A Comparison Study for DNA Motif Modeling on Protein Binding Microarray

A Comparison Study for DNA Motif Modeling on Protein Binding Microarray

Ka-Chun Wong, Yue Li, Chengbin Peng, Hau-San Wong

***Abstract*—Transcription Factor Binding Sites (TFBSs) are relatively short (5-15 bp) and degenerate. Identifying them is a computationally challenging task. In particular, Protein Binding Microarray (PBM) is a high-throughput platform that can measure the DNA binding preference of a protein in a comprehensive and unbiased manner; for instance, a typical PBM experiment can measure binding signal intensities of a protein to all possible DNA k-mers (k=8∼10). Since proteins can often bind to DNA with different binding intensities, one of the major challenges is to build motif models which can fully capture the quantitative binding affinity data.**

To learn DNA motif models from the non-convex objective function landscape, several optimization methods are compared and applied to the PBM motif model building problem. In particular, representative methods from different optimization paradigms have been chosen for modeling performance comparison on hundreds of PBM datasets. The results suggest that the multimodal optimization methods are very effective for capturing the binding preference information from PBM data. In particular, we observe a general performance improvement using di-nucleotide modeling over mono-nucleotide modeling. In addition, the models learned by the best-performing method are applied to two independent applications: PBM probe rotation testing and ChIP-Seq peak sequence prediction, demonstrating its biological applicability.

***Index Terms*—Transcription Factor Binding Site, Genetic Algorithm, Crowding, Ranking, Protein Binding Microarray**

I. Introduction

THE DNA binding of various modulatory transcription factors (TF) onto cis-regulatory DNA elements near genes in human and other eukaryotes is one of the gene regulation mechanism. Binding of different combinations of TFs may result in a gene being expressed in different tissues or at different developmental stages. To fully understand a gene's function, it is essential to identify the TFs that regulate the gene and the corresponding TF binding sites (TFBS). Traditionally, these regulatory sites were determined by labor-intensive experiments such as DNA footprinting or gel-shift assays. Various computational approaches have been developed to predict TF binding sites *in silico*. Detailed comparisons can be found in the survey by Tompa et al. [1]. TFBS are relatively short (5-15 bp) and highly degenerate sequence motifs, which makes their effective identification a computationally challenging task. A number of high-throughput experimental technologies were developed recently to determine protein-DNA binding such as protein binding microarray (PBM) [2], chromatin immunoprecipitation (ChIP) followed by microarray or sequencing (ChIP-Chip or ChIP-Seq) [3], [4], microfluidic affinity analysis [5], and protein microarray assays [6], [7]. In contrast, unfortunately, it is still difficult and time-consuming to extract the high-resolution 3D protein-DNA (e.g. TF-TFBS) complex structures with X-Ray Crystallography [8] or Nuclear Magnetic Resonance (NMR) spectroscopic analysis [9].

The technology of Chromatin immunoprecipitation (ChIP) followed by microarray or sequencing (ChIP-Chip [3] and ChIP-Seq [4]) measures the binding occupancy of a particular TF to the nucleotide sequences of co-regulated genes on a genome-wide basis *in vivo*, but at low resolution. Further processing are needed to extract precise TFBSs [10]. On the other hand, *in vitro* techniques such as protein binding microarray (PBM) [2], microfluidic affinity analysis [5], and protein microarray assays [6], [7] enable us to measure the DNA sequence binding of TFs *in vitro* completely. In particular, the protein binding microarray (PBM) was developed to measure the binding preference of a protein to a complete set of k-mers *in vitro* [2]. The PBM data resolution is unprecedentedly high, comparing with the other traditional techniques. It has also been shown to be largely consistent with those generated by *in vivo* genome-wide location analysis (ChIP-Chip and ChIP-Seq) [2]. As a result, researchers have applied this technique onto many transcription factors, and a large amount of PBM data has been being accumulated and deposited to the UniProbe database [11].

In recent years, the Encyclopedia of DNA Elements (ENCODE) project has revealed genome-wide TFBS locations on human genomes in 2011 [12]. The drastically decreasing cost of sequencing enables the 1000 Genomes Project to be completed, resulting in an integrated map of genetic variation from 1,092 human genomes published in 2012 [13]. Those massive genomic data call for accurate DNA motif modeling techniques in TFBS sequence pattern recognition.

Traditional approaches that rely on cut-offs are no longer adequate. Robust and probabilistic methods were developed to take into account those quantitative affinity data. In light of that, Seed and Wobble has been proposed as a seed-based approach using rank statistics [2]. RankMotif++ was proposed to maximize the log likelihood of their probabilistic model of binding preferences [14]. MatrixREDUCE was proposed to

K.C. Wong is affiliated with Department of Computer Science, City University of Hong Kong, Hong Kong. e-mail: kc.w@cityu.edu.hk. Y. Li is affiliated with Computer Science and Artificial Intelligence Laboratory (CSAIL) at Massachusetts Institute of Technology, Cambridge, MA 02139, United States of America C. Peng is affiliated with Extreme Computing Research Center, King Abdullah University of Science and Technology, Jeddah, Thuwal, Saudi Arabia. H.S. Wong is affiliated with Department of Computer Science, City University of Hong Kong, Hong Kong.

Manuscript received April 19, 2099; revised December 27, 2099.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2443782, IEEE/ACM Transactions on Computational Biology and Bioinformatics

perform forward variable selections to minimize the sum of squared deviations [15]. MDScan was proposed to combine two search strategies together, namely word enumeration and position-specific weight matrix updating [10]. PREGO was proposed to maximize the Spearman rank correlation between the predicted and the actual binding intensities of ChIP-Chip data [16]. Herd clustering was proposed for multiple TFBS motif elucidation on PBM data [17].

Given a set of DNA sequences, PBM can be used to measure their binding signal intensities for a given DNA-binding protein. Specifically, each probe sequence is associated with a normalized signal intensity value. The higher the normalized signal intensity, the stronger is the binding preference of the DNA-binding protein to the corresponding probe sequences. It should be noted that the actual mathematical relationship between the real binding affinity and the normalized signal intensity is unknown since it still depends on specific experimental settings [14]. Given such data, our goal is to uncover a motif model which can summarize and represent the DNA binding preference of the DNA-binding protein. The most common motif models are matrix models which assume independence between adjacent motif positions, justified by the experimental and theoretical statistical mechanical study [18]. Although a recent attempt has been made to generalize matrix models, the insertion and deletion operations between adjacent nucleotide positions are still challenging [19]. In this work, we describe our efforts in comparing different approaches to learn DNA motif models for ranking motif instances on PBM data.

II. Problem Description

For each PBM dataset, we are given a set of DNA sequences $\{seq_1, seq_2, ..., seq_n\}$ and the corresponding normalized signal intensity values $\{I_1, I_2, ..., I_n\}$ (e.g. Array #1). Following the PBM data analysis convention, we refer to such type of input dataset as an array in this manuscript. To extract informative motif data, a sliding window of length k is used to scan each DNA sequence (and its reverse complement) in order to count and record the normalized signal intensity values for each k-mer. Once all the DNA sequences are scanned, a list of normalized signal intensity values is obtained for each k-mer that is present in those DNA sequences. The median of the list is calculated as the median signal intensity μ_m for each k-mer s_m. Among those k-mers, some are motif instances (positive k-mers) while the others are just background k-mers. Robust estimate procedures proposed in RankMotif++ [14] can then be applied to learn the positive k-mers. Nonetheless, it has been pointed out that such a robust procedure may not be suitable for all proteins [20]. In light of that, the highly ranked k-mers are regarded as the positive k-mers in this study.

After a set of positive k-mers were selected, they are aligned using a multiple sequence alignment method. The aligned k-mers are then input for training a motif matrix model to represent the binding preferences of the DNA-binding protein of interest, using evolutionary algorithms.

In this study, we aim at evolving motif models which can truly capture the information from PBM data, reflecting the true binding sequence preferences of DNA-binding proteins. In particular, we seek to evolve matrix models to accurately rank the median signal intensities of positive k-mers. In summary, the problem is formulated as follows:

Input: A set of aligned DNA k-mers with their median signal intensities $D = \{(s_1, \mu_1), (s_2, \mu_2), (s_3, \mu_3), ..., (s_M, \mu_M)\}$ of length L where s_m is the mth aligned DNA k-mer with its median signal intensity μ_m. Each aligned DNA k-mer s_m can be represented as $s_m = s_{m1}s_{m2}...s_{mL}$ where s_{mp} is the p-th nucleotide of s_m:

$$s_{mp} \in \{A, C, G, T, -\}$$
$$\forall m \in \{1, 2, ..., M\}, \forall p \in \{1, 2, ..., L\}$$

Output: A matrix model Θ trained to represent the input aligned k-mers D such that the objective function, Spearman rank correlation coefficient between the predicted scores $\{S_\Theta(s_m), \forall m \in \{1, 2, ..., M\}\}$ and the actual median binding intensities$\{\mu_m, \forall m \in \{1, 2, ..., M\}\}$ of the input aligned k-mers D, is maximized:

$$\arg\max_{\Theta} f(\Theta, D) = \frac{\sum_{m=1}^{M}((X_m - \bar{X})(Y_m - \bar{Y}))}{\sqrt{\sum_{m=1}^{M}(X_m - \bar{X})^2(Y_m - \bar{Y})^2}}$$

$s.t.$

$$\sum_{i \in \{A,C,G,T,-\}} \Theta_{ij} = 1 \ \forall j \in \{1, 2, ..., L\}$$
$$0 \leq \Theta_{ij} \leq 1 \ \forall i \in \{A, C, G, T, -\}, \ \forall j \in \{1, 2, ..., L\}$$

where X_m is the rank of $S_\Theta(s_m)$ and Y_m is the rank of μ_m. $\bar{X}$ and $\bar{Y}$ are the average ranks and the function $S_\Theta(s_m)$ is defined as follows:

$$S_\Theta(s_m) = log\frac{\prod_{j=1}^{L}\prod_{i=\{A,C,G,T,-\}}\Theta_{ij}^{[s_{mj}=i]}}{\prod_{j=1}^{L}\prod_{i=\{A,C,G,T,-\}}B_i^{[s_{mj}=i]}}$$

where B_i is the occurring fraction of the ith nucleotide in all the background sequences [21].

III. Methodology

We note that past literature usually focus on recognizing motif consensus patterns which is not our major theme here. Our main focus is to learn models to fit regressions on the ranks of the top k-mers from PBM. To solve the problem, we apply and compare different optimization methods on PBM benchmark datasets.

In this section, we briefly review and describe the optimization methods which we have applied to the proposed problem (See "Problem Description"), including interior point method [22], genetic algorithm [23], differential evolution [24], crowding genetic algorithm [25], and crowding differential evolution [25]. Those methods are selected to represent different algorithmic paradigms. For instance, interior point method represents the line search optimization paradigm [26]; genetic algorithm represents the nature-inspired optimization paradigm [27]; differential evolution represents the stochastic

beam search optimization paradigm [28]; crowding differential evolution and crowding genetic algorithm represent the multimodal optimization paradigm [29].

A. Interior Point Method (IPM)

Assuming function convexity, line search numerical optimization techniques have been proved successful in different applications , for example, gradient descent and Newton's method [26]. In particular, the Interior Point Method (IPM) is considered as the state-of-the-arts numerical optimization technique to optimize nonlinear functions with linear constraints [22]. Briefly, instead of traversing the surface of feasible regions, it traverses from the interior points of feasible regions to reach an optimal solution. Such a strategy enables the interior point method to handle multiple sparse linear constraints effectively, making it suitable for this study since the proposed porblem here also has multiple sparse linear constraints (See "Problem Description").

B. Genetic Algorithm (GA)

Nonetheless, the existing real world problems are seldom convex [23]. To circumvent the issue, drawing inspiration from the nature, Genetic Algorithm (GA) is proposed. Comparing to traditional algorithms, its parallel search capability and stochastic nature enable it to excel in search performance in a unique way [30]. A genetic algorithm usually starts with a randomly initialized population. The population then evolves across several generations. In each generation, fit individuals are selected to become parent individuals. They cross-over with each other to generate new individuals, which are subsequently called offspring individuals. Randomly selected offspring individuals then undergo certain mutations. After that, the algorithm selects the optimal individuals for survival to the next generation according to the survival selection scheme designed in advance. For instance, under the overlapping population scheme [31], both parent and offspring populations participate in the survival selection. Otherwise, only the offspring population will participate in the survival selection. The selected individuals then survive to the next generation. Such a procedure is repeated again until certain termination condition is met [32]. In this study, we follow the unified approach proposed by De Jong to implement GA [31].

C. Differential Evolution (DE)

Differential Evolution (DE) was first proposed by Price and Storn in the 1990s [33]. It demonstrated great potential for real function optimization in the subsequent contests [24]. For each individual in a generation, the algorithm randomly selects three individuals to form a trial vector. One individual forms a base vector, whereas the value difference between the other two individuals forms a difference vector. The sum of those two vectors forms a trial vector, which recombines with the individual to form an offspring. Replacing the typical crossover and mutation operation by this trial vector generation, manual parameter tuning of crossover and mutation is no longer needed. It can provide differential evolution a self-organizing ability and high adaptability for choosing suitable step sizes which demonstrated its potential for continuous optimization in the past contests [27]. A self-organizing ability is granted for moving toward the optima. A high adaptability is achieved for optimizing different landscapes [34]. With such self-adaptability, differential evolution is considered as one of the most powerful evolutionary algorithms for real function optimization. For example, mechanical engineering design [35] and nuclear reactor core design [36].

D. Crowding DE (CDE) and Crowding GA (CGA)

Although the above, people pointed out that most of the existing algorithms can be easily trapped in local optima. To extend their capabilities, Thomsen [25] incorporated crowding techniques [37] into differential evolution (CDE) and genetic algorithm (CGA) for diversity-preserving and multimodal optimization. For all offspring in each generation, they can only replace the most similar individuals. Although an intensive computation is accompanied, it can effectively transform them into new and effective algorithms specialized for multimodal optimization [38]. To determine the dissimilarity (or distance) between two individuals, the dissimilarity measurement proposed by Goldberg and Richardson [39] and Li et al. [40] is adopted. The distance between two individuals is based on their Euclidean distance. The smaller the distance, the more similar they are and vice versa.

IV. Benchmarking

A. Data Sources

We have adopted the PBM datasets from [14]. Specifically, the PBM datasets focus on five proteins of interest. For each protein, we have two array sets of DNA probe sequences, i.e. array $\#1$ and array $\#2$. Each DNA probe sequence on the array is associated with a normalized signal intensity value. The higher the value, the higher is the binding preference of a DNA-binding protein to that DNA sequence. For each DNA-binding protein, the two arrays (data replicates) are both used for performance comparison. In particular, we seek to examine how each method can rank the aligned input k-mers accurately (i.e. Spearman rank correlation coefficient in "Problem Description"). For the sake of completeness, the PBM microarray data provided in the comprehensive mouse PBM dataset repository [41] have also been adopted.

B. Parameter Setting

All methods are implemented in MATLAB codes with the default floating point number representation. Progressive multiple alignment is adopted (MATLAB function: multialign with terminal gaps adjusted option); each pair-wise alignment is done with the NUC44 scoring matrix [42]. After that, pair-wise distances between sequences are computed by counting the proportion of sites at which each pair of sequences are similar and different using NUC44 (ignoring gaps). Assuming equal variance and independence of evolutionary distance estimates, the guide tree is calculated by the neighbor-joining method. For all evolutionary computation methods, population type is overlapping [31]. Population size is set to 50. For

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2443782, IEEE/ACM Transactions on Computational Biology and Bioinformatics

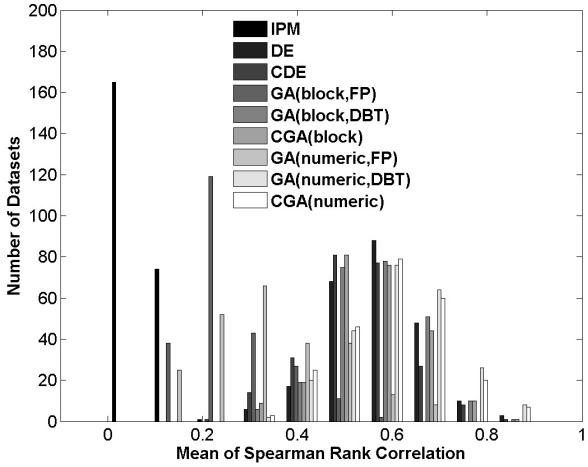
population initialization, half of individuals (i.e. 25 individuals) are set to the top 25 aligned k-mers while the other half are randomly generated. On each dataset, we have run each method 30 times to estimate their overall performance. Regarding about the termination condition, as mentioned in the previous study [38], different algorithms perform different operations in one generation, it is unfair to set the termination condition as the number of generations. Alternatively, it is also unfair to adopt CPU time because it substantially depends on the implementation techniques for different algorithms. For instance, the sorting techniques to find elitists and the programming languages used. In contrast, objective function evaluation is always the performance bottleneck [43] [1]. Thus the termination condition is set to 1000 and 10000 objective function evaluations in this study. Uniform nucleotide background distribution is adopted to compute the fitness function since the k-mers are from PBM. Numerical finite difference method is adopted to approximate gradients for the interior point method. For all genetic algorithm methods, block crossover and intermediate crossover can be chosen. The crossover probability is set to 0.8 while mutation probability is set to 0.05. Gaussian mutation with step size 0.5 is applied [31]. Parent selection is set to stochastic uniform selection. For GA, deterministic binary tournament and roulette-wheel (a.k.a. fitness proportional) selection can be chosen for survival selection; For CGA, crowding selection is used for survival selection. For all differential evolution methods, crossover probability is set to 0.8 while the trial vector coefficient F is set to 0.9. For crowding-based methods, crowding factor is set to the population size, avoiding replacement errors.

