Indian J. Genet., 74(1): 73-80 (2014)

Application of Gibbs sampling methodology for identification of transcription factor binding sites in MADS box family genes in *Arabidopsis thaliana*

Prabina Kumar Meher, Tanmaya Kumar Sahu, A. R. Rao[1]* and S. D. Wahi

Indian Agricultural Statistics Research Institute, New Delhi 110 012

(Received: August 2013; Revised: November 2013; Accepted: December 2013)

Abstract

The Omic revolution has generated voluminous genome sequence data. The discovery of genomic elements like genes, splice sites, regulatory motifs and transcription factor binding sites have become thrust areas of bioinformatics. The transcription factors are the proteins that bind to the transcription factor binding sites on the genome to regulate the gene expression. Thus, the identification of transcription factor binding sites and their genomic co-ordinates has been a prime interest in genomic research to understand the underlying mechanism of gene expression. Various experimental and computational approaches have been used to detect these sites. In this paper, Gibbs sampling has been applied to identify transcription factor binding sites and is discussed in terms of its parameters, model and procedures using the sequence data of *Arabidopsis thaliana*.

Key words: Gene expression, transcription factors, statistical techniques, genomic elements

Introduction

The functional and structural annotation of the gene is not enough to understand the gene expression completely. The presence of regulatory motifs usually in the intergenic region also plays an important role in controlling the gene expression. Transcription Factor Binding Sites (TFBS) are such type of regulatory motifs to which the transcription factors bind and regulate the gene expression either by upregulating or by downregulating the transcription process. Transcription factors act as activators, repressors, enhancers or silencers in the cell. Accordingly, there are specific TFBS for the transcription factors. Thus, presence of a specific type of transcription factor binding sites on genome indicates the kind of transcription factor will bind to it, thereby, indicating its role in controlling gene expression. As the combinatorial presence and absence of TFBSs is responsible for the complexity of gene regulation in every living organism [1]. The interest in the identification of TFBSs has grown dramatically with the arrival of microarray gene expression and transcriptome data. Several experimental techniques such as chromatin immune-precipitation have been developed for the identification of TFBSs, but these experiments are expensive and time consuming. Now a days, the availability of computer hardware installed with less expensive and free bioinformatics software has enabled scientists to run MatchTM [2] or HMMgene and HMMPro [3] to predict the location of TFBS.

Most TFBS classification algorithms compute numerical score reflecting the degree to which a given sequence site matches a given motif. Moreover, the underlying scoring model is either a zero order Markov model or simply a position weight matrix (PWM) model. The PWM model, to some extent, has been successfully applied to the problem of TFBS classification. Although, other scoring models are able to improve the accuracy of the PWM model, they are not as prevalent as the simple PWM model found in the classification of TFBSs [4]. Hence, most of the TFBS classification algorithms that are developed today still rely on the PWM model that provides a relatively good performance. However, PWM suffers from the drawback of assumption on independence

*Corresponding author's e-mail: rao.cshl.work@gmail.com
[1]Present address: Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, New Delhi 110 012
Published by the Indian Society of Genetics & Plant Breeding, F2, First Floor, NASC Complex, PB#11312, IARI, New Delhi 110 012
Online management by indianjournals.com

between positions in a motif. To overcome such drawback, many statistical techniques based efficient algorithms have been developed. Gibbs sampling is one such statistical technique which can be used for genome sequence analysis, in particular, for finding TFBS that are conserved in nature. Statistically, Gibbs sampling is a generalized probabilistic algorithm used to generate sequence of samples from a joint probability distribution of two or more random variables [5]. The use of Gibbs sampling as a statistical technique was first described in restoration of images [6]. However, in the arena of bioinformatics, Gibbs sampling was initially used in multiple sequence alignment [7]. Besides, application of Gibbs sampling strategy has not been fully explored to identify TFBS in plant species.

The model plant *Arabidopsis* has been used in many fields of research including genetics, evolution, population genetics and plant development to understand the biological systems in plants. MADS box transcription factors in the plants have an important role in controlling the development of male and female gametophyte; embryo and seed; root, flower and fruit [8-9]. Some MADS-box genes of flowering plants have homeotic functions like HOX genes of animals [10]. The floral homeotic MADS-box genes participate in the determination of floral organ identity [11]. In *Arabidopsis thaliana* the MADS box genes SOC1 [12] and Flowering Locus C (FLC) [13] have an important role in the integration of molecular flowering time pathways. Hence, understanding the transcription process of these genes and particularly, identification of TFBS of MADS box genes is important. The main aim of this paper is to identify TFBS of MADS box genes in *Arabidopsis thaliana* by applying Gibbs sampling methodology as well as to give insights into the identification of TFBS in the agriculturally important crops.

Materials and methods

TFBS sequences were collected from the JASPAR database [14] for AGL3 gene of *Arabidopsis thaliana*. The collected TFBS sequences including flanking sequences in fasta format are reported in **Supplementary dataset 1**. The AGL3 (P29383) genes belong to the α-helix class and MADS family.

Gibbs sampling procedure

Gibbs sampling is a Monte Carlo Markov Chain (MCMC) procedure for estimation of the joint distribution as well as marginal distribution when the full conditional distribution of all the concerned random variables are available. In Gibbs sampling procedure, samples are drawn in iterative process from the full conditional distribution and samples collected in this way are guaranteed to converge to the true joint distribution.

Consider a set of random variables $x_1, x_2, ..., x_V$ whose marginal distribution is unknown but the full conditional distribution $p(x_i \mid x_j; j \neq i)$ i = 1, 2, ..., V, is known. Then, the Gibbs sampler draws samples of the random variables in the following manner.

$$
\begin{aligned}
&draw\ x_1^{(t+1)}\ \ from\ p(x_1 \mid x_2 = x_2^{(t)}, ..., x_V = x_V^{(t)})\\
&draw\ x_2^{(t+1)}\ \ from\ p(x_2 \mid x_1 = x_1^{(t+1)},\\
&\qquad\qquad x_3 = x_3^{(t)}, ..., x_V = x_V^{(t)})\\
&\vdots\\
&draw\ x_i^{(t+1)}\ \ from\ p(x_i \mid x_1 = x_1^{(t+1)}, ..., x_{i-1}\\
&\qquad\qquad = x_{i-1}^{(t+1)}, x_{i+1} = x_{i+1}^{(t)}, ..., x_V = x_V^{(t)})\\
&\vdots\\
&draw\ x_V^{(t+1)}\ \ from\ p(x_V \mid x_1 = x_1^{(t+1)}, ..., x_{V-1} = x_{V-1}^{(t+1)})
\end{aligned}
$$

where t denotes the number of iterations, when $t \rightarrow \infty$, the distribution of $(x_1^{(t)}, x_2^{(t)}, ..., x_V^{(t)})$ converges to (x_1, x_2, ..., x_v) which is the joint distribution and equivalently the distribution of $x_i^{(t)}$ converges to $p(x_i)$ i = , 1, ..., V, which is the marginal distribution.

