GAPWM: a genetic algorithm method for optimizing a position weight matrix

Leping Li1,*, Yu Liang1 and Robert L. Bass2
1Biostatistics Branch and 2Computational Biology Facility, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27709, USA

Received on December 19, 2006; revised on February 21, 2007; accepted on February 27, 2007
Advance Access publication March 6, 2007
Associate Editor: Martin Bishop

ABSTRACT
Motivation: Position weight matrices (PMWs) are simple models commonly used in motif-finding algorithms to identify short functional elements, such as cis-regulatory motifs, on genes. When few experimentally verified motifs are available, estimation of the PWM may be poor. The resultant PWM may not reliably discriminate a true motif from a false one. While experimentally identifying such motifs remains time-consuming and expensive, low-resolution binding data from techniques such as ChIP-on-chip and ChIP-PET have become available. We propose a novel but simple method to improve a poorly estimated PWM using ChIP data.

Methodology: Starting from an existing PWM, a set of ChIP sequences, and a set of background sequences, our method, GAPWM, derives an improved PWM via a genetic algorithm that maximizes the area under the receiver operating characteristic (ROC) curve. GAPWM can easily incorporate prior information such as base conservation. We tested our method on two PMWs (Oct4/Sox2 and p53) using three recently published ChIP data sets (human Oct4, mouse Oct4, and human p53).

Results: GAPWM substantially increased the sensitivity/specificity of a poorly estimated PWM and further improved the quality of a good PWM. Furthermore, it still functioned when the starting PWM contained a major error. The ROC performance of GAPWM compared favorably with that of MEME and others. With increasing availability of ChIP data, our method provides an alternative for obtaining high-quality PMWs for genome-wide identification of transcription factor binding sites.

Availability: The C source code and all data used in this report are available at http://dir.niehs.nih.gov/dirbb/gapwm
Contact: li3@niehs.nih.gov
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

For a transcription factor with a set of experimentally verified binding-motif sequences, a position weight matrix (PWM) may be constructed from these known sequences by calculating the proportions of sequences for which each specific base, A, C, G and T, occurs at each position in a set of aligned motif sequences (for a review, see Stormo 2000). Once a PWM is constructed, it can be used to scan sequences for putative binding sites using a sliding window of length of the PWM to score how well each sequence segment in the window matches the PWM. A site is declared when the score passes a predefined cutoff. Well-known PWM-based methods include the MatInspector (Cartharius et al., 2005; Quandt et al., 1995) and Match (Kel et al., 2003). Complex tools combining PWM and pattern matching have also been reported (Matys et al., 2006). For a recent review on methods for identifying regulatory elements, see Wasserman and Sandelin (2004).

The commonly used PMWs assume that the positions within a motif are mutually independent, i.e. a motif sequence follows a product of multinomial distributions. Thus, the observed frequencies of A, C, G and T in each column are the maximum likelihood (ML) estimates of the distribution of the multinomial random variable for that column, regardless of the contents of nearby columns. Many PMWs are available in the TRANSFAC (Knuppel et al., 1994) and JASPER (Sandelin et al., 2004) databases; however, the number of known transcription factor binding sites in these databases is typically small. Thus, the ML estimates in many PMWs may be poor, as the estimators are vulnerable to overfitting when based on insufficient data. The resultant PWM models may be ineffective in distinguishing a true motif from a random segment. A further complication arises from the choice of cut point for declaring a site to be a motif. A less stringent cut point results in a large number of false positives whereas a more stringent cut point eliminates true positives. This problem makes the genome-wide motif identification using PMWs less useful.

