me1, H3K27me2, H3K27me3, H3K36me1, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me1, H3K79me2, H3K79me3, H3K9me1, H3K9me2, H3K9me3, H3R2me1, H3R2me2, H4K20me1, H4K20me3, H4R3me2.
- **Histone variant**
  Data for the occurrence of the histone variant H2A.Z were taken from Barski *et al.* [21].
- **PolII binding**
  Data for RNA Polymerase II binding were taken from Barski *et al.* [21].
- **Histone Acetylation**
  Histone acetylation data from Wang *et al.* [22].
- **Transcription Start Site Distance**
  The distance to the nearest transcription start site annotated in AceView [23].
- **Phylogenetic conservation**
  The degree of Phlyogenetic conservation observed for each position as quantified by the `phastCons` program [24], “masked” for exons (which are expected to be conserved for other reasons) as described in Whitington *et al.* [14]

C. Dataset Format

The dataset is represented in a simple tabular text format with one line corresponding to one position in the genome. The fields are: Chromosome number, Position in chromosome, strand (+ or -), PWM_Score, TFBS_Status, AceView, H2AZ, H2BK5me1, H3K27me1, H3K27me2, H3K27me3, H3K36me1, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K79me1, H3K79me2, H3K79me3, H3K9me1, H3K9me2, H3K9me3, H3R2me1, H3R2me2, H4K20me1, H4K20me3, H4R3me2, PolII, H2AK5ac, H2AK9ac, H2BK120ac, H2BK12ac, H2BK20ac, H2BK5ac, H3K14ac, H3K18ac, H3K23ac, H3K27ac, H3K36ac, H3K4ac, H3K9ac, H4K12ac, H4K16ac, H4K5ac, H4K8ac, H4K91ac, phastCons. This dataset will be made avaiable upon request.

V. Classifiers

The datasets used in our experiments contain a huge number of genomic loci with features described in the previous section. Although, following the procedure of Whitington *et al.* [14], a number of loci were removed out by requiring a minimum PWM p-value, more than 100,000 loci remained in the data set. The number of training examples for each transcription factor is shown in table I. For such large-scale datasets, the application of state-of-the-art machine learning algorithms, such as support vector machines, is difficult due to the computation time required when training on many examples. Therefore, we applied the faster naïve Bayes and random forests classifers.

TABLE I
THE NUMBER OF POSITIVE AND NEGATIVE EXAMPLES FOR EACH OF FIVE TRANSCRIPTION FACTORS.

	positive	negative
GABP	3809	488099
NRSF	1668	375683
SRF	409	381059
cMyc	137	4741
sp1	146	21239

A. Naïve Bayes

A naïve Bayes classifier is a simple classifier based on Bayes' theorem [25]. The probabilistic model $p(C|F_1, \dots, F_n)$ for a classifier over a class variable C and n feature variables $F_1, \dots, F_n$ can be rewritten by using Bayes' theorem:

$$p(C|F_1, \dots, F_n) = \frac{p(C)p(F_1, \dots, F_n|C)}{p(F_1, \dots, F_n)}. \qquad (1)$$

Assuming conditional independence of each feature, we can factorize (1) as the following:

$$p(F_1, \dots, F_n|C) = p(F_1|C) \cdots p(F_n|C) \qquad (2)$$

The prior $p(C)$ and the likelihood of each feature $p(F_i|C)$ are estimated from given training data. After that, we can calculate the posterior class probability of C for given data with features $F_1, \dots, F_n$ by using (1) and (2).

1) Representation of Conditional Probabilities: As described in the previous section, we employed four types of features: PWM p-value as computed by MAST [26], distance to nearest annotated transcription start site (TSS distance), ChIP-seq read counts for various ChIP-seq experiments and phylogenetic conservation score. In order to estimate $p(F_i|C)$ for each feature, we assume that PWM p-value and phylogenetic conservation score are distributed according to the beta distribution, which is a conjugate prior for the binomial probability distribution. We assume that TSS distance and ChIP-seq read counts are distributed according to the normal distribution and the Poisson distribution respectively. The parameters of the distributions were estimated by maximum likelihood estimation from the given data for each class respectively (TFBS or non-TFBS).

