Chapter 15
Evolutionary Optimization of Transcription Factor Binding Motif Detection

Zhao Zhang, Ze Wang, Guoqin Mai, Youxi Luo, Miaomiao Zhao and Fengfeng Zhou

Abstract All the cell types are under strict control of how their genes are transcribed into expressed transcripts by the temporally dynamic orchestration of the transcription factor binding activities. Given a set of known binding sites (BSs) of a given transcription factor (TF), computational TFBS screening technique represents a cost efficient and large scale strategy to complement the experimental ones. There are two major classes of computational TFBS prediction algorithms based on the tertiary and primary structures, respectively. A tertiary structure based algorithm tries to calculate the binding affinity between a query DNA fragment and the tertiary structure of the given TF. Due to the limited number of available TF tertiary structures, primary structure based TFBS prediction algorithm is a necessary complementary technique for large scale TFBS screening. This study proposes a novel evolutionary algorithm to randomly mutate the weights of different positions in the binding motif of a TF, so that the overall TFBS prediction accuracy is optimized. The comparison with the most widely used algorithm, Position Weight Matrix (PWM), suggests that our algorithm performs better or the same level in all the performance measurements, including sensitivity, specificity, accuracy and Matthews correlation coefficient. Our data also suggests

Zhao Zhang and Miaomiao Zhao have been contributed equally to this paper.

Z. Zhang · Z. Wang
School of Computer Science and Software Engineering, Tianjin Polytechnic University,
Tianjin, China

Z. Zhang · G. Mai · Y. Luo · M. Zhao · F. Zhou (✉)
Shenzhen Institutes of Advanced Technology and Key Laboratory for Health Informatics,
Chinese Academy of Sciences, Shenzhen, Guangdong, China
e-mail: FengfengZhou@gmail.com; ff.zhou@siat.ac.cn

Y. Luo
School of Science, Hubei University of Technology, Wuhan, Hubei, China

D. Wei et al. (eds.), *Advance in Structural Bioinformatics*, Advances in
Experimental Medicine and Biology 827, DOI 10.1007/978-94-017-9245-5_15

that it is necessary to remove the widely used assumption of independence between motif positions. The supplementary material may be found at: http://www.healthinformaticslab.org/supp/ .

Keywords Binding sites · Transcription factor · Position weight matrix · Motif

15.1 Introduction

Transcription of genic regions into RNA molecules is the first step of the biological central dogma, and is dynamically controlled by various transcription factors (TFs) [1]. A TF regulates a gene's transcription through its dynamic binding to a short (5–20 bps) DNA sequence upstream to the regulated gene. This DNA sequence is the TF's binding site (TFBS), which is usually highly specific to this TF and is called a motif [2]. Mutations within TFBSs will change the host's transcription regulatory network, and lead to species specific phenotypes or genetic diseases [3].

There are two major high-throughput strategies to screen the binding sites of a TF in the host genome. Firstly, various high-throughput experimental techniques were developed to screen the TFBSs under the given cell culture conditions, including DNase I footprinting [4], electrophoretic mobility shift assay [5], ChIP-on-chip [6] and ChIP-Seq [7], etc. The dynamic landscape of the transcription regulatory network may be elucidated through these screening techniques. But they are usually costly and labor-intensive, and can only detect the binding sites of one TF under one cell culture condition at a time. Considering the 2,886 transcription factors curated in the human DNA-binding domain (DBD) database [8], and the dynamic nature of transcription regulation, it can be anticipated that the transcription regulatory landscape is significantly under-estimated.

Computational TFBS screening techniques have been used to infer the comprehensive list of TFBSs. The majority of *in silico* TFBS screening techniques assumes that the binding sites of a given TF have a fixed length, and calculates the similarity score of a query DNA sequence compared with the local oligo-nucleotide frequency patterns in the known TFBSs [9]. The computational techniques include the position weigh matrix (PWM) [10], WebLogo [11], and position specific pairwise score [12], etc. The introduction of TF's structural information will greatly reduce the false positive rates, as demonstrated by Facelli [13], Saito et al. [14]. But there are only 300 unique human TF structures in the PDB database [ref], and the limited availability of the experimentally detected TF structures restricts the extensive application of these methods [15].

This study hypothesizes that positions contribute differently to the motif scoring based on their nucleotide frequency patterns, and formulates the position contribution as a weight for the position. The vector of weights for different motif positions were randomly mutated by an evolutionary algorithm, with the optimization goal to maximize the overall accuracy. The prediction performance suggests that our algorithm performs similarly or better than the position specific scoring strategies.