Sequence data and multinomial model

Multinomial model has been successfully used to capture the variability on the DNA sequences. Let a given set of sequences be R_{1s}, R_{2s},..., R_{Ks}, which can be written in matrix form as

$$
\mathbf{R}^s = \begin{bmatrix}
\text{sequence-R}_{1s} & r_{11} & r_{12} & \ldots & r_{1L_1} \\
\text{sequence-R}_{2s} & r_{21} & r_{22} & \ldots & r_{2L_2} \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
\text{sequence-R}_{Ks} & r_{K1} & r_{K2} & \ldots & r_{KL_K}
\end{bmatrix}
$$

where r_{kl} (k = 1, 2 ,..., K and l = 1, 2 ,..., L_K) is the residue (nucleotide bases) that takes values from an

alphabet with P (=4, for DNA *i.e*., A, T, G, C) different letters. L_k denotes the length of the k^{th} sequence. Consider a segment of same fixed length J within each sequences then these elements (segments) can be assumed to be independent observations from a product multinomial model that describes the residue frequency for each positions j (j = 1, 2 ,…, J) within an element or motif and consists of parameters $\boldsymbol{\theta}_j = (\theta_{1j}, \theta_{2j}, ..., \theta_{Pj})'$, j = 1, 2, …, J and $\boldsymbol{\theta}_0 = (\theta_{10}, \theta_{20}, ..., \theta_{P0})'$ describe the frequency of nucleotides of remaining positions (background region) which is assumed to be an independent observation from multinomial model.

Multinomial distribution with Dirichlet prior

Let $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, ..., \boldsymbol{\theta}_J)$ be a P × J matrix where $\boldsymbol{\theta}_j$ is a probability vector of length P, that is $\boldsymbol{\theta}_j = (\theta_{1j}, \theta_{2j}, ..., \theta_{Pj})'$ where $\theta_{ij} \geq 0$ and $|\theta_j| = \sum_{i=1}^{P} \theta_{ij} = 1.$ Then an integer matrix $\boldsymbol{M}_{PxJ} = (\boldsymbol{m}_2, \boldsymbol{m}_2, ..., \boldsymbol{m}_j)$ where $\boldsymbol{m}_j = (\boldsymbol{m}_{1j}, \boldsymbol{m}_{2j}, ..., m_{Pj})'$ is said to follow a product multinomial distribution with parameter $\boldsymbol{\Theta}$ that is **M**~Product Multinomial(PM) ($\boldsymbol{\Theta}$ $|\boldsymbol{m}_1|, |\boldsymbol{m}_2|, ..., |\boldsymbol{m}_j|$) if each $\boldsymbol{m}_j$ follows multinomial distribution $(\boldsymbol{\theta}_j; |\mathbf{m}_j| = \sum_{i=A,T,G,C} m_{ij})$ and $\boldsymbol{m}_j$ are mutually independent (represents the number of times i^{th} type letter occurs in j^{th} position where i = 1, 2, …, P and j = 1, 2, … , J). Dirichlet distribution is used as conjugate prior of multinomial distribution [15], in a similar way Product Dirichlet (PD) distribution can be used as a conjugate prior of $\boldsymbol{\Theta}$, that is $\boldsymbol{\Theta}$ ~PD(**B**) where $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_J)$ is a P x J matrix and $\boldsymbol{\beta}_j = (\beta_{1j}, \beta_{2j}, ..., \beta_{pj})'$, if $\boldsymbol{\theta}_j$ are independent P-dimensional Dirichlet random variables with distribution Dir ($\boldsymbol{\beta}_j$) ; j = 1, 2, …, J.

Likelihood of parameters of motif (segments) region

Let $r_{k1}, r_{k2}, ..., r_{kj}$ be the specified segment for the k^{th} sequence, where r_{kj}s are the mutually independent alphabets. By considering segments of all the sequences, it can be arranged as follows:

$$\mathbf{R} = \begin{array}{c} \\ R_1 \\ R_2 \\ ... \\ R_K \end{array} \begin{array}{c} \begin{array}{cccc} R_{.1} & R_{.2} & ... & R_{.J} \end{array} \\ \begin{bmatrix} r_{11} & r_{12} & ... & r_{1J} \\ r_{21} & r_{22} & ... & r_{2J} \\ ... & ... & ... & ... \\ r_{K1} & r_{K2} & ... & r_{KJ} \end{bmatrix} \end{array}$$

Let $R_{.j} = (r_{1j}, r_{2j}, ..., r_{kj})$ and $\boldsymbol{h}(R_{.j}) = (m_1, m_2, ..., m_p)'$, where m_i is the number of times i^{th} type letter observed in R_j. In other words, $\boldsymbol{h}(R_{.j})$ is a vector of order 4 × 1, where the elements of the vector are the number of times of the occurrence of the nucleotides A, T, G, C. Then the likelihood of $\boldsymbol{\Theta}$ can be written as

$$f(R_1, R_2, ..., R_K | \boldsymbol{\Theta}) \alpha \prod_{j=1}^{J} (\boldsymbol{\theta}_j)^{\mathbf{h}(R_j)} \quad (1)$$

where

$$\{\mathbf{h}(R_{.1}), \mathbf{h}(R_{.2}), ..., \mathbf{h}(R_{.J})\}$$
$$\sim PM(\boldsymbol{\Theta}; K, K, ..., K).$$

Posterior distribution of $\boldsymbol{\Theta}$

According to Bayes theorem, the posterior distribution of $\boldsymbol{\Theta}$ can be written as

$$f(\boldsymbol{\Theta} | R_1, R_2, ..., R_K) = \frac{f(R_1, R_2, ..., R_K | \boldsymbol{\Theta}).g(\boldsymbol{\Theta})}{\sum f(R_1, R_2, ..., R_K | \boldsymbol{\Theta}).g(\boldsymbol{\Theta})} \quad (2)$$

where $g(\boldsymbol{\Theta})$ is prior distribution of $\boldsymbol{\Theta}$ and $f(R_1, R_2, ..., R_K | \boldsymbol{\Theta})$ is the likelihood function of $\boldsymbol{\Theta}$.