B. Random Forest

Random forest [27] is a classifier which, although more sophisticated than naïve Bayes, also requires relatively little computation time for training. A random forest is an ensemble classifier that consists of many decision trees and predicts by majority voting amongst its constituent trees. We employed our own ruby implementation of the random

TABLE II

Average AUC and its standard deviation from 10-fold cross-validation is shown for each combination of classifier and transcriptions factor. The "PWM baseline" classifier only used the PWM score.

	cMyc	sp1	GABP
PWM baseline	0.551 ± 0.009	0.689 ± 0.007	0.765 ± 0.001
Naïve Bayes	0.668 ± 0.098	0.770 ± 0.064	0.964 ± 0.004
Random Forest	**0.808** ± 0.033	**0.869** ± 0.046	**0.971** ± 0.003
	NRSF	SRF	
PWM baseline	**0.956** ± 0.002	0.783 ± 0.004	
Naïve Bayes	0.950 ± 0.017	0.926 ± 0.015	
Random Forest	0.948 ± 0.019	**0.928** ± 0.014	

TABLE III

Average AUC and its standard deviation from 10-fold cross-validation is shown for prediction of NRSF sites with naïve Bayes trained on: all features, just the PWM p-value, just H3K4me3 data, those two combined, and those two plus the phastCons measure of evolutionary conservation.

Features	AUC
all	0.950 ± 0.017
PWM	0.955 ± 0.018
H3K4me3	0.723 ± 0.017
PWM+H3K4me3	0.951 ± 0.019
PWM+H3K4me3+phastCons	0.950 ± 0.021

TABLE IV

Summarization of information from figures 2, 3. The features which proved essential for maximum classification accuracy are summarized for each transcription factor classified with naïve Bayes and random forest classifiers are shown. The factors are listed in order of size of the effect of hiding them from classifier, however the number of factors shown is merely suggestive and not based on a statistical significance test. Boldface indicates especially informative features.

	Naïve Bayes	Random Forest
GABP	PolII, H3K4me3	PolII, H3K4me3
NRSF	**PWM score**	**PWM score**
SRF	H3K4me3, H2AZ, H3K4me1, PolII	H2AZ, H3K4me1
cMyc	AceView	H2BK5me1, PolII, H3K27me1, phastCons, H4K16ac
sp1	AceView	AceView, H3K27me2

forest classifier. We trained random forests consisting of 200 naïve Bayes trees, each of which is a kind of decision tree consisting of single feature naïve Bayes classifiers assigned to each node. For each tree, 20 features were randomly selected and trained on 100 positive and negative examples each. Finally, random forests classify a given data by majority voting among 200 naïve Bayes trees. The parameter values used for the random forests (200, 100 and 20) were chosen arbitrarily. We leave the optimization of these for future work.

VI. Results

A. Prediction Accuracy

The cross-validation prediction accuracy as AUC (so-called Area Under the Curve) for the three classification on five transcription factors is shown in table II.

B. Feature Selection

The effect of leave-one-out feature selection on AUC, using the naïve Bayes and random forests classifiers are shown in figures 2, 3 and the import features are summarized in table IV.

VII. Discussion

For four out of the five transcription factors tested, the machine learning classifiers generally gave equal or superior results to the H3K4me filter method of Whitington *et al.* [14], especially when high sensitivity (missing few true positives) is required. The sole exception to this is NRSF, for which no improvement is seen over simply using the PWM score alone. We further confirmed this result by trying various hand picked feature combinations with naïve Bayes (table III). Actually, NRSF is not really a transcription factor, but is instead a negative regulator [28] with an unusually wide and informative PWM, consistent with the high AUC it achieves by itself (0.956).