C. Performance Comparison

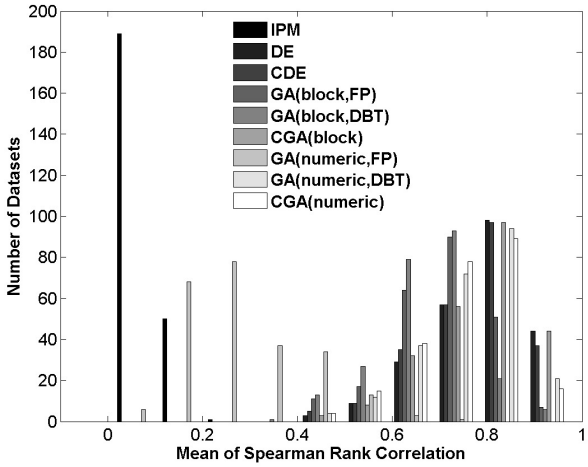
Having set the parameters, we have run each method on each dataset for 30 times. For each run, the best individual model is selected as the final model which is then evaluated its ability to rank the input aligned k-mers. For each dataset, we have calculated the mean and standard deviation of the 30 runs for each method. Especially, a naming scheme is proposed to denote different versions of GA and CGA. GA(block,FP) denotes the GA with block crossovers and fitness proportional selection; GA(block,DBT) denotes the GA with block crossovers and deterministic binary tournament; CGA(block) denotes the CGA with block crossovers and crowding selection; GA(numeric,FP) denotes the GA with intermediate crossovers and fitness proportional selection; GA(numeric,DBT) denotes the GA with intermediate crossovers and deterministic binary tournament; CGA(numeric) denotes the CGA with intermediate crossovers and crowding selection.

1) Mono-nucleotide modeling: The results at 1000 objective function evaluations are depicted in Fig. 1. From the results at 1000 fitness function evaluations, we can have several observations. (1) The numeric crossover operator (i.e. intermediate crossover) is found beneficial to the overall performance. If we compare the GAs with block crossovers to the GAs with intermediate crossovers, we can observe an overall increase in the mean performance if the intermediate crossover operator is used. (2) Although both CDE and CGA(numeric) both adopt numeric operators, CGA(numeric) methods show better performance than CDE at 1000 objective function evaluations. (3) Fitness proportional selection is worse than the other selections. If we compare the GAs with fitness proportional selection to the GAs with deterministic binary tournament and crowding selection, a decrease in mean performance is observed. A possible explanation is that fitness proportional selection is stochastic, resulting in replacement errors [34]. On the other hand, the other selection methods are

[1]For example, over ten hours are needed to evaluate a calculation in computational fluid dynamics [44]



(a) 1000 objective function evaluations (mono-nucleotide modeling)



(b) 10000 objective function evaluations (mono-nucleotide modeling)

Fig. 1. Performance histograms of different methods applied to mono-nucleotide motif modeling at 1000 and 10000 objective function evaluations. The vertical axis denotes the number of datasets falling into each mean performance bin, while the horizontal axis denotes the mean performance bins (Spearman rank correlation). GA(block,FP) denotes the GA with block crossovers and fitness proportional selection; GA(block,DBT) denotes the GA with block crossovers and deterministic binary tournament; CGA(block) denotes the CGA with block crossovers and crowding selection; GA(numeric,FP) denotes the GA with intermediate crossovers and fitness proportional selection; GA(numeric,DBT) denotes the GA with intermediate crossovers and deterministic binary tournament; CGA(numeric) denotes the CGA with intermediate crossovers and crowding selection

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2443782, IEEE/ACM Transactions on Computational Biology and Bioinformatics

deterministic such that replacement errors can be avoided. (4) GA(numeric,DBT) and CGA(numeric) perform the best, while IPM performs the worst on the datasets. It is surprising because IPM is one of the state-of-the-art numerical optimization method but even the simplest method (DE) can still performs better than it, indicating the non-convexity of the objective function. Another possible reason is that IPM spend too many function evaluations to compute numerical gradients.

On the other hand, we observe that most of the methods have not reached their maximal performance for ranking k-mers on the datasets at 1000 objective function evaluations; for instance, only less than 10% datasets can achieve the mean of Spearman rank correlation coefficients larger than 0.8. Thus we have relaxed the termination condition to be ten-fold (10000 evaluations). The results at 10000 objective function evaluations are also depicted in Fig. 1. From the results, we also have several observations. (1) The overall performance of most methods are enhanced after we have relaxed the termination condition except IPM and GA(numeric,FP). The performance of IPM and GA(numeric,FP) have been inhibited due to function non-convexity and replacement errors respectively as described in the previous text. (2) The best-performing methods are DE, CDE, and CGA(block), while the worst one is still IPM. (3) In contrast to the observation at 1000 evaluations, CDE show better performance than CDE at 10000 evaluations, reflecting the relative long-term competitiveness of CDE. (4) With the relaxed termination condition, the long-term advantages of block crossover operators can now be observed. The GA methods with block crossovers demonstrate better performance than the previous setting (1000 evaluations). They can even show comparable results with the other methods.

To investigate the relationships between different methods under different termination conditions (1000 and 10000 objective function evaluations), we have drawn a scatter plot to visualize and compare the methods' performance at 1000 and 10000 evaluations in Fig. 2. From the figure, we can observe that most of the methods perform better at 10000 evaluations than the reciprocal cases at 1000 evaluations (except IPM and GA(numeric,FP)). It is expected since the methods are given more evaluations for their convergence. IPM does not perform well because it assumes function convexity which is not realistic and applicable to the problem here. On the other hand, the GA(numeric,FP) involves the combination of intermediate crossover operators and fitness proportional selection. The intermediate crossover operators are known to promote incremental convergence which can deepen the effects of replacement errors induced by fitness proportional selection [45].

2) Di-nucleotide modeling: In the previous section, we have focused on mono-nucleotide motif modeling which is the most common motif model. It assumes independence between adjacent motif positions, as justified by the experimental and theoretical statistical mechanical study [18]. Nonetheless, it is well-known that adjacent nucleotide dependency exists in some DNA motifs [20]. Thus a recent attempt has been made to generalize the model to handle di-nucleotide representations [19]. In this study, we try to apply and compare the methods

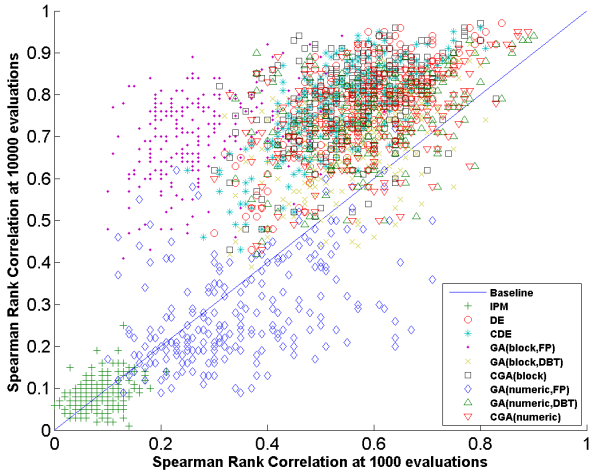


Fig. 2. Scatter plot for comparing the performance values at 1000 function evaluations and the performance values at 10000 function evaluations under mono-nucleotide motif modeling. Each dot denotes a single method's performance on a single dataset. The vertical axis is the mean performance at 10000 evaluations, while the horizontal axis is the mean performance at 1000 evaluations. The solid line denotes the baseline on which the performance at 1000 and that at 10000 evaluations are the same.

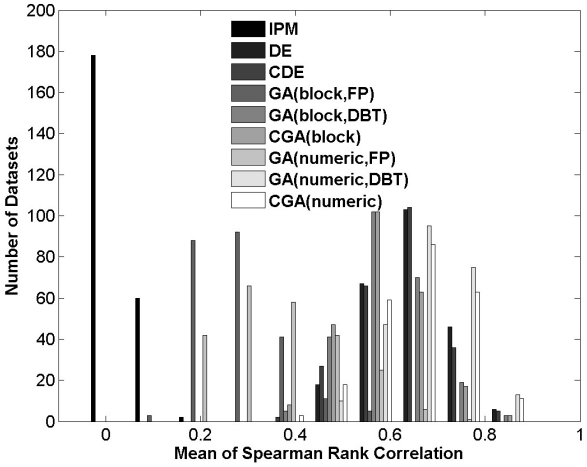
to build di-nucleotide motif models to solve the PBM motif ranking problem.

The di-nucleotide modeling results at 1000 objective function evaluations are depicted in Fig. 3. From the figure, we can observe the phenomenon similar to the reciprocal mono-nucleotide modeling results in Fig. 1; for instance, GA(numeric,DBT) and CGA(numeric) still performs the best while IPM is the worst one. Block crossovers still don't have enough evaluations to unleash its long-term competitiveness. Fitness proportion selection is still not found beneficial to solve the problem. CGA performs better than CDE, although their methodology are similar to each other.

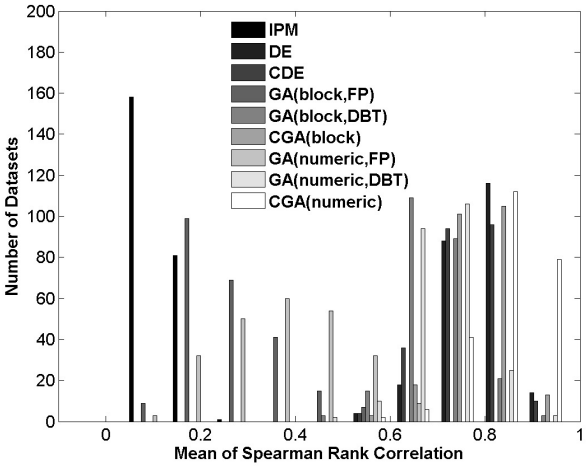
Similar to mono-nucleotide modeling, we observe that most of the methods have not reached their maximal performance for ranking k-mers on the datasets at 1000 objective function evaluations. Thus we have relaxed the termination condition to be ten-fold (10000 evaluations). The results at 10000 objective function evaluations are depicted on Fig. 3 (details can be found in supplementary materials). From Fig. 3, we can observe that the results are largely consistent with the reciprocal mono-nucleotide modeling results in Fig. 1 with some differences which are described one by one as follows: (1) GA(block,FP) performs even poorer than GA(numeric,FP) which is not observed in mono-nucleotide modeling, implying that the replacement errors induced by fitness proportional selection are disastrous for di-nucleotide modeling. (2) CGA(numeric) becomes the best-performing method for di-nucleotide modeling, beating DE and CDE which are the best methods for mono-nucleotide modeling at 10000 evaluations. In particular, CGA(numeric) can achieve good ranking performance (Spearman rank correlation coefficients≥ 0.75) on about 90% of the datasets.

To investigate the relationships between different methods under different termination conditions (1000 and 10000 ob-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2443782, IEEE/ACM Transactions on Computational Biology and Bioinformatics



(a) 1000 objective function evaluations (di-nucleotide modeling)



(b) 10000 objective function evaluations (di-nucleotide modeling)

Fig. 3. Performance histograms of different methods applied to di-nucleotide motif modeling at 1000 and 10000 objective function evaluations. The vertical axis denotes the number of datasets falling into each mean performance bin, while the horizontal axis denotes the mean performance bins (Spearman rank correlation). GA(block,FP) denotes the GA with block crossovers and fitness proportional selection; GA(block,DBT) denotes the GA with block crossovers and deterministic binary tournament; CGA(block) denotes the CGA with block crossovers and crowding selection; GA(numeric,FP) denotes the GA with intermediate crossovers and fitness proportional selection; GA(numeric,DBT) denotes the GA with intermediate crossovers and deterministic binary tournament; CGA(numeric) denotes the CGA with intermediate crossovers and crowding selection

jective function evaluations), we have drawn a scatter plot to visualize and compare the methods' performance at 1000 and 10000 evaluations in Fig. 4. From the figure, we can observe that a sharper improvement trend than the improvement trend observed under mono-nucleotide modeling in Fig. 2. One of the possible explanations is that the degree of freedom under di-nucleotide modeling is much bigger than that under mono-nucleotide modeling. It can enable diverse optimization methods to find their own ways to optimize the objective function smoothly.

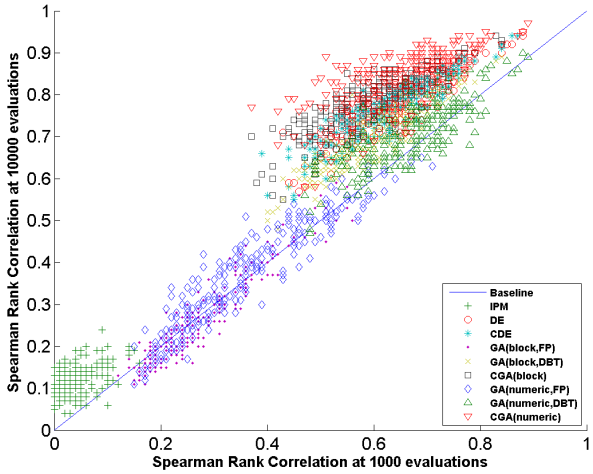


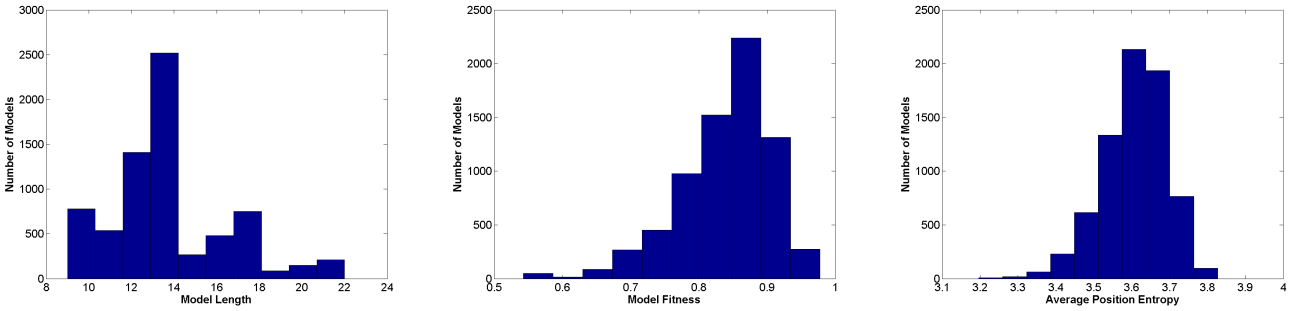
Fig. 4. Scatter plot for comparing the performance values at 1000 function evaluations and the performance values at 10000 function evaluations under di-nucleotide motif modeling. Each dot denotes a single method's performance on a single dataset. The vertical axis is the mean performance at 10000 evaluations, while the horizontal axis is the mean performance at 1000 evaluations. The solid line denotes the baseline on which the performance at 1000 and that at 10000 evaluations are the same.

3) Modeling Comparison: Since we have applied the methods under different modeling settings: mono-nucleotide modeling and di-nucleotide modeling, it is interesting to check whether there is any performance improvement after using di-nucleotide modeling (quadratic model complexity [19]) over mono-nucleotide modeling (linear model complexity [19]). Thus we have drawn a scatter plot to compare the performance values of different methods under mono-nucleotide modeling and those under di-nucleotide modeling in Fig. 5. It can be observed that all of the methods, except GA(block,FP), can achieve better performance under di-nucleotide modeling than mono-nucleotide modeling, justifying the quadratic increase in model complexity.