From equation (1), the expression (2) can be written as

$$f(\boldsymbol{\Theta}|\mathrm{R}_1,\mathrm{R}_2,...,\mathrm{R}_K)\alpha\prod_{j=1}^{J}(\boldsymbol{\theta}_j)^{\mathbf{h}(\mathrm{R}_j)}.\prod_{j=1}^{J}(\boldsymbol{\theta}_j)^{\boldsymbol{\beta}_j} \alpha\prod_{j=1}^{J}(\boldsymbol{\theta}_j)^{\mathbf{h}(\mathrm{R}_j)+\boldsymbol{\beta}_j} \quad (3)$$

That is if the prior of $\boldsymbol{\Theta}$ is PD(**B**) then the posterior distribution of $\boldsymbol{\Theta}$ is PD(**B+H**) where

$$\mathbf{H}=\{\mathbf{h}(\mathrm{R}_{.1}),\mathbf{h}(\mathrm{R}_{.2}),...,\mathbf{h}(\mathrm{R}_{.J})\}$$

and

$$\mathbf{B}=(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,...,\boldsymbol{\beta}_J).$$

Likelihood of parameters of motif and background region

For any random alignment of segments of all the sequences denoted as *A*, the following notations are defined:

a_k: starting position of the k^{th} sequence, k = 1, 2, …, K and $a_k \in 1(1)L_k - J + 1$. A = {(1,a_1), (2, a_2), …, (K,a_K)} : denote the collection of starting positions of K sequences. {A} = {(k, a_k + j – 1); k = 1, 2, ..., K and j = 1, 2, ..., J} : denote the collection of starting positions of motif region in K sequences.

$\boldsymbol{R}_{\{A\}}$ = {r_{k,a_k+j-1} for k = 1, 2, ..., K and j = 1, 2, ..., J} : denote the collection of residues in {A}. {a_k} = {(k, a_k + j – 1); j = 1, 2, ..., J} : denote the collection of J consecutive positions in the k^{th} sequence.

$A_{(j)}$ = {k, a_k + j – 1); k = 1, 2, ..., K} : denote the set of j^{th} position of all elements.

Let all the residues outside the motif (aligning) region are independently drawn from a common multinomial model with parameter $\boldsymbol{\theta}_0=(\theta_{10},\theta_{20},...,\theta_{p0})'$ and the residues frequencies for the motif are independently drawn from product multinomial with parameter $\boldsymbol{\Theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,...,\boldsymbol{\theta}_J)$. Then ($J$ + 1) × P dimensional vector parameters completely describe the data. By treating the alignment A as missing the complete data likelihood of the parameters can be written as

$$f(\mathbf{R},\mathrm{A}|\boldsymbol{\theta}_0,\boldsymbol{\Theta})\alpha\,(\boldsymbol{\theta}_0)^{\mathbf{h}(\mathbf{R}_{\{A\}^C})}\prod_{j=1}^{J}(\boldsymbol{\theta}_j)^{\mathbf{h}(\mathbf{R}_{A(j)})} \quad (4)$$

Predictive distribution of aligned segments

Taking the prior distribution of $\boldsymbol{\theta}_0$ as Dirichlet distribution Dir($\boldsymbol{\alpha}$) where $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,...,\alpha_p)$ and prior distribution of $\boldsymbol{\Theta}$ as product Dirichlet distribution PD(**B**) where $\mathbf{B}=(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,...,\boldsymbol{\beta}_J)$ and $\boldsymbol{\beta}_j=(\beta_{1j},\beta_{2j},...,\beta_{pj})'$, the predictive distribution of A can be obtained using Bayes theorem as follows:

$$\begin{aligned} f(\mathrm{A}|\mathbf{R}) &\alpha\, f(\mathbf{R},\mathrm{A}) \\ &\alpha\iint f(\mathbf{R},\mathrm{A}|\boldsymbol{\theta}_0,\boldsymbol{\Theta})g(\boldsymbol{\theta}_0)g(\boldsymbol{\Theta})\,d\boldsymbol{\theta}_0\,d\boldsymbol{\Theta} \\ &\alpha\iint(\boldsymbol{\theta}_0)^{\mathbf{h}(\mathbf{R}_{\{A\}^C})}\prod_{j=1}^{J}(\boldsymbol{\theta}_j)^{\mathbf{h}(\mathbf{R}_{A(j)})} \\ &\qquad g(\boldsymbol{\theta}_0)\,g(\boldsymbol{\Theta})\,d\boldsymbol{\theta}_0\,d\boldsymbol{\Theta} \\ &\alpha\,\Gamma\{\mathbf{h}(\mathbf{R}_{\{A\}^C})+\boldsymbol{\alpha}\}.\prod_{j=1}^{J}\Gamma\{\mathbf{h}(\mathbf{R}_{A(j)})+\boldsymbol{\beta}_j\} \end{aligned} \quad (5)$$

Posterior distribution of individual segment

Let $A_{[-k]}$ denote the set of starting positions of elements in all sequences excluding k^{th} sequence. Then $\mathbf{h}(\mathbf{R}_{\{A\}^c})=\mathbf{h}(\mathbf{R}_{\{A_{[-k]}\}^c})-\mathbf{h}(\mathbf{R}_{\{a_k\}})$ and $\mathbf{h}(\mathbf{R}_{A(j)})=\mathbf{h}(\mathbf{R}_{A_{[-k]}(j)})+\mathbf{h}(r_{k,a_{k+j-1}})$ is a vector of (p – 1). Here, $\mathbf{h}(r_{k,a_{k+j-1}})$ is a vector of (p – 1) zeros and a value of one corresponding to residual at position a_{k+j-1} in the k^{th} sequence. Now the predictive distribution of the starting position of the element in the k^{th} sequence is $f(a_k|A_{[-k]},\mathbf{R})$. Now,

$$
\begin{aligned}
f(\mathrm{A}\,|\mathbf{R}) &= f(\mathrm{A}_{[-\mathrm{k}]}, \mathrm{a}_{\mathrm{k}}\,|\mathbf{R}) \\
&= f(\mathrm{A}_{[-\mathrm{k}]}\,|\mathbf{R}).\, f(\mathrm{a}_{\mathrm{k}}\,|\mathrm{A}_{[-\mathrm{k}]}, \mathbf{R}).\, f(\mathbf{R}) \\
\Rightarrow f(\mathrm{a}_{\mathrm{k}}\,|\mathrm{A}_{[-\mathrm{k}]}, \mathbf{R}) &\;\alpha\; \frac{f(\mathrm{A}\,|\mathbf{R})}{f(\mathrm{A}_{[-\mathrm{k}]}\,|\mathbf{R})} \\
&\alpha \prod_{\mathrm{j}=1}^{\mathrm{J}} \left(\frac{\hat{\boldsymbol{\theta}}_{\mathrm{j}[\mathrm{k}]}}{\hat{\boldsymbol{\theta}}_{0[\mathrm{k}]}} \right)^{\mathbf{h}(\mathrm{r}_{\mathrm{k},\, \mathrm{a}_{\mathrm{k}+\mathrm{j}-1}})}
\end{aligned}
\tag{6}
$$

where $\hat{\boldsymbol{\theta}}_{j[k]}$ are the posterior mean of $\boldsymbol{\theta}_j$ conditioned on $\mathbf{R}$ and current alignment $A_{[-K]}$, and $\hat{\boldsymbol{\theta}}_{0[k]}$ is the posterior mean of $\boldsymbol{\theta}_0$ based on the current non-motif position $\mathbf{R}_{\{A_{[-k]}\}^c}$.

