Bioinformatics, 33(14), 2017, 2097–2105
doi: 10.1093/bioinformatics/btx115
Advance Access Publication Date: 1 March 2017
Original Paper

OXFORD

Sequence analysis

Computational modeling of *in vivo* and *in vitro* protein-DNA interactions by multiple instance learning

Zhen Gao and Jianhua Ruan*

Department of Computer Science, University of Texas at San Antonio, San Antonio, TX 78249, USA

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Received on June 30, 2016; revised on February 4, 2017; editorial decision on February 22, 2017; accepted on February 27, 2017

Abstract

Motivation: The study of transcriptional regulation is still difficult yet fundamental in molecular biology research. While the development of both *in vivo* and *in vitro* profiling techniques have significantly enhanced our knowledge of transcription factor (TF)-DNA interactions, computational models of TF-DNA interactions are relatively simple and may not reveal sufficient biological insight. In particular, supervised learning based models for TF-DNA interactions attempt to map sequence-level features (*k*-mers) to binding event but usually ignore the location of *k*-mers, which can cause data fragmentation and consequently inferior model performance.

Results: Here, we propose a novel algorithm based on the so-called multiple-instance learning (MIL) paradigm. MIL breaks each DNA sequence into multiple overlapping subsequences and models each subsequence separately, therefore implicitly takes into consideration binding site locations, resulting in both higher accuracy and better interpretability of the models. The result from both *in vivo* and *in vitro* TF-DNA interaction data show that our approach significantly outperform conventional single-instance learning based algorithms. Importantly, the models learned from *in vitro* data using our approach can predict *in vivo* binding with very good accuracy. In addition, the location information obtained by our method provides additional insight for motif finding results from ChIP-Seq data. Finally, our approach can be easily combined with other state-of-the-art TF-DNA interaction modeling methods.

Availability and Implementation: http://www.cs.utsa.edu/~jruan/MIL/

Contact: jianhua.ruan@utsa.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Modeling transcription factor (TF) binding to DNA is a fundamental yet challenging step towards deciphering transcriptional regulatory networks. While *in vivo* TF-DNA interactions can be profiled at whole-genome scale by chromatin immunoprecipitation followed by sequencing (ChIP-Seq), the sequencing reads obtained from ChIP-Seq do not precisely represent transcription factor binding sites (TFBS) due to the low resolution of ChIP-Seq experiments and the dynamic nature of TF-DNA interactions involving co-factors and chromatin architecture (Jothi *et al.*, 2008). On the other hand, protein-binding microarray (PBM) can measure *in vitro* binding of a transcription factor to an exhaustive enumeration of short nucleic acid sequences arranged on a probe array (Mukherjee *et al.*, 2004). Since common confounding factors present in ChIP-based experiments are eliminated, PBM data delivers an excellent information source to develop TFBS models in a direct manner. However, it is unknown whether models learned from *in vitro* data can be used to accurately predict *in vivo* TF-DNA binding.

Despite the vast amount of *in vivo* and *in vitro* TF-DNA interaction data accumulated, modeling methods have remained

relatively simple and suffer from low modeling accuracy and/or inability to reveal the underlying biophysical mechanism of binding. Most computational approaches to modeling TF-DNA interactions can be categorized as supervised or unsupervised. Given a set of DNA sequences bound by a certain TF, unsupervised approaches attempt to identify common subsequence patterns as TFBS, often represented by a position-specific weight matrix (PWM) (Berg and von Hippel, 1987; Foat *et al.*, 2006; Stormo, 1990), or a consensus sequence (Badis *et al.*, 2009; Zhao *et al.*, 2005). While easy to interpret and theoretically well connected to thermodynamic models of TF-DNA interactions (Stormo, 1990), the unsupervised model usually makes strong assumption about the nature of TF-DNA interaction, such as the fixed length of the binding site and the independent contribution from neighboring nucleotides, which makes modeling of certain TFBS difficult (e.g. those with variable lengths or structural motifs) (Badis *et al.*, 2009; Maerkl and Quake, 2007; Nutiu *et al.*, 2011; Zhao *et al.*, 2005). In contrast, supervised methods do not assume the existence of a common consensus, and instead attempt to use classification or regression models that rely on multiple (sometimes thousands of) features to separate positive sequences (TF-bound sequences) from negative sequences (non-bound sequences). Most commonly used features are oligonucleotide sequences up to a certain length (k-mers) (Bussemaker *et al.*, 2001; Conlon *et al.*, 2003; Roven and Bussemaker, 2003; Weirauch *et al.*, 2013), but other types of features including epigenetic modifications or structural properties of DNA can also be used (Bauer *et al.*, 2010; Cuellar-Partida *et al.*, 2012; Boyle *et al.*, 2011; Gao and Ruan, 2015; Pique-Regi *et al.*, 2011; Siwo *et al.*, 2016; Zhang *et al.*, 2015). Not surprisingly, supervised models are usually more accurate at predicting TF-DNA interactions (Weirauch *et al.*, 2013). In addition, the model is not limited by the common assumptions associated with consensus or PWM-based methods and therefore can be used for different types of TF-DNA interactions (Weirauch *et al.*, 2013). On the other hand, these models often do not provide enough information to generate a consensus motif or PWM for visualization or biological interpretation.

In these conventional TF-DNA interaction models (supervised or unsupervised), locational information of candidate binding sites is usually ignored; as a result, longer binding sites models are usually preferred as they have higher information content (hence they are less likely to appear in the genome bymotif chance). On the other hand, longer models will result in larger search space for optimization in the case of non-supervised models, and larger feature space in the case of supervised models. In particular, when k-mer frequencies were used as features in supervised models, longer k-mer representation can result in the so-called data fragmentation problem (Pagallo and Haussler, 1990) as a true binding site can be represented by the union of multiple longer sequences, resulting in poor modeling performance.