D. Model Analysis

After the runs (30 runs for each PBM dataset), we have learned thousands of models. It is interesting for us to investigate how the models have been distributed. In particular, as elaborated in the previous sections, we are very interested in the models learned by CGA(numeric) since CGA(numeric) is the best method for di-nucleotide modeling among the methods tested. Thus we have plotted the model lengths, ranking performance (Spearman rank correlation coefficient), and position entropies of the models learned by CGA(numeric) at 10000 objective function evaluations as shown in Fig. 6. It can be observed that the average position entropies of the models tend to be be centered around 3.6 (Maximum is 4), reflecting that the motif models learned are complex from the information theory view. The performance of the models is quite satisfactory since the mode of Spearman rank correlation coefficient is near 0.9. On the other hand, it can be observed that the model lengths are usually around $12 \sim 14$ nt which are consistent with the existing TFBS knowledge [20].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2443782, IEEE/ACM Transactions on Computational Biology and Bioinformatics



(a) Model Length (b) Spearman Rank Correlation (c) Position Information Entropy

Fig. 6. Model analysis histograms of CGA(numeric) at 10000 objective function evaluations. The vertical axis denotes the number of models falling into each horizontal bin, while the horizontal axis denotes the measurements of interest.

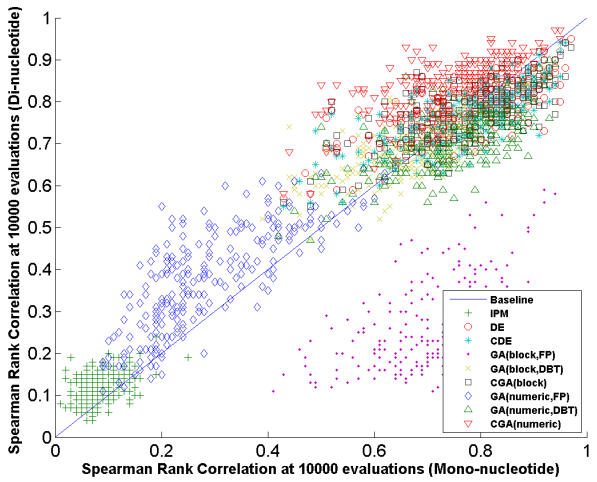


Fig. 5. Scatter plot for comparing the performance values of different methods under mono-nucleotide modeling and those under di-nucleotide modeling. Each dot denotes a single method's performance on a single dataset. The vertical axis denotes the mono-nucleotide modeling performance at 10000 evaluations, while the horizontal axis denotes the di-nucleotide modeling performance at 10000 evaluations. The solid line denotes the baseline on which the mono-nucleotide modeling performance value is the same as the di-nucleotide modeling performance value.

E. Sensitivity Analysis

In the previous comparison, we have just adopted a single multiple sequence alignment method for the PBM k-mer alignment. One may wonder if the choice of multiple sequence alignment method could affect the resultant DNA motif k-mer ranking performance (Spearman rank correlation coefficient). Therefore, we have conducted a performance sensitivity analysis on different uses of multiple sequence alignment method with CGA(numeric) which is the best modeling optimization method we have found so far. The results are tabulated in Table I. Interestingly, it can be observed that MUSCLE is the best performing multiple sequence alignment method which is consistent with the existing benchmark studies [46].

F. Parameter Analysis

On the other hand, algorithmic parameter choice is one of the determining factors for the modeling performance of the optimization methods tested. Of the parameters used, population size is the most influential parameter in this study since such a parameter is involved in all the optimization methods we have compared except IPM. Therefore, we have conducted a parameter analysis on population size settings. Similar to the previous section, CGA(numeric) is chosen for investigation because it is the best optimization method concluded from the previous comparisons. The results are tabulated in Table II. The results indicate that small population sizes are beneficial to the current DNA motif modeling problem. It is expected because it has just been reported that a single consensus DNA motif k-mer pattern is more energetically favorable than its neighborhood patterns [47]. Thus the population optimizations methods in this study require few individuals for evolutionarily capturing that single consensus DNA motif k-mer pattern. In addition, we observe that the effect of population size parameter is less pronounced in di-nucleotide modeling than mono-nucleotide modeling. A possible explanation is that di-nucleotide modeling needs more candidates for encoding that consensus DNA motif k-mer pattern to address its increased model complexity than mono-nucleotide modeling.

V. Applications

A. PBM Rotation Testing

In PBM technology, we usually have two PBM array datasets for each protein of interest. To test the accuracy of the models learned using CGA(numeric), we can apply each model to its corresponding replicate alternatively. For instance, a model learned on array $\#1$ can be applied to rank the probes on array $\#2$ and vice versa. Especially, we are interested in the abilities of the models learned using CGA(numeric) to predict positive probes among all the available probes in an array dataset. Thus we have adopted the traditional measure to define positive probes on each dataset [14]. Mathematically, we define a positive probe to be the probe y whose normalized signal intensity $m_y > mi + 4\sigma$ where mi and σ are the median and the median absolute deviation (MAD) of all the probe normalized intensities in the same dataset divided by 0.6745 (the MAD of the unit normal distribution) respectively. Following that definition, we have tested the models learned using CGA(numeric) at 10000 evaluations on the previous

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2443782, IEEE/ACM Transactions on Computational Biology and Bioinformatics

TABLE I
SENSITIVITY ANALYSIS ON DIFFERENT USES OF MULTIPLE SEQUENCE ALIGNMENT METHODS UNDER DIFFERENT MODELING STRATEGIES. THE TERMINATION CONDITION IS RELAXED TO 10000 OBJECTIVE FUNCTION EVALUATIONS FOR COMPREHENSIVE COMPARISONS. ENTRIES DENOTE THE SPEARMAN RANK CORRELATION COEFFICIENTS BETWEEN THE ACTUAL MEDIAN BINDING INTENSITIES OF THE INPUT ALIGNED K-MERS AND THE TENTATIVE SCORES PREDICTED BY CGA(NUMERIC) ON DIFFERENT PBM DATASETS (ON THE TOP ROW).

Modeling Type	MSA	Cbf1_deBruijn_v1	Cbf1_deBruijn_v2	Ceh-22_deBruijn_v1	Ceh-22_deBruijn_v2	Oct-1_deBruijn_v1
mono-nucleotide modeling	multialign	0.90 ± 0.00	$\mathbf{0.89 \pm 0.01}$	0.69 ± 0.01	0.56 ± 0.01	0.71 ± 0.01
	MUSCLE	0.89 ± 0.01	0.86 ± 0.01	$\mathbf{0.84 \pm 0.01}$	$\mathbf{0.81 \pm 0.01}$	$\mathbf{0.92 \pm 0.00}$
	ClustalW	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.89 \pm 0.01}$	0.61 ± 0.01	0.59 ± 0.02	0.62 ± 0.01
di-nucleotide modeling	multialign	0.90 ± 0.00	$\mathbf{0.90 \pm 0.01}$	0.75 ± 0.02	0.57 ± 0.02	0.67 ± 0.01
	MUSCLE	$\mathbf{0.92 \pm 0.01}$	0.87 ± 0.01	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.85 \pm 0.01}$	$\mathbf{0.94 \pm 0.00}$
	ClustalW	0.91 ± 0.00	$\mathbf{0.90 \pm 0.01}$	0.72 ± 0.01	0.70 ± 0.01	0.66 ± 0.02
Modeling Type	MSA	Oct-1_deBruijn_v2	Rap1_deBruijn_v1	Rap1_deBruijn_v2	Zif268_deBruijn_v1	Zif268_deBruijn_v2
mono-nucleotide modeling	multialign	0.67 ± 0.01	0.79 ± 0.01	0.80 ± 0.01	0.76 ± 0.01	0.68 ± 0.01
	MUSCLE	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.80 \pm 0.01}$	$\mathbf{0.92 \pm 0.00}$	$\mathbf{0.80 \pm 0.01}$	$\mathbf{0.89 \pm 0.00}$
	ClustalW	0.65 ± 0.01	0.68 ± 0.01	0.73 ± 0.01	0.67 ± 0.01	0.59 ± 0.02
di-nucleotide modeling	multialign	0.71 ± 0.01	0.81 ± 0.01	0.81 ± 0.01	0.77 ± 0.01	0.70 ± 0.02
	MUSCLE	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.95 \pm 0.00}$	$\mathbf{0.95 \pm 0.00}$	$\mathbf{0.82 \pm 0.01}$	$\mathbf{0.92 \pm 0.00}$
	ClustalW	0.69 ± 0.01	0.78 ± 0.01	0.79 ± 0.01	0.72 ± 0.01	0.72 ± 0.02

TABLE II
PARAMETER ANALYSIS ON POPULATION SIZE UNDER DIFFERENT MODELING STRATEGIES. THE TERMINATION CONDITION IS RELAXED TO 10000 OBJECTIVE FUNCTION EVALUATIONS FOR COMPREHENSIVE COMPARISONS. ENTRIES DENOTE THE SPEARMAN RANK CORRELATION COEFFICIENTS BETWEEN THE ACTUAL MEDIAN BINDING INTENSITIES OF THE INPUT ALIGNED K-MERS AND THE TENTATIVE SCORES PREDICTED BY CGA(NUMERIC) ON DIFFERENT PBM DATASETS (ON THE TOP ROW).

Modeling Type	PopSize	Cbf1_deBruijn_v1	Cbf1_deBruijn_v2	Ceh-22_deBruijn_v1	Ceh-22_deBruijn_v2	Oct-1_deBruijn_v1
mono-nucleotide modeling	25	$\mathbf{0.91 \pm 0.00}$	$\mathbf{0.91 \pm 0.00}$	$\mathbf{0.70 \pm 0.01}$	$\mathbf{0.56 \pm 0.00}$	$\mathbf{0.73 \pm 0.01}$
	50	0.90 ± 0.00	0.89 ± 0.01	0.69 ± 0.01	$\mathbf{0.56 \pm 0.01}$	0.71 ± 0.01
	100	0.88 ± 0.01	0.87 ± 0.01	0.65 ± 0.01	0.53 ± 0.01	0.66 ± 0.01
di-nucleotide modeling	25	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.75 \pm 0.02}$	$\mathbf{0.57 \pm 0.02}$	$\mathbf{0.67 \pm 0.02}$
	50	$\mathbf{0.90 \pm 0.00}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{0.75 \pm 0.02}$	$\mathbf{0.57 \pm 0.02}$	$\mathbf{0.67 \pm 0.01}$
	100	0.89 ± 0.00	$\mathbf{0.90 \pm 0.00}$	0.73 ± 0.01	0.56 ± 0.01	0.65 ± 0.01
Modeling Type	PopSize	Oct-1_deBruijn_v2	Rap1_deBruijn_v1	Rap1_deBruijn_v2	Zif268_deBruijn_v1	Zif268_deBruijn_v2
mono-nucleotide modeling	25	$\mathbf{0.67 \pm 0.01}$	$\mathbf{0.80 \pm 0.01}$	$\mathbf{0.83 \pm 0.01}$	$\mathbf{0.77 \pm 0.01}$	$\mathbf{0.71 \pm 0.01}$
	50	$\mathbf{0.67 \pm 0.01}$	0.79 ± 0.01	0.80 ± 0.01	0.76 ± 0.01	0.68 ± 0.01
	100	0.65 ± 0.01	0.76 ± 0.01	0.77 ± 0.01	0.74 ± 0.01	0.63 ± 0.02
di-nucleotide modeling	25	0.70 ± 0.03	$\mathbf{0.81 \pm 0.01}$	$\mathbf{0.82 \pm 0.01}$	$\mathbf{0.77 \pm 0.02}$	$\mathbf{0.71 \pm 0.02}$
	50	$\mathbf{0.71 \pm 0.01}$	$\mathbf{0.81 \pm 0.01}$	0.81 ± 0.01	$\mathbf{0.77 \pm 0.01}$	0.70 ± 0.02
	100	0.69 ± 0.01	0.79 ± 0.01	0.80 ± 0.01	0.74 ± 0.02	0.68 ± 0.01

benchmark PBM datasets. In particular, we use a sliding window (with the same length as the learned model of the protein of interest) to scan each sequence and adopt the maximal score as the score of each sequence. Mathematically, given a DNA sequence $D = d_1 d_2 d_3 ... d_T$ and the corresponding model learned M, we compute its predicted score $B_M(D)$ as :

$$B_M(D) = \max_p S_M(d_p d_{p+1} d_{p+2} ... d_{p+L-1})$$
$$\forall p \in \{1, 2, ..., T - L + 1\}$$

where $S_M(d_p d_{p+1} d_{p+2} ... d_{p+L-1})$ is the function S previously described.

The Area Under Curve (AUC) of Receiver Operating Characteristics (ROC) curve is adopted as the performance metric to estimate the models' accuracies which are depicted in Fig. 7. From the results, it can be observed that the models learned using CGA(numeric) at 10000 objective function evaluations usually show good performance in predicting positive probes on another array dataset. In particular, most of their AUC values are above 0.5 which is the baseline performance value, reflecting the usefulness of the models learned using CGA(numeric). It also implies some data consistency exists between different PBM array datasets for each protein of interest.

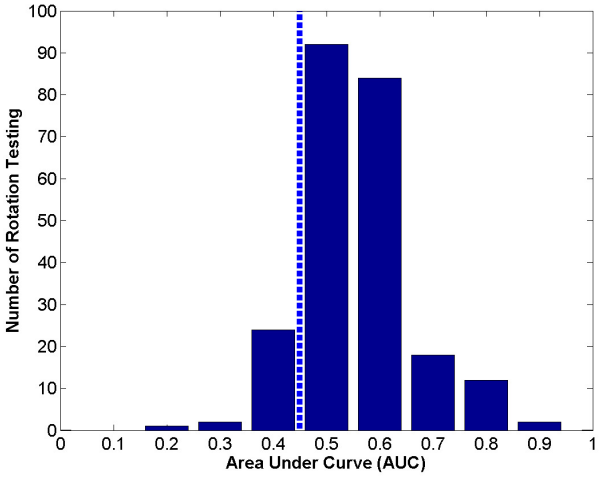
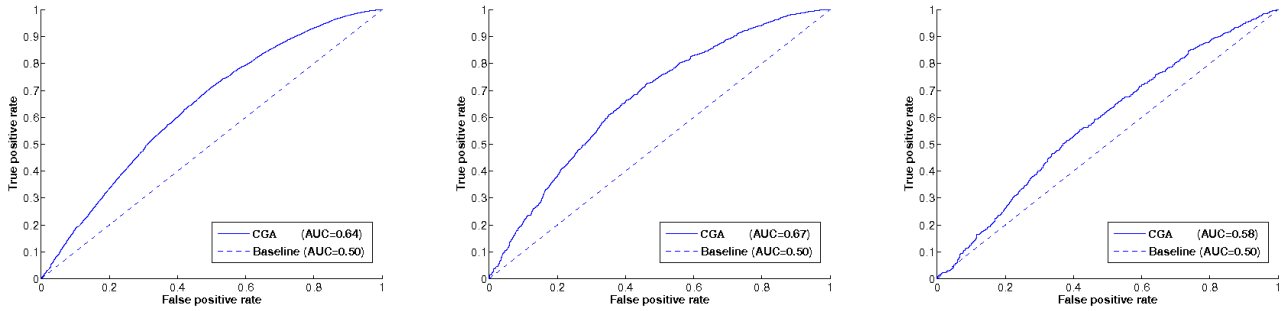


Fig. 7. Area Under Curve (AUC) Value Distribution after PBM rotation testing. The vertical dotted line is the baseline borderline (AUC≥ 0.5).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2443782, IEEE/ACM Transactions on Computational Biology and Bioinformatics



(a) Egr1 (ChIP-Seq) (b) Max (ChIP-Seq) (c) Zbtb3 (ChIP-Seq)

Fig. 8. Peak Sequence Prediction Performance on the ENCODE peak sequence datasets (K562 cell line). Different thresholds are cut at the prediction scores $B_M(D)$ to observe the performance trade-off between sensitivity and false positive rate on (a) Egr1 ChIP-Seq dataset (b) Max ChIP-Seq dataset (c) Zbtb3 ChIP-Seq dataset.

B. ChIP-Seq Peak Sequence Predictions

To demonstrate the utility of the models we have learned further, we have followed the past literature to apply the models to predict ChIP-Seq peak sequences among a set of sequences [48]. In particular, we have checked which models we have learned from the datasets are also found in the ENCODE database (K562 cell line), resulting in the following proteins of interests; Namely, Egr1, Mafk, and Max. The Egr1protein is called Early Growth Response protein 1, a zinc finger protein encoded by Egr1 gene in mammalian genomes [49]. It is a nuclear protein which function is to regulate cell differentiation and mitogenesis, which is also suggested to be involved in different cancers [50]. In contrast, the Mafk protein is a relatively unknown basic leucine zipper (bZIP) transcription factor responsible for developmentally regulated expression of the globin genes [51]. The third one is the MAX protein which is a member of the basic helix-loop-helix leucine zipper (bHLHZ) family of transcription factors. It can interact with several other proteins including the oncoprotein, Myc [52].