Sampling step

$f(\mathrm{a}_{\mathrm{k}} = |\mathrm{A}_{[-\mathrm{k}]}, \mathbf{R})$ gives the probability of occurrence of motif position, conditioned on the observation **R** and current alignment $A_{[-K]}$, whose starting point is a_k in the k^{th} sequence where a_k takes values from 1 to $L_k - J + 1$ and is given by

$$
\mathrm{P}_{\mathrm{x}} = \mathrm{P}(\mathrm{a}_{\mathrm{k}} = \mathrm{x}) = \prod_{\mathrm{j}=1}^{\mathrm{J}} \left(\frac{\hat{\boldsymbol{\theta}}_{\mathrm{j}[\mathrm{k}]}}{\hat{\boldsymbol{\theta}}_{0[k]}} \right)^{\mathbf{h}(\mathrm{r}_{\mathrm{k},\, \mathrm{x}+\mathrm{j}-1})}; \qquad x = 1(1)\,\mathrm{L}_{\mathrm{k}} - \mathrm{J}+1 \tag{7}
$$

where

$$
\hat{\boldsymbol{\theta}}_{\mathrm{j}[\mathrm{k}]} = (\hat{\theta}_{1\mathrm{j}[\mathrm{k}]}, \hat{\theta}_{2\mathrm{j}[\mathrm{k}]}, ..., \hat{\theta}_{\mathrm{Pj}[\mathrm{k}]})'
$$

and

$$
\hat{\boldsymbol{\theta}}_{0[\mathrm{k}]} = (\hat{\theta}_{10[\mathrm{k}]}, \hat{\theta}_{20[\mathrm{k}]}, ..., \hat{\theta}_{\mathrm{P}0[\mathrm{k}]})'
$$

$$
\hat{\theta}_{\mathrm{Pj}[\mathrm{k}]} = \mathrm{q}_{\mathrm{Pj}[\mathrm{k}]} = \frac{\mathrm{c}_{\mathrm{pj}[\mathrm{k}]} + \mathrm{b}_{\mathrm{p}}}{\mathrm{K}\text{-}1+\mathrm{B}};\; p = 1, 2, ..., P;
$$

$$
\hat{\theta}_{\mathrm{p}0[\mathrm{k}]} = \mathrm{q}_{\mathrm{p}0[\mathrm{k}]} = \frac{\mathrm{c}_{\mathrm{p}0[\mathrm{k}]} + \mathrm{b}_{\mathrm{p}}}{\sum_{\mathrm{p}=1}^{\mathrm{P}} \mathrm{c}_{\mathrm{p}0} + \mathrm{B}}; \quad \mathrm{B} = \sum_{\mathrm{p}} \mathrm{b}_{\mathrm{p}} \tag{8}
$$

$c_{pj[k]}$ is the observed count of p^{th} residue in position j for the element having starting point a_k in the k^{th} sequence, $c_{p0[k]}$ is the observed count of p^{th} residue in background, b_p can be considered as the pseudo counts of p^{th} residue. After calculating probability of occurrence of all possible segments of width *J* in the selected (k^{th}) sequence, a new position is chosen by randomly sampling over the set of standardized probability weight.

Convergence criteria

The predictive distribution and sampling will be repeated for each of the sequences (k = 1, 2 ,…, K). Once each of the sequence has been sampled, a new alignment position is found. Now for this alignment, Maximum A Posterior probability (MAP) is calculated. The MAP value is measured relative to an empty alignment by taking the difference between the log of the probability of the alignment and the log of the probability of an empty alignment. The MAP can be computed as

$$
\mathrm{F} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \sum_{\mathrm{p}=1}^{\mathrm{P}} \mathrm{c}_{\mathrm{pj}} \log \left(\frac{\hat{\theta}_{\mathrm{pj}}}{\hat{\theta}_{\mathrm{p}0}} \right) = \sum_{\mathrm{j}=1}^{\mathrm{J}} \sum_{\mathrm{p}=1}^{\mathrm{P}} \mathrm{c}_{\mathrm{pj}} \log \left(\frac{\mathrm{q}_{\mathrm{pj}}}{\mathrm{q}_{\mathrm{p}0}} \right) \tag{9}
$$

A value greater than zero indicates, the alignment is more likely to occur. The process is continued till the estimated MAP is converged.

Implementation of the Gibbs sampling

The three steps of Gibbs sampling *viz*., *Initialization*, *Predictive update* and *Sampling* were followed for the identification of TFBS in the MAD box genes of *Arabidopsis thaliana*. In the *initialization step*, a random alignment of sequence of length 10bp was made (shown as un-bold font in the Supplementary dataset 1).

In the *Predictive update* step, one sequence was selected and the motif un-bold font, obtained from the random alignment, observed within the sequence was placed in the background and the residue counts were updated. In order to alleviate the problem of zero count, pseudo counts were introduced in the observed counts.

Then in the *sampling* step, a new motif position for the selected sequence was determined randomly after looking at all possible starting positions of the motif within the selected sequence, using the updated residue counts obtained from the previous step.

The *predictive update* and *sampling* steps were repeated for each of the sequence. Then a final alignment was obtained and the alignment probability was tested by using the MAP criteria. The whole process was repeated till the MAP was converged. The convergence in the MAP was obtained by using the Gibbs Motif Sampler (http://ccmbweb.ccv.brown.edu/gibbs/gibbs.html).

Results and discussion

Initially, from the sequence dataset provided in the Supplementary dataset 1, the number of A's, T's, G's and C's are found as 860, 781, 441 and 501 respectively. The observed counts obtained after the random alignment (10bp length, shown in un-bold font) of the sequences are shown in Table 1. Then the first sequence of **Supplementary dataset 1** *i.e.*, ACAACCATATATAGTAGCCACTGAAT was selected (any sequence can also be selected) and the motif within the sequence *i.e.*, CATATATAGT was placed in the background. The updated residue counts are shown in Table 1. Here, 10% of the total number of each nucleotide was set as pseudo counts. So b_1 = 86, b_2 = 50.1, b_3 = 44.1 and b_4 = 78.1, which results in B = 258.3. The counts and frequencies of nucleotides updated with the pseudo counts are presented in Table 1. The residues frequency matrix is further updated based on equation (8) and given in Table 1.