While experimentally identifying and validating a binding site is time-consuming, chromatin immunoprecipitation (ChIP) provides an alternative for identifying direct target genes via isolating the DNA fragments bound by the protein of interest. When coupled with microarray (ChIP-on-chip) (Kim and Ren, 2006) or other technologies such as paired end tags (ChIP-PET) (Ng et al., 2005), regions where a protein binds can be identified throughout the entire genome. Such regions are typically 500–1500-bp long with an embedded 6–30 bp functional element and are considered low-resolution binding sites. Recently, ChIP-on-chip and ChIP-PET have been successfully used to identify target genes of Oct4, Sox2 and...
Nanog (Boyer et al., 2005; Loh et al., 2006) and p53 (Wei et al., 2006) transcription factors. In addition, the ENCODE project aiming to identify all functional elements in human genome (ENCODE Project Consortium, 2005) has also generated a large amount of low-resolution binding data. Because these techniques identify regions rather than specific motifs bound by a protein, the resultant data are rarely used directly to construct a PWM. Rather, de novo motif discovery programs such as AlignACE (Hughes et al., 2000), BioProspector (Liu et al., 2001), CONSENSUS (Hertz and Stormo, 1999), MDScan (Liu et al., 2002), MEME (Bailey and Elkan, 1994) motifSampler (Thijs et al., 2001), MUSA (Mendes et al., 2006) and Weeder (Pavesi et al., 2004) are typically used to identify over-represented motifs from such data. The performance of several of these methods has recently been assessed (Tompa et al., 2005). Recently, Ji et al. (2006) applied a Gibbs sampler to some of the ChIP data sets. Although these methods are reasonably successful, finding all of the over-represented motifs in a set of sequences with an average length of 1 kb remains a formidable computational challenge.

Herein we present a method, GAPWM, for improving an existing poorly estimated PWM using ChIP data. This task is relatively simple compared to de novo motif discovery. First, de novo methods seek PWMs for multiple motifs, but our aim is to enhance a single PWM. Second, most de novo methods use a sampling strategy that typically starts at random positions. We use a PWM as the starting point, thus, the sampling space is dramatically reduced. We employ a genetic algorithm (GA), a widely used approach (Goldberg, 1989), as the optimization tool. Earlier, we developed a GA/KNN approach for selecting discriminative features for sample classification (Li et al., 2001). Recently, GA has also been used to optimize hidden Markov models (Won et al., 2004). Wei and Jensen (2006) replaced the Metropolis sampling step in BioOptimizer by a genetic algorithm. To our knowledge, GA has not been used to improve PWMs. We show that our method, GAPWM, can dramatically improve the quality of a PWM.

2 DATA

2.1 Sequence data

We used three sets of ChIP data in our analysis: human Oct4 (Boyer et al., 2005), mouse Oct4 (Loh et al., 2006) and human p53 (Wei et al., 2006). All were downloaded from the UCSC genome browser using custom tracks. The numbers of sequences in human Oct4, mouse Oct4 and human p53 are 603, 1083 and 542, respectively, with average lengths 695, 1187 and 1481 bp, respectively. The 1083 mouse Oct4 loci had overlap PET size of 4 or above; but we only used loci with an overlap PET size of 5 or greater, reducing the number of Oct4 loci to 367. The repetitive elements in human Oct4 data were masked by N’s. To investigate the behavior of GA optimization using sequence data containing repetitive elements, we chose not to mask the repetitive elements in both mouse Oct4 and human p53 ChIP data.

We randomly divided each ChIP data set into a training set and a test set of roughly equal size. We repeated this random splitting M times to generate M sets of training and test data. Next, we randomly selected 1500 sequences of equal length to the training data from the coding regions of the human and mouse reference genes as the background sequences, respectively. This procedure was also repeated M times. For human Oct4, mouse Oct4 and human p53 ChIP data, M = 3, 2 and 1, respectively.

2.2 Initial PWM models

Our GAPWM method starts with an initial PWM. The starting PWM can be constructed from a set of binding sites taken from databases such as TRANSFAC and JASPER or from a consensus-binding site.

The initial PWM for Oct4/Sox2 was created from 11 known human and mouse Oct4/Sox2 binding sites. The p53 is known to bind to two half sites, each of which has the canonical consensus sequence of 5'-PuPuPuPuA(T)/C(T/A)GPyPyPy-3', where Pu and Py identify purine (A,G) and pyrimidine bases (C,T), respectively. We set the base frequency to 1 if the base matches the corresponding base in the consensus sequence; otherwise to 0. All starting PWMs were standardized so that the column sum is equal to 1. The number of columns (motif length) for Oct4/Sox2 and p53 are 17 and 20, respectively.

3 METHODS

3.1 Scoring a motif

We adapted the methodology of Match (Kel et al., 2003) to score a sequence segment. Let \( m = (x_1, x_2, \ldots, x_w) \) be a sequence segment of length w, where \( x_i \) is the base at position \( i \) in \( m \). The score of segment \( m \) is calculated as follows,

\[
s(m) = s(x_1, x_2, \ldots, x_w) = \sum_{i=1}^{w} \log_{10} 4 \cdot f_{i,x} - \sum_{i=1}^{w} \log_{10} 4 \cdot f_{i,x}^{\min} - \sum_{i=1}^{w} \log_{10} 4 \cdot f_{i,x}^{\max}
\]

where \( f_{i,x} \) is the observed frequency of base \( x \) at position \( i \) and \( f_{i,x}^{\min} \) and \( f_{i,x}^{\max} \) are the minimal and maximal base frequencies at position \( i \), respectively. \( R(i) \) is the information at position \( i \), defined as

\[
R(i) = \frac{1}{\log_{10} 4} \left( \sum_{x} f_{i,x} \cdot \log_{10}(f_{i,x}) + \log_{10} 4 \right)
\]

For a sequence of length \( L \), the algorithm computes \( s(m) \) for each of the \( 2^L \cdot (L - w + 1) \) sites by moving a sliding window of length \( w \) through the sequence for both the forward and reverse complementary strands.

3.2 The genetic algorithm for PWMs (GAPWM)

3.2.1 Initialization

We made 100 'clones' of the starting PWM to form a 'population'. Different population sizes can be used.

3.2.2 Mutation

All except the best performing PWM were subject to 'mutation'. No 'crossover' operation was applied. For each PWM to be mutated, \( n \) (the number of columns subject to mutation) was randomly generated from the following distribution, \( P(n = k) = 1/2^k \), \( k = 2, \ldots, w \) and \( P(n = 1) = 1 - \sum_{k=2}^{w} 1/2^k \). For Oct4/Sox2, \( E(n) \sim 2 \), indicating that on average two positions/columns were changed.
Next, $n$ columns were randomly selected with equal probability from the $w$ positions. For each column chosen, one of the four bases was then randomly selected and a small value, $d$, was added to or subtracted from $f_{i,j}$ with equal probability. Mutation was performed independently for each selected column. After mutation, we set negative values to 0 and standardized each column.

We typically set the value of $d$ to 0.05 for the first 100 generations and to 0.02 for the remaining generations. A larger mutation early in the GA run allows the algorithm to explore a larger sampling space, whereas a smaller mutation allows the search to thoroughly sample the local space and find the maximum. Also, we sometimes impose constraints on certain positions in the PWM. For Oct4/Sox2, positions 6–9 (geat) and 11–14 (aaca) are high conserved in the known binding sites. We fixed such positions (no mutations were allowed) during the first 100 generations. We also introduced additional constraints in which only certain bases in a column were allowed to mutate. For example, in the 11 known Oct4/Sox2-binding sites, positions 2, 4, 5 and 15 consist of a’s and t’s only. Mutations in these positions were restricted to these two bases during the initial 100 generations. Such constraints are useful for obtaining good starting values for columns with no observed data (e.g. base frequency = 0.25 for all four bases). This scenario arises when a motif is extended beyond its current length. Mutations were unconstrained for the remaining generations.

### 3.2.3 Fitness evaluation

Each PWM was evaluated for fitness, that is, to determine how well it distinguishes true motifs from background motifs. We used the segment-scoring procedure outlined in Section 3.1. For a sequence of length $L$, we scanned both forward and reverse complementary strands using a sliding window of size of $w$. The score for the entire sequence was defined as the highest score among all $2(L-w+1)$ scores. These scores, one per sequence, were ranked separately for the training set and the background set. We assumed that each ChIP sequence in the training set contains at least one motif and each background sequence contains no motifs. Thus, given a threshold, one can calculate ‘true’ positive and ‘false’ positive rates, defined as the proportions of sequences scoring at or above the threshold in training and background sets, respectively. Finally, we defined the fitness score of each PWM as the area under the receiver operating characteristic (ROC) curve across a range of false positive rates from 0 to 0.1. Essentially, GAPWM aims to maximize the sensitivity across a range of specificity from 100 to 90%. This range of specificity is biologically meaningful and computationally efficient, as GAPWM does not need to integrate the whole area under the ROC curve.

### 3.2.4 Selection

We used roulette wheel-based selection to form the next generation of PWMs from the current generation. The single highest scoring PWM was always guaranteed to pass onto the next generation. The remaining PWMs were selected with probabilities proportional to their fitness values with replacement. The larger the fitness value, the better chance of being selected.

The newly formed population of PWMs was subject to a new round of mutation, fitness evaluation and selection. This cycle was continued until a predefined number of generations (e.g. 1500) had been reached. Alternatively, convergence of the PWM can also be used as the stopping criterion. Typically, for each optimization, we performed eight independent GA runs as GA is a stochastic search algorithm and does not guarantee finding the global optimum in a single run.

### 4 RESULTS

#### 4.1 Optimization of a poorly estimated PWM

##### 4.1.1 Human Oct4 ChIP data

We performed eight independent GA runs on the same training and background data with the same parameter settings (population size = 100, maximal number of generations = 1500) but different random seeds. For each run, we saved the single highest scoring PWM from each generation, based on the training set and its background set, 1500 in total. We then applied the best PWM from each generation of the training set to the test set and its background set and computed fitness scores for the test sets. The results for all eight runs were comparable (Supplementary Fig. S1). If the results differ substantially, one should choose the run that resulted in the highest overall fitness scores for both the training set and test sets as the best run. For fairness, we chose an average performing run (No. 3) for the subsequent comparisons. The fitness score for this run more than doubled after ~500 generations and stabilized (Supplementary Fig. S1). Compared to the training set, the fitness scores for the test set fluctuated. This result is expected as the test set and its background set were not used in optimizing the PWM. Nonetheless, a substantial improvement in fitness score is evident for both the training data and the test data.

Next, we averaged the single highest scoring PWMs from generations 500 to 1500 to obtain an averaged PWM (referred to as the PWM from GA). We computed the ROC curves for the PWM from GA and the starting PWM using the same parameter settings and constraints. Eight independent runs were carried out. In all runs, the GA-optimized Oct4 PWMs outperformed the starting PWM (Fig. 1a). The ROC curves for all eight averaged PWMs are provided in Supplementary Figure S1. We then qualitatively assessed the performances of PWMs using these ROC curves. We reasoned that the larger the area under the ROC, the better the performance of a PWM. The area for the PWM from GA was substantially larger than that for the starting PWM (Fig. 1a), suggesting that the PWM from GA performed much better than the starting PWM.

To see if our GA-optimized PWM can still recognize the 11 known binding sites from which the initial PWM were constructed, we applied the PMW to the 11 known sites and 5000 17-bp sites that were randomly selected from the regions of known human reference genes. We found that only one of the 5000 background sites scored higher than one of the 11 known sites (data not shown). This result suggests that the GA optimization did not converge to a different (but perhaps wrong) global optimum.

To test the robustness of the GA optimization, we repeated the entire procedure two more times. In each pass, we started with a random division of the 603 ChIP sequences into training and test sets and a random selection of two independent background sequence sets. In all three passes, the PWMs from GA performed comparably to each other and consistently outperformed the starting PWM (Supplementary Figs S2 and S3).

Lastly, we optimized the Oct4 PWM (without the corresponding Sox2 PWM) using the same parameter settings and constraints. Eight independent runs were carried out. In all runs, the GA-optimized Oct4 PWMs outperformed the starting PWM (not shown) but performed worse than the GA-optimized Oct4/Sox2 PWM. The performance of a typical GA-optimized Oct4 PWM (hOct4_GA) is shown in Figure 1a.

##### 4.1.2 Mouse Oct4 ChIP-PET data

Applying the same parameter settings and constraints as we used in the human Oct4 data to the mouse Oct4 data, we performed
eight independent GA runs on the same mouse training data. As with the human Oct4 data, the fitness score doubled for the training data and nearly doubled for the test data after 1000 generations (Supplementary Fig. S4). For each run, we averaged the single highest scoring PWMs from generations 500–1500 to obtain an averaged PWM and compared its performance with that of the starting PWM using the test set and its background set. As before, we selected a typical run (No. 3) for performance comparisons. The ROC analysis comparing the PWM from GA to the starting PWM (Fig. 1b) indicated that GA was capable of improving the starting PWM using ChIP sequences containing repetitive elements.

We also repeated the entire above procedure one more time. The results of the second eight independent runs were comparable (Supplementary Fig. S5).

### 4.1.3 Human p53 ChIP data
We did four independent runs on human p53 ChIP data. In all four runs, the fitness score increased sharply within 50 generations and stabilized (Supplementary Fig. S6). Thus, we stopped the runs before they reached the predefined 1500 generations. Unlike the human and mouse Oct4 ChIP data, the ROC curves for both the PWM from GA and the starting PWM were steep and quickly reached to 0.9 (Fig. 2a), suggesting that (1) nearly all of the ChIP sequences contain at least one p53 motif and the background sequences contain none or few motifs and (2) the two PWMs are nearly equally effective in finding these motifs.

### 4.2 Comparison with EM-based algorithms
In principle, de novo motif discovery programs should identify the Oct4/Sox2 and p53-binding sites in the corresponding ChIP data. Using a highly stringent criterion by requiring PET overlap size of 7 or greater, Loh et al. (2006) selected 90 mouse Oct4 loci and applied the Weeder algorithm to these sequences. A total of 108 Oct4/Sox2-binding sites were identified and a PWM was constructed from these binding sites. The authors subsequently improved this PWM using an expectation-maximization (EM) procedure described in their Supplementary Material. The EM-optimized PWM, which we will refer to as the PWM from NMICA, showed substantial improvement over the initial PWM from Weeder as assessed by ROC curves.

MEME is a well-known and efficient de novo motif discovery program that also uses an EM procedure. We downloaded the MEME source code and applied it to all training sets (three human Oct4, two mouse Oct4 and one human p53—see Section 2.1 for details). In all MEME runs, the repetitive elements in the training sets were masked (if not done so in GA optimization). We used the ZOOPS model (zero or one occurrence per sequence), provided a consensus sequence of catttgcataacaatgg as the starting point, and a fifth order Markov chain from the coding sequences of all human and mouse reference genes as the background models for human and mouse training data, respectively. Using these settings, MEME was able to find the binding sites in five of the six data sets (human Oct4 splits 1 and 2, mouse Oct4 splits 1 and 2, and human p53 split 1), resulting in five PWMs. We will refer to these PWMs as the PWMs from MEME.

We computed the ROC curves for the PWMs from MEME and NMICA using our test sets and their corresponding background sets. We compared the ROC curves for the PWMs from MEME and NMICA with the ROC curves for the PWMs from GA (Figs 1 and 2). In all cases, the ROC curves for the PWMs from GA are the left-most curves, suggesting that the PWMs from GA performed better than PWMs from MEME and NMICA.

### 4.3 Optimization of a suboptimal PWM
MEME is an effective motif-discovery program especially when a consensus sequence is used to approximate the
motif-start positions. For the first human Oct4 training set (split 1), MEME found 101 Oct4/Sox2 motifs in 302 sequences. This result is comparable to our estimate of the proportion of number of Oct4/Sox2-binding sites in this data set (35% true positive rate at 5% false positive rate). Thus, we believe that the PWM from MEME was near optimal, if not optimal. To investigate if GA can further improve such a PWM, we used the PWM from MEME as the starting PWM. We performed eight independent GA runs on the same training data from which the PWM from MEME was obtained and its background set. All parameter settings and constraints were the same as before. In all eight runs, the fitness score increased by ~20% from its starting value. The ROC curves for the PWM from GA and the starting PWM from MEME are shown in Figure 2b. GA further improved the PWM from MEME. The results for all eight runs are shown in Supplementary Figure S7.

4.4 Robustness to an error in starting PWM
So far, we showed that our method can improve a poorly estimated, but correct, PWM. To test if our method can still function when the starting PWM contains a major error, we changed a conserved base at position 8 from ‘A’ to ‘T’ in the starting Oct4/Sox2 PWM by setting $f_{8,a} = 0$ and $f_{8,t} = 1$. We specifically chose this position as all available evidence suggests that this position is the most conserved position among all 17 positions. We then carried out eight independent GA runs on the first human Oct4 training data (split 1) and its background set for 2000 generations. All other parameter settings and constraints were the same as before. Positions 6–9 (gcac) and 11–14 (acaa) including 8 were fixed during the first 100 generations. In all runs, $f_{8,a}$ increased while $f_{8,t}$ decreased. The frequencies of the other two bases remained close to 0. Among the eight runs, the PWM having the highest fitness score also had the largest $f_{8,a} = 0.636$ (peaked at generation of 1500) (Supplementary Fig. S8). The optimized PWM closely resembles the optimized PWM from the original starting PWM. This result demonstrates that many small mutations can result in large changes in base frequency, such as from 0 to 0.636 and that GA may be capable of jumping out from a sub-optimal local maximum.

5 DISCUSSION
PWMs are simple motif models that have been widely used in motif-identification algorithms. Despite their popularity, users can be frustrated when a large number of hits are produced, many of which are potentially false positives. This problem can arise when insufficient motif data are used to construct a PWM and/or the independence assumption between positions is violated. Between these two, lack of data is more serious and cannot be solved computationally. Methods that address the dependency between positions have been reported (Ellrott et al., 2002; Frith et al., 2004; Huang et al., 2006; Naughton et al., 2006; Thijs et al., 2001; Zhou and Liu, 2004). Although the dependent models are in principle superior, more data are needed to adequately model the dependency.

While experimental identification and validation of binding sites remains time-consuming and expensive, low-resolution binding-site data have become available. Using these data, we propose a GA method for improving a PWM, GAPWM. We tested our method on three data sets with multiple divisions into a training set and a test set. We showed that our method can improve sensitivity/specificity of a poorly estimated PWM when evaluated using ROC curves.

GAPWM starts with a PWM that can be constructed from a single binding site or a consensus sequence. It maximizes the area under the ROC curve across a range of false positive rates from 0 to 0.1, thus, maximizing the true positive rate. This maximization is done by adjusting the parameters in the PWM.

![Figure 2](https://academic.oup.com/bioinformatics/article-abstract/23/10/1188/196564)
‘intelligently’ through many small steps using a genetic algorithm. As a result, our method attempts to improve any starting PWM by making small incremental changes. However, many de novo motif-discovery methods often fail to find the motifs that are not significantly enriched, i.e. with more than 50% of the sequences being noisy. Many de novo methods also fail to find true motifs if repetitive elements are not masked. We ran several de novo algorithms using their default settings and only MEME (with a consensus motif) and Weeder (NMICA, motif length limited to 12) were able to find the Oct4/Sox2 motifs. Moreover, we showed that our method may further improve a good PWM from the motifs identified by de novo motif-discovery algorithms such as MEME. Although our method may not substantially improve such a PWM, even a small improvement may be beneficial when used in a genome-wide scan. Lastly, we showed that our method can potentially tolerate errors in the starting PWM. It ‘corrected’ an error we introduced in a starting PWM in which we set the observed frequency of a conserved base to 0 and the frequency of a non-observed based to 1. This result demonstrates that (1) our method may be capable of jumping out from a local maximum and (2) small incremental mutations in GA can, after many generations, have a large effect on a PWM. Despite all of the above, our method is likely to fail when all elements in the starting PWM are 0.25; the subsequent sampling space is too large to obtain a good starting position.

In our work, we assumed that each sequence contains at least one true motif in the ChIP data set and no motif in the background set. This assumption may be violated as ChIP may identify regions from indirect DNA–protein interactions. In addition, false positives can arise from other sources. Unfortunately, one does not know which sequences contain true motifs and the locations of the motifs. Thus, it is impossible to plot the true ROC curves for ChIP data. However, under our assumption, performance comparisons of different PWMs using the ROC curves should be meaningful. If we had known the true binding-site information, we believe that the ROC curves would have been much steeper.

The above assumption limits the steepness of the ROC curves. Despite the assumption, one might wonder why the ROC curve for the Oct4/Sox2 PWM is not steeper after the GA optimization. For human Oct4 ChIP data, the ‘true’ positive rate was only 34% at a ‘true’ negative rate of 5%. We initially reasoned that this result may be largely due to the fact that the ChIP Oct4 data set contains loci bound by only the Oct4 protein and our PWM is a model for joint un-gapped Oct4 and Sox2-binding sites. The negative sequences from the ChIP experiment would be needed. Among different regions of the genome, coding regions mostly likely contain the fewest binding sites. For a ROC scoring-based approach, a ‘true’ negative sequence set would be needed. Among different regions of the genome, coding regions mostly likely contain the fewest binding sites. The negative sequences from the ChIP experiment would also be excellent choices.

Overtraining is often a concern, especially for small training data sets. Fortunately, ChIP data sets are typically large (the number of sequences is in the hundreds to thousands). In most runs, we observed that the fitness scores for the training sets were slightly higher than those for the test sets. Also, one would usually expect that the fitness scores for the training data may continue to increase slowly as the run continues. However, fitness scores for the test set may not change or may even decrease, especially for a long run. Thus, like many other optimization problems, one should monitor the performance of the target function not only on the training set but also on the test set. A model performing the best for both the training set and test set may be chosen as the final model. Fortunately, in most cases, the performances for both training set and test set were consistent. Averaging over many generations based on test set performance may also alleviate overtraining.

In conclusion, we have proposed a novel simple method, GAPWM, for improving a PWM using ChIP data. We showed that our method can substantially improve the performance of a poorly estimated PWM and further improve a good PWM from a de novo motif-discovery method such as MEME. With increasing availability of ChIP data, GAPWM should prove
useful for improving PWMs for genome-wide identification of transcription factor binding sites.

ACKNOWLEDGEMENTS

We are grateful to Grace Kissling, David Umbach and Lee Pedersen for their comments and carefully reading the manuscript. We thank Connie Chen for downloading many of the sequence data sets used in this work and Rongheng Lin for discussion. We also thank the Computational Biology Facility at NIEHS for computing time. This research was supported by Intramural Research Program of the NIH, National Institute of Environmental Health Sciences.

Conflict of Interest: none declared.

REFERENCES