Following the convention in the past literature [48], we have obtained the ChIP-Seq peak sequences from their BED files in the ENCODE database. After that, we randomly sample equal amounts of sequences with the same lengths as the peak sequences such that a ChIP-Seq peak sequence dataset is obtained for each protein of interest. In each dataset, half of the sequences are the peak sequences from the ENCODE database while the other half are the background sequences randomly sampled.

After the ChIP-Seq peak sequence datasets are obtained, similar to the previous section, we applied the models learned after 10000 objective function evaluations to perform the binary classification tasks on the ChIP-Seq peak sequence datasets using $B_M(D)$. Having scanned the datasets, we checked our predictions with the actual peak labels. The results are depicted as Receiver Operating Characteristic (ROC) curves in Fig. 8. It can be observed that the models learned are found beneficial for the proteins of interests in predicting peak sequences (above baseline), demonstrating its biological applicability. Nonetheless, we would like to note that the performance can be improved further by incorporating additional biological features from other information sources such as histone marks, DNA methylation, nucleosome occupancy, and DNA double helix shape types.

VI. Discussion

Gene expression is primarily regulated by the DNA binding of various modulatory transcription factors (TF) onto cis-regulatory DNA elements near genes. To fully understand gene functions, it is essential to identify the binding preference of TFs to their corresponding DNA binding sites (TFBS).

To elucidate TFBSs, we have introduced and described the Protein Binding Microarray (PBM) DNA motif model building problem. Such a problem is slightly different from the previous motif discovery problems in the sense that we seek to build DNA motif models which can recover the binding preference of TFs quantitatively, instead of pure TFBS sequence pattern discovery. To tackle the problem, different optimization methods have been applied to learn mono-nucleotide matrix models and di-nucleotide models on more than 200 datasets. From the results, it can be observed that the stochastic beam search method (i.e. DE) and multimodal optimization methods (i.e. CDE and CGA(numeric)) performed very well in building mono-nucleotide matrix model for ranking the DNA motif instances correctly. For di-nucleotide modeling, CGA(numeric), representing the natured-inspired multimodal optimization paradigm, has been shown to be the best method among the methods tested for capturing the binding preference of proteins on those datasets. In addition, we have observed a general performance improvement trend after adopting di-nucleotide modeling over mono-nucleotide modeling. Model analysis has been conducted to analyse the statistical properties of the models learned by CGA(numeric). Lastly, the models have been applied to two different biological problems; namely, PBM probe rotation testing and ChIP-Seq peak sequence prediction. The testing and prediction results independently validate and demonstrate the biological applicability of the models learned.

In the future, the prediction of the measured binding affinities will be an interesting direction if the PBM technology can be improved to be more noise-free and platform-independent than the current form.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their time. The authors would also like to thank Morris Lab and Bulyk Lab for making their PBM data publicly available. The work described in this paper was fully supported by a grant from City University of Hong Kong (Project No. 7200444/CS).

REFERENCES

[1] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. D. Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, , and Z. Zhu, "Assessing computational tools for the discovery of transcription factor binding sites," *Nature Biotechnology*, vol. 23, no. 1, pp. 137–144, 2005.

[2] M. F. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep, and M. L. Bulyk, "Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities," *Nat. Biotechnol.*, vol. 24, pp. 1429–1435, Nov 2006.

[3] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young, "Genome-wide location and function of DNA binding proteins," *Science*, vol. 290, no. 5500, pp. 2306–2309, Dec 2000.

[4] 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.

[5] P. M. Fordyce, D. Gerber, D. Tran, J. Zheng, H. Li, J. L. DeRisi, and S. R. Quake, "De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis," *Nat. Biotechnol.*, vol. 28, no. 9, pp. 970–975, Sep 2010.

[6] S. Hu, Z. Xie, A. Onishi, X. Yu, L. Jiang, J. Lin, H. S. Rho, C. Woodard, H. Wang, J. S. Jeong, S. Long, X. He, H. Wade, S. Blackshaw, J. Qian, and H. Zhu, "Profiling the human protein-DNA interactome reveals ERK2 as a transcriptional repressor of interferon signaling," *Cell*, vol. 139, no. 3, pp. 610–622, Oct 2009.

[7] S. W. Ho, G. Jona, C. T. Chen, M. Johnston, and M. Snyder, "Linking DNA-binding proteins to their recognition sequences by using protein microarrays," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, no. 26, pp. 9940–9945, Jun 2006.

[8] M. S. Smyth and J. H. Martin, "x ray crystallography." *Molecular pathology : MP*, vol. 53, no. 1, pp. 8–14, February 2000. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/10884915

[9] P. M. Mohan and R. V. Hosur, "Structure-function-folding relationships and native energy landscape of dynein light chain protein: nuclear magnetic resonance insights," *J. Biosci.*, vol. 34, pp. 465–479, Sep 2009.

[10] X. S. Liu, D. L. Brutlag, and J. S. Liu, "An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments," *Nat. Biotechnol.*, vol. 20, pp. 835–839, Aug 2002.

[11] K. Robasky and M. L. Bulyk, "UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions," *Nucleic Acids Res.*, vol. 39, pp. D124–128, Jan 2011.

[12] E. Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep 2012.

[13] G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, and et al., "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, Nov 2012.

[14] X. Chen, T. R. Hughes, and Q. Morris, "RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors," *Bioinformatics*, vol. 23, pp. i72–79, Jul 2007.

[15] B. C. Foat, S. S. Houshmandi, W. M. Olivas, and H. J. Bussemaker, "Profiling condition-specific, genome-wide regulation of mRNA stability in yeast," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 102, pp. 17 675–17 680, Dec 2005.

[16] A. Tanay, "Extensive low-affinity transcriptional interactions in the yeast genome," *Genome Res.*, vol. 16, pp. 962–972, Aug 2006.

[17] K. C. Wong, C. Peng, Y. Li, and T. M. Chan, "Herd Clustering: a synergistic data clustering approach using collective intelligence," *Applied Soft Computing (In press)*, Dec 2014.

[18] O. G. Berg and P. H. von Hippel, "Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters," *J. Mol. Biol.*, vol. 193, no. 4, pp. 723–750, Feb 1987.

[19] G. D. Stormo, "Maximally efficient modeling of dna sequence motifs at all levels of complexity." *Genetics*, vol. 187, no. 4, pp. 1219–1224, 2011. [Online]. Available: http://dx.doi.org/10.1534/genetics.110.126052

[20] K. C. Wong, T. M. Chan, C. Peng, Y. Li, and Z. Zhang, "DNA motif elucidation using belief propagation," *Nucleic Acids Res.*, vol. 41, no. 16, p. e153, Sep 2013.

[21] A. M. Moses and S. Sinha, "Regulatory motif analysis," *Bioinformatics: Tools and Applications (Edwards D, Stajich J,Hansen D) Springer Biomedical and Life Sciences collection*, pp. 137–163, 2009.

[22] M. Wright, "The interior-point revolution in optimization: history, recent developments, and lasting consequences," *Bulletin of the American mathematical society*, vol. 42, no. 1, pp. 39–56, 2005.

[23] D. E. Goldberg, *Genetic algorithms in search, optimization and machine learning*. Boston, MA: Kluwer Academic Publishers, 1989.

[24] K. V. Price, "Differential evolution vs. the functions of the 2nd ICEO," in *Evolutionary Computation, 1997., IEEE International Conference on*, Indianapolis, IN, USA, Apr. 1997, pp. 153–157.

[25] R. Thomsen, "Multimodal optimization using crowding-based differential evolution," in *Evolutionary Computation, 2004. CEC2004. Congress on*, vol. 2, Jun. 2004, pp. 1382–1389.

[26] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2009.

[27] H. Bersini, M. Dorigo, S. Langerman, G. Seront, and L. Gambardella, "Results of the first international contest on evolutionaryoptimisation (1st ICEO)," in *Evolutionary Computation, 1996., Proceedings of IEEE International Conference on*, Nagoya, Japan, May 1996, pp. 611–615.

[28] J. Tvrdík, "Adaptation in differential evolution: A numerical comparison," *Applied Soft Computing*, vol. 9, no. 3, pp. 1149–1155, June 2009. [Online]. Available: http://dx.doi.org/10.1016/j.asoc.2009.02.010

[29] K. Deb and D. E. Goldberg, "An investigation of niche and species formation in genetic function optimization," in *Proceedings of the third international conference on Genetic algorithms*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1989, pp. 42–50.

[30] K. C. Wong, K. S. Leung, and M. H. Wong, "An evolutionary algorithm with species-specific explosion for multimodal optimization," in *GECCO '09: Proceedings of the 11th Annual conference on Genetic and evolutionary computation*. New York, NY, USA: ACM, 2009, pp. 923–930.

[31] K. A. De Jong, *Evolutionary Computation. A Unified Approach*. Cambridge, MA, USA: MIT Press, 2006.

[32] K. C. Wong, K. S. Leung, and M. H. Wong, "Protein structure prediction on a lattice model via multimodal optimization techniques," in *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation*, ser. GECCO '10. New York, NY, USA: ACM, 2010, pp. 155–162. [Online]. Available: http://doi.acm.org/10.1145/1830483.1830513

[33] R. Storn and K. Price, "Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, December 1997. [Online]. Available: http://www.springerlink.com/content/x555692233083677/

[34] K. C. Wong, K. S. Leung, and M. H. Wong, "Effect of spatial locality on an evolutionary algorithm for multimodal optimization," in *EvoApplications (1)*, 2010, pp. 481–490.

[35] J. Lampinen and I. Zelinka, "Mechanical engineering design optimization by differential evolution," *New ideas in optimization*, pp. 127–146, 1999.

[36] W. F. Sacco, N. Henderson, A. C. Rios-Coelho, M. M. Ali, and C. M. N. A. Pereira, "Differential evolution algorithms applied to nuclear reactor core design," *Annals of Nuclear Energy*, June 2009. [Online]. Available: http://dx.doi.org/10.1016/j.anucene.2009.05.007

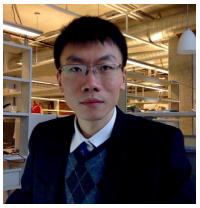
[37] K. A. De Jong, "An analysis of the behavior of a class of genetic adaptive systems." Ph.D. dissertation, University of Michigan, Ann Arbor, MI, USA, 1975.

[38] K. C. Wong, C. H. Wu, R. K. P. Mok, C. Peng, and Z. Zhang, "Evolutionary multimodal optimization using the principle of locality," *Information Sciences*, vol. 194, pp. 138–170, 2012.

[39] D. E. Goldberg and J. Richardson, "Genetic algorithms with sharing for multimodal function optimization," in *Proceedings of the Second International Conference on Genetic algorithms and their application*. Hillsdale, NJ, USA: L. Erlbaum Associates Inc., 1987, pp. 41–49.

[40] J. P. Li, M. E. Balazs, G. T. Parks, and P. J. Clarkson, "A species conserving genetic algorithm for multimodal function optimization," *Evol. Comput.*, vol. 10, no. 3, pp. 207–234, 2002.
[41] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C. F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk, "Diversity and complexity in DNA recognition by transcription factors," *Science*, vol. 324, no. 5935, pp. 1720–1723, Jun 2009.
[42] R. Durbin, S. R. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Jul. 1998. [Online]. Available: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0521629713
[43] Y. S. Ong, P. B. Nair, and A. J. Keane, "Evolutionary optimization of computationally expensive problems via surrogate modeling," *AIAA Journal*, vol. 41, no. 4, pp. 687 –696, 2003.
[44] Y. Jin, "A comprehensive survey of fitness approximation in evolutionary computation," *Soft Comput.*, vol. 9, no. 1, pp. 3–12, 2005.
[45] K. C. Wong, C. Peng, M. H. Wong, and K. S. Leung, "Generalizing and learning protein-dna binding sequence representations by an evolutionary algorithm," *Soft Comput.*, vol. 15, no. 8, pp. 1631–1642, Aug. 2011. [Online]. Available: http://dx.doi.org/10.1007/s00500-011-0692-5
[46] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, 2004.
[47] M. Levo, E. Zalckvar, E. Sharon, A. C. Dantas Machado, Y. Kalma, M. Lotam-Pompan, A. Weinberger, Z. Yakhini, R. Rohs, and E. Segal, "Unraveling determinants of transcription factor binding outside the core binding site," *Genome Res.*, Mar 2015.
[48] J. O. Yanez-Cuna, C. D. Arnold, G. Stampfel, L. M. Bory?, D. Gerlach, M. Rath, and A. Stark, "Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features," *Genome Res.*, vol. 24, no. 7, pp. 1147–1156, Jul 2014.
[49] G. Thiel and G. Cibelli, "Regulation of life and death by the zinc finger transcription factor egr-1," *Journal of cellular physiology*, vol. 193, no. 3, pp. 287–292, 2002.
[50] V. Baron, E. D. Adamson, A. Calogero, G. Ragona, and D. Mercola, "The transcription factor egr1 is a direct regulator of multiple tumor suppressors including tgfβ1, pten, p53, and fibronectin," *Cancer gene therapy*, vol. 13, no. 2, pp. 115–124, 2005.
[51] T. Oyake, K. Itoh, H. Motohashi, N. Hayashi, H. Hoshino, M. Nishizawa, M. Yamamoto, and K. Igarashi, "Bach proteins belong to a novel family of btb-basic leucine zipper transcription factors that interact with mafk and regulate transcription through the nf-e2 site." *Molecular and cellular biology*, vol. 16, no. 11, pp. 6083–6095, 1996.
[52] D. E. Ayer, L. Kretzner, and R. N. Eisenman, "Mad: a heterodimeric partner for max that antagonizes myc transcriptional activity," *Cell*, vol. 72, no. 2, pp. 211–222, 1993.

Ka-Chun Wong received his B.Eng. in Computer Engineering from United College, Chinese University of Hong Kong in 2008. He has also received his M.Phil. degree from the Department of Computer Science and Engineering at the same university in 2010. He received his PhD degree from the Department of Computer Science, University of Toronto in 2014. He assumed his duty as assistant professor at City University of Hong Kong in 2015. His research interests include Bioinformatics, Computational Biology, Evolutionary Computation, Data Mining, Machine Learning, and Interdisciplinary Research.



Yue Li received a B.Sc. from the University of Saskatchewan with Honours in Bioinformatics and Minors in Statistics in 2010 and a M.Sc. from University of Toronto in 2012. After that, he completed his PhD in Computational Biology in the same research group at University of Toronto (2014). He is now a postdoctoral associate from Prof. Manolis Kellis research group at Computer Science and Artificial Intelligence Laboratory (CSAIL) at Massachusetts Institute of Technology. He is mainly interested in developing machine-learning methods and bioinformatics tools to reveal meaningful patterns involving genetics, epigenetics, expression dynamics that are associated with complex human diseases.

Chengbin Peng received his Bachelor degree in Computer Science from Zhejiang University in 2007. He has also received his Master degree in Computer Science from the same university in 2010. Since then, he is a PhD candidate in the Extreme Computing Research Center, KAUST with financial supports from KAUST PhD fellowship, provost's award, and teaching assistantships. His research interests include Data Mining Algorithms and Complex Networks.

Hau-San Wong received the BSc and MPhil degrees in electronic engineering from the Chinese University of Hong Kong, and the PhD degree in electrical and information engineering from the University of Sydney. He has also held research positions in the University of Sydney and Hong Kong Baptist University. He is currently an associate professor in the Department of Computer Science, City University of Hong Kong. His research interests include multimedia information processing, multimodal human-computer interaction, and machine learning. He is a member of the IEEE.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2443782, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Supplementary Materials

A Comparison Study for DNA Motif Modeling on Protein Binding Microarray

May 23, 2015

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2443782, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Table S1: Di-nucleotide modeling performance comparison if the termination condition is relaxed to 10000 objective function evaluations. Entries denote the Spearman rank correlation coefficients between the actual median binding intensities of the input aligned k-mers and the tentative scores predicted by different methods (on the top row) on different PBM datasets (on the leftmost column).