In the *sampling* step, the length of the selected sequence shown in *predictive update* step is 26 and the width of the motif is 10, so there are 17 possible starting sites (26-10+1=17). For each of these 17 possible segments (x), a weight A_x is estimated according to ratio $A_x = \frac{Q_x}{I_x}$, where $Q_x = \prod_{j=1}^{J} q_{j,r_j}$ and $I_x = \prod_{j=1}^{J} q_{0,r_j}$; q_{j,r_j} is the motif residue frequency, q_{j,r_j} is background residue frequency respectively, r_j refers to the residue located at position *j* of the segment *x*, *J* is the width of the segment. The computed weights are given in the Table 2. A new motif position was chosen randomly over the set of weight A_x and it was found as TATATAGTAG.

The *predictive update* and *sampling* steps were repeated for each of the sequences till a convergence in MAP was reached. It was observed that on an average around 500 iterations were required to reach

Table 1. Observed Residue Counts (ORC) in random alignment, Updated Residues Counts (URC), Updated Residue Counts based on Pseudo Counts (URC-PC) and Updated Residue Frequency (URF) for the nucleotides A, C, G & T, both in background and aligned regions, where 0^{th} position indicates background region and rest 1-10 positions indicate the aligned region.

Nucleotide positions	Counts	A	C	G	T
0	ORC	517	342	288	466
	URC	521	343	289	470
	URC-PC	607	393.1	333.1	548.1
	URF	0.3226	0.209	0.1771	0.2913
1	ORC	36	22	16	23
	URC	36	21	16	23
	URC-PC	122	71.1	60.1	101.1
	URF	0.3443	0.2007	0.1696	0.2854
2	ORC	39	18	14	26
	URC	38	18	14	26
	URC-PC	124	68.1	58.1	104.1
	URF	0.35	0.1922	0.164	0.2938
3	ORC	34	12	19	32
	URC	34	12	19	31
	URC-PC	120	62.1	63.1	109.1
	URF	0.3387	0.1753	0.1781	0.3079
4	ORC	34	15	10	38
	URC	33	15	10	38
	URC-PC	119	65.1	54.1	116.1
	URF	0.3359	0.1837	0.1527	0.3277
5	ORC	31	15	13	38
	URC	31	15	13	37
	URC-PC	117	65.1	57.1	115.1
	URF	0.3302	0.1837	0.1612	0.3249
6	ORC	30	11	16	40
	URC	29	11	16	40
	URC-PC	115	61.1	60.1	118.1
	URF	0.3246	0.1725	0.1696	0.3333
7	ORC	36	15	12	34
	URC	36	15	12	33
	URC-PC	122	65.1	56.1	111.1
	URF	0.3443	0.1837	0.1583	0.3136
8	ORC	33	18	17	29
	URC	32	18	17	29
	URC-PC	118	68.1	61.1	107.1
	URF	0.3331	0.1922	0.1725	0.3023
9	ORC	36	17	18	26
	URC	36	17	17	26
	URC-PC	122	67.1	61.1	104.1
	URF	0.3443	0.1894	0.1725	0.2938
10	ORC	34	16	18	29
	URC	34	16	18	28
	URC-PC	120	66.1	62.1	106.1
	URF	0.3387	0.1866	0.1753	0.2995

Table 2. Estimated weights (A_x) of all consecutive segments of length 10 for the selected sequence ACAACCATATATAGTAGCCACTGAAT (TFBS_AT.1, Supplementary dataset 1)

S.No.	x	Q_x	I_x	A_x	Normalized A_x
1	ACAACCATAT	2.56049E-07	2.70671E-06	0.094597844	0.619655084
2	CAACCATATA	2.70904E-06	2.70671E-06	1.000860324	0.286607396
3	AACCATATAT	4.58371E-07	3.77256E-06	0.121501517	0.592751411
4	ACCATATATA	4.27168E-07	3.77256E-06	0.113230229	0.601022699
5	CCATATATAG	3.15086E-06	2.07105E-06	1.521383034	0.807130106
6	CATATATAGT	4.13495E-06	2.88658E-06	1.43247058	0.718217652
7	ATATATAGTA	7.30274E-07	4.45556E-06	0.163901765	0.550351163
8	TATATAGTAG	3.34971E-06	2.446E-06	1.369465651	0.655212723
9	ATATAGTAGC	2.11415E-07	5.11214E-07	0.413553781	0.300699147
10	TATAGTAGCC	1.16509E-06	1.13696E-06	1.024744404	0.310491476
11	ATAGTAGCCA	1.07689E-07	1.25912E-06	0.08552688	0.628726048
12	TAGTAGCCAC	1.38805E-06	8.15737E-07	1.70158763	0.987334702
13	AGTAGCCACT	1.05644E-07	8.15737E-07	0.129506963	0.584745965
14	GTAGCCACTG	2.96267E-07	4.47821E-07	0.661575979	0.052676949
15	TAGCCACTGA	1.18504E-06	8.15737E-07	1.452720909	0.738467981
16	AGCCACTGAA	6.53364E-08	9.03387E-07	0.072323855	0.641929073
17	GCCACTGAAT	6.39006E-07	8.15737E-07	0.783348433	0.069095505

Table 3. Position wise estimated probabilities from frequencies of nucleotide residues in the final alignment

Position	0	1	2	3	4	5	6	7	8	9	10
Nucleotide											
A	0.3	0.086	0.039	0.745	0.513	0.875	0.699	0.726	0.123	0.513	0.039
T	0.315	0.027	0.324	0.148	0.408	0.083	0.241	0.129	0.816	0.064	0.064
C	0.194	0.872	0.612	0.064	0.017	0.027	0.045	0.045	0.045	0.017	0.036
G	0.191	0.015	0.025	0.043	0.062	0.015	0.015	0.099	0.015	0.405	0.86

convergence in MAP. The frequencies of residues and the final alignment obtained after reaching convergence is given in Table 3 and **Supplementary dataset 2**, respectively. In Supplementary dataset 2, the aligned region in capital letters indicates the presence of TFBS in each sequence. The 3rd and 5th columns are the flanking regions of TFBS whereas 2nd and 6th columns denote the starting and ending locations within each sequence.