In this work, we propose a novel supervised method to model TF-DNA interactions based on the so-called multiple-instance learning (MIL), which overcomes the limits in conventional approaches, and improves the modeling and predicting of both *in vivo* and *in vitro* TF-DNA bindings. MIL was initially proposed in the mid-1990s' to deal with uncertainty in instance labels and has found many noteworthy applications in computational biology and information retrieval (Andrews *et al.*, 2002; Auer, 1997; Gao and Ruan, 2015; Maron and Lozano-Pérez, 1998). For example, we recently proposed a MIL-based method to model *in vitro* TF-DNA interactions using DNA structural features (Gao and Ruan, 2015). In a traditional single-instance learning (SIL) based classification model for TF-DNA interaction, each sequence is called an instance, represented by a feature vector, and is labeled positive or negative based on its TF-binding status. On the contrary, the input of MIL is a set of bags with more than one instance in each bag, but the labels are associated with bags rather than individual instances. A bag with a positive label contains at least one positive instance (whose identity is unknown), while all instances in the negative bag should all have negative labels. The MIL framework fits the TFBS modeling scenario very well, as it is commonly assumed that a DNA sequence that is not bound by a TF (negative sequence) should have no binding site, while a DNA sequence that can be bound by a TF (positive sequence) should contain one or more binding sites of that TF, and, the exact location of the binding site within the sequence is typically not known (although some preference might exist). Therefore, it is fairly intuitive to consider each DNA sequence as a bag, and any subsequence that can be a potential binding site as an instance.

Using MIL, our method models the TF-DNA interaction as a function of k-mer occurrences in different regions of a DNA sequence, and outputs a sub-sequence that is most likely bound by a TF or its co-factors. We test our method on multiple ChIP-Seq and PBM datasets, and compare its performance with SIL-based approaches (Annala *et al.*, 2011). Evaluation results show that our method outperforms the competing SIL-based methods in general, and, improves a state-of-the-art approach notably. In addition, our method provides additional insights for TFBS prediction, and can even be applied to predict *in vivo* TF-DNA bindings using models build on *in vitro* data. Due to the simplicity of our MIL method and the superior performance showed in the evaluation, we believe that many existing methods for predicting TF-DNA interactions (most of which are SIL based) can be combined with the key idea proposed in this paper to achieve a more accurate model.

2 Materials and methods

2.1 *In vivo* and *in vitro* TF-DNA interaction data

2.1.1 ENCODE ChIP-seq data

ChIP-Seq data on 495 human TFs is acquired from ENCODE (Hoffman *et al.*, 2012) in NarrowPeak format, which provides peak location and statistics. For each TF, up to top m ($m \in \{1000, 2000, 3000, 5000, 10\,000, 15\,000\}$) peaks with the highest signal values is selected, and 500 bp centered at each peak summit is retrieved as positive samples. For each positive sample (ChIP-Seq peak sequence), two corresponding control samples (negative sequences) of the same length (500 bp) are generated—one is generated randomly using a second order Markov model, and the other is obtained by shifting the location of the peak sequence on the genome by 5000 bp. We focus on four human embryonic stem cell TFs (Gata2, Gata3, Mef2C and Nanog) and select top 3000 peaks of each TF for in-detailed analysis, unless otherwise specified.

2.1.2 Protein-binding microarray data

The *in vitro* protein-binding microarray (PBM) data is obtained from Weirauch *et al.* (2013). Two entirely different array designs, each of which contains 40 000 unique 35 bp probe sequences, are used for eighty-six mouse TFs. In these arrays, all patterns of 10-mer, and 32 copies of every non-palindromic 8-mer are included, giving an impartial analysis of TF binding preferences. This dataset has been used in the DREAM (Dialogue for Reverse Engineering Assessments and Methods) competition (Stolovitzky *et al.*, 2007) and is accessible on the DREAM5 competition website. In our study, we select 3000 probes with the highest binding signals as positive samples (sequences) and 3000 probes with the lowest binding

signals as control samples (negative sequences) for each of the eighty-six TFs. Data from the two different designs of array are used for training and testing in turn.

2.1.3 Cebpb ChIP-seq time series data in the mouse liver regeneration

To test if the *in vitro* models learned from PBM data can be used to predict *in vivo* binding, Cebpb ChIP-Seq data is obtained from (Jakobsen *et al.*, 2013), which is a time series data in the process of liver regeneration elicited by partial liver hepatectomy in mouse. The liver tissue is gathered at eight time points (0, 3, 8, 16, 24, 36, 48 and 168 h) covering several cell stages. To choose positive samples, peaks with score less than 150 are filtered out and we retrieve 500 bp centered at each peak summit as positive samples. The number of peaks per time point ranges from 1458 to 4150, and the median number of peaks is 1847. Similar as for human ENCODE ChIP-Seq data, for each positive sample, two corresponding control sequences with the same length (500 bp) are generated—one randomly generated using a first order Markov model and the other by shifting the location of the peak sequence on the genome by 5000 bp.

2.2 Multiple-instance learning model for TF-DNA binding

SIL-based supervised learner deals with instances that consist of feature vector-label pairs, and, the modeling task is to abstract useful information to map the feature vectors to labels. A crucial difference between MIL and SIL is that, in MIL, labels are associated with bags rather than instances, and each bag contains more than one instance. While the instances in the same bag do not have their own labels, depending on the MIL algorithm, it is assumed that they will contribute to the label of the bag in a probabilistic or binary manner.

The MIL algorithm we use in this study is 'wrapper' (Frank and Xu, 2003)—a simple MIL algorithm that allows us to compare between MIL-based models and SIL-based models in a relatively unbiased manner, since the wrapper approach can freely use any SIL-based learners as base learners. In contrast, a non-wrapper based MIL algorithm does not use SIL learners directly, which makes it hard to underpin the source of the performance difference between the two.

In the task of modeling TF-DNA binding, it is reasonable to assume that a binding sequence with high binding signal value or affinity contains at least one 'binding region'—a subsequence that contains TFBS(s), while a non-binding sequence with low affinity does not have any binding regions. Therefore, we consider each subsequence (with length c) as an instance, and, all possible subsequences (with starting point shift s) in the whole sequence (with length l) as a bag. Bags are labeled positive or negative according to the binding signal value (or affinity). Each instance is then mapped to a feature vector, such as k-mer appearance. The number of instances per bag n, is therefore $\lceil (l-c)/s \rceil + 1$, where c and s are the two parameters of our MIL model. The wrapper-based approach essentially converts the learning problem of MIL to a SIL-based learning problem. At the training step, each instance is initially assigned the label of the bag that they belong to, and is assigned a weight proportional to the inverse of the size of its bag. Any conventional classification algorithm that can handle weighted instances can then be applied to learn a model using these instances. (Optionally, the labels and weights can be adjusted iteratively in the learning process. However, we find that doing so does not improve the evaluation results.) In prediction time, the wrapper algorithm simply combines (averages) the predicted scores for all instances in the same bag to give the score of a bag. Besides the predicted bag scores, our application also outputs the predicted instance scores, and the top subsequences corresponding to the top scoring instances.