	IPM	DE	CDE	GA(block,FP)	GA(block,DBT)	CGA(block)	GA(numeric,FP)	GA(numeric,DBT)	CGA(numeric)
Arid3a_3875.1_v1_deBruijn	0.15 ± 0.13	0.68 ± 0.02	0.66 ± 0.01	0.32 ± 0.08	0.58 ± 0.03	0.68 ± 0.02	0.37 ± 0.09	0.60 ± 0.03	0.75 ± 0.02
Arid3a_3875.1_v2_deBruijn	0.11 ± 0.14	0.82 ± 0.02	0.80 ± 0.01	0.39 ± 0.13	0.73 ± 0.03	0.81 ± 0.01	0.48 ± 0.14	0.74 ± 0.03	0.85 ± 0.01
Arid3a_3875.2_v1_deBruijn	0.14 ± 0.14	0.74 ± 0.01	0.73 ± 0.01	0.27 ± 0.08	0.67 ± 0.03	0.74 ± 0.01	0.40 ± 0.13	0.68 ± 0.04	0.79 ± 0.01
Arid3a_3875.2_v2_deBruijn	0.11 ± 0.13	0.67 ± 0.02	0.65 ± 0.02	0.16 ± 0.08	0.58 ± 0.04	0.71 ± 0.02	0.32 ± 0.14	0.60 ± 0.03	0.80 ± 0.02
Arid5a_3770.2_v1_deBruijn	0.09 ± 0.14	0.84 ± 0.02	0.83 ± 0.01	0.39 ± 0.10	0.74 ± 0.04	0.82 ± 0.02	0.47 ± 0.16	0.76 ± 0.05	0.89 ± 0.01
Arid5a_3770.2_v2_deBruijn	0.11 ± 0.10	0.73 ± 0.01	0.72 ± 0.01	0.26 ± 0.05	0.66 ± 0.03	0.73 ± 0.01	0.39 ± 0.14	0.68 ± 0.03	0.78 ± 0.01
Ascl2_2654.2_v1_deBruijn	0.15 ± 0.10	0.71 ± 0.01	0.70 ± 0.01	0.37 ± 0.09	0.64 ± 0.04	0.72 ± 0.01	0.45 ± 0.14	0.63 ± 0.05	0.73 ± 0.01
Ascl2_2654.2_v2_deBruijn	0.15 ± 0.12	0.77 ± 0.01	0.75 ± 0.01	0.39 ± 0.09	0.73 ± 0.04	0.80 ± 0.01	0.44 ± 0.11	0.71 ± 0.04	0.81 ± 0.01
Atf1_3026.3_v1_deBruijn	0.09 ± 0.11	0.80 ± 0.02	0.78 ± 0.03	0.19 ± 0.10	0.60 ± 0.04	0.75 ± 0.02	0.30 ± 0.19	0.63 ± 0.04	0.85 ± 0.02
Atf1_3026.3_v2_deBruijn	0.09 ± 0.14	0.85 ± 0.02	0.85 ± 0.02	0.29 ± 0.15	0.76 ± 0.04	0.84 ± 0.02	0.60 ± 0.08	0.76 ± 0.05	0.92 ± 0.01
Bbx_3753.1_v1_deBruijn	0.11 ± 0.12	0.69 ± 0.02	0.66 ± 0.01	0.22 ± 0.08	0.65 ± 0.03	0.70 ± 0.01	0.32 ± 0.13	0.66 ± 0.02	0.72 ± 0.01
Bbx_3753.1_v2_deBruijn	0.15 ± 0.13	0.77 ± 0.02	0.72 ± 0.02	0.21 ± 0.09	0.67 ± 0.04	0.76 ± 0.02	0.30 ± 0.14	0.69 ± 0.04	0.84 ± 0.02
Bcl6b_0961.2_v1_deBruijn	0.10 ± 0.13	0.71 ± 0.02	0.67 ± 0.02	0.30 ± 0.07	0.62 ± 0.03	0.70 ± 0.02	0.36 ± 0.11	0.63 ± 0.04	0.79 ± 0.02
Bcl6b_0961.2_v2_deBruijn	0.20 ± 0.17	0.82 ± 0.01	0.80 ± 0.02	0.37 ± 0.10	0.75 ± 0.03	0.82 ± 0.01	0.41 ± 0.14	0.78 ± 0.03	0.86 ± 0.01
Bhlhb2_1274.3_v1_deBruijn	0.19 ± 0.11	0.94 ± 0.01	0.94 ± 0.01	0.59 ± 0.14	0.88 ± 0.04	0.94 ± 0.01	0.64 ± 0.09	0.90 ± 0.02	0.97 ± 0.01
Bhlhb2_1274.3_v2_deBruijn	0.16 ± 0.09	0.92 ± 0.01	0.91 ± 0.01	0.53 ± 0.15	0.83 ± 0.03	0.90 ± 0.01	0.63 ± 0.08	0.85 ± 0.02	0.94 ± 0.01
Bhlhb2_4971.1_v1_deBruijn	0.19 ± 0.13	0.95 ± 0.01	0.94 ± 0.01	0.58 ± 0.12	0.88 ± 0.03	0.93 ± 0.01	0.65 ± 0.12	0.89 ± 0.02	0.97 ± 0.00
Bhlhb2_4971.1_v2_deBruijn	0.16 ± 0.13	0.92 ± 0.01	0.92 ± 0.01	0.56 ± 0.13	0.84 ± 0.03	0.91 ± 0.01	0.62 ± 0.10	0.86 ± 0.03	0.95 ± 0.01
E2F2_1022.2_v1_deBruijn	0.07 ± 0.10	0.69 ± 0.02	0.66 ± 0.02	0.20 ± 0.07	0.60 ± 0.05	0.70 ± 0.02	0.28 ± 0.16	0.64 ± 0.04	0.77 ± 0.02
E2F2_1022.2_v2_deBruijn	0.15 ± 0.16	0.81 ± 0.02	0.77 ± 0.02	0.21 ± 0.09	0.80 ± 0.03	0.85 ± 0.01	0.24 ± 0.10	0.83 ± 0.03	0.89 ± 0.02
E2F2_1022.4_v1_deBruijn	0.07 ± 0.09	0.76 ± 0.03	0.73 ± 0.03	0.13 ± 0.07	0.60 ± 0.05	0.74 ± 0.02	0.24 ± 0.15	0.61 ± 0.04	0.84 ± 0.02
E2F2_1022.4_v2_deBruijn	0.12 ± 0.15	0.77 ± 0.02	0.76 ± 0.01	0.27 ± 0.09	0.69 ± 0.04	0.76 ± 0.01	0.36 ± 0.18	0.72 ± 0.03	0.82 ± 0.01
E2F3_3752.1_v1_deBruijn	0.14 ± 0.10	0.79 ± 0.02	0.78 ± 0.02	0.24 ± 0.07	0.66 ± 0.03	0.76 ± 0.01	0.34 ± 0.13	0.68 ± 0.04	0.85 ± 0.01
E2F3_3752.1_v2_deBruijn	0.09 ± 0.08	0.70 ± 0.03	0.69 ± 0.02	0.17 ± 0.05	0.62 ± 0.04	0.72 ± 0.02	0.34 ± 0.15	0.62 ± 0.04	0.77 ± 0.02
E2F3_3752.2_v1_deBruijn	0.19 ± 0.13	0.73 ± 0.02	0.70 ± 0.02	0.24 ± 0.08	0.67 ± 0.03	0.73 ± 0.02	0.20 ± 0.09	0.67 ± 0.03	0.79 ± 0.01
E2F3_3752.2_v2_deBruijn	0.13 ± 0.13	0.76 ± 0.02	0.71 ± 0.02	0.30 ± 0.07	0.69 ± 0.04	0.77 ± 0.02	0.32 ± 0.10	0.71 ± 0.04	0.79 ± 0.04
Egr1_2580.1_v1_deBruijn	0.14 ± 0.16	0.84 ± 0.02	0.83 ± 0.02	0.20 ± 0.11	0.73 ± 0.06	0.85 ± 0.02	0.39 ± 0.21	0.74 ± 0.04	0.90 ± 0.01
Egr1_2580.1_v2_deBruijn	0.11 ± 0.10	0.72 ± 0.03	0.71 ± 0.02	0.17 ± 0.13	0.64 ± 0.04	0.74 ± 0.02	0.31 ± 0.16	0.65 ± 0.05	0.83 ± 0.02
Egr1_2580.2_v1_deBruijn	0.06 ± 0.14	0.80 ± 0.02	0.78 ± 0.02	0.16 ± 0.16	0.68 ± 0.05	0.79 ± 0.02	0.31 ± 0.19	0.73 ± 0.04	0.89 ± 0.01
Egr1_2580.2_v2_deBruijn	0.14 ± 0.10	0.78 ± 0.02	0.76 ± 0.02	0.15 ± 0.10	0.65 ± 0.04	0.78 ± 0.02	0.29 ± 0.18	0.66 ± 0.04	0.85 ± 0.01
Ehf_3056.2_v1_deBruijn	0.06 ± 0.12	0.72 ± 0.02	0.71 ± 0.03	0.16 ± 0.10	0.54 ± 0.04	0.68 ± 0.02	0.41 ± 0.13	0.57 ± 0.03	0.76 ± 0.02
Ehf_3056.2_v2_deBruijn	0.14 ± 0.13	0.86 ± 0.01	0.85 ± 0.02	0.30 ± 0.08	0.73 ± 0.05	0.84 ± 0.01	0.51 ± 0.15	0.77 ± 0.04	0.91 ± 0.01
Elf3_3876.1_v1_deBruijn	0.08 ± 0.12	0.70 ± 0.01	0.69 ± 0.02	0.18 ± 0.10	0.61 ± 0.03	0.68 ± 0.01	0.34 ± 0.16	0.62 ± 0.04	0.75 ± 0.01
Elf3_3876.1_v2_deBruijn	0.11 ± 0.10	0.72 ± 0.03	0.69 ± 0.02	0.23 ± 0.07	0.63 ± 0.04	0.71 ± 0.02	0.32 ± 0.13	0.64 ± 0.04	0.80 ± 0.01
Eomes_0921.4_v1_deBruijn	0.13 ± 0.14	0.82 ± 0.02	0.81 ± 0.02	0.22 ± 0.11	0.68 ± 0.05	0.78 ± 0.02	0.52 ± 0.11	0.69 ± 0.04	0.86 ± 0.01
Eomes_0921.4_v2_deBruijn	0.07 ± 0.07	0.66 ± 0.02	0.65 ± 0.02	0.17 ± 0.09	0.52 ± 0.04	0.63 ± 0.03	0.33 ± 0.13	0.56 ± 0.04	0.78 ± 0.02
Esrra_2190.2_v1_deBruijn	0.09 ± 0.12	0.76 ± 0.01	0.74 ± 0.01	0.15 ± 0.12	0.68 ± 0.03	0.75 ± 0.01	0.41 ± 0.16	0.70 ± 0.02	0.81 ± 0.01
Esrra_2190.2_v2_deBruijn	0.11 ± 0.12	0.78 ± 0.03	0.74 ± 0.03	0.15 ± 0.10	0.60 ± 0.03	0.70 ± 0.03	0.20 ± 0.12	0.63 ± 0.04	0.85 ± 0.02
Foxa2_2830.2_v1_deBruijn	0.10 ± 0.11	0.84 ± 0.01	0.82 ± 0.01	0.36 ± 0.14	0.66 ± 0.04	0.79 ± 0.01	0.55 ± 0.06	0.68 ± 0.06	0.89 ± 0.01
Foxa2_2830.2_v2_deBruijn	0.14 ± 0.15	0.80 ± 0.01	0.78 ± 0.02	0.23 ± 0.14	0.64 ± 0.04	0.77 ± 0.01	0.42 ± 0.13	0.65 ± 0.04	0.82 ± 0.01
Foxj1_3125.2_v1_deBruijn	0.07 ± 0.11	0.71 ± 0.02	0.67 ± 0.03	0.15 ± 0.10	0.65 ± 0.03	0.72 ± 0.01	0.19 ± 0.12	0.65 ± 0.03	0.76 ± 0.02
Foxj1_3125.2_v2_deBruijn	0.10 ± 0.10	0.79 ± 0.02	0.77 ± 0.02	0.17 ± 0.12	0.71 ± 0.04	0.79 ± 0.02	0.28 ± 0.21	0.73 ± 0.03	0.86 ± 0.01
Foxj3_0982.2_v1_deBruijn	0.13 ± 0.13	0.83 ± 0.02	0.81 ± 0.01	0.41 ± 0.11	0.70 ± 0.04	0.80 ± 0.01	0.54 ± 0.11	0.72 ± 0.03	0.86 ± 0.01
Foxj3_0982.2_v2_deBruijn	0.13 ± 0.14	0.87 ± 0.02	0.86 ± 0.01	0.41 ± 0.17	0.79 ± 0.04	0.88 ± 0.01	0.54 ± 0.15	0.78 ± 0.04	0.93 ± 0.01
Foxk1_2323.4_v1_deBruijn	0.12 ± 0.11	0.88 ± 0.01	0.88 ± 0.01	0.40 ± 0.11	0.74 ± 0.04	0.86 ± 0.02	0.53 ± 0.14	0.75 ± 0.04	0.92 ± 0.01
Foxk1_2323.4_v2_deBruijn	0.17 ± 0.13	0.86 ± 0.01	0.85 ± 0.01	0.40 ± 0.15	0.79 ± 0.03	0.85 ± 0.01	0.57 ± 0.14	0.79 ± 0.03	0.92 ± 0.01
Foxl1_2809.2_v1_deBruijn	0.16 ± 0.14	0.83 ± 0.02	0.82 ± 0.02	0.34 ± 0.09	0.70 ± 0.05	0.82 ± 0.02	0.48 ± 0.13	0.70 ± 0.05	0.88 ± 0.02
Foxl1_2809.2_v2_deBruijn	0.11 ± 0.12	0.82 ± 0.03	0.81 ± 0.02	0.25 ± 0.11	0.67 ± 0.06	0.81 ± 0.02	0.47 ± 0.13	0.69 ± 0.05	0.90 ± 0.02
Gabpa_2829.2_v1_deBruijn	0.09 ± 0.10	0.85 ± 0.02	0.83 ± 0.02	0.31 ± 0.16	0.69 ± 0.04	0.82 ± 0.02	0.51 ± 0.13	0.70 ± 0.05	0.90 ± 0.02
Gabpa_2829.2_v2_deBruijn	0.11 ± 0.12	0.79 ± 0.02	0.79 ± 0.02	0.23 ± 0.10	0.65 ± 0.04	0.75 ± 0.02	0.43 ± 0.16	0.67 ± 0.05	0.88 ± 0.02
Gata3_1024.3_v1_deBruijn	0.12 ± 0.15	0.76 ± 0.02	0.74 ± 0.02	0.20 ± 0.11	0.72 ± 0.04	0.77 ± 0.01	0.31 ± 0.18	0.73 ± 0.03	0.82 ± 0.01
Gata3_1024.3_v2_deBruijn	0.14 ± 0.12	0.82 ± 0.02	0.80 ± 0.02	0.20 ± 0.10	0.72 ± 0.04	0.81 ± 0.02	0.36 ± 0.22	0.75 ± 0.04	0.90 ± 0.01
Gata3_4964.2_v1_deBruijn	0.07 ± 0.07	0.69 ± 0.03	0.65 ± 0.03	0.19 ± 0.09	0.59 ± 0.04	0.70 ± 0.02	0.18 ± 0.08	0.62 ± 0.03	0.77 ± 0.02
Gata3_4964.2_v2_deBruijn	0.16 ± 0.13	0.81 ± 0.02	0.79 ± 0.02	0.24 ± 0.07	0.76 ± 0.04	0.81 ± 0.02	0.30 ± 0.13	0.77 ± 0.04	0.86 ± 0.01
Gata5_3768.1_v1_deBruijn	0.13 ± 0.14	0.86 ± 0.02	0.84 ± 0.02	0.19 ± 0.11	0.75 ± 0.05	0.86 ± 0.02	0.39 ± 0.24	0.78 ± 0.04	0.90 ± 0.01
Gata5_3768.1_v2_deBruijn	0.09 ± 0.11	0.59 ± 0.02	0.58 ± 0.02	0.15 ± 0.08	0.50 ± 0.04	0.59 ± 0.02	0.20 ± 0.12	0.52 ± 0.03	0.68 ± 0.02
Gata6_3769.1_v1_deBruijn	0.14 ± 0.13	0.82 ± 0.01	0.81 ± 0.01	0.26 ± 0.15	0.74 ± 0.04	0.82 ± 0.01	0.45 ± 0.18	0.77 ± 0.03	0.87 ± 0.01
Gata6_3769.1_v2_deBruijn	0.15 ± 0.13	0.82 ± 0.02	0.79 ± 0.02	0.36 ± 0.05	0.74 ± 0.03	0.81 ± 0.01	0.40 ± 0.09	0.73 ± 0.03	0.83 ± 0.02
Gcm1_3732.1_v1_deBruijn	0.10 ± 0.09	0.69 ± 0.02	0.68 ± 0.02	0.16 ± 0.07	0.58 ± 0.06	0.72 ± 0.03	0.40 ± 0.13	0.59 ± 0.06	0.79 ± 0.02
Gcm1_3732.1_v2_deBruijn	0.14 ± 0.13	0.84 ± 0.02	0.82 ± 0.02	0.29 ± 0.10	0.70 ± 0.04	0.80 ± 0.02	0.37 ± 0.11	0.70 ± 0.04	0.88 ± 0.02
Glis2_1757.2_v1_deBruijn	0.08 ± 0.09	0.72 ± 0.02	0.71 ± 0.02	0.12 ± 0.08	0.60 ± 0.05	0.70 ± 0.02	0.23 ± 0.15	0.65 ± 0.05	0.83 ± 0.02
Glis2_1757.2_v2_deBruijn	0.08 ± 0.14	0.79 ± 0.02	0.77 ± 0.02	0.26 ± 0.13	0.70 ± 0.04	0.79 ± 0.02	0.50 ± 0.14	0.73 ± 0.03	0.88 ± 0.01
Gm397_1753.4_v1_deBruijn	0.08 ± 0.10	0.83 ± 0.02	0.82 ± 0.02	0.28 ± 0.12	0.72 ± 0.04	0.82 ± 0.02	0.44 ± 0.16	0.73 ± 0.05	0.90 ± 0.02
Gm397_1753.4_v2_deBruijn	0.13 ± 0.10	0.81 ± 0.02	0.81 ± 0.02	0.30 ± 0.14	0.68 ± 0.05	0.80 ± 0.02	0.47 ± 0.12	0.69 ± 0.04	0.88 ± 0.01
Gmeb1_1745.2_v1_deBruijn	0.12 ± 0.09	0.66 ± 0.02	0.63 ± 0.01	0.13 ± 0.10	0.55 ± 0.04	0.63 ± 0.02	0.15 ± 0.11	0.58 ± 0.03	0.70 ± 0.01
Gmeb1_1745.2_v2_deBruijn	0.12 ± 0.15	0.80 ± 0.01	0.78 ± 0.01	0.37 ± 0.14	0.71 ± 0.04	0.79 ± 0.01	0.52 ± 0.12	0.73 ± 0.04	0.84 ± 0.01
Hbp1_2241.2_v1_deBruijn	0.08 ± 0.11	0.70 ± 0.02	0.68 ± 0.02	0.15 ± 0.12	0.62 ± 0.04	0.71 ± 0.02	0.17 ± 0.10	0.59 ± 0.05	0.76 ± 0.01
Hbp1_2241.2_v2_deBruijn	0.16 ± 0.21	0.74 ± 0.01	0.72 ± 0.01	0.44 ± 0.04	0.70 ± 0.02	0.74 ± 0.01	0.42 ± 0.05	0.70 ± 0.02	0.76 ± 0.01
Hic1_2816.2_v1_deBruijn	0.06 ± 0.11	0.78 ± 0.02	0.76 ± 0.02	0.20 ± 0.10	0.63 ± 0.05	0.76 ± 0.02	0.40 ± 0.19	0.63 ± 0.05	0.83 ± 0.01
Hic1_2816.2_v2_deBruijn	0.17 ± 0.13	0.72 ± 0.01	0.70 ± 0.01	0.34 ± 0.08	0.62 ± 0.03	0.70 ± 0.01	0.38 ± 0.12	0.65 ± 0.02	0.74 ± 0.01
Hnf4a_2640.2_v1_deBruijn	0.06 ± 0.13	0.76 ± 0.02	0.73 ± 0.02	0.21 ± 0.13	0.74 ± 0.04	0.80 ± 0.01	0.19 ± 0.14	0.74 ± 0.04	0.84 ± 0.01
Hnf4a_2640.2_v2_deBruijn	0.04 ± 0.08	0.63 ± 0.02	0.61 ± 0.02	0.11 ± 0.07	0.52 ± 0.04	0.61 ± 0.02	0.17 ± 0.11	0.54 ± 0.04	0.68 ± 0.02
Hoxa3_2783.2_v1_deBruijn	0.13 ± 0.15	0.82 ± 0.02	0.79 ± 0.02	0.45 ± 0.07	0.74 ± 0.03	0.80 ± 0.01	0.51 ± 0.09	0.75 ± 0.04	0.84 ± 0.01
Hoxa3_2783.2_v2_deBruijn	0.14 ± 0.14	0.82 ± 0.02	0.81 ± 0.02	0.22 ± 0.09	0.73 ± 0.04	0.81 ± 0.02	0.39 ± 0.18	0.77 ± 0.04	0.91 ± 0.02
IRC900814_3520.1_v1_deBruijn	0.13 ± 0.12	0.74 ± 0.01	0.73 ± 0.01	0.19 ± 0.14	0.69 ± 0.03	0.74 ± 0.01	0.26 ± 0.14	0.69 ± 0.03	0.77 ± 0.01
IRC900814_3520.1_v2_deBruijn	0.15 ± 0.18	0.86 ± 0.02	0.85 ± 0.02	0.21 ± 0.14	0.75 ± 0.04	0.84 ± 0.02	0.32 ± 0.13	0.76 ± 0.03	0.89 ± 0.01