It is observed from the results that, though the sequences provided to the Gibbs sampler are of unequal lengths, the sampler has identified the motif from each sequence and aligned them accurately, by leaving the unaligned portion of the sequences as flanking segments (**Supplementary dataset 2**). In 96 out of 97 sequences considered under study, TFBS elements were predicted with the probability 1 in either forward or reverse strand. However, for the sequence number 8 the TFBS is detected in both forward and reverse strand. Therefore the probability of TFBS on sequence no 8 is shared by both the strands as 0.24(+) and 0.76(–). This implies that the probability of the presence of TFBS on the reverse stand of sequence 8 is higher than the forward stand. The demonstration of Gibbs sampling presented in this paper using MADS [illegible] *Arabidopsis thaliana* may serve as a guide to identify TFBS of different gene families of Rice, Wheat, Maize etc. Though Gibbs sampler works efficiently in pattern recognition problems it works with certain limitations in TFBS identification, i.e., it requires a minimum of 30 sequences and each sequence with at least one transcription factor to provide good prediction accuracy.

From the results, it can be concluded that the accurate prediction of TFBS can help understand the regulatory mechanism of genes. Thus, to address the biological problems such as biotic and abiotic stresses, identification of the TFBS present in 1K upstream region of different families of genes in agriculturally important species will be highly beneficial.

Acknowledgements

This study is a part of Ph. D. thesis of P. K. Meher, PG School, IARI. Authors acknowledge the INSPIRE fellowship of Department of Science and Technology, New Delhi and IARI Fellowship. The authors also acknowledge the computational facilities of SCGL, developed under NAIP grant NAIP/Comp-4/C4/C-30033/2008-09.

References

1. **Wingender E.** 2001. The TRANSFAC system on gene expression regulation. Nucleic Acids Res., **29**: 281-283.
2. **Kel A. E., Gobling E., Reuter I., Cheremushkin E., Kel-Margoulis O. V. and Wingender E.** 2003. MATCH: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res., **31**: 3576-3579.
3. **Baldi P. and Brunak S.** 2001. Bioinformatics: The Machine Learning Approach. 2nd edition. The MIT Press, Cambridge, MA.
4. **Fickett J. W. and Hatzigeorgiou A. G.** 1997. Eukaryotic Promoter Recognition. Genome Research, **7**: 861-878.
5. **Casella G. and George E. I.** 1992. Explaining the Gibbs Sampler. J. Am. Statist. Assoc., **46:** 167-174.
6. **Geman S. and Geman D.** 1984. Stochastic relaxation, Gibbs distribution, and the Bayes restoration of images. IEEE Trans. Pattern Anal. Mach. Intell., **6**: 721-741.
7. **Lawrence C. E., Altschul S. F., Boguski M. S., Liu J. S., Neuwald A. F. and Wootton J. C.** 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignments. Science, **262**(5131): 208-214.
8. **Becker A. and Theissen G.** 2003. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. Mol. Phylo. Evol., **29**: 464-489.
9. **Gramzow L. and Theissen G. A.** 2010. Hitchhiker's guide to the MADS world of plants. Genome Biol., **11**: 214.
10. **Schwarz-Sommer Z., Huijser P., Nacken W., Saedler H. and Sommer H.** 1990. Genetic control of flower development by homeotic genes in Antirrhinum majus. Science, **250**: 931-936.
11. **Coen E. S. and Meyerowitz E. M.** 1991. The war of the whorls: genetic interactions controlling flower development. Nature, **353**(6339): 31-37.
12. **Onouchi H., Igeño M. I., Périlleux C., Graves K. and Coupland G.** 2000. "Mutagenesis of plants overexpressing CONSTANS demonstrates novel interactions among Arabidopsis flowering-time genes". Plant Cell, **12**: 885-900.
13. **Michaels S. D. and Amasino R. M.** 1999. "FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering". Plant Cell, **11**: 949-956.
14. **Kotz S., Balakrishnan N. and Johnson N. L.** 2000. Continuous Multivariate Distributions. Volume 1: Models and Applications. New York: Wiley. ISBN 0-471-18387-3.
15. **Sandelin A., Alkema W., Engström P., Wasserman W. W. and Lenhard B.** 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res., **1**: 32(Database issue): D91-4.

Supplementary dataset 1. Random initial alignments of sequences of length 10 (font in un-bold face) collected from 1K upstream region of sequences of AGL3 gene in *Arabidopsis thaliana*.

>TFBS_AT.1
ACAACCATATATAGT**AGCCACTGAAT**
>TFBS_AT.2
CCACCCCATATAT**AGTAGTGCGGGTGGTG**
>TFBS_AT.3
CCATAAATAGATAGGCAGACTGTC**GCTGT**
>TFBS_AT.4
GTAAACATACCAT**AAATAGGA**
>TFBS_AT.5
TTCAAGAAACTGCCATAAATAGCGAT
>TFBS_AT.6
TAGAGGTTTTTGTGCCATAAA**TAGGT**
>TFBS_AT.7
CCCCATAAATAGGAATATCG**GGATGA**
>TFBS_AT.8
TTGCCATTAATAGA**TTATACCATATATGG**
>TFBS_AT.9
TATCAACAACGATACCAACC**CATATATGG**
>TFBS_AT.10
TTTCCAAATATAGAAGGTGTGGAAA**G**
>TFBS_AT.11
TCCAAATATAG**TAAAATCGAGTCGCGGAT**
>TFBS_AT.12
GACTGGGGCCCAAA**TATAGCATGTTC**
>TFBS_AT.13
ATCATTAGCTTTTACTT**ACCATAAATGG**
>TFBS_AT.14
ATTCTTTTGCCATAAATGGTAACTCG
>TFBS_AT.15
CCATAAATGGCAAGTCTGTCGAA**TAACGG**
>TFBS_AT.16
CCCATAAATGGCA**GGGTATTAGCACG**
>TFBS_AT.17
CCAAAAATAGATATGTGTCGT**AACAGCTT**
>TFBS_AT.18
CCAAAAATAGGGG**GACAATGGAAGTGGGG**
>TFBS_AT.19
CCAAAAATAGG**CCAGACGTGTTTACAACG**
>TFBS_AT.20
CCAAAAATAGTTAAATAA**TGTCATACATT**
>TFBS_AT.21
CTACACCTTCCAAAAATAGTA**ATCT**
>TFBS_AT.22
TTGCCAAATATGGG**GTTAGAGTGTTC**
>TFBS_AT.23
GTCTTTACCAAAAATGGTGATCC**TGT**
>TFBS_AT.24
TTGCCAAAAATGGAGCGT**TTACCAAT**
>TFBS_AT.25
ATCCACCATTTATAGAAAGTTC**AGGAGGC**
>TFBS_AT.26
GCATAAGAGAACATT**CCATTTATAGG**
>TFBS_AT.27
TCAACCCCATTTATAGCCA**CGTCAGT**
>TFBS_AT.28
CATCCATTAATAG**TAGCCATAATGGCG**
>TFBS_AT.29
GGAGTAGGCCCATTAATA**GTATCTTT**
>TFBS_AT.30
CCATTAATAGCATACAAAATCG**ACTCAAG**
>TFBS_AT.31
CCAATTATAGAAAGCTGTGGC**TGGTCGTC**
>TFBS_AT.32
AACTATTATTTCTCACAT**TCCATTAATGG**
>TFBS_AT.33
ATGCTTTACCAAT**AATAGAGCGCAA**

