J. Basic. Appl. Sci. Res., 2(7)7276-7286, 2012

© 2012, TextRoad Publication

ISSN 2090-4304
Journal of Basic and Applied
Scientific Research
www.textroad.com

EPWM: An Extended Position Weight Matrix for Motif Representation in Biological Sequences

Mohammad-Reza Sadeghi[1], Fatemeh Zare-Mirakabad[1*], Maryam Tahmasebi[1], and Mehdi Sadeghi[2]

[1] Faculty of Mathematics and Computer Science, Amirkabir University of Tecnology, Tehran, Iran.
[2]National Institute of Genetic Engineering and Biotechnology, Tehran, Iran.

ABSTRACT

Motif discovery is one of the fundamental problems in the signal detection and gene regulation. Motif discovery in biology is equivalent motif search and de novo motif finding problems in computer science. The challenging problem for both of these cases is Motif Representation (MR). A Common Position Weight Matrix (CPWM) is a simple MR model in which each position is independent from the other positions. However, the CPWM model is not an appropriate MR model. In fact, biological experiments show that the structural information is extremely important in motif discovery. The structural information is included in an MR model by considering dependent and conserved positions. Recently, some MR models have been introduced based on this assumption. These MR models are used only for the motif search problem.

In this paper, we design a new MR model based on information theory. This model can be used for the de novo motif finding and motif search problems. We extract some known motifs from JASPAR and TRANSFAC databases to search for common features among them. Based on these features, a new MR model is constructed called EPWM. A jackknife test is used to show the EPWM model is more successful than the other MR models for the motif search problem. The jackknife test with each MR model (the EPWM model and the other MR models) is implemented and performed on the JASPAR and TRANSFAC databases. To verify the efficiency of the EPWM model for the de novo motif finding problem, we implement the EPWM model and the other MR models in the Gibbs sampling method. Finally, the Gibbs sampling method is performed on the JASPAR and TRANSFAC databases. The results show that the EPWM model gives more accurate prediction than the other MR models for the motif search and de novo motif finding problems.

***KEYWORDS*:** motif finding, motif representation, mutual information, joint information content.

1. INTRODUCTION

DNA-binding proteins called Transcription Factors (TFs) are involved in transcription regulation. These factors bind to specific positions in promoter regions for gene regulation. Identifying Transcription Factor Binding Sites (TFBSs) in promoter regions is a major challenge for molecular biologists. A large number of computational methods have been proposed for finding TFBSs in a set of promoter regions.

In computational methods, a set of common TFBSs detected by a TF is called a motif. The occurrence of the motif in a promoter region is called a motif instance. Thus, the motif instance is a subsequence which occurs in the promoter region sequence.

In computer scienc, there existe two problems encountered for TFBS discovery or motif discovery which are motif search and de novo motif finding (Soleimani-damaneh, 2011), (Zörnig, 2011). In the first one, a motif M with length ℓ and a sequence S are given as inputs, and the goal is to find some subsequences of S with the highest similarity to M. In the second one, given a sample set of sequences and the length ℓ of an unknown motif, the goal is to find all subsequences of the length ℓ that occur in the sample set with the highest similarity.

In both problems, motif representation is significantly important. One of the best statistical MR models is the Position Weight Matrix (PWM). The Common PWM (CPWM) assumes that all positions are independent of each other (Bailey & Elkan, 1995), (Hughes et al., 2000), (Lawrence et al., 1993). This MR model is easy to implement because it needs only a few parameters. MEME (Bailey & Elkan,

***Corresponding Author:** Fatemeh Zare-Mirakabad Faculty of Mathematics and Computer Science, Amirkabir University of Tecnology, Tehran, Iran.Tel: +98-2166460948; Email address: f.zare@aut.ac.ir

1995), AlignACE (Hughes et al.,2000), Gibbs-Sampler (Lawrence et al., 1993), an immune genetic algorithm (Luo & Wang, 2010), YMF (Sinha & Tompa, 2003) and GADPAF (Zare-Mirakabad et al., 2009) are based on the CPWM model for motif finding problem. However, recent experimental evidences identify that the structure of a motif is important (Kim, 2010), (Tomovic & Oakeley, 2007). Some models consider position dependecies to represnt the structure of a motif in the MR model (Benos et al., 2002), (Bulyk et al., 2002). The related methods include Bayesian networks (Barash et al., 2003), Markov chain optimization (Ellrott et al., 2002), non-parametric (King & Roth, 2003), hidden Markov (Marinescu et al., 2005), permuted Markov (Zhao et al., 2005), and generalized weight matrix (Zhou & Liu, 2005). However, they require more complicated mathematical tools, more parameters for estimation and more experimental data than the common models (Barash et al., 2003), (Ellrott et al., 2002), (King & Roth, 2003).

Recently, two other MR models have been described by Tomovic (Tomovic & Oakeley, 2007) and Zare (Zare-Mirakabad et al., 2009) based on the position dependencies. The MR models based on the position dependencies predict binding sites with lower false positive rates. Tomovic's model is applicable in the motif search problem. In this problem, a motif M with length ℓ and a sequence S are given as inputs. At first, the model finds some dependency between positions