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2443782, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Table S2: Di-nucleotide modeling performance comparison if the termination condition is relaxed to 10000 objective function evaluations. Entries denote the Spearman rank correlation coefficients between the actual median binding intensities of the input aligned k-mers and the tentative scores predicted by different methods (on the top row) on different PBM datasets (on the leftmost column).

	IPM	DE	CDE	GA(block,FP)	GA(block,DBT)	CGA(block)	GA(numeric,FP)	GA(numeric,DBT)	CGA(numeric)
Irf3_3985.1_v1_deBruijn	0.07 ± 0.11	0.70 ± 0.02	0.68 ± 0.02	0.28 ± 0.15	0.62 ± 0.03	0.71 ± 0.02	0.42 ± 0.12	0.62 ± 0.04	0.78 ± 0.02
Irf3_3985.1_v2_deBruijn	0.13 ± 0.10	0.75 ± 0.02	0.72 ± 0.02	0.17 ± 0.11	0.66 ± 0.03	0.74 ± 0.02	0.21 ± 0.12	0.69 ± 0.04	0.85 ± 0.02
Irf4_3476.1_v1_deBruijn	0.13 ± 0.11	0.87 ± 0.01	0.86 ± 0.01	0.37 ± 0.19	0.79 ± 0.04	0.86 ± 0.01	0.61 ± 0.11	0.80 ± 0.04	0.91 ± 0.01
Irf4_3476.1_v2_deBruijn	0.11 ± 0.13	0.82 ± 0.02	0.81 ± 0.01	0.23 ± 0.13	0.70 ± 0.04	0.79 ± 0.01	0.50 ± 0.17	0.72 ± 0.05	0.87 ± 0.01
Irf5_3874.1_v1_deBruijn	0.18 ± 0.16	0.86 ± 0.02	0.85 ± 0.01	0.29 ± 0.11	0.75 ± 0.03	0.85 ± 0.02	0.51 ± 0.16	0.77 ± 0.04	0.93 ± 0.01
Irf5_3874.1_v2_deBruijn	0.12 ± 0.13	0.86 ± 0.01	0.85 ± 0.01	0.37 ± 0.19	0.72 ± 0.06	0.84 ± 0.02	0.56 ± 0.07	0.71 ± 0.05	0.90 ± 0.01
Irf6_3803.1_v1_deBruijn	0.09 ± 0.11	0.82 ± 0.03	0.81 ± 0.02	0.16 ± 0.09	0.69 ± 0.05	0.81 ± 0.02	0.48 ± 0.16	0.71 ± 0.06	0.91 ± 0.01
Irf6_3803.1_v2_deBruijn	0.15 ± 0.12	0.81 ± 0.01	0.79 ± 0.02	0.26 ± 0.14	0.76 ± 0.04	0.82 ± 0.02	0.31 ± 0.17	0.77 ± 0.03	0.86 ± 0.01
Isgf3g_2853.2_v1_deBruijn	0.08 ± 0.10	0.79 ± 0.03	0.77 ± 0.02	0.19 ± 0.10	0.67 ± 0.03	0.77 ± 0.02	0.30 ± 0.19	0.71 ± 0.04	0.88 ± 0.01
Isgf3g_2853.2_v2_deBruijn	0.13 ± 0.14	0.86 ± 0.02	0.83 ± 0.02	0.33 ± 0.15	0.76 ± 0.03	0.85 ± 0.01	0.50 ± 0.16	0.78 ± 0.04	0.92 ± 0.01
Jundm2_0911.3_v1_deBruijn	0.16 ± 0.13	0.80 ± 0.02	0.77 ± 0.02	0.25 ± 0.11	0.69 ± 0.03	0.76 ± 0.02	0.28 ± 0.13	0.69 ± 0.03	0.82 ± 0.01
Jundm2_0911.3_v2_deBruijn	0.12 ± 0.10	0.79 ± 0.02	0.78 ± 0.02	0.18 ± 0.09	0.68 ± 0.04	0.77 ± 0.02	0.21 ± 0.09	0.70 ± 0.04	0.87 ± 0.01
Klf7_0974.2_v1_deBruijn	0.07 ± 0.15	0.78 ± 0.01	0.75 ± 0.02	0.17 ± 0.11	0.69 ± 0.04	0.78 ± 0.02	0.22 ± 0.12	0.72 ± 0.03	0.85 ± 0.01
Klf7_0974.2_v2_deBruijn	0.14 ± 0.13	0.83 ± 0.01	0.81 ± 0.01	0.16 ± 0.15	0.74 ± 0.04	0.81 ± 0.02	0.34 ± 0.15	0.74 ± 0.03	0.88 ± 0.01
Lef1_3504.1_v1_deBruijn	0.04 ± 0.10	0.76 ± 0.03	0.76 ± 0.02	0.11 ± 0.08	0.62 ± 0.06	0.76 ± 0.02	0.31 ± 0.18	0.65 ± 0.06	0.85 ± 0.02
Lef1_3504.1_v2_deBruijn	0.10 ± 0.11	0.76 ± 0.02	0.76 ± 0.02	0.17 ± 0.10	0.61 ± 0.05	0.73 ± 0.02	0.28 ± 0.17	0.62 ± 0.06	0.83 ± 0.01
Mafb_2914.2_v1_deBruijn	0.18 ± 0.19	0.88 ± 0.01	0.86 ± 0.01	0.42 ± 0.13	0.82 ± 0.03	0.88 ± 0.01	0.56 ± 0.14	0.85 ± 0.02	0.92 ± 0.01
Mafb_2914.2_v2_deBruijn	0.16 ± 0.15	0.80 ± 0.03	0.77 ± 0.02	0.21 ± 0.10	0.72 ± 0.04	0.81 ± 0.02	0.26 ± 0.12	0.74 ± 0.03	0.87 ± 0.01
Mafk_3106.2_v1_deBruijn	0.09 ± 0.10	0.74 ± 0.02	0.73 ± 0.01	0.19 ± 0.08	0.63 ± 0.04	0.72 ± 0.02	0.34 ± 0.17	0.64 ± 0.04	0.82 ± 0.02
Mafk_3106.2_v2_deBruijn	0.11 ± 0.10	0.71 ± 0.02	0.69 ± 0.02	0.21 ± 0.09	0.65 ± 0.04	0.73 ± 0.02	0.27 ± 0.14	0.66 ± 0.03	0.78 ± 0.02
Max_3863.1_v1_deBruijn	0.13 ± 0.16	0.82 ± 0.02	0.82 ± 0.02	0.45 ± 0.07	0.78 ± 0.02	0.84 ± 0.01	0.52 ± 0.09	0.81 ± 0.04	0.88 ± 0.01
Max_3863.1_v2_deBruijn	0.14 ± 0.16	0.90 ± 0.01	0.89 ± 0.01	0.56 ± 0.14	0.83 ± 0.04	0.90 ± 0.01	0.60 ± 0.13	0.87 ± 0.03	0.95 ± 0.01
Max_3864.1_v1_deBruijn	0.16 ± 0.16	0.78 ± 0.01	0.78 ± 0.02	0.46 ± 0.08	0.76 ± 0.02	0.80 ± 0.02	0.50 ± 0.08	0.79 ± 0.02	0.85 ± 0.01
Max_3864.1_v2_deBruijn	0.17 ± 0.20	0.84 ± 0.02	0.82 ± 0.02	0.46 ± 0.11	0.81 ± 0.03	0.87 ± 0.01	0.58 ± 0.13	0.86 ± 0.02	0.90 ± 0.01
Mtf1_2377.2_v1_deBruijn	0.15 ± 0.15	0.81 ± 0.01	0.79 ± 0.01	0.41 ± 0.14	0.75 ± 0.03	0.81 ± 0.01	0.50 ± 0.14	0.75 ± 0.03	0.85 ± 0.01
Mtf1_2377.2_v2_deBruijn	0.14 ± 0.18	0.84 ± 0.01	0.82 ± 0.01	0.26 ± 0.10	0.77 ± 0.02	0.83 ± 0.01	0.43 ± 0.20	0.78 ± 0.03	0.89 ± 0.01
Myb_1047.3_v1_deBruijn	0.10 ± 0.14	0.78 ± 0.01	0.77 ± 0.01	0.21 ± 0.10	0.69 ± 0.04	0.79 ± 0.01	0.50 ± 0.12	0.69 ± 0.05	0.84 ± 0.01
Myb_1047.3_v2_deBruijn	0.10 ± 0.11	0.73 ± 0.01	0.71 ± 0.02	0.35 ± 0.09	0.64 ± 0.03	0.71 ± 0.01	0.47 ± 0.12	0.65 ± 0.03	0.76 ± 0.01
Mybl1_1717.2_v1_deBruijn	0.15 ± 0.19	0.81 ± 0.01	0.80 ± 0.01	0.29 ± 0.09	0.74 ± 0.02	0.81 ± 0.01	0.50 ± 0.09	0.74 ± 0.04	0.84 ± 0.00
Mybl1_1717.2_v2_deBruijn	0.17 ± 0.14	0.80 ± 0.01	0.80 ± 0.02	0.42 ± 0.10	0.71 ± 0.04	0.78 ± 0.02	0.52 ± 0.10	0.72 ± 0.04	0.87 ± 0.01
Myf6_3824.2_v1_deBruijn	0.15 ± 0.12	0.70 ± 0.01	0.69 ± 0.01	0.28 ± 0.09	0.62 ± 0.03	0.69 ± 0.01	0.42 ± 0.12	0.63 ± 0.03	0.71 ± 0.01
Myf6_3824.2_v2_deBruijn	0.06 ± 0.12	0.71 ± 0.02	0.68 ± 0.02	0.17 ± 0.10	0.61 ± 0.04	0.69 ± 0.02	0.21 ± 0.10	0.64 ± 0.04	0.74 ± 0.02
Nkx3-1_2923.2_v1_deBruijn	0.11 ± 0.11	0.73 ± 0.02	0.71 ± 0.02	0.17 ± 0.10	0.63 ± 0.04	0.71 ± 0.02	0.21 ± 0.13	0.64 ± 0.04	0.79 ± 0.02
Nkx3-1_2923.2_v2_deBruijn	0.14 ± 0.12	0.77 ± 0.02	0.76 ± 0.02	0.25 ± 0.07	0.61 ± 0.06	0.75 ± 0.02	0.29 ± 0.11	0.63 ± 0.05	0.85 ± 0.01
Nr2f2_2192.2_v1_deBruijn	0.19 ± 0.19	0.84 ± 0.01	0.82 ± 0.01	0.32 ± 0.12	0.80 ± 0.02	0.85 ± 0.01	0.34 ± 0.12	0.81 ± 0.03	0.87 ± 0.02
Nr2f2_2192.2_v2_deBruijn	0.15 ± 0.14	0.80 ± 0.02	0.79 ± 0.02	0.20 ± 0.14	0.75 ± 0.03	0.80 ± 0.01	0.28 ± 0.15	0.76 ± 0.03	0.86 ± 0.01
Osr1_3033.2_v1_deBruijn	0.07 ± 0.14	0.81 ± 0.02	0.79 ± 0.02	0.33 ± 0.16	0.69 ± 0.04	0.79 ± 0.02	0.51 ± 0.14	0.70 ± 0.05	0.87 ± 0.01
Osr1_3033.2_v2_deBruijn	0.14 ± 0.12	0.80 ± 0.02	0.78 ± 0.02	0.20 ± 0.06	0.68 ± 0.05	0.79 ± 0.01	0.27 ± 0.13	0.69 ± 0.04	0.87 ± 0.01
Osr2_1727.2_v1_deBruijn	0.11 ± 0.13	0.74 ± 0.01	0.73 ± 0.01	0.19 ± 0.04	0.62 ± 0.03	0.72 ± 0.02	0.27 ± 0.13	0.64 ± 0.04	0.80 ± 0.02
Osr2_1727.2_v2_deBruijn	0.10 ± 0.12	0.85 ± 0.02	0.84 ± 0.02	0.20 ± 0.06	0.73 ± 0.04	0.84 ± 0.02	0.30 ± 0.18	0.74 ± 0.04	0.91 ± 0.01
Plagl1_0972.2_v1_deBruijn	0.15 ± 0.12	0.87 ± 0.01	0.86 ± 0.02	0.53 ± 0.11	0.78 ± 0.03	0.86 ± 0.02	0.56 ± 0.07	0.77 ± 0.03	0.90 ± 0.01
Plagl1_0972.2_v2_deBruijn	0.11 ± 0.12	0.84 ± 0.01	0.83 ± 0.01	0.40 ± 0.11	0.75 ± 0.05	0.83 ± 0.01	0.51 ± 0.14	0.76 ± 0.04	0.88 ± 0.01
Rara_1051.2_v1_deBruijn	0.13 ± 0.12	0.78 ± 0.02	0.77 ± 0.01	0.22 ± 0.10	0.67 ± 0.03	0.77 ± 0.02	0.52 ± 0.14	0.69 ± 0.05	0.85 ± 0.01
Rara_1051.2_v2_deBruijn	0.09 ± 0.09	0.79 ± 0.02	0.78 ± 0.02	0.21 ± 0.13	0.67 ± 0.04	0.77 ± 0.02	0.42 ± 0.19	0.69 ± 0.04	0.84 ± 0.02
Rfx3_3961.1_v1_deBruijn	0.15 ± 0.21	0.81 ± 0.01	0.80 ± 0.01	0.30 ± 0.19	0.79 ± 0.02	0.82 ± 0.00	0.33 ± 0.19	0.77 ± 0.02	0.82 ± 0.01
Rfx3_3961.1_v2_deBruijn	0.18 ± 0.15	0.80 ± 0.01	0.77 ± 0.02	0.23 ± 0.11	0.73 ± 0.03	0.80 ± 0.01	0.27 ± 0.08	0.74 ± 0.04	0.82 ± 0.01
Rfx3_4970.2_v1_deBruijn	0.12 ± 0.17	0.71 ± 0.01	0.70 ± 0.01	0.28 ± 0.14	0.66 ± 0.02	0.72 ± 0.01	0.34 ± 0.16	0.66 ± 0.02	0.73 ± 0.01
Rfx3_4970.2_v2_deBruijn	0.19 ± 0.24	0.90 ± 0.01	0.89 ± 0.01	0.50 ± 0.09	0.89 ± 0.02	0.92 ± 0.01	0.63 ± 0.13	0.88 ± 0.02	0.92 ± 0.01
Rfx4_3761.1_v1_deBruijn	0.15 ± 0.17	0.77 ± 0.01	0.77 ± 0.01	0.28 ± 0.16	0.74 ± 0.02	0.77 ± 0.01	0.28 ± 0.15	0.72 ± 0.02	0.78 ± 0.01
Rfx4_3761.1_v2_deBruijn	0.10 ± 0.18	0.85 ± 0.01	0.84 ± 0.01	0.32 ± 0.16	0.78 ± 0.03	0.84 ± 0.01	0.33 ± 0.14	0.78 ± 0.03	0.86 ± 0.01
Rfxdc2_3516.1_v1_deBruijn	0.10 ± 0.10	0.78 ± 0.01	0.75 ± 0.02	0.18 ± 0.11	0.65 ± 0.05	0.76 ± 0.02	0.17 ± 0.09	0.65 ± 0.04	0.82 ± 0.02
Rfxdc2_3516.1_v2_deBruijn	0.15 ± 0.16	0.82 ± 0.01	0.80 ± 0.01	0.26 ± 0.16	0.69 ± 0.04	0.80 ± 0.01	0.31 ± 0.14	0.71 ± 0.04	0.83 ± 0.01
Rxra_1035.2_v1_deBruijn	0.16 ± 0.11	0.67 ± 0.01	0.65 ± 0.01	0.19 ± 0.12	0.62 ± 0.03	0.67 ± 0.01	0.24 ± 0.11	0.64 ± 0.02	0.71 ± 0.01
Rxra_1035.2_v2_deBruijn	0.20 ± 0.16	0.80 ± 0.01	0.79 ± 0.02	0.26 ± 0.16	0.78 ± 0.03	0.82 ± 0.01	0.46 ± 0.15	0.79 ± 0.02	0.85 ± 0.01
Sfpi1_1034.2_v1_deBruijn	0.17 ± 0.14	0.83 ± 0.01	0.82 ± 0.01	0.22 ± 0.11	0.74 ± 0.03	0.83 ± 0.01	0.44 ± 0.19	0.78 ± 0.02	0.88 ± 0.01
Sfpi1_1034.2_v2_deBruijn	0.07 ± 0.12	0.71 ± 0.02	0.70 ± 0.02	0.20 ± 0.08	0.66 ± 0.03	0.71 ± 0.01	0.43 ± 0.17	0.66 ± 0.04	0.76 ± 0.01
Sfpi1_1034.3_v1_deBruijn	0.08 ± 0.11	0.80 ± 0.02	0.79 ± 0.01	0.28 ± 0.14	0.75 ± 0.02	0.81 ± 0.02	0.36 ± 0.20	0.75 ± 0.04	0.87 ± 0.01
Sfpi1_1034.3_v2_deBruijn	0.15 ± 0.13	0.77 ± 0.01	0.76 ± 0.01	0.17 ± 0.14	0.70 ± 0.03	0.76 ± 0.01	0.25 ± 0.13	0.73 ± 0.03	0.83 ± 0.01
Six6_2267.4_v1_deBruijn	0.10 ± 0.15	0.80 ± 0.01	0.80 ± 0.01	0.36 ± 0.15	0.71 ± 0.04	0.78 ± 0.01	0.56 ± 0.10	0.72 ± 0.04	0.83 ± 0.01
Six6_2267.4_v2_deBruijn	0.10 ± 0.12	0.78 ± 0.01	0.75 ± 0.01	0.19 ± 0.12	0.73 ± 0.03	0.78 ± 0.01	0.27 ± 0.20	0.74 ± 0.03	0.83 ± 0.01
Smad3_3805.1_v1_deBruijn	0.19 ± 0.21	0.87 ± 0.01	0.85 ± 0.01	0.36 ± 0.12	0.79 ± 0.04	0.87 ± 0.01	0.48 ± 0.13	0.80 ± 0.02	0.89 ± 0.01
Smad3_3805.1_v2_deBruijn	0.13 ± 0.15	0.79 ± 0.02	0.78 ± 0.02	0.18 ± 0.08	0.63 ± 0.04	0.75 ± 0.01	0.30 ± 0.13	0.66 ± 0.05	0.85 ± 0.02
Sox1_2631.2_v1_deBruijn	0.07 ± 0.07	0.71 ± 0.03	0.66 ± 0.03	0.13 ± 0.08	0.60 ± 0.06	0.70 ± 0.02	0.11 ± 0.09	0.61 ± 0.04	0.77 ± 0.03
Sox1_2631.2_v2_deBruijn	0.12 ± 0.14	0.61 ± 0.01	0.60 ± 0.01	0.25 ± 0.09	0.55 ± 0.02	0.61 ± 0.01	0.26 ± 0.07	0.56 ± 0.02	0.64 ± 0.01
Sox11_2266.2_v1_deBruijn	0.09 ± 0.09	0.70 ± 0.02	0.68 ± 0.02	0.14 ± 0.08	0.58 ± 0.04	0.70 ± 0.04	0.21 ± 0.13	0.58 ± 0.05	0.82 ± 0.02
Sox11_2266.2_v2_deBruijn	0.08 ± 0.10	0.71 ± 0.03	0.67 ± 0.03	0.13 ± 0.08	0.58 ± 0.05	0.70 ± 0.02	0.17 ± 0.09	0.61 ± 0.05	0.78 ± 0.02
Sox12_3957.1_v1_deBruijn	0.08 ± 0.10	0.77 ± 0.02	0.74 ± 0.02	0.16 ± 0.09	0.69 ± 0.05	0.77 ± 0.03	0.21 ± 0.11	0.69 ± 0.04	0.86 ± 0.02
Sox12_3957.1_v2_deBruijn	0.09 ± 0.08	0.82 ± 0.02	0.81 ± 0.01	0.28 ± 0.15	0.77 ± 0.05	0.83 ± 0.02	0.49 ± 0.17	0.76 ± 0.04	0.90 ± 0.01
Sox13_1718.2_v1_deBruijn	0.09 ± 0.13	0.81 ± 0.01	0.80 ± 0.01	0.22 ± 0.12	0.70 ± 0.04	0.80 ± 0.02	0.39 ± 0.21	0.74 ± 0.03	0.89 ± 0.01
Sox13_1718.2_v2_deBruijn	0.15 ± 0.14	0.86 ± 0.02	0.83 ± 0.01	0.21 ± 0.11	0.77 ± 0.05	0.86 ± 0.02	0.34 ± 0.18	0.80 ± 0.04	0.93 ± 0.01
Sox14_2677.2_v1_deBruijn	0.13 ± 0.15	0.77 ± 0.02	0.73 ± 0.02	0.32 ± 0.06	0.67 ± 0.04	0.77 ± 0.02	0.35 ± 0.10	0.67 ± 0.04	0.76 ± 0.02
Sox14_2677.2_v2_deBruijn	0.10 ± 0.11	0.80 ± 0.02	0.79 ± 0.03	0.24 ± 0.09	0.67 ± 0.05	0.79 ± 0.02	0.42 ± 0.19	0.69 ± 0.04	0.87 ± 0.01
Sox15_3457.1_v1_deBruijn	0.05 ± 0.08	0.70 ± 0.02	0.65 ± 0.02	0.16 ± 0.07	0.63 ± 0.03	0.68 ± 0.01	0.19 ± 0.09	0.65 ± 0.03	0.76 ± 0.02
Sox15_3457.1_v2_deBruijn	0.06 ± 0.12	0.73 ± 0.03	0.72 ± 0.03	0.22 ± 0.11	0.62 ± 0.04	0.74 ± 0.03	0.32 ± 0.19	0.67 ± 0.04	0.87 ± 0.02
Sox17_2837.2_v1_deBruijn	0.07 ± 0.09	0.69 ± 0.02	0.66 ± 0.02	0.14 ± 0.10	0.59 ± 0.04	0.69 ± 0.02	0.17 ± 0.08	0.63 ± 0.05	0.77 ± 0.01
Sox17_2837.2_v2_deBruijn	0.13 ± 0.17	0.82 ± 0.02	0.80 ± 0.01	0.33 ± 0.14	0.76 ± 0.03	0.83 ± 0.01	0.47 ± 0.17	0.75 ± 0.03	0.87 ± 0.01