>TFBS_AT.34
GGTCAGTTAGATCC**AATTATGGAATG**
>TFBS_AT.35
GCATTCCAAAATTAGTAACGATATC**T**
>TFBS_AT.36
CCTCCTTTCCAAAATTAGTT**GAGAAG**
>TFBS_AT.37
CTTTGCCAAAATTAGC**TATTCTGAC**
>TFBS_AT.38
ACGCATGCACCACATATAGTAACG**TG**
>TFBS_AT.39
AGCGCCCCATTTTTAGGG**TTTAAGCT**
>TFBS_AT.40
CCATTTTTAGTAATTTA**ATACAACGCCGC**
>TFBS_AT.41
CCATTTTTAGTATGGAACCGCCGTGAGT
>TFBS_AT.42
AAATTACCATAAGTGGTAAT**GCACACAC**
>TFBS_AT.43
CAATTACTATATATAGCGTGTTTGTC
>TFBS_AT.44
CCATACATGGTAAAAT**GTACCGAAACACT**
>TFBS_AT.45
CCGAAAGTACTATA**AATAGTAATCCA**
>TFBS_AT.46
GTTTTCGCCTTT**CTATAAATAGTACC**
>TFBS_AT.47
GGTACCTAATATAGTAAT**CAGCTCTG**
>TFBS_AT.48
TCTTTAATTACTTGCCTAATATAG**CT**
>TFBS_AT.49
GGATGCATCCCTAATATA**GTTAATAA**
>TFBS_AT.50
ATCTTTACCAAAACT**AGTTAATTCGA**
>TFBS_AT.51
TGCTAAATATAG**AACATCTCCAAATA**
>TFBS_AT.52
CCTTCTAACTAAATATAGAA**GTGATA**
>TFBS_AT.53
ATTACCATAAACAGAA**ATCAGTGGAT**
>TFBS_AT.54
CCTAAAAATAGATCGAATG**TGTGCTC**
>TFBS_AT.55
TACTAAAAATAGATC**ATGAGCTACGA**
>TFBS_AT.56
TAGTCACTTGATTTCCATA**CCTAAAATGG**
>TFBS_AT.57
ATCAACAGTCTACAA**TATCCTAAAATGG**
>TFBS_AT.58
CAAACATTCCAGATTTAGAAT**GGTTA**
>TFBS_AT.59
ACTTCTTTCCTTTTAT**AGCTGAGTGC**
>TFBS_AT.60
CCGCTAAGCCTTGCCTTTTATAG**CA**
>TFBS_AT.61
TGCCCAAAAAAAGAAAGT**TGTCAGAC**
>TFBS_AT.62
CCCAAATAAGGAAAG**CTCTCTGGAC**
>TFBS_AT.63
AACCACACACC**CAGAATTAGTAAG**
>TFBS_AT.64
TGGTTCATTTAAGACTTTACCA**TTTCTGG**
>TFBS_AT.65
TGCGATGGAATAGTACT**AAATTTAGG**
>TFBS_AT.66
TGCGGAAGAGGGTTCCCGATATA**GATA**

>TFBS_AT.67
TCCCAAAAATACTCATATGT**CGGGCT**
>TFBS_AT.68
CCATATTGGGTAAA**GATTGCTTTTTAGCA**
>TFBS_AT.69
TAGTGGTTAACTA**CCCGTTTTTAGTA**
>TFBS_AT.70
CCAATTTATAGTTCACTTTCGTGATGAG**AA**
>TFBS_AT.71
TACTCGCTTTTCTTACTAAAA**GTAGA**
>TFBS_AT.72
TACTAAAAGTAGTAGT**TTGTCTGCA**
>TFBS_AT.73
CCATTTTAAGGAATTTACGA**TCTAGTGAA**
>TFBS_AT.74
TCCCATTTTAAGACCAACTTCTC**ATT**
>TFBS_AT.75
GATCCAGTAGATTCCAT**TTATGTACG**
>TFBS_AT.76
CCTGTAAATAGTAACAAGGT**GCATCG**
>TFBS_AT.77
CGATCAATATGTTACCATTTTGGGGT
>TFBS_AT.78
GAGCGAGACCTTCCC**AATAATTAGTAA**
>TFBS_AT.79
AAGTGACTATATTACTCA**AAATAGAA**
>TFBS_AT.80
TTACCAAATTGCCAATTTAGGCTAA**A**
>TFBS_AT.81
TTACCAAATTGCCAATTTAGG**CTAAA**
>TFBS_AT.82
GAAGATACCTAA**TACGGAAATTTTCC**
>TFBS_AT.83
GCAGACTTGCTATATTAA**GCTAATAT**
>TFBS_AT.84
CGCTTTCTTACTA**TAAATGTTTACTA**
>TFBS_AT.85
GGTTCAAGATTTTTCCTTTTA**TGTAC**
>TFBS_AT.86
TTTATAGTGTCCCTTTTT**CGGTAAGT**
>TFBS_AT.87
ATCCTACGATGCTTTCTAAAA**GAAGG**
>TFBS_AT.88
CCCTAGCAATTTTTTAC**TATATTTGT**
>TFBS_AT.89
GGAAGAATTTACTATAATTGTACA**TG**
>TFBS_AT.90
CTCTAATGGCCTTACTACTT**AAAGCA**
>TFBS_AT.91
GCCGAGTCCGGAA**AATTTCCGAAAAATG**
>TFBS_AT.92
ATTTATTTCCTAAAGTGTAA**CTAAC**
>TFBS_AT.93
GACCTGTATCCTTTTC**TACTTTTGTC**
>TFBS_AT.94
ATATGCCCCTCACAAGTTACCAA**TTA**
>TFBS_AT.95
AAAATGGTAATTTCTC**GGGACAGG**
>TFBS_AT.96
CTCAGTGCACACAGACATTCCA**AATA**
>TFBS_AT.97
GATTAGGATTCGTTTGTTT**CCAAATA**

Supplementary dataset 2. The final alignment of TFBS in upstream sequences of AGL3 genes of *Arabidopsis thaliana*