15.2 Materials and Methods

15.2.1 Data Resources

The proposed algorithm is applied to the following seven transcription factors (TFs), i.e. Ebox, Myc, P53, Q6MAZ, Q601MAZ, V_SREBP_Q3-SREBP (abbreviated as Q3), and V_SREBP2_Q6-SREBP2 (abbreviated as Q6). The known binding sites of these seven transcription factors were manually collected from the database TRANSFAC in August 2012 [16]. Only those binding sites without an "N" letter were kept for further analysis. The target gene sequences and their promoter regions were extracted from the database ENSEMBL [17].

15.2.2 Motif Screening Problem

The mathematical model of the transcription factor binding site (TFBS) screening problem (sTFBS) is formulated as follows. For a given transcription factor (TF), its known fixed-length binding sites are defined to be the positive dataset $P = \{M_1, M_2, \ldots, M_n\}$, where $|M_i| = L$. A negative dataset $N = \{B_1, B_2, \ldots, B_m\}$ is randomly extracted from the promoter regions of the genes regulated by the given TF, where $|B_j| = L$, B_j has no "N" letters and B_j does not overlap with M_i. Considering the promoter region is much larger than a TFBS, we set $m = 10 \times n$. A TFBS screening model is denoted as the classification function $f(X) \in \{P, N\}$, where $X \in P \cup N$.

Firstly, a similarity score between two fixed-length DNA fragments $V = \{v_1, v_2, \ldots, v_L\}$ and $U = \{u_1, u_2, \ldots, u_L\}$ is defined to be $Score(V, U) = (w_1 \times S(v_1, u_1) + w_2 \times S(v_2, u_2) + \cdots + w_L \times S(v_L, u_L))$, where the weight vector $W = \langle w_1, w_2, \ldots, w_L \rangle$ is the pre-calculated combination pattern, and $w_i \in [0, 1]$. The nucleotide similarity score matrix $S(v_i, u_i)$ is defined to be 2 if $v_i = u_i$, 1 for A versus G or C versus T, and -1 for the other pairs [18]. The combination pattern $W = \langle w_1, w_2, \ldots, w_L \rangle$ will be optimized by an evolutionary algorithm, as described in the next section.

This study chose the simple nearest neighbor algorithm as the classification model $f(X)$.

```
Algorithm SNN
Input: The positive and negative datasets are $P=\{M_1, M_2, …, M_n\}$ and $N=\{B_1, B_2, …, B_m\}$, respectively, where $|M_i|=|B_j|=L$. The query sequence is $Q$, where $|Q|=L$.
Procedure:
MaxScoreP=$Score(Q, M_1)$; MaxScoreN=$Score(Q, B_1)$;
for(i=1;i<=n;i++)
{
    CurrentScore= $Score(Q, M_i)$;
    if( CurrentScore>MaxScoreP ) { MaxScoreP=CurrentScore; }
}
for($i$=1;$i$<=$m$;$i$++)
{
    CurrentScore= $Score(Q, B_i)$;
    if( CurrentScore>MaxScoreN ) { MaxScoreN=CurrentScore; }
}
if( MaxScoreP>MaxScoreN ) return $P$;
else return $N$;
```

Position Weight Matrix (PWM) algorithm assumes that positions in a fixed-length motif are independent to each other and calculates how a query sequence is similar to the set of known motif occurrences [10, 19]. Firstly, a position conservation factor M_i is calculated as $M_i = \sum_{b \in \{A,T,C,G\}} (f_i(b)/N - P_0(b))^2/P_0(b)$, $i = 1, 2, \ldots, L$, where $f_i(b)$ is the observed frequencies of nucleotide b at position i in the set of known motif occurrences, and $P_0(b)$ is the background frequency of nucleotide b. Then the position probability matrix (PPM) is calculated as:

$$PPM = \begin{pmatrix} P_1(A) & P_2(A) & \cdots & P_n(A) \\ P_1(T) & P_2(T) & \cdots & P_n(T) \\ P_1(C) & P_2(C) & \cdots & P_n(C) \\ P_1(G) & P_2(G) & \cdots & P_n(G) \end{pmatrix},$$

where $P_j(b) = \{f_j(b) + s(b)\} / \left\{N + \sum_{b \in \{A,T,C,G\}} s(b)\right\}$, and $s(b) = P_0(b)\sqrt{N}$ is a smoothing factor.

Then the position weight matrix (PWM) is calculated as

$$PWM = \begin{pmatrix} w_1(A) & w_2(A) & \cdots & w_n(A) \\ w_1(T) & w_2(T) & \cdots & w_n(T) \\ w_1(C) & w_2(C) & \cdots & w_n(C) \\ w_1(G) & w_2(G) & \cdots & w_n(G) \end{pmatrix},$$

where $w_i(b) = \ln\{P_i(b)/P_0(b)\}$.