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2015.2443782, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Table S3: Di-nucleotide modeling performance comparison if the termination condition is relaxed to 10000 objective function evaluations. Entries denote the Spearman rank correlation coefficients between the actual median binding intensities of the input aligned k-mers and the tentative scores predicted by different methods (on the top row) on different PBM datasets (on the leftmost column).

	IPM	DE	CDE	GA(block,FP)	GA(block,DBT)	CGA(block)	GA(numeric,FP)	GA(numeric,DBT)	CGA(numeric)
Sox18_3506.1_v1_deBruijn	0.10 ± 0.09	0.81 ± 0.02	0.77 ± 0.02	0.22 ± 0.09	0.71 ± 0.04	0.81 ± 0.02	0.22 ± 0.08	0.72 ± 0.04	0.89 ± 0.01
Sox18_3506.1_v2_deBruijn	0.09 ± 0.14	0.83 ± 0.01	0.81 ± 0.01	0.27 ± 0.15	0.74 ± 0.04	0.82 ± 0.01	0.39 ± 0.19	0.75 ± 0.04	0.87 ± 0.01
Sox21_3417.1_v1_deBruijn	0.10 ± 0.14	0.78 ± 0.02	0.76 ± 0.02	0.30 ± 0.11	0.74 ± 0.03	0.80 ± 0.02	0.41 ± 0.15	0.77 ± 0.04	0.83 ± 0.02
Sox21_3417.1_v2_deBruijn	0.11 ± 0.09	0.75 ± 0.02	0.74 ± 0.02	0.20 ± 0.08	0.65 ± 0.04	0.76 ± 0.02	0.42 ± 0.19	0.66 ± 0.04	0.85 ± 0.02
Sox30_2781.2_v1_deBruijn	0.12 ± 0.10	0.76 ± 0.02	0.75 ± 0.02	0.23 ± 0.15	0.65 ± 0.05	0.75 ± 0.01	0.45 ± 0.15	0.67 ± 0.04	0.82 ± 0.01
Sox30_2781.2_v2_deBruijn	0.11 ± 0.14	0.83 ± 0.02	0.82 ± 0.01	0.32 ± 0.15	0.74 ± 0.04	0.85 ± 0.02	0.50 ± 0.15	0.75 ± 0.04	0.92 ± 0.01
Sox4_2941.2_v1_deBruijn	0.11 ± 0.11	0.77 ± 0.02	0.75 ± 0.02	0.27 ± 0.11	0.66 ± 0.04	0.74 ± 0.02	0.35 ± 0.15	0.71 ± 0.04	0.85 ± 0.01
Sox4_2941.2_v2_deBruijn	0.08 ± 0.11	0.75 ± 0.02	0.72 ± 0.03	0.16 ± 0.07	0.62 ± 0.05	0.74 ± 0.02	0.25 ± 0.17	0.66 ± 0.04	0.83 ± 0.02
Sox5_3459.1_v1_deBruijn	0.09 ± 0.11	0.80 ± 0.02	0.79 ± 0.02	0.21 ± 0.11	0.64 ± 0.06	0.79 ± 0.02	0.38 ± 0.19	0.68 ± 0.05	0.89 ± 0.01
Sox5_3459.1_v2_deBruijn	0.09 ± 0.12	0.80 ± 0.02	0.77 ± 0.02	0.19 ± 0.09	0.75 ± 0.03	0.82 ± 0.02	0.46 ± 0.17	0.77 ± 0.03	0.88 ± 0.01
Sox7_3460.1_v1_deBruijn	0.08 ± 0.12	0.80 ± 0.02	0.79 ± 0.02	0.20 ± 0.11	0.68 ± 0.04	0.78 ± 0.02	0.37 ± 0.16	0.70 ± 0.03	0.86 ± 0.01
Sox7_3460.1_v2_deBruijn	0.11 ± 0.10	0.81 ± 0.02	0.80 ± 0.01	0.23 ± 0.15	0.73 ± 0.03	0.81 ± 0.01	0.38 ± 0.21	0.77 ± 0.03	0.89 ± 0.01
Sox7_4972.2_v1_deBruijn	0.14 ± 0.13	0.81 ± 0.02	0.79 ± 0.02	0.25 ± 0.14	0.72 ± 0.04	0.80 ± 0.02	0.24 ± 0.10	0.75 ± 0.03	0.89 ± 0.01
Sox7_4972.2_v2_deBruijn	0.11 ± 0.12	0.69 ± 0.02	0.66 ± 0.02	0.14 ± 0.10	0.64 ± 0.04	0.70 ± 0.02	0.24 ± 0.17	0.67 ± 0.04	0.81 ± 0.02
Sox8_1733.2_v1_deBruijn	0.16 ± 0.14	0.84 ± 0.02	0.83 ± 0.02	0.24 ± 0.11	0.76 ± 0.03	0.84 ± 0.01	0.40 ± 0.21	0.78 ± 0.03	0.91 ± 0.01
Sox8_1733.2_v2_deBruijn	0.14 ± 0.16	0.88 ± 0.01	0.88 ± 0.01	0.44 ± 0.18	0.82 ± 0.02	0.89 ± 0.01	0.61 ± 0.16	0.85 ± 0.02	0.91 ± 0.01
Sp100_2947.2_v1_deBruijn	0.11 ± 0.12	0.83 ± 0.02	0.81 ± 0.02	0.23 ± 0.09	0.75 ± 0.04	0.83 ± 0.02	0.31 ± 0.17	0.77 ± 0.05	0.89 ± 0.01
Sp100_2947.2_v2_deBruijn	0.11 ± 0.15	0.82 ± 0.01	0.79 ± 0.01	0.40 ± 0.12	0.73 ± 0.03	0.81 ± 0.01	0.54 ± 0.13	0.76 ± 0.03	0.87 ± 0.01
Sp4_1011.2_v1_deBruijn	0.09 ± 0.13	0.88 ± 0.02	0.86 ± 0.01	0.19 ± 0.13	0.78 ± 0.06	0.86 ± 0.02	0.53 ± 0.20	0.76 ± 0.04	0.94 ± 0.01
Sp4_1011.2_v2_deBruijn	0.15 ± 0.14	0.81 ± 0.02	0.79 ± 0.02	0.21 ± 0.12	0.72 ± 0.04	0.82 ± 0.02	0.39 ± 0.18	0.74 ± 0.05	0.86 ± 0.01
Spdef_0905.2_v1_deBruijn	0.10 ± 0.12	0.80 ± 0.01	0.80 ± 0.01	0.20 ± 0.10	0.67 ± 0.05	0.77 ± 0.02	0.44 ± 0.18	0.68 ± 0.05	0.86 ± 0.01
Spdef_0905.2_v2_deBruijn	0.17 ± 0.16	0.84 ± 0.01	0.83 ± 0.01	0.23 ± 0.14	0.76 ± 0.03	0.83 ± 0.01	0.40 ± 0.22	0.79 ± 0.02	0.90 ± 0.01
Srf_3509.1_v1_deBruijn	0.13 ± 0.12	0.76 ± 0.03	0.73 ± 0.02	0.23 ± 0.11	0.67 ± 0.04	0.76 ± 0.02	0.24 ± 0.12	0.70 ± 0.03	0.84 ± 0.01
Srf_3509.1_v2_deBruijn	0.13 ± 0.14	0.77 ± 0.02	0.75 ± 0.02	0.18 ± 0.12	0.67 ± 0.04	0.76 ± 0.02	0.17 ± 0.15	0.70 ± 0.04	0.84 ± 0.01
Sry_2833.2_v1_deBruijn	0.08 ± 0.09	0.74 ± 0.03	0.70 ± 0.03	0.17 ± 0.10	0.66 ± 0.05	0.75 ± 0.02	0.14 ± 0.11	0.67 ± 0.05	0.83 ± 0.02
Sry_2833.2_v2_deBruijn	0.10 ± 0.11	0.79 ± 0.02	0.75 ± 0.02	0.15 ± 0.09	0.74 ± 0.03	0.81 ± 0.02	0.19 ± 0.11	0.77 ± 0.03	0.89 ± 0.01
Tbp_pr781.1_v1_deBruijn	0.12 ± 0.11	0.80 ± 0.01	0.79 ± 0.01	0.41 ± 0.10	0.67 ± 0.05	0.77 ± 0.02	0.51 ± 0.07	0.66 ± 0.05	0.83 ± 0.01
Tbp_pr781.1_v2_deBruijn	0.14 ± 0.14	0.74 ± 0.01	0.73 ± 0.01	0.38 ± 0.05	0.65 ± 0.03	0.73 ± 0.01	0.39 ± 0.10	0.65 ± 0.04	0.78 ± 0.01
Tcf1_2666.2_v1_deBruijn	0.15 ± 0.16	0.87 ± 0.01	0.86 ± 0.01	0.28 ± 0.16	0.85 ± 0.03	0.89 ± 0.01	0.38 ± 0.15	0.85 ± 0.02	0.92 ± 0.01
Tcf1_2666.2_v2_deBruijn	0.07 ± 0.11	0.77 ± 0.02	0.77 ± 0.01	0.45 ± 0.11	0.69 ± 0.04	0.80 ± 0.02	0.50 ± 0.11	0.70 ± 0.05	0.87 ± 0.02
Tcf1_2666.3_v1_deBruijn	0.16 ± 0.14	0.80 ± 0.02	0.79 ± 0.02	0.44 ± 0.10	0.70 ± 0.03	0.78 ± 0.01	0.50 ± 0.10	0.72 ± 0.04	0.86 ± 0.01
Tcf1_2666.3_v2_deBruijn	0.08 ± 0.12	0.73 ± 0.02	0.72 ± 0.01	0.34 ± 0.08	0.64 ± 0.04	0.72 ± 0.02	0.43 ± 0.09	0.65 ± 0.05	0.84 ± 0.02
Tcf3_3787.1_v1_deBruijn	0.07 ± 0.10	0.79 ± 0.02	0.78 ± 0.02	0.21 ± 0.11	0.66 ± 0.03	0.75 ± 0.02	0.39 ± 0.19	0.68 ± 0.04	0.85 ± 0.02
Tcf3_3787.1_v2_deBruijn	0.13 ± 0.11	0.83 ± 0.01	0.82 ± 0.01	0.26 ± 0.11	0.72 ± 0.03	0.81 ± 0.01	0.39 ± 0.14	0.74 ± 0.03	0.88 ± 0.01
Tcf7_0950.2_v1_deBruijn	0.07 ± 0.10	0.76 ± 0.02	0.75 ± 0.02	0.16 ± 0.07	0.64 ± 0.03	0.73 ± 0.02	0.32 ± 0.18	0.67 ± 0.04	0.83 ± 0.02
Tcf7_0950.2_v2_deBruijn	0.11 ± 0.13	0.70 ± 0.01	0.69 ± 0.01	0.20 ± 0.09	0.62 ± 0.03	0.70 ± 0.02	0.42 ± 0.13	0.63 ± 0.04	0.77 ± 0.01
Tcf7l2_3461.1_v1_deBruijn	0.12 ± 0.12	0.77 ± 0.02	0.74 ± 0.02	0.19 ± 0.09	0.68 ± 0.05	0.78 ± 0.02	0.24 ± 0.14	0.68 ± 0.04	0.86 ± 0.02