A. phastCons data helps predict cMyc binding sites

In general, the classifiers were not able to utilize the phastCons data to improve classification, except in the case of cMyc, in which the random forest method accuracy drops significantly when this feature is withheld (figure 3). The earlier filter based work of Whitington and the naïve Bayes classifier presented here are unable to effectively utilize the phastCons data. Presumably the correlation structure of the features requires a reasonably sophisticated classifier to be exploited effectively.

B. Contribution of TSS information

AceView is a direct, but general source of information about TSSs. PolII and H3K4me3 should correlate with TSSs, including the appropriate tissue specific one, since their measurements and the TFBS dataset used were both measured in Jurkat cells. As summarized in table IV, AceView has essential predictive value for cMyc and sp1, while GABP and SRF prediction is improved by added PolII and H3K4me3 information. It is tempting to speculate that Jurkat cell specific expression regulation explain this difference, but we note that the GABP and SRF have many more positive sites in the gold standard and this difference should be controlled for before making any conclusions.

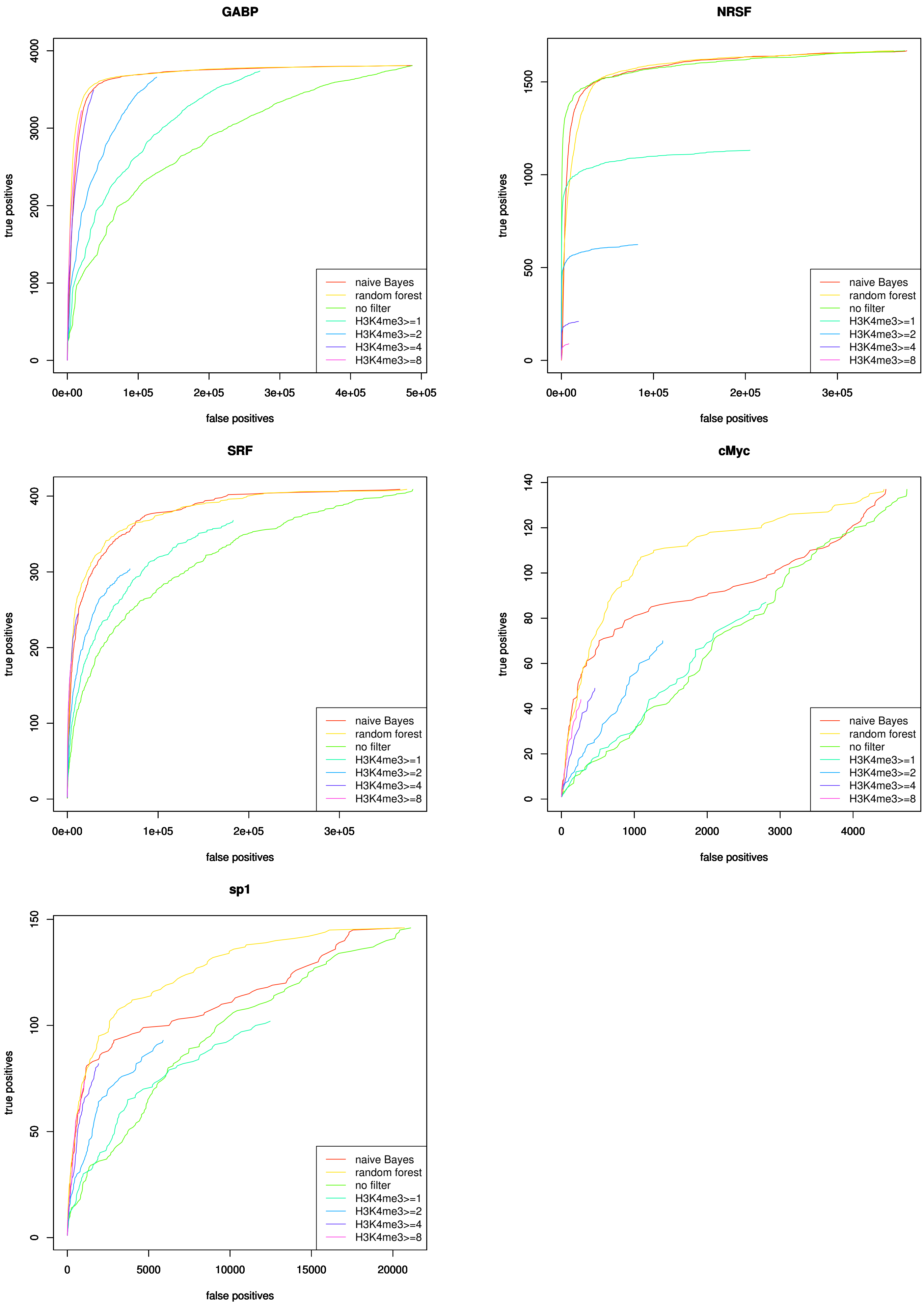

Fig. 1. Comparison of several classifiers, including simple filters with different thresholds for H3K4me3, and just the PWM score alone (“no filter”). Each plot shows the prediction performance for the listed transcription factor. These ROC-like plots show the tradeoff in the number of true positives versus the number of false positives achievable by each method.