The standardized similarity score of a query sequence Q is defined to be

$$S(Q) = \frac{\sum_{i=1}^{L} M_i w_i(Q_i) - \sum_{i=1}^{L} M_i \min\{w_i(b)\}}{\sum_{i=1}^{L} M_i \max\{w_i(b)\} - \sum_{i=1}^{L} M_i \min\{w_i(b)\}},$$

where Q_i is the i^{th} nucleotide in Q, and $b \in \{\text{A, T, C, G}\}$. For a cutoff S_0, only if $S(Q) \geq S_0$, Q is defined as a binding motif of the transcription factor.

15.2.3 Prediction Performance Measurements and Evaluation

Given the positive dataset $P = \{M_1, M_2, \ldots, M_n\}$, and the negative dataset $N = \{B_1, B_2, \ldots, B_m\}$, where $|M_i| = |B_j| = L$. M_i is a true positive or false negative if $SNN(M_i) = P$ or N, respectively, whereas B_j is a true negative or false positive if $SNN(B_j) = N$ or P, respectively. For the classification model $SNN(X)$, the numbers of true positives, false negatives, true negatives and false positives are abbreviated as TP, FN, TN and FP, respectively. The classification performance of the model is measured by sensitivity (Sn), specificity (Sp), accuracy (Ac) and Matthews correlation coefficient (MCC) [20, 21], which are defined as follows. $Sn = TP/(TP + FN)$, $Sp = TN/(TN + FP)$, $Ac = (Sn + Sp)/2$, and $MCC = (TP \times TN - FP \times FN)/sqrt((TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN))$, where $sqrt(t)$ is the squared root of t.

A line plot will be generated for the evolutionarily optimized combination pattern $W = \langle w_1, w_2, \ldots, w_L \rangle$ for the comparison with the WebLogo plot. TFBS screening algorithms usually use the visual technique WebLogo to demonstrate the DNA compositions at each position in the TFBS, and a higher plotted position suggests a larger information content [11]. An initial weight vector $W^0 = \langle w_1^0, w_2^0, \ldots, w_L^0 \rangle$ is generated from a transcription factor's WebLogo plot, by scaling the information content at position i to [0, 1] as w_i^0.

Two validation strategies are adopted to evaluate the classification algorithm SNN's prediction performance. Firstly, the algorithm SNN is investigated for its leave-one-out (LOO) cross validation performance, i.e. iteratively choosing one data entry and investigating its prediction by the classification model trained on the rest data sets. The LOO validation strategy has been widely used to measure how a TFBS or other functional element prediction algorithm performs [22, 23]. To further investigate the dataset dependency of the proposed SNN algorithm, this study conducted 3-fold cross validation (3FCV) strategy [24–26]. The basic idea is to randomly split the positive and negative datasets into 3 equal-size subsets $\{P_1, P_2, P_3\}$ and $\{N_1, N_2, N_3\}$, respectively. The prediction results are iteratively investigated for $\{P_i, N_i\}$ using the SNN trained on $P \backslash P_i$ and $N \backslash N_i$, where $i = 1$, 2, and 3. A self validation (denoted as Self) is also used to evaluate the self consistency, which is to evaluate how a classification model performs on the training dataset.

15.2.4 Evolutionary Optimization Algorithm

This study proposed an evolutionary optimization algorithm to screen for the weight vector with the best overall accuracy *Ac* of the algorithm *SNN*, as shown in Fig. 15.1. The basic idea of an evolutionary optimization algorithm (EOA) is to simulate the natural selection process [27, 28]. Each generation of individuals produce children through the operations of crossing and mutation from a pair of parents. A fitness function is defined to describe how each children fit the natural selection pressure. A better fitness leads to a higher chance to survive into the next generation. The population size is usually fixed to a constant value [11, 29–37].