Tcf7l2_3461.1_v2_deBruijn	0.11 ± 0.15	0.81 ± 0.02	0.80 ± 0.02	0.24 ± 0.07	0.70 ± 0.04	0.79 ± 0.02	0.40 ± 0.18	0.72 ± 0.04	0.88 ± 0.02
Tcfap2a_2337.3_v1_deBruijn	0.10 ± 0.10	0.57 ± 0.01	0.55 ± 0.01	0.15 ± 0.06	0.48 ± 0.04	0.56 ± 0.01	0.22 ± 0.10	0.48 ± 0.04	0.58 ± 0.02
Tcfap2a_2337.3_v2_deBruijn	0.12 ± 0.13	0.76 ± 0.01	0.74 ± 0.01	0.30 ± 0.07	0.69 ± 0.03	0.75 ± 0.01	0.50 ± 0.13	0.68 ± 0.03	0.77 ± 0.01
Tcfap2b_3988.1_v1_deBruijn	0.06 ± 0.12	0.69 ± 0.01	0.67 ± 0.01	0.26 ± 0.07	0.63 ± 0.03	0.68 ± 0.01	0.29 ± 0.15	0.63 ± 0.02	0.69 ± 0.01
Tcfap2b_3988.1_v2_deBruijn	0.13 ± 0.13	0.71 ± 0.00	0.70 ± 0.01	0.21 ± 0.10	0.63 ± 0.02	0.69 ± 0.01	0.37 ± 0.18	0.62 ± 0.04	0.71 ± 0.01
Tcfap2c_2912.2_v1_deBruijn	0.15 ± 0.13	0.58 ± 0.01	0.56 ± 0.01	0.23 ± 0.07	0.47 ± 0.03	0.55 ± 0.02	0.24 ± 0.07	0.47 ± 0.03	0.58 ± 0.01
Tcfap2c_2912.2_v2_deBruijn	0.16 ± 0.15	0.75 ± 0.01	0.73 ± 0.01	0.26 ± 0.08	0.67 ± 0.03	0.73 ± 0.01	0.38 ± 0.15	0.66 ± 0.03	0.75 ± 0.01
Tcfap2e_3713.1_v1_deBruijn	0.11 ± 0.16	0.78 ± 0.01	0.77 ± 0.01	0.29 ± 0.12	0.68 ± 0.03	0.78 ± 0.01	0.46 ± 0.16	0.70 ± 0.04	0.80 ± 0.01
Tcfap2e_3713.1_v2_deBruijn	0.24 ± 0.15	0.74 ± 0.01	0.74 ± 0.01	0.47 ± 0.07	0.69 ± 0.03	0.75 ± 0.01	0.52 ± 0.06	0.69 ± 0.02	0.78 ± 0.01
Tcfe2a_3865.1_v1_deBruijn	0.13 ± 0.11	0.65 ± 0.01	0.64 ± 0.01	0.18 ± 0.08	0.59 ± 0.02	0.64 ± 0.01	0.29 ± 0.15	0.61 ± 0.01	0.66 ± 0.01
Tcfe2a_3865.1_v2_deBruijn	0.19 ± 0.13	0.79 ± 0.02	0.77 ± 0.01	0.27 ± 0.10	0.80 ± 0.02	0.81 ± 0.01	0.31 ± 0.15	0.83 ± 0.01	0.86 ± 0.01
Zbtb12_2932.2_v1_deBruijn	0.12 ± 0.11	0.88 ± 0.02	0.88 ± 0.01	0.28 ± 0.17	0.75 ± 0.04	0.85 ± 0.02	0.55 ± 0.13	0.76 ± 0.04	0.94 ± 0.01
Zbtb12_2932.2_v2_deBruijn	0.15 ± 0.13	0.88 ± 0.02	0.87 ± 0.02	0.31 ± 0.16	0.78 ± 0.03	0.88 ± 0.01	0.50 ± 0.17	0.77 ± 0.05	0.94 ± 0.01
Zbtb3_1048.2_v1_deBruijn	0.12 ± 0.15	0.82 ± 0.01	0.81 ± 0.01	0.42 ± 0.15	0.80 ± 0.03	0.83 ± 0.01	0.54 ± 0.14	0.77 ± 0.04	0.86 ± 0.01
Zbtb3_1048.2_v2_deBruijn	0.11 ± 0.11	0.73 ± 0.01	0.71 ± 0.01	0.26 ± 0.15	0.64 ± 0.03	0.72 ± 0.01	0.48 ± 0.11	0.65 ± 0.03	0.76 ± 0.01
Zbtb7b_1054.2_v1_deBruijn	0.10 ± 0.13	0.77 ± 0.02	0.75 ± 0.02	0.23 ± 0.11	0.71 ± 0.04	0.78 ± 0.02	0.26 ± 0.13	0.74 ± 0.03	0.86 ± 0.01
Zbtb7b_1054.2_v2_deBruijn	0.14 ± 0.12	0.81 ± 0.02	0.78 ± 0.01	0.24 ± 0.11	0.74 ± 0.03	0.80 ± 0.01	0.43 ± 0.18	0.77 ± 0.04	0.90 ± 0.01
Zfp105_2634.2_v1_deBruijn	0.11 ± 0.08	0.71 ± 0.02	0.67 ± 0.02	0.16 ± 0.09	0.69 ± 0.04	0.74 ± 0.01	0.13 ± 0.12	0.70 ± 0.04	0.79 ± 0.01
Zfp105_2634.2_v2_deBruijn	0.11 ± 0.14	0.81 ± 0.02	0.78 ± 0.01	0.22 ± 0.13	0.76 ± 0.04	0.83 ± 0.01	0.28 ± 0.21	0.78 ± 0.02	0.87 ± 0.01
Zfp128_2806.2_v1_deBruijn	0.17 ± 0.11	0.83 ± 0.01	0.81 ± 0.01	0.29 ± 0.16	0.75 ± 0.02	0.82 ± 0.01	0.45 ± 0.17	0.75 ± 0.02	0.85 ± 0.01
Zfp128_2806.2_v2_deBruijn	0.15 ± 0.14	0.84 ± 0.01	0.82 ± 0.01	0.29 ± 0.15	0.75 ± 0.02	0.81 ± 0.01	0.30 ± 0.15	0.77 ± 0.02	0.88 ± 0.01
Zfp161_2858.2_v1_deBruijn	0.08 ± 0.15	0.86 ± 0.02	0.84 ± 0.01	0.28 ± 0.15	0.75 ± 0.05	0.83 ± 0.01	0.41 ± 0.21	0.79 ± 0.03	0.90 ± 0.01
Zfp161_2858.2_v2_deBruijn	0.12 ± 0.15	0.90 ± 0.01	0.88 ± 0.01	0.49 ± 0.14	0.84 ± 0.03	0.90 ± 0.01	0.57 ± 0.16	0.84 ± 0.04	0.93 ± 0.01
Zfp187_2626.2_v1_deBruijn	0.11 ± 0.12	0.75 ± 0.02	0.74 ± 0.02	0.21 ± 0.10	0.60 ± 0.04	0.72 ± 0.02	0.38 ± 0.18	0.63 ± 0.04	0.80 ± 0.02
Zfp187_2626.2_v2_deBruijn	0.09 ± 0.14	0.77 ± 0.02	0.76 ± 0.02	0.21 ± 0.10	0.65 ± 0.04	0.75 ± 0.01	0.43 ± 0.15	0.66 ± 0.03	0.82 ± 0.01
Zfp281_0973.2_v1_deBruijn	0.09 ± 0.10	0.71 ± 0.02	0.69 ± 0.02	0.14 ± 0.09	0.63 ± 0.04	0.71 ± 0.02	0.12 ± 0.10	0.68 ± 0.04	0.81 ± 0.02
Zfp281_0973.2_v2_deBruijn	0.07 ± 0.07	0.72 ± 0.02	0.68 ± 0.03	0.15 ± 0.10	0.62 ± 0.04	0.72 ± 0.02	0.18 ± 0.14	0.64 ± 0.05	0.84 ± 0.02
Zfp410_3034.2_v1_deBruijn	0.09 ± 0.11	0.81 ± 0.02	0.81 ± 0.02	0.28 ± 0.18	0.66 ± 0.04	0.79 ± 0.03	0.52 ± 0.11	0.68 ± 0.06	0.90 ± 0.01
Zfp410_3034.2_v2_deBruijn	0.11 ± 0.13	0.87 ± 0.02	0.87 ± 0.01	0.27 ± 0.16	0.80 ± 0.02	0.86 ± 0.01	0.49 ± 0.21	0.81 ± 0.03	0.92 ± 0.01
Zfp691_0895.2_v1_deBruijn	0.13 ± 0.20	0.88 ± 0.01	0.87 ± 0.01	0.37 ± 0.14	0.83 ± 0.02	0.88 ± 0.01	0.45 ± 0.22	0.86 ± 0.02	0.93 ± 0.01
Zfp691_0895.2_v2_deBruijn	0.20 ± 0.19	0.83 ± 0.02	0.81 ± 0.02	0.37 ± 0.14	0.80 ± 0.02	0.84 ± 0.01	0.38 ± 0.10	0.81 ± 0.02	0.88 ± 0.01
Zfp740_0925.2_v1_deBruijn	0.08 ± 0.08	0.84 ± 0.03	0.82 ± 0.03	0.18 ± 0.11	0.73 ± 0.04	0.82 ± 0.02	0.27 ± 0.17	0.77 ± 0.04	0.92 ± 0.01
Zfp740_0925.2_v2_deBruijn	0.06 ± 0.08	0.79 ± 0.02	0.76 ± 0.02	0.15 ± 0.09	0.72 ± 0.04	0.79 ± 0.02	0.29 ± 0.16	0.74 ± 0.04	0.89 ± 0.02
Zic1_0991.2_v1_deBruijn	0.13 ± 0.11	0.68 ± 0.02	0.65 ± 0.02	0.14 ± 0.10	0.62 ± 0.04	0.70 ± 0.02	0.16 ± 0.11	0.65 ± 0.03	0.77 ± 0.01
Zic1_0991.2_v2_deBruijn	0.09 ± 0.10	0.59 ± 0.02	0.56 ± 0.02	0.12 ± 0.09	0.53 ± 0.03	0.60 ± 0.02	0.15 ± 0.08	0.56 ± 0.03	0.68 ± 0.02
Zic2_2895.2_v1_deBruijn	0.11 ± 0.09	0.78 ± 0.02	0.76 ± 0.02	0.19 ± 0.09	0.69 ± 0.04	0.78 ± 0.02	0.21 ± 0.14	0.74 ± 0.03	0.86 ± 0.02
Zic2_2895.2_v2_deBruijn	0.10 ± 0.10	0.75 ± 0.02	0.74 ± 0.03	0.23 ± 0.11	0.67 ± 0.04	0.76 ± 0.02	0.42 ± 0.15	0.69 ± 0.04	0.83 ± 0.02
Zic3_3119.2_v1_deBruijn	0.16 ± 0.13	0.76 ± 0.02	0.74 ± 0.01	0.21 ± 0.13	0.70 ± 0.03	0.76 ± 0.01	0.24 ± 0.13	0.72 ± 0.03	0.81 ± 0.01
Zic3_3119.2_v2_deBruijn	0.13 ± 0.12	0.84 ± 0.01	0.81 ± 0.02	0.21 ± 0.10	0.73 ± 0.04	0.83 ± 0.02	0.22 ± 0.14	0.76 ± 0.04	0.88 ± 0.01
Zscan4_2667.2_v1_deBruijn	0.14 ± 0.15	0.87 ± 0.02	0.85 ± 0.02	0.53 ± 0.08	0.77 ± 0.04	0.86 ± 0.01	0.55 ± 0.12	0.76 ± 0.04	0.90 ± 0.02
Zscan4_2667.2_v2_deBruijn	0.09 ± 0.08	0.84 ± 0.02	0.84 ± 0.02	0.40 ± 0.15	0.69 ± 0.05	0.80 ± 0.01	0.46 ± 0.11	0.67 ± 0.04	0.87 ± 0.01