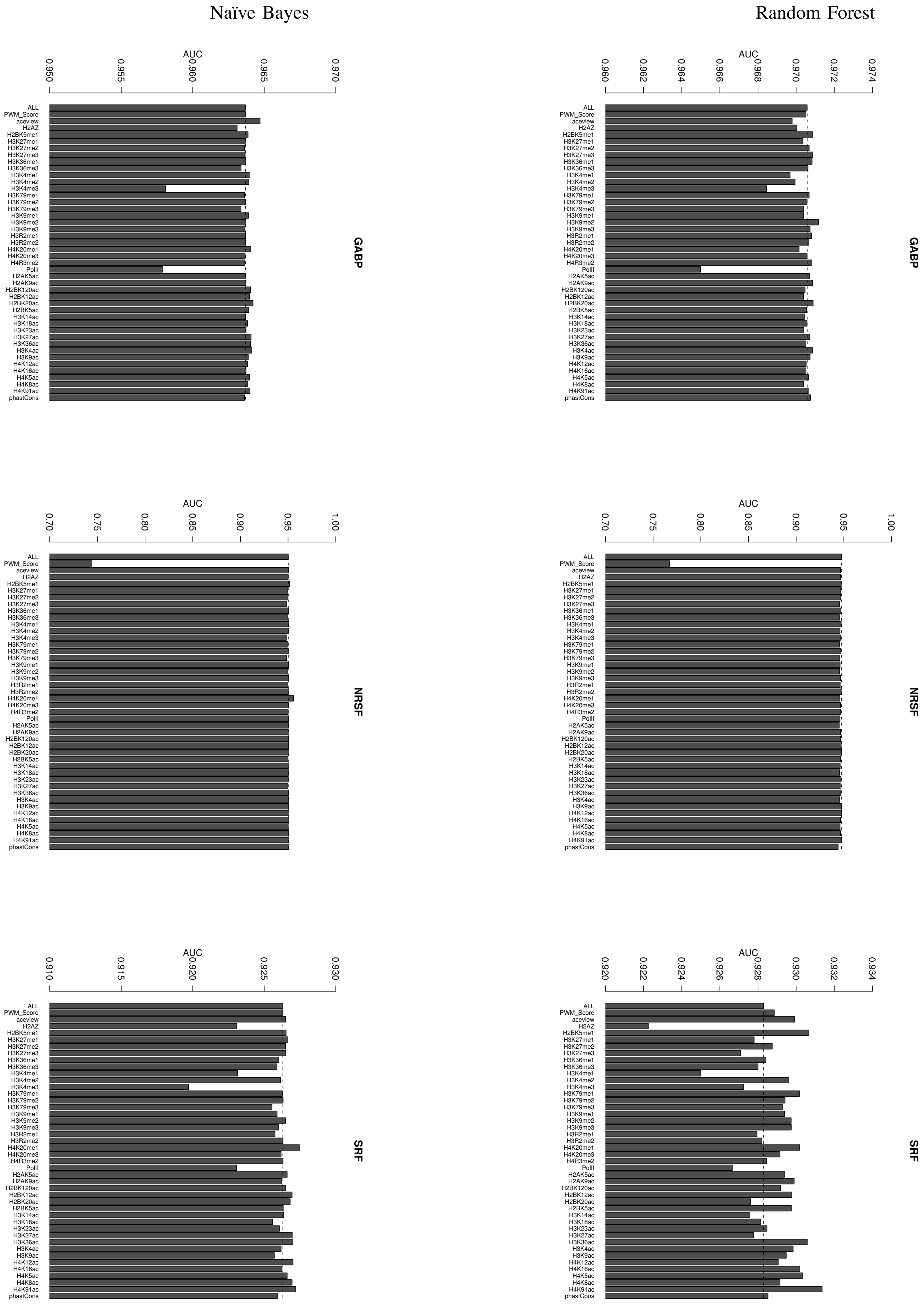

Fig. 2. Effect on prediction accuracy upon removing single features is shown for each transcription factor. In each graph, each column represents a feature, and its height the accuracy obtained when trained on all other features. Results with the naïve Bayes and random forest classifiers are shown for transcription factors GABP, NRSF and SRF.