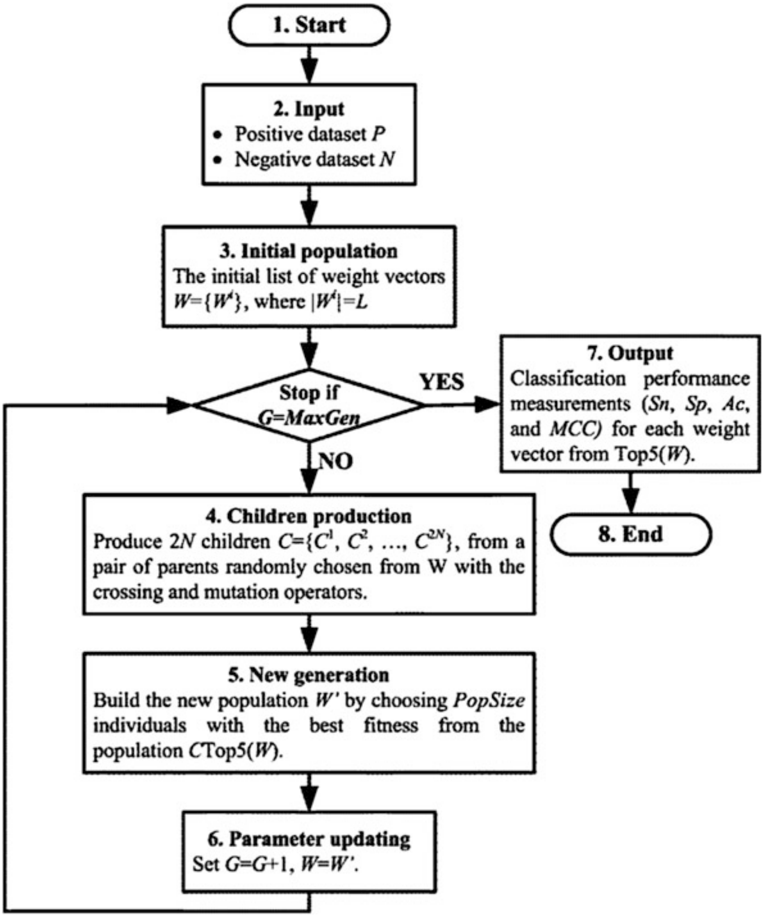
1. Start
2. Input
• Positive dataset P
• Negative dataset N
3. Initial population
The initial list of weight vectors $W=\{W^i\}$, where $|W^i|=L$
Stop if $G=MaxGen$
YES
NO
7. Output
Classification performance measurements (Sn, Sp, Ac, and MCC) for each weight vector from Top5(W).
8. End
4. Children production
Produce $2N$ children $C=\{C^1, C^2, \ldots, C^{2N}\}$, from a pair of parents randomly chosen from W with the crossing and mutation operators.
5. New generation
Build the new population W' by choosing $PopSize$ individuals with the best fitness from the population CTop5(W).
6. Parameter updating
Set $G=G+1$, $W=W'$.

Fig. 15.1 Procedure of the evolutionary optimization algorithm. 5 weight vectors with the best accuracies *Ac* will be output

The initial population W consists of *PopSize* individual weight vectors, *i.e.* W^i, where $i \in \{1, 2, \ldots, PopSize\}$. Each individual W^i is an L-dimension vector $W^i = \langle W_0^i, W_1^i, \ldots, W_L^i \rangle$, where W_j^i is a random value between 0 and 1.

MaxGen generations of natural mutation and selection are conducted to find the fittest weight vectors. For a given weight vector W^i, an SNN classification model is built, and the overall classification accuracy *Ac* with the 4-fold cross validation is defined to be the fitness function $Ac(W^i)$, as used in step 5. For the population of weight vectors *W*, *Top5*(*W*) consists of 5 weight vectors with the best fitness in the population. The final top 5 weight vectors together with the performance measurements of their classification models are output.

15.3 Results and Discussion

15.3.1 Best Parameters for EOA

There are two parameters for the evolutionary algorithm EOA, i.e. the population size *PopSize* and the generation number *MaxGen*. Previous studies suggested that $PopSize = 100$ performs well for the evolutionary optimization problems with individual vector size ~ 10 [38]. So we firstly fix $PopSize = 100$, and investigate how the optimization goal, *Ac*, changes with the increased number of generations, i.e. *MaxGen*. The parameter *MaxGen* is set between 0 and 5,000, and the step size is 100. Q6MAZ and Q3 quickly reach the peak *Ac* value 1.00 after just $MaxGen = 200$ generations of optimizations, as shown in Fig. 15.2a. The TF genes Ebox, Myc and P53 also reach very high *Ac* values (>97 %) at just $MaxGen = 200$. If we choose the Ac value at MaxGen = 5,000 as the final result, all the six investigated TFs reach this peak value at MaxGen = 3,000, as shown in Fig. 15.2a.

We further investigate how the parameter *PopSize* impacts the optimization performance of EOA, as shown in Fig. 15.2 and Supplementary Figure S1. By choosing $PopSize \in \{20, 40, 60, 80, 100, 120, 140, 160, 200\}$, the overall accuracy Ac is calculated for generation $G \in \{0, 100, 200, \ldots, 4{,}900, 5{,}000\}$ of EOA on each of the six TFs. Figure 15.2 shows that the TFBS prediction problem of Q6 is the most difficult to be optimized, and reaches the peak values at generations 3,800, 3,000 and 2,600 for $PopSize = 60$, 100 and 140, respectively. All the other five TFs reach the peak *Ac* values before the optimization generation 3,000. Similar patterns can be observed for other population sizes *PopSize*, as in Supplementary Figure S1.

Considering that the running time of the evolutionary algorithm EOA increases linearly with the product $PopSize \times MaxGen$, and the above data, this study will set $PopSize = 100$ and $MaxGen = 3{,}000$ for the following experiments.