(1) Sequence number	(2) Left end location	(3) Left flanking	(4) Motif element	(5) Right flanking	(6) Right end location	(7) Probability of element	(8) Forward motif (+) or reverse complement (–)
1	5	acaa	CCATATATAG	tagcc	14	1	+
2	6	ccacc	CCATATATAG	tagtg	15	1	+
3	1		CCATAAATAG	atagg	10	1	+
4	10	acata	CCATAAATAG	ga	19	1	+
5	13	aactg	CCATAAATAG	cgat	22	1	+
6	15	ttgtg	CCATAAATAG	gt	24	1	+
7	3	cc	CCATAAATAG	gaata	12	1	+
8	4	ttg	CCATTAATAG	attat	13	0.24	+
8	29		CCATATATGG	tataa	20	0.76	-
9	20	ccaac	CCATATATGG		29	1	+
10	4	ttt	CCAAATATAG	aaggt	13	1	+
11	2	t	CCAAATATAG	taaaa	11	1	+
12	10	ggggc	CCAAATATAG	catgt	19	1	+
13	19	actta	CCATAAATGG		28	1	+
14	10	ttttg	CCATAAATGG	taact	19	1	+
15	1		CCATAAATGG	caagt	10	1	+
16	2	c	CCATAAATGG	cagggg	11	1	+
17	1		CCAAAAATAG	atatg	10	1	+
18	1		CCAAAAATAG	gggga	10	1	+
19	1		CCAAAAATAG	gccag	10	1	+
20	1		CCAAAAATAG	ttaaa	10	1	+
21	10	acctt	CCAAAAATAG	taatc	19	1	+
22	4	ttg	CCAAATATGG	ggtta	13	1	+
23	8	cttta	CCAAAAATGG	tgatc	17	1	+
24	4	ttg	CCAAAAATGG	agcgt	13	1	+
25	15	acttt	CTATAAATGG	tggat	6	1	-
26	25	c	CTATAAATGG	aatgt	16	1	-
27	16	cgtgg	CTATAAATGG	ggttg	7	1	-
28	4	cat	CCATTAATAG	tagcc	13	1	+
29	10	taggc	CCATTAATAG	tatct	19	1	+
30	1		CCATTAATAG	catac	10	1	+
31	10	gcttt	CTATAATTGG		1	1	-
32	20	acatt	CCATTAATGG		29	1	+
33	9	cttta	CCAATAATAG	agcgc	18	1	+
34	22	catt	CCATAATTGG	atcta	13	1	-
35	6	gcatt	CCAAAATTAG	taacg	15	1	+
36	9	ccttt	CCAAAATTAG	ttgag	18	1	+
37	6	ctttg	CCAAAATTAG	ctatt	15	1	+

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
38	10	atgca	CCACATATAG	taacg	19	1	+
39	16	aaacc	CTAAAAATGG	ggcgc	7	1	-
40	10	aatta	CTAAAAATGG		1	1	-
41	10	ccata	CTAAAAATGG		1	1	-
42	7	aatta	CCATAAGTGG	taatg	16	1	+
43	7	aatta	CTATATATAG	cgtgt	16	1	+
44	1		CCATACATGG	taaaa	10	1	+
45	10	aagta	CTATAAATAG	taatc	19	1	+
46	13	ccttt	CTATAAATAG	tacc	22	1	+
47	5	ggta	CCTAATATAG	taatc	14	1	+
48	15	acttg	CCTAATATAG	ct	24	1	+
49	10	gcatc	CCTAATATAG	ttaat	19	1	+
50	8	cttta	CCAAAACTAG	ttaat	17	1	+
51	3	tg	CTAAATATAG	aacat	12	1	+
52	9	tctaa	CTAAATATAG	aagtg	18	1	+
53	5	atta	CCATAAACAG	aaatc	14	1	+
54	2	c	CTAAAAATAG	atcga	11	1	+
55	3	ta	CTAAAAATAG	atcat	12	1	+
56	20	ccata	CCTAAAATGG		29	1	+
57	19	aatat	CCTAAAATGG		28	1	+
58	9	acatt	CCAGATTTAG	aatgg	18	1	+
59	18	ctcag	CTATAAAAGG	aaaga	9	1	-
60	23	tg	CTATAAAAGG	caagg	14	1	-
61	4	tgc	CCAAAAAAAG	aaagt	13	1	+
62	2	c	CCAAATAAGG	aaagc	11	1	+
63	11	cacac	CCAGAATTAG	taag	20	1	+
64	29		CCAGAAATGG	taaag	20	1	-
65	16	tagta	CTAAATTTAG	g	25	1	+
66	15	gggtt	CCCGATATAG	ata	24	1	+
67	3	tc	CCAAAAATAC	tcata	12	1	+
68	10	cttta	CCCAATATGG		1	1	-
69	24	ta	CTAAAAACGG	gtagt	15	1	-
70	11	gtgaa	CTATAAATTG	g	2	1	-
71	16	tctta	CTAAAAGTAG	a	25	1	+
72	3	ta	CTAAAAGTAG	tagtt	12	1	+
73	10	aattc	CTTAAAATGG		1	1	-
74	12	ttggt	CTTAAAATGG	ga	3	1	-
75	23	cgt	ACATAAATGG	aatct	14	1	-
76	2	c	CTGTAAATAG	taaca	11	1	+
77	24	ac	CCCAAAATGG	taaca	15	1	-
78	15	cttcc	CAATAATTAG	taa	24	1	+
79	15	tatta	CTCAAAATAG	aa	24	1	+

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
80	21	tttag	CCTAAATTGG	caatt	12	1	-
81	21	tttag	CCTAAATTGG	caatt	12	1	-
82	8	agata	CCTAATACGG	aaatt	17	1	+
83	19	attag	CTTAATATAG	caagt	10	1	-
84	11	tctta	CTATAAATGT	ttact	20	1	+
85	24	gt	ACATAAAAGG	aaaaa	15	1	-
86	21	actta	CCGAAAAAGG	gacac	12	1	-
87	16	gcttt	CTAAAAGAAG	g	25	1	+
88	26		ACAAATATAG	taaaa	17	1	-
89	12	attta	CTATAATTGT	acatg	21	1	+
90	24	tg	CTTTAAGTAG	taagg	15	1	-
91	19	aattt	CCGAAAAATG		28	1	+
92	9	tattt	CCTAAAGTGT	aacta	18	1	+
93	25	g	ACAAAAGTAG	aaaag	16	1	-
94	7	tatgc	CCCTCACAAG	ttacc	16	1	+
95	17	tgtcc	CGAGAAATTA	ccatt	8	1	-
96	11	tgcac	ACAGACATTC	caaat	20	1	+
97	17	tggaa	ACAAACGAAT	cctaa	8	1	-