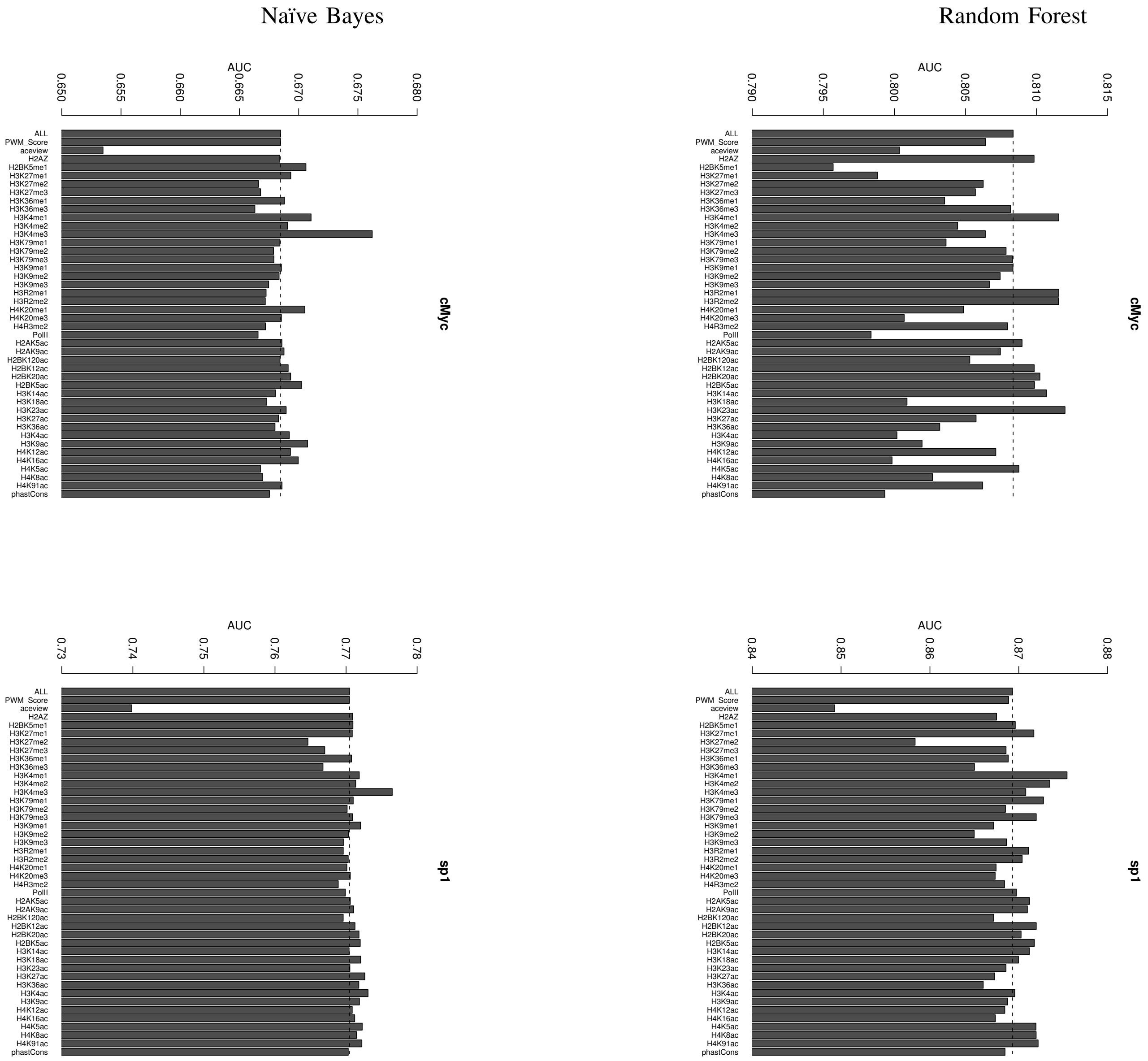

Fig. 3. Results shown for transcription factors cMyc and sp1. Otherwise the same as figure 2

C. Future Work

1) Linear SVMs: LIBLINEAR [29] is a support vector machine (SVM) implementation specialized for linear kernels. In its most recent release [30], it is now fast enough to handle datasets of the size used in this study. SVMs often outperform other classifiers in cross-validation, but this is not necessarily true when the kernel is constrained to be linear. Thus it would be interesting to see if a linear SVM could outperform the random forest classifier used here.

2) Evolutionary Conservation: We initially hoped that information on evolutionary conservation would be more useful than seen here. However, we note that many alternative methods to phastCons have been proposed to measure the evolutionary conservation of candidate TFBSs [8], [9], and suggest testing some of these alternative strategies in future work.

D. Summary

We have demonstrated that TFBS prediction accuracy can be significantly improved by using machine learning classifiers with histone modification and other auxillary features. Furthermore we showed which features are essential for maximally accurate prediction of each transcription factor.

References

[1] G. Z. Hertz, G. W. Hartzell III, and G. D. Stormo, “Identification of consensus patterns in unaligned DNA sequences known to be functionally related,” *CABIOS*, vol. 6, no. 2, pp. 81–92, 1990.