In order to find a suitable base classifier, we test logistic regression (Le Cessie and Van Houwelingen, 1992), SVM (Keerthi *et al.*, 2001; Platt, 1999), C4.5 decision tree (Quinlan, 1993), random forest (Breiman, 2001) and linear regression (Hellevik, 2009) for this 0/1 dichotomy dependent variable modeling task (on multiple datasets including both *in vivo* and *in vitro* data). Surprisingly, the experimental result shows that linear regression outperforms all the other classifiers significantly except SVM, which has about the same accuracy as linear regression but runs much slower. While traditionally logistic regression is preferred over linear regression for dichotomy dependent variable modeling task, it has been shown recently that in most practical cases linear regression can do as well as or even better than logistic regression, especially in high-dimensional data (Hellevik, 2009). Therefore, we choose linear regression as the base classifier in this work. In addition, this choice of base classifier facilitates a fair evaluation as one of the competing methods, TeamD (see below), is also based on linear regression.

We have implemented a workflow which takes as input the positive and negative sequences, and outputs the top subsequences (instances) for the positive sequences. This is available on our website (http://www.cs.utsa.edu/~jruan/MIL/). The package also includes utility functions to generate motif models, perform k-fold cross-validation evaluation, and output AUCs and instance scores. The trained models on 86 mouse PBM data and 495 ENCODE data are also provided.

2.3 Competing methods

The competing algorithms include TeamD and a widely used consensus-based pattern-driven TFBS modeling approach. Both algorithms are SIL-based.

TeamD (Annala *et al.*, 2011) is known for its best performance in predicting PBM data of mouse TFs in the fifth DREAM (Dialogue on Reverse Engineering Assessment and Methods) challenge (Weirauch *et al.*, 2013). TeamD is a SIL-based method and uses k-mer appearances as features. Evaluations in this paper uses the same setting of k-mer length as in the original TeamD in DREAM5, where $k = [4, 8] \in \mathbb{Z}$. We have also tried $k = [4, 7] \in \mathbb{Z}$ when the original parameter $k = [4, 8] \in \mathbb{Z}$ fails to complete due to excessive running time. TeamD efficiently selects features by filtering out the 6mer to 8mer features with low variance, and, performs normalization and a few transformation tricks on the data. TeamD uses linear-regression as the base learner. The original SIL-based TeamD is denoted as *SIL-TeamD*.

Another comparison method is a traditional consensus-based pattern-driven TFBS modeling approach with the simple k-mer counting data as features. The k-mer length parameter is a single value rather than a combination of multiple k-mer lengths as in TeamD, and, no data transformation nor feature selection is performed. We denote it as *SIL-counting*.

To convert the two competing methods into their corresponding MIL version, each ChIP-Seq sequence or PBM probe sequence with length l is separated into n overlapping sub-sequences (instances) with length c and with instance shift s (see section 2.2). The instances are then mapped to feature-vectors using the same way as the two SIL-based competing methods. In later discussions, the MIL version of TeamD is denoted as *MIL-TeamD*, and, the MIL version

of the traditional consensus-based approach is denoted as *MIL-counting*.

2.4 Motif finding and co-factor analysis on ChIP-seq data

To identify consensus motifs within the subsequences predicted to be bound by the TF, we use the HOMER suite which includes a *de novo* motif discovery algorithm for finding 8-20 bp motifs in large-scale genomic data such as ChIP-Seq data (Heinz *et al.*, 2010). We choose HOMER over other motif analysis tools for its usage of both positive and negative sequences in motif finding and significance testing, and built-in comparison with existing TFBS.

As the TFBS identified from ChIP-Seq can also belong to other co-factors, we attempt to assign a binding factor for each of the top five motifs (PWMs) identified by HOMER. As HOMER only returns the best match TFBS for each predicted motif, and many of the existing TFBS in the HOMER database are derived from ChIP-Seq data, we use another TFBS database and search tool, Cis-BP (Weirauch *et al.*, 2014), to identify potential binding factors for the motifs found by HOMER. Cis-BP contains binding sites of more than 600 TFs; many of the TFBS in its database are derived from *in vitro* binding assays such as PBM or HT-SELEX, and therefore can reasonably represent *bona fide* DNA pattern recognized by the TF. We rely on the default statistical significance threshold of Cis-BP for motif matching. We pool all possible binding factors for the predicted motifs, and compare the list of possible binding factors with the list of proteins that have physical protein-protein interactions with the ChIP-ed TF. The common TFs between the two lists are considered as potential co-factors of the ChIP-ed TF because (1) they physically interact, and (2) their binding sites are both enriched in the ChIP-Seq peak sequences. Protein-protein interactions for human and mouse are obtained from the STRING database (Szklarczyk *et al.*, 2014).

3 Results and discussion

3.1 MIL outperforms SIL on ChIP-seq data

A comparison between SIL and MIL on the ChIP-Seq data of four individual TFs is shown in Figure 1. Figure 1a compares the results from the TeamD approach with or without MIL. Evaluation is done with 5-fold cross validation and prediction accuracy is measured by the area under the receiver operating characteristic curve (AUC). For the MIL-based model, the instance length c is set to 250, and the instance shift range s is set to 50. (As will be shown later, varying these parameters does not affect the main conclusion of the comparison.)

As shown in Figure 1a, our MIL-TeamD approach outperforms SIL-TeamD significantly for all four TFs (P-value $\leq 10^{-4}$, paired t-test, average AUC 0.914 versus 0.784 with gain +0.129). Supplementary Figure S1 shows the AUC of MIL-TeamD versus SIL-TeamD for 495 ChIP-ed TFs from ENCODE, where MIL-TeamD consistently outperforms SIL-TeamD (AUC = 0.921 versus 0.747, P-value$\leqslant 10^{-90}$, paired t-test).

Figure 1b shows the comparison between SIL-counting and MIL-counting, with the k-mer length setting $k = 5$. As shown, MIL-counting outperforms SIL-counting on all four TFs, with an average AUC gain of 0.077 (P-value$\leqslant 0.02$, paired t-test), confirming that the improvement of MIL over SIL can be generalized to different types of classifiers.

The average run time for a ChIP-Seq dataset is 11 min 10 s (±10 s) by MIL-TeamD and 2 min 40 s (±10 s) by SIL-TeamD on a 2.5 GHz Intel Core i5 16 GB RAM computer. For the counting based methods, it takes 30 s (±3 s) by SIL-counting and 90 s (±5 s) by MIL-counting.