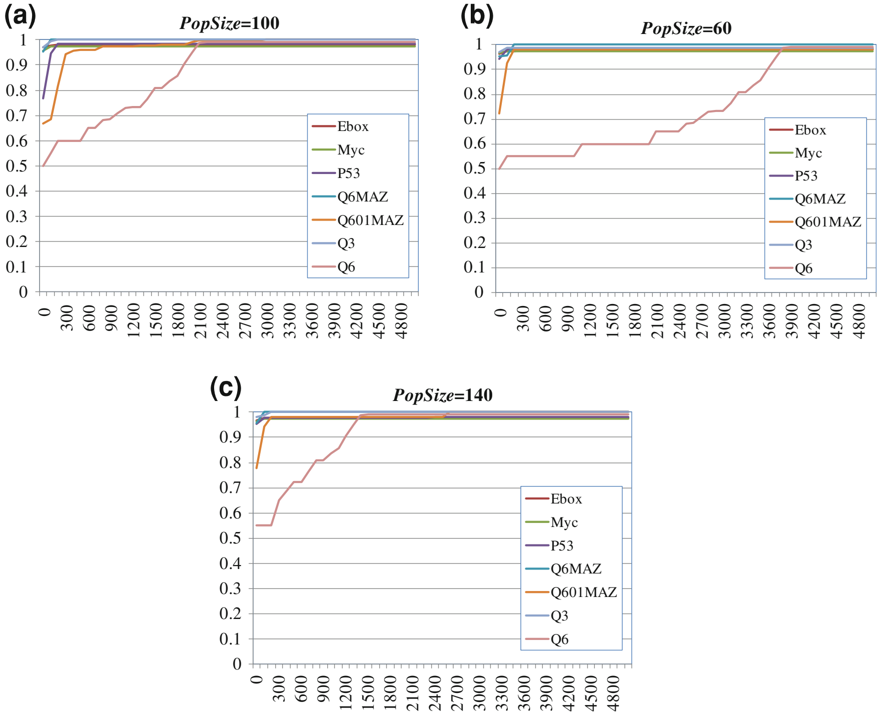

Fig. 15.2 Distributions of overall classification accuracy, *Ac*, for different generation numbers. The population sizes *PopSize* are fixed to **a** 100, **b** 60 and **c** 140, respectively

15.3.2 Comparison of PWM and SNN(W0)

We firstly compare the widely used PWM algorithm with the SNN algorithm. WebLogo is also widely used to demonstrate the information content or conservation at each position of a motif [11]. The higher a position is, the larger information content this position has, as shown in Fig. 15.3. And the binding sites of all the seven TFs do show significant patterns in information content of some motif positions. So we hypothesize that the information content from WebLogo plot may represent well the weight of each motif position for the SNN algorithm, and the weight vector is denoted as W_0.

Both PWM and SNN score the similarity of a query DNA sequence to the known TFBSs, and this study chooses the cutoff score with $Sn \sim Sp$ for the comparison. In general, the SNN(W_0) algorithm performs similarly well or slightly worse compared with the PWM algorithm, as shown in Table 15.1. Both algorithms produce ~90 % or larger overall accuracy *Ac* for the TFBS motif screening problem, and the TF Q3 even receives 100 % accurate separation of the positive and negative data entries from both algorithms under the two validation strategies.

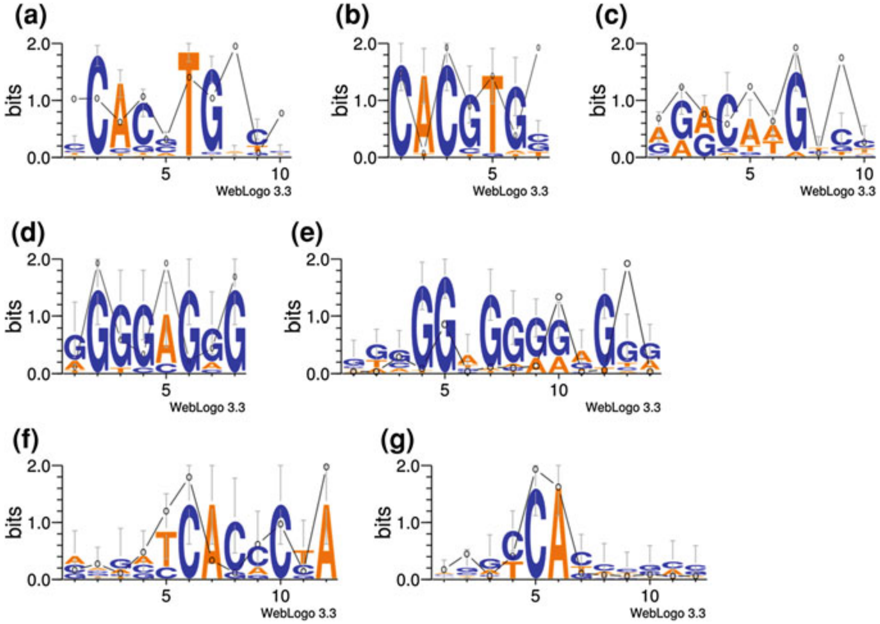

Fig. 15.3 WebLogo plots for the TFs. **a** Ebox, **b** Myc, **c** P53, **d** Q6MAZ, **e** Q601MAZ, **f** Q3 and **g** Q6. *The line plot* is for the evolutionarily optimized weight vector by the SNN + EOA algorithms for each TF