[2] T. Bailey and C. Elkan, "Unsupervised learning of multiple motifs in biopolymers," *Machine Learning*, vol. 21, pp. 51–80, 1995.

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

[4] E. Wingener, "The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation," *Briefings in Bioinformatics*, vol. 9, no. 4, pp. 326–332, 2008.

[5] G. M.-H. Esperanza Benítez-Bellón and J. Collado-Vides, "Evaluation of thresholds for the detection of binding sites for regulatory proteins in escherichia coli K12 DNA," *Bioinformatics*, vol. 16, no. 1, pp. 16–23, 2000.

[6] M. L. Bulyk, "Computational prediction of transcription-factor binding site locations," *Genome Biology*, vol. 5, p. 201, 2003.

[7] W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory modules," *Nature Review Genetics*, vol. 5, pp. 276–287, 2004.

[8] E. C. E and A. Kel, "Whole genome human / mouse phylogenetic footprinting of potential transcription regulatory signals," in *Pacific Symposium on Biocomputing*, vol. 8, 2003, pp. 291–302.

[9] J. Hawkins, C. Grant, W. S. Noble, and T. L. Bailey, "Assessing phylogenetic motif models for predicting transcription factor binding sites," *Bioinformatics*, vol. 25 ISMB, pp. i339–i347, 2009.

[10] M. C. Frith, M. C. Li, and Z. Weng, "Cluster-Buster: finding dense clusters of motifs in DNA sequences," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3666–3668, 2003.