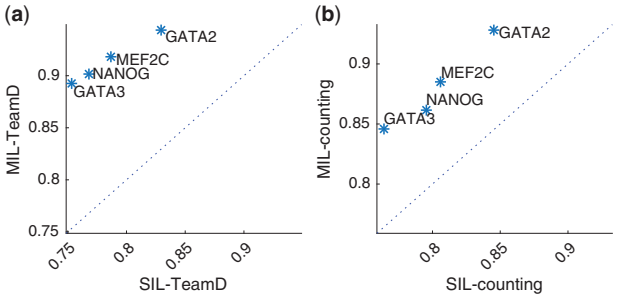

Fig. 1. AUC score comparison between MIL and SIL on ChIP-Seq data of four TFs. (**a**) $k = [4, 8] \in \mathbb{Z}$ for both SIL-TeamD and MIL-TeamD. (**b**) $k = 5$ for both models. For MIL based models, the settings of instance length c is 250 and instance shift range s is 50

Figure 2 shows the effect of the instance length parameter, c, and the shift parameter, s, on classification accuracy. As shown in Figure 2a, both MIL-TeamD and MIL-counting achieve their peak performance when c ranges from 200 to 300 bp. Compared to MIL-counting, the performance of MIL-TeamD is more stable, with comparable accuracy for c as small as 50, and as large as 350. Both MIL methods significantly outperforms their SIL counterparts for almost all values of c. As c approaches 500, the number of instances decreases, so does the prediction accuracy, and eventually the MIL-approach becomes equivalent to the SIL approach when $c = 500$, as only one instance is present. The settings of the shift parameter, s, and the resulting number of instances, n, for each MIL model in Figure 2a are shown in Figure 2b. The value of s is set manually to allow enough overlap between neighboring instances without unnecessary redundancy. For the models with c from 100 to 500, s is set to 50, while for the models with $c \in \{20, 30, 40, 50, 75\}$, s is set to 10. Although n varies in a wide range (from 1 to 49) as s and c varies, AUC of MIL-TeamD does not change significantly and remains high when $n \geqslant 4$ ($c \leqslant 350$). Figure 2c and d shows the parameter analysis of MIL-counting on s when $k = 5$. As can be seen, although n varies in a wide range as s varies from 10 to 50 for each c, AUC only changes slightly, especially for c in the range of 200–300. As shown in Section 2.2, the number of instances, n, is determined by s and c. In general, using smaller c for MIL- counting can cause the feature vector to be very sparse and therefore reduce accuracy; for MIL-TeamD the built-in feature selection seems to be helpful in such cases. Small changes of s do not make a big difference if there is sufficient overlap between neighboring instances so that the real binding sites are contained in at least one instance.

To investigate the impact of the number of positive sequences on model accuracy, we also performed model cross-validation using between 1000 and 15 000 peaks for the four TFs, while keeping the other parameters the same as in Figure 1. As shown in Supplementary Figure S2, AUC of the two MIL-based approaches are robust with respect to the number of positive sequences, while the two SIL-based approaches suffers significantly when the number of sequences used is less than 2000. Repeating the analysis on randomly sampled 100 TFs from the ENCODE dataset shows the same trend (Supplementary Fig. S3). While the highest AUC for the MIL-based approaches is achieved at 2000 sequences, we choose to build our model based on 3000 sequences to have sufficiently diverse sequences for binding sites analysis in order to reveal the complex sequence pattern recognized by the TF and its co-regulators.

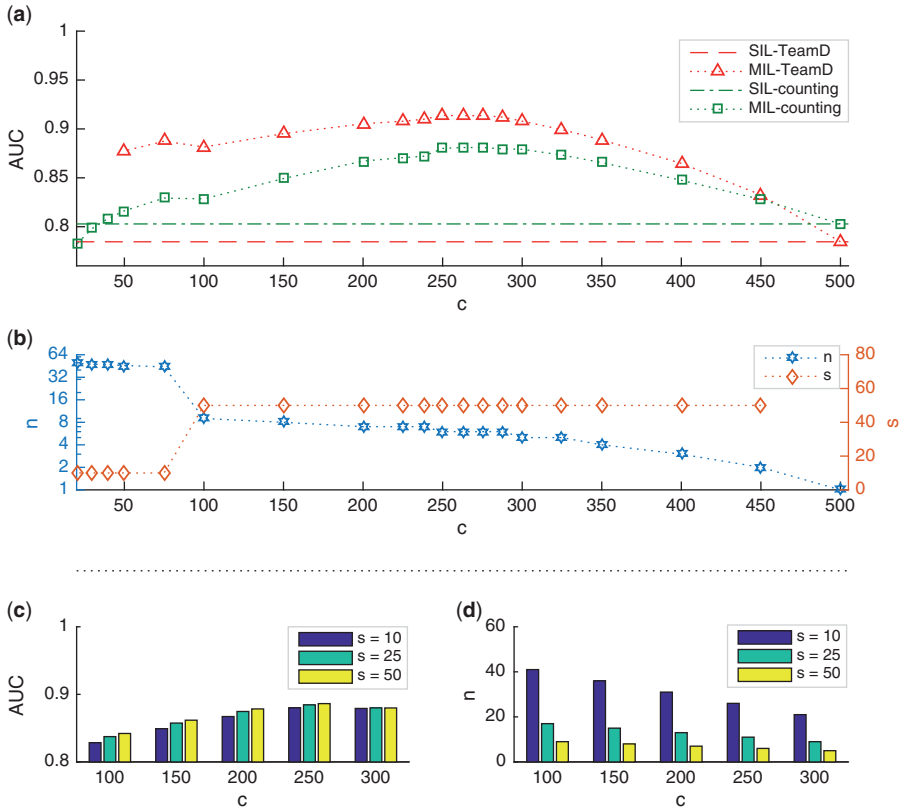

Fig. 2. Effect of instance length *c* and instance shift range *s* on MIL performance. (**a**) AUC varies as a function of *c*. *k*-mer length $k=5$ for SIL-counting and MIL-counting, $k=[4,8]\in\mathbb{Z}$ for SIL-TeamD and MIL-TeamD. (**b**) No. of instances per bag (*n*) varies as *c* and *s* change for the MIL models in (a). For $k\geqslant 100$, *s* is set to be 50, while for $k\leqslant 75$, *s* is set to 10. (**c**) AUC of MIL-counting as a function of *s*. *k* is fixed at 5. (**d**) *n* varies as *s* and *c* change for the MIL-counting model in (c)

Figure 3 shows the influence of *k*-mer length on the performance of both SIL and MIL-based approaches. As shown, for SIL-counting, the best result is obtained with $k=5$ followed by $k=4$ and $k=6$, and much lower for other values of *k*. MIL-counting is always better than SIL-counting for all values of *k*, suggesting that the improvement is independent of the features. For smaller *k*, the poor performance of SIL (and MIL) is due to the low information content of shorter *k*-mers as the expected occurrence of short *k*-mers is high for any *k*-mer even in random sequences. On the other hand, the dramatic decrease of the performance of SIL with larger *k* can be attributed to both overfitting and data fragmentation—as *k* increases, the number of features increases exponentially, and the number of instances containing a particular *k*-mer decreases substantially. Note that SIL-TeamD (which uses all *k*-mers up to $k=5$ and then performs feature selection for *k* between 6 and 8) is even slightly worse than SIL-counting with $k=5$ (AUC = 0.785 and 0.803, respectively), suggesting that the inclusion of longer *k*-mers, even with the use of feature selections, does not provide significant benefit in SIL. In contrast, MIL-counting with $k=6$ and $k=7$ results in much higher AUC than SIL-counting with $k=6$ and $k=7$, respectively, suggesting that overfitting and feature segmentation is of less concern in MIL. Taken together, the results suggest that the performance gain of MIL-TeamD over SIL-TeamD is due to the increased number of training instances in MIL and the localized utilization of both short and long *k*-mers in subsequences (instances) from ChIP-Seq data.

3.2 Location information identified by MIL provides additional insights for motif finding

In vivo TF-DNA interaction is complex and often involves multiple co-regulators. On the other hand, ChIP-Seq experiment has a resolution limit and cannot pinpoint the exact TF binding sites even with sophisticated peak calling. As MIL predicts a score for each instance (subsequence), and the top-scoring instances do not always overlap with the instances centered at the peak, we hypothesize that the predicted high-scoring instances are putative TF binding subsequences, which may contain the binding sites for the TFs or its co-regulators.

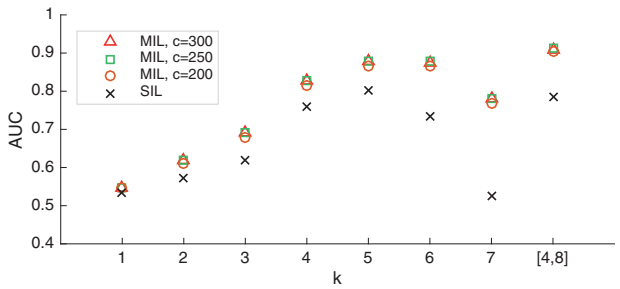

Fig. 3. Effect of *k*-mer length on MIL performance. Instance-shift, *s*, is set to 50bp for all MIL models. Note that MIL with *k* between 1 and 7 refers to MIL-counting while $k=[4,8]$ refers to MIL-TeamD

3.2.1 MIL helps identify in vivo binding locations

In Figure 2a we have shown that MIL achieves its best performance when $c=250$. Here, in order to have enough instances for better fine-tuning of the binding site locations and analysis of binding motifs, we set MIL parameters $c=75$ and $s=5$, which results in comparable AUC but more (and shorter) instances. The number of

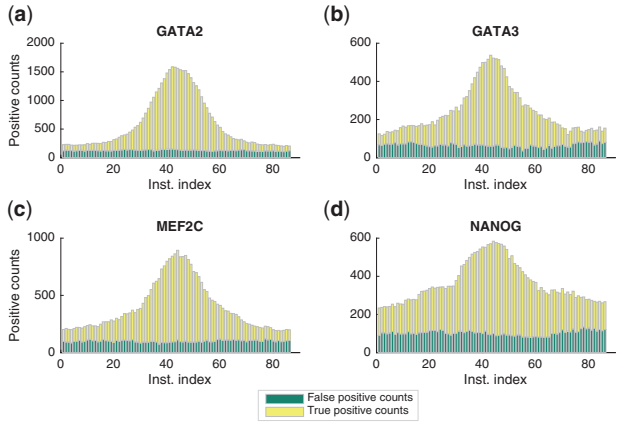

Fig. 4. Number of predicted positive instances (putative TF binding sub-sequences) at each instance location. Here $c=75$, $s=5$

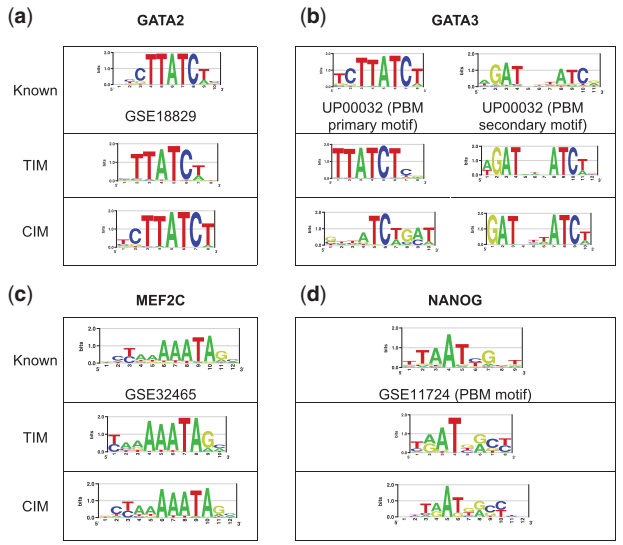

Fig. 5. Comparison between literature motifs and best match top instance motifs ('TIM') and best match center instance motifs ('CIM'). Motifs were identified by HOMER which was separately run on top (TI) and central (CI) sequences

instances per bag is 86 ($\lceil(500-75)/5\rceil+1$). For convenience, we index the instances from 1 to 86. We modify the MIL program to extract the prediction score for each instance, and define an instance as predicted positive if the prediction score is above a certain threshold t. For each instance location, we count the total number of predicted positive instances in the positive and negative sequences, respectively. Here we choose $t=0.6$ as the cutoff for positive predictions; other values of t have also been tested and the results are consistent. Generally, the instances in the positive sequences have a much higher probability of being predicted as positive than those in the negative sequences. Here we denote the high-scoring instances within the 3000 positive sequences as true positive instances, and denote the high-scoring instances within the 6000 negative sequences (3000 random generated negative sequences; 3000 peak-shifting control sequences) as false positive instances. As shown in Figure 4, true positive instances are more likely to be found near the center, which is expected as the positive sequences are obtained from the ChIP-Seq data by retrieving sequences centered on the ChIP-Seq signal peaks. In contrast, false positive instances are always uniformly distributed in all possible locations.

It is worth noting that, while most of the high-scoring instances are located near the center region of the positive sequences, still a significant number of high-scoring instances are located in the non-center region, especially for Gata3 and Nanog, where more than half of the high-scoring instances are located in the non-center region. In addition, the locational distribution of high-scoring instances is not identical to the distribution of ChIP-Seq signal values (Supplementary Fig. S4). For example, Mef2c has a much sharper peak in high-scoring instance location distribution than ChIP-seq value distribution, while for Nanog it is the opposite. Such difference may be due to technical reasons of ChIP-Seq/Peak calling; alternatively, it may suggest that the high-scoring instances contain additional sequence signals for the TF-DNA interaction, such as binding sites for co-factors.

3.2.2 MIL finds known TFBS and co-factor binding sites

To investigate the potential of using MIL-predicted positive instances in identifying TFBS, we perform motif finding on two types of instances separately. For each ChIP-sequence, which consists of 86 overlapping subsequences (instances), the top instance (TI) is defined as the one with the highest prediction score by the MIL model, while the center instance (CI) is simply the 75 bp subsequence at the center of the ChIP-sequence. Supplementary Figure S5 shows the number of top instances at each instance location for the 3000 positive sequences, which shows that the distribution is very similar to the locational distribution of high-scoring instances shown in Figure 4.

HOMER motif finding algorithm (Heinz *et al.*, 2010) was applied to the TI and CI sequences separately—each set has 3000 positive sequences and the length of each sequence is 75. In addition, 3000 negative sequences for HOMER are randomly sampled within the 6000 negative bags (3000 randomly generated and 3000 peak-shifted). We also applied HOMER on a revised version of TI—by simply removing all the TIs that overlaps at least 25 bp with CI. We denote this set of revised TI as TFI (Top Flanking Instance).

A comparison of the known motifs for each ChIP-ed TF and the best matched TFBS found by HOMER from TI and CI is shown in Figure 5. The top 5 motifs predicted by HOMER on TFI, TI and CI sequences as well as the best matched motif for each predicted motif are shown in Supplementary Figure S6. (Complete motif analysis results by HOMER are available at http://www.cs.utsa.edu/~jruan/MIL/.) Since TI sequences and CI sequences only have partial overlaps (Supplementary Fig. S5), we expect the motifs found from CI and TI to be different. Note that for Gata3 there are two known motifs derived from PBM data (Hume *et al.*, 2015), where the primary motif is similar to the binding site of Gata2, and the secondary motif is significantly different. As shown in Figure 5, all known motifs of the four TFs can be found on TI, including both the primary and secondary motifs for Gata3. In contrast, on CI, HOMER cannot find an exact match to the Gata3 primary motif. Interestingly, using the TFI sequences, HOMER successfully identified the primary motif but not the secondary motif. Furthermore, for GATA2 and Mef2c, despite the large extent of overlap between the TI and CI sequences, HOMER identified the known motifs from the small number of TFI sequences.

To investigate whether some of the predicted motifs that do not match to the canonical PWM of the ChIPed TF can be the binding sites of possible co-factors, we map these motifs to the binding sites of TFs that physically interact with the ChIPed TF; these TFs are then predicted as putative co-factors of the ChIPed TF (see Section 2.4). As shown in Figure 6, using TI we are able to identify

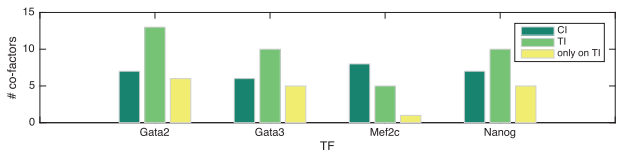

Fig. 6. Number of putative co-factors found on CI and TI

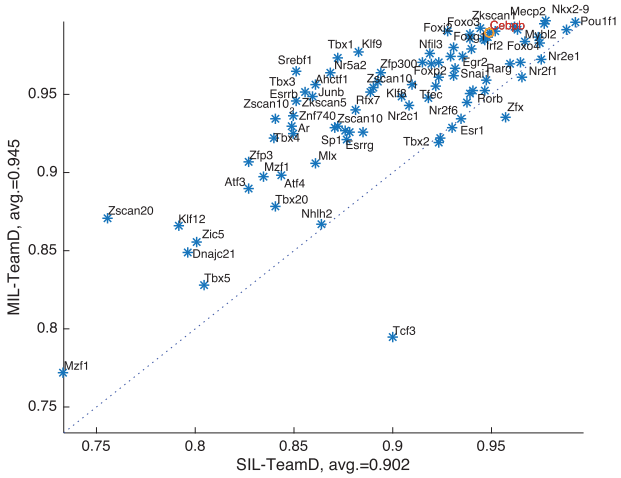

Fig. 7. AUC comparison between SIL-TeamD and MIL-TeamD on PBM data for eighty-six mouse TFs. *x*-axis shows the predicted AUC by SIL-TeamD, and the *y*-axis shows the predicted AUC by MIL-TeamD with $c=8$, $s=2$, $k=[4,7]$. The highlighted TF Cebpb will be discussed in the next section about using *in vitro* model to predict *in vivo* bindings

significantly more putative co-factors for Gata2, Gata3 and Nanog than using CI. Supplementary Table S1 gives the names of the potential co-factors predicted from the CI, TI and TFI sequences, which shows that almost all co-factors found on CI can also be found on TI, but TI/TFI often suggest new co-factors not found on CI. For example, on Nanog both CI and TI contain binding sites for the well-known Sox family transcription factors, but TI (as well as TFI) also predicts binding sites for Kruppel-Like Factor family, which is known to play important roles in the self-renewal of mouse embryonic stem cells and cellular reprogramming (Jeon *et al.*, 2016; Schmidt and Plath, 2012), and physically interact with Nanog (Szklarczyk *et al.*, 2014).

3.3 MIL outperforms SIL on PBM data

In vitro experiments, such as PBM, measure the TF-DNA interaction in a more direct manner, compared to *in vivo* experiment such as ChIP-Seq. As a result of the differences in the experiment setup and the sequence length (35 bp for PBM versus a few hundred bp for ChIP-Seq), the parameter setting of MIL on PBM is also different—the instance length of MIL on PBM can be set much shorter (such as 8) and works as a motif model (where the *k*-mer appearance data can be used as feature of the motif model), while SIL does not have an explicit motif model. In addition, the location information of each putative motif can be maintained by the bag-instance setting of MIL automatically.

To investigate the capability of MIL models on *in vitro* data, we apply MIL-TeamD and SIL-TeamD to the PBM data of 86 mouse TFs (see Materials and Methods). A TF by TF comparison between SIL-TeamD and MIL-TeamD is shown in Figure 7. Both SIL- and MIL-TeamD use the setting $k=[4,7]$, $c=8$ and $s=2$. A performance improvement by MIL-TeamD over SIL-TeamD can be observed for all except eight of the 86 TFs. The average AUC for MIL-TeamD is 0.945, while that for SIL-TeamD is 0.902 (P-value $\leq 10^{-7}$, paired t-test). For several TFs that are designated as 'hard to model' (Weirauch *et al.*, 2013), such as Tbx5, Mzf1 and Junb, MIL-TeamD improves the prediction accuracy considerably. The binding sites of Junb are known to be very complex and show non-canonical patterns, including multiple variable-length consensuses such as TGACGT[C/T]A and TGA[G/C]TCA (Li *et al.*, 2011; Wang *et al.*, 2011). The AUC for Junb by MIL-TeamD is 0.949, while it is 0.859 by the SIL-TeamD (+0.090 gain), which shows the remarkable improvement of MIL for modeling complex binding sites over SIL. Note that the model built on Cebpb PBM data will be used for modeling Cebpb ChIP-Seq data in the next section (Cebpb is highlighted in Fig. 7).

3.4 *In vitro* models built by MIL can predict *in vivo* bindings

To test whether the *in vitro* models learned from PBM data by our method can predict *in vivo* binding, we download the time series ChIP-Seq data for the well-known Cebpb TF, in various stages of liver regeneration, and test whether the *in vivo* Cebpb binding can be predicted using the model built from the *in vitro* Cebpb PBM data. The positive and negative sequences are described in section 2.1. At training time, a single SIL-TeamD model is built on the *in vitro* PBM data of Cebpb. The PBM data of Cebpb is the same as which in the last section. To fill the gap between the short probe sequence length of PBM data (35 bp) and the long peak sequence length of ChIP-Seq data (500 bp), for each probe sequence, we map the entire 35-mer to TeamD feature vector rather than separate the 35-mer to multiple pieces, and then build a linear regression model from the feature vectors. At predicting time, SIL-TeamD maps each ChIP-Seq peak sequence to a single feature vector, while MIL-TeamD separates each peak sequence into multiple sub-sequences using the parameters $c=75$ and $s=15$, then each sub-sequence is mapped to a feature vector. The same Cebpb PBM model is then applied to both the SIL-TeamD ChIP-Seq feature vectors and the MIL-TeamD ChIP-Seq feature vectors, and the prediction results of the two models are compared.

As shown in Figure 8, MIL-TeamD outperforms SIL-Team significantly in all time points. While the average AUC for MIL-TeamD (0.626) is still relatively low compared to the cases where the training and testing data are both *in vivo* or *in vitro*, it is much better than random guessing and represents a significant improvement over SIL-TeamD (average AUC again = 0.048, P-value $\leq 10^{-8}$, paired t-test).

3.4.1 Co-factor analysis reveals dynamic interaction between Cebpb and other proteins during liver regeneration

We perform motif analysis on the TI and CI sequences for each time point. Using HOMER motif finding tool, we are not able to find an exact match to the known Cebpb motif from either TI or CI sequences, likely due to the very short consensus of the motif, which consists of only a dinucleotide core (Supplementary Fig. S7). Analysis of the locations of predicted positive instances reveals an almost uniform distribution (Supplementary Fig. S8), suggesting that there is no strong consensus for Cebpb. Scanning the TI and CI sequences suggested a similar enrichment of the known consensus on CI and TI (Supplementary Fig. S9).

On the other hand, HOMER returns many strong motifs that match to the binding sites of other known TFs. We therefore use Cis-BP and STRING to investigate the possible co-factors of Cebpb

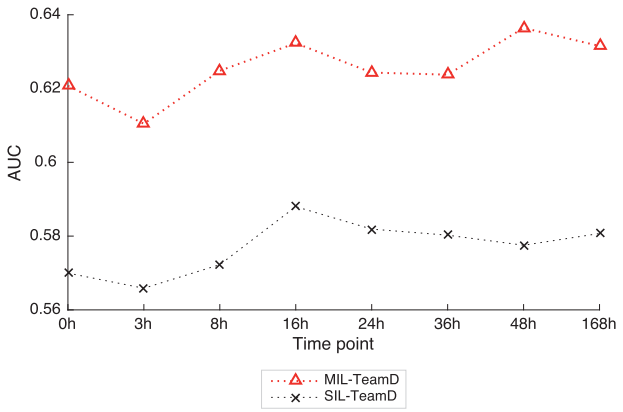

Fig. 8. Comparison between SIL-TeamD and MIL-TeamD using model trained on *in vitro* Cebpb protein binding microarray data to predict *in vivo* Cebpb binding

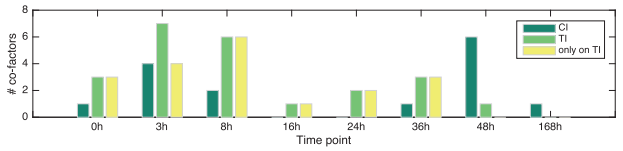

Fig. 9. Number of co-factors found on CI and TI of Cebpb ChIP-Seq data during liver regeneration

(see Section 2.4). Figure 9 shows the number of co-factors identified from CI and TI, as well as the number of co-factors that are found from TI but not CI. The co-factor names are shown in Supplementary Table S2. As shown in Figure 9, TI identifies more co-factors than CI for the first six time points, while CI identifies more co-factors for the last two time points. Interestingly, RelA, a subunit of the NF-kB complex, is identified as a co-factor of Cebpb in either CI or TI, for all the time points except at 0 h. The binding site of RelA is present in TI but not CI from 3 h to 36 h, in both TI and CI at 48 h, and only in CI at 168 h. It is known that Cebpb and RelA form a heteromeric complex as a transcription activator in liver (Ray *et al.*, 1995). The diverse co-factors identified from different time points reveal the dynamic nature of Cebpb and co-factor interaction, and the shift of the RelA binding sites from TI to CI sequences in the time series may suggest that RelA plays an important role in the temporal regulation of the differential recruitment of other co-factors during liver regeneration.

4 Conclusion

In this article, we have proposed a novel TFBS modeling algorithm based on multiple-instance learning. Our algorithm breaks each single DNA sequence into multiple overlapping subsequences and models each subsequence individually, which implicitly takes *k*-mer locations into consideration and can utilize both long and short *k*-mers as features. Performance evaluations on multiple *in vivo*/*in vitro* datasets show that our MIL-based method significantly outperforms SIL-based TFBS modeling algorithms, including a state-of-the-art algorithm. The binding locations predicted by our approach provide more insights for TF-DNA bindings. Moreover, models learned from *in vitro* data using our algorithm can predict *in vivo* binding with high accuracy, further validating our approach. Given the promising evaluation results and the simplicity of the MIL method, we believe that the MIL framework can be combined with other existing algorithms for more accurate modeling of TF-DNA interactions.

Funding

This work was supported in part by grants from the National Science Foundation (award number IIS-1218201, ABI-1565076) and the National Institutes of Health (award number G12MD007591).

Conflict of Interest: none declared.

References

Andrews,S. *et al.* (2002) Support vector machines for multiple-instance learning. In: *Advances in Neural Information Processing Systems*, MIT Press Cambridge, MA, pp. 561–568.

Annala,M. *et al.* (2011) A linear model for transcription factor binding affinity prediction in protein binding microarrays. *PloS One*, **6**, e20059.

Auer,P. (1997) On learning from multi-instance examples: Empirical evaluation of a theoretical approach. *ICML*, **97**, 21–29.

Badis,G. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.

Bauer,A.L. *et al.* (2010) Using sequence-specific chemical and structural properties of DNA to predict transcription factor binding sites. *PLoS Comput. Biol.*, **6**,

Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–743.

Boyle,A.P. *et al.* (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.*, **21**, 456–464.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Bussemaker,H.J. *et al.* (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–174.

Conlon,E.M. *et al.* (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 3339–3344.

Cuellar-Partida,G. *et al.* (2012) Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, **28**, 56–62.

Foat,B.C. *et al.* (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by matrixreduce. *Bioinformatics*, **22**, e141–e149.

Frank,E. and Xu,X. (2003) *Applying propositional learning algorithms to multi-instance data*. Technical Report, University of Waikato, Department of Computer Science.

Gao,Z. and Ruan,J. (2015) A structure-based multiple-instance learning approach to predicting in vitro transcription factor-DNA interaction. *BMC Genomics*, **16**, S3.

Heinz,S. *et al.* (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol. Cell*, **38**, 576–589.

Hellevik,O. (2009) Linear versus logistic regression when the dependent variable is a dichotomy. *Qual. Quan.*, **43**, 59–74.

Hoffman,M.M. *et al.* (2012) Integrative annotation of chromatin elements from encode data. *Nucleic Acids Res.*, gks1284.

Hume,M.A. *et al.* (2015) Uniprobe, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *J. Biol. Chem.*, **43**, (Database issue), D117–D122.

Jakobsen,J.S. *et al.* (2013) Temporal mapping of cebpa and cebpb binding during liver regeneration reveals dynamic occupancy and specific regulatory codes for homeostatic and cell cycle gene batteries. *Genome Res.*, **23**, 592–603.

Jeon,H. *et al.* (2016) Comprehensive identification of kruppel-like factor family members contributing to the self-renewal of mouse embryonic stem cells and cellular reprogramming. *PLoS ONE*, **11**, e0150715.

Jothi,R. *et al.* (2008) Genome-wide identification of in vivo protein–DNA binding sites from chip-seq data. *Nucleic Acids Res.*, **36**, 5221–5231.

Keerthi,S.S. *et al.* (2001) Improvements to platt's smo algorithm for SVM classifier design. *Neural Comput.*, **13**, 637–649.

Le Cessie,S. and Van Houwelingen,J.C. (1992) Ridge estimators in logistic regression. *Appl. Stat.*, **41**, 191–201.

Li,M. *et al.* (2011) c-jun binding site identification in k562 cells. *J. Genet. Genomics*, **38**, 235–242.

Maerkl,S.J. and Quake,S.R. (2007) A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, **315**, 233–237.

Maron,O. and Lozano-Pérez,T. (1998) A framework for multiple-instance learning. *Adv. Neural Inf. Process. Syst.*, MIT Press Cambridge, MA, pp. 570–576.

Mukherjee,S. *et al.* (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, **36**, 1331–1339.

Nutiu,R. *et al.* (2011) Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat. Biotechnol.*, **29**, 659–664.

Pagallo,G. and Haussler,D. (1990) Boolean feature discovery in empirical learning. *Mach. Learn.*, **5**, 71–99.

Pique-Regi,R. *et al.* (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.

Platt,J.C. (1999) 12 fast training of support vector machines using sequential minimal optimization. *Adv. Kernel Methods*, 185–208.

Quinlan,J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Ray,A. *et al.* (1995) Concerted participation of nf-kappa b and c/ebp heteromer in lipopolysaccharide induction of serum amyloid a gene expression in liver. *J. Biol. Chem.*, **270**, 7365–7374.

Roven,C. and Bussemaker,H.J. (2003) Reduce: an online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. *Nucleic Acids Res.*, **31**, 3487–3490.

Schmidt,R. and Plath,K. (2012) The roles of the reprogramming factors oct4, sox2 and klf4 in resetting the somatic cell epigenome during induced pluripotent stem cell generation. *Genome Biol.*, **13**, 251.

Siwo,G. *et al.* (2016) Prediction of fine-tuned promoter activity from DNA sequence. *F1000Research*, **5**, (158.

Stolovitzky,G. *et al.* (2007) Dialogue on reverse-engineering assessment and methods. *Ann. N. Y. Acad. Sci.*, **1115**, 1–22.

Stormo,G.D. (1990) [13] consensus patterns in DNA. *Methods Enzymol.*, **183**, 211–221.

Szklarczyk,D. *et al.* (2014) String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, gku1003.

Wang,W.-M. *et al.* (2011) Binding site specificity and factor redundancy in activator protein-1-driven human papillomavirus chromatin-dependent transcription. *J. Biol. Chem.*, **286**, 40974–40986.

Weirauch,M.T. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.

Weirauch,M.T. *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.

Zhang,Y. *et al.* (2015) Most+: a de novo motif finding approach combining genomic sequence and heterogeneous genome-wide signatures. *BMC Genomics*, **16**, 1.

Zhao,X. *et al.* (2005) Finding short DNA motifs using permuted Markov models. *J. Comput. Biol.*, **12**, 894–906.