The biggest difference between the two algorithms is for the TFBS motif screening problem of Myc, where SNN(W_0) performs 5.01 and 5.48 % worse in *Ac* than PWM using the LOO and 3FCV validations, respectively. So our first hypothesis about the usage of W_0 is reasonable but may need further optimization.

15.3.3 Comparison of PWM and SNN + EOA

The next hypothesis is that there may exist a weight vector $W = \langle w_1, w_2, \ldots, w_L \rangle$ with increased *Ac* value for the SNN algorithm. Besides the position independent measurements, e.g. PWM or WebLogo, there is no available knowledge about how to optimize the weight vector. So we choose to use the evolutionary optimization algorithm to search for a weight vector with optimal overall accuracy Ac by just random mutations in the weight vectors, as described in Sect. 15.2.4.

After the optimization of *MaxGen* = 3,000 generations of *PopSize* = 100 individuals (weight vectors), the motif screening algorithm SNN outperforms the PWM algorithm in any performance measurements for all the seven TFs, as shown in Table 15.2. The PWM algorithm achieves 100 % accuracy for the LOO validation of Q6MAZ and both LOO and 3FCV validations of Q3, and the

Table 15.1 Prediction performances of the algorithms PWM and SNN(W_0)

		LOO					3FCV				
	Method	Cutoff	*Sn*	*Sp*	*Ac*	*MCC*	Cutoff	*Sn*	*Sp*	*Ac*	*MCC*
Ebox	PWM	0.6500	0.9664	0.9454	0.9559	0.7660	0.6400	0.9580	0.9420	0.9500	0.7312
	SNN(W_0)	1.3830	0.9832	0.8521	0.9176	0.5761	1.4720	0.9664	0.9134	0.9399	0.6789
Myc	PWM	0.5900	0.9524	0.9429	0.9477	0.7450	0.5600	0.9524	0.9333	0.9429	0.7187
	SNN(W_0)	1.0310	1.0000	0.7952	0.8976	0.5108	1.0310	0.9524	0.8281	0.8881	0.5176
P53	PWM	0.7500	0.8478	0.9674	0.9076	0.7590	0.6800	0.9130	0.9196	0.9163	0.6594
	SNN(W_0)	1.3640	1.0000	0.8913	0.9457	0.6235	1.3220	1.0000	0.9000	0.9500	0.6708
Q6MAZ	PWM	0.7600	1.0000	1.0000	1.0000	1.0000	0.7400	1.0000	0.9917	0.9959	0.9568
	SNN(W_0)	1.3900	1.0000	0.9667	0.9833	0.8515	1.3900	1.0000	0.9667	0.9833	0.8515
Q601MAZ	PWM	0.7100	0.9630	0.9963	0.9796	0.9593	0.6600	0.9630	0.9778	0.9704	0.8722
	SNN(W_0)	1.3210	0.9259	0.9778	0.9519	0.8497	1.3000	0.9630	0.9741	0.9685	0.8572
Q3	PWM	0.6600	1.0000	1.0000	1.0000	1.0000	0.6600	1.0000	1.0000	1.0000	1.0000
	SNN(W_0)	1.4000	1.0000	1.0000	1.0000	1.0000	1.4000	1.0000	1.0000	1.0000	1.0000
Q6	PWM	0.8800	1.0000	0.9550	0.9775	0.8050	0.8300	1.0000	0.9550	0.9775	0.8050
	SNN(W_0)	1.4400	1.0000	0.9500	0.9750	0.7889	1.4400	1.0000	0.9550	0.9775	0.8050

The leave-one-out (LOO) validation and 3-fold cross validation (3FCV) strategies are used for the seven TFs