[11] E. Guccione, F. Martinato, G. Finocchiaro, L. Luzi, L. Tizzoni, V. Dal-lOlio, G. Zardo, C. Nervi, L. Bernard, and B. Amati, "Myc-binding-site recognition in the human genome is determined by chromatin context," *Nature Cell Biology*, vol. 8, no. 7, pp. 764–770, 2006.

[12] ENCODE Project Consortium, "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project," *Nature*, vol. 447, pp. 799–816, 2007.

[13] T. Kouzarides, "Chromatin modifications and their function," *Cell*, vol. 128, no. 4, pp. 693–705, 2007.

[14] T. Whitington, A. C. Perkins, and T. L. Bailey, "High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites," *Nucleic Acid Research*, vol. 37, no. 1, pp. 14–25, 2008.

[15] A. Barski, R. Jothi, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, and K. Zhao, "Chromatin poises miRNA- and protein-coding genes for expression," *19*, vol. 19, pp. 1742–1751, 2009.

[16] L. Narlikar, R. Gordân, and A. Hartemink, "A nucleosome-guided map of transcription factor binding sites in yeast," *PLoS Computational Biology*, vol. 3, no. 11, p. e215, 2007.

[17] X. Wang, Z. Xuan, X. Zhao, Y. Li, and M. Zhang, "High-resolution human core-promoter prediction with CoreBoost_HM," *Genome Research*, vol. 19, no. 2, pp. 266–75, 2009.

[18] K. Won, I. Chepelev, B. Ren, and W. Wang, "Prediction of regulatory elements in mammalian genomes using chromatin signatures," *BMC Bioinformatics*, vol. 9, p. 547, 2008.

[19] K.-J. Won, S. Agarwal, L. Shen, R. Shoemaker, B. Ren, and W. Wang, "An integrated approach to identifying cis-regulatory modules in the human genome," *PLoS ONE*, vol. 4, no. 5, p. e5501, 2009.

[20] A. Valouev, D. S. Johnson, A. Sundquist, C. Medina, E. Anton, S. Batzoglou, R. M. Myers, and A. Sidow, "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data." *Nature Methods*, vol. 5, no. 9, pp. 829–834, 2008.

[21] A. Barski, S. Cuddapah, K. Cui, T. Roh, D. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, "High-resolution profiling of histone methylations in the human genome," *Cell*, vol. 129, no. 4, pp. 823–837, 2007.

[22] Z. Wang, C. Zang, J. Rosenfeld, D. Schones, A. Barski, S. Cuddapah, K. C. K, T. Roh, W. Peng, M. Zhang, and K. Zhao, "Combinatorial patterns of histone acetylations and methylations in the human genome," *Nature Genetics*, vol. 40, no. 7, pp. 897–903, 2008.

[23] D. Thierry-Mieg and J. Thierry-Mieg, "AceView: a comprehensive cDNA-supported gene and transcripts annotation," *Genome Biology*, vol. 7, no. Suppl 1, p. S12, 2006.

[24] A. Siepel, G. Bejerano, J. Pedersen, A. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. Hillier, S. Richards, G. Weinstock, R. Wilson, R.A.Gibbs, W. Kent, W. Miller, and D. Haussler, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Research*, vol. 15, pp. 1034–1050, 2005.

[25] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, pp. 103–137, 1997.

[26] T. L. Bailey and M. Gribskov, "Combining evidence using p-values: application to sequence homology searches," *Bioinformatics*, vol. 14, pp. 48–54, 1998.

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[28] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-wide mapping of in vivo protein-DNA interactions." *Science*, vol. 316, no. 5830, pp. 1497–1502, Jun 2007. [Online]. Available: http://dx.doi.org/10.1126/science.1141319

[29] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[30] "LIBLINEAR – A Library for Large Linear Classification," 2009. [Online]. Available: www.csie.ntu.edu.tw/~cjlin/liblinear/