Table 15.2 Prediction performances of the algorithms PWM and SNN + EOA

		3FCV					LOO				
	Method	Cutoff	*Sn*	*Sp*	*Ac*	*MCC*	Cutoff	*Sn*	*Sp*	*Ac*	*MCC*
Ebox	PWM	0.6400	0.9580	0.9420	0.9500	0.7312	0.6500	0.9664	0.9454	0.9559	0.7660
	SNN + EOA	1.5000	0.9748	0.9445	0.9597	0.7639	1.6040	0.9832	0.9773	0.9803	0.8825
Myc	PWM	0.5600	0.9524	0.9333	0.9429	0.7187	0.5900	0.9524	0.9429	0.9477	0.7450
	SNN + EOA	1.6320	0.9524	0.9667	0.9596	0.8224	1.8000	0.9524	0.9952	0.9738	0.9476
P53	PWM	0.6800	0.9130	0.9196	0.9163	0.6594	0.7500	0.8478	0.9674	0.9076	0.7590
	SNN + EOA	1.4870	1.0000	0.9413	0.9707	0.7702	1.6000	1.0000	0.9717	0.9859	0.8704
Q6MAZ	PWM	0.7400	1.0000	0.9917	0.9959	0.9568	0.7600	1.0000	1.0000	1.0000	1.0000
	SNN + EOA	1.7500	1.0000	1.0000	1.0000	1.0000	1.7500	1.0000	1.0000	1.0000	1.0000
Q601MAZ	PWM	0.6600	0.9630	0.9778	0.9704	0.8722	0.7100	0.9630	0.9963	0.9796	0.9593
	SNN + EOA	1.9000	1.0000	1.0000	1.0000	1.0000	1.9000	1.0000	1.0000	1.0000	1.0000
Q3	PWM	0.6600	1.0000	1.0000	1.0000	1.0000	0.6600	1.0000	1.0000	1.0000	1.0000
	SNN + EOA	1.5000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Q6	PWM	0.8300	1.0000	0.9550	0.9775	0.8050	0.8800	1.0000	0.9550	0.9775	0.8050
	SNN + EOA	1.9000	1.0000	0.9850	0.9925	0.9223	1.9010	1.0000	0.9850	0.9925	0.9223

The leave-one-out (LOO) validation and 3-fold cross validation (3FCV) strategies are used for the seven TFs

SNN + EOA algorithm achieves such perfect classification. For the other transcription factors, SNN + EOA outperforms PWM by 0.97–7.83 % in overall accuracy *Ac*. The measurements MCC ∊ [−1, 1] evaluates how the prediction results match the positive and negative datasets, and a larger MCC means a better prediction. Besides the two TFs Q6MAZ and Q3 that both algorithms perform equally well, SNN + EOA improves the MCC of PWM algorithm by 0.0327–0.2026. The PWM algorithm does not perform well on the dataset of the well-known tumor suppressor P53, as in Table 15.2. It only achieves *Sn* = 84.78 % and *Sp* = 96.74 % for the LOO validation of P53, and the overall accuracy is only 90.76 %. SNN + EOA achieves a slightly better specificity (*Sp* = 97.17 %) and a much better sensitivity (*Sn* = 100 %). A similar improvement is also achieved by SNN + EOA for the 3FCV validation of P53.

It's also interesting to observe that the weight vector achieving the best prediction performance does not match the position independent measurement WebLogo, as shown in Fig. 15.3. For the tumor suppressor P53, the optimized weight vector does not agree with WebLogo at positions 4, 5 and 9, as shown in Fig. 15.3c. The information content at position 4 is larger than that at position 5, but their weights in the optimized vector weighs the two positions reversely. And although the information content at position 9 only ranks 8th, position 9 has the second largest weight. Similar discrepancy exists for all the seven investigated TFs, as in Fig. 15.3, and suggests that a concerted weighing of different positions is necessary for motif screening and other similar problems.

Acknowledgments This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13040400), Shenzhen Peacock Plan (KQCX20130628112914301), Shenzhen Research Grant (ZDSY20120617113021359), China 973 program (2010CB732606), the MOE Humanities Social Sciences Fund (No.13YJC790105) and Doctoral Research Fund of HBUT (No. BSQD13050). Computing resources were partly provided by the Dawning supercomputing clusters at SIAT CAS.

References

1. Crick F (1970) Central dogma of molecular biology. Nature 227(5258):561–563
2. Ameur A, Rada-Iglesias A, Komorowski J, Wadelius C (2009) Identification of candidate regulatory SNPs by combination of transcription-factor-binding site prediction, SNP genotyping and haploChIP. Nucleic Acids Res 37(12):e85
3. Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8(3):206–216
4. Galas DJ, Schmitz A (1978) DNAase footprinting a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res 5(9):3157–3170
5. Dent C, Latchman D (1993) The DNA mobility shift assay. In: Transcription factors: a practical approach, pp 1–3
6. Pillai S, Chellappan SP (2009) ChIP on chip assays: genome-wide analysis of transcription factor binding and histone modifications. In: Chromatin protocols. Springer, Berlin, pp 341–366

7. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316(5830):1497–1502
8. Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA (2008) DBD–taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res 36(Database issue):D88–D92
9. Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics 16(1):16–23
10. Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S, Grosse I (2005) Identification of transcription factor binding sites with variable-order Bayesian networks. Bioinformatics 21(11):2657–2666
11. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14(6):1188–1190
12. Quader S, Huang CH (2012) Effect of positional dependence and alignment strategy on modeling transcription factor binding sites. BMC Res Notes 5:340
13. Gorin AA, Zhurkin VB, Wilma K (1995) *B*-DNA twisting correlates with base-pair morphology. J Mol Biol 247(1):34–48
14. Oshchepkov DY, Vityaev EE, Grigorovich DA, Ignatieva EV, Khlebodarova TM (2004) SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. Nucleic Acids Res 32(suppl 2):W208–W212
15. Rose PW, Bi C, Bluhm WF, Christie CH, Dimitropoulos D, Dutta S, Green RK, Goodsell DS, Prlic A, Quesada M et al (2013) The RCSB Protein Data Bank: new resources for research and education. Nucleic Acids Res 41(Database issue):D475–D482
16. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K et al (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34(Database issue):D108–D110
17. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S et al (2013) Ensembl 2013. Nucleic Acids Res 41(Database issue):D48–D55
18. String Alignment using Dynamic Programming.(http://www.biorecipes.com/DynProgBasic/code.html)
19. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res 31(13):3576–3579
20. Xue Y, Zhou F, Zhu M, Ahmed K, Chen G, Yao X (2005) GPS: a comprehensive www server for phosphorylation sites prediction. Nucleic Acids Res 33(Web Server issue):W184–W187
21. Zhou FF, Xue Y, Chen GL, Yao X (2004) GPS: a novel group-based phosphorylation predicting and scoring method. Biochem Biophys Res Commun 325(4):1443–1448
22. Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, Crawford GE, Furey TS (2013) Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Res 23(5):777–788
23. Zhou Q, Liu JS (2004) Modeling within-motif dependence for transcription factor binding site predictions. Bioinformatics 20(6):909–916
24. Cheng C, Ung M, Grant GD, Whitfield ML (2013) Transcription factor binding profiles reveal cyclic expression of human protein-coding genes and non-coding RNAs. PLoS Comput Biol 9(7):e1003132
25. Zhou F, Xu Y (2010) cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. Bioinformatics 26(16):2051–2052

26. Qian J, Lin J, Luscombe NM, Yu H, Gerstein M (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. Bioinformatics 19(15):1917–1926
27. Potts JC, Giddens TD, Yadav SB (1994) The development and evaluation of an improved genetic algorithm based on migration and artificial selection. IEEE Trans Syst Man Cybern 24(1):73–86
28. Tam KY (1992) Genetic algorithms, function optimization, and facility layout design. Eur J Oper Res 63(2):322–346
29. Anastassopoulos G, Adamopoulos A, Galiatsatos D, Drosos G (2013) Feature extraction of osteoporosis risk factors using artificial neural networks and genetic algorithms. Stud Health Technol Inform 190:186–188
30. Santiso EE, Musolino N, Trout BL (2013) Design of linear ligands for selective separation using a genetic algorithm applied to molecular architecture. J Chem Inf Model 53(7):1638–1660
31. Chen JB, Chuang LY, Lin YD, Liou CW, Lin TK, Lee WC, Cheng BC, Chang HW, Yang CH (2013) Genetic algorithm-generated SNP barcodes of the mitochondrial D-loop for chronic dialysis susceptibility. Mitochondrial DNA
32. Sale M, Sherer EA (2013) A genetic algorithm based global search strategy for population pharmacokinetic/pharmacodynamic model selection. Brit J Clin Pharmacol
33. Yoon Y, Kim YH (2013) An efficient genetic algorithm for maximum coverage deployment in wireless sensor networks. IEEE Trans Cybern
34. Azadnia AH, Taheri S, Ghadimi P, Mat Saman MZ, Wong KY (2013) Order batching in warehouses by minimizing total tardiness: a hybrid approach of weighted association rule mining and genetic algorithms. Sci World J 2013:246578
35. Chuang LY, Cheng YH, Yang CH, Yang CH (2013) Associate PCR-RFLP assay design with SNPs based on genetic algorithm in appropriate parameters estimation. IEEE Trans Nanobiosci 12(2):119–127
36. Khotanlou H, Afrasiabi M (2012) Feature selection in order to extract multiple sclerosis lesions automatically in 3D brain magnetic resonance images using combination of support vector machine and genetic algorithm. J Med Signals Sens 2(4):211–218
37. Kou J, Xiong S, Fang Z, Zong X, Chen Z (2013) Multiobjective optimization of evacuation routes in stadium using superposed potential field network based ACO. Comput Intell Neurosci 2013:369016
38. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6(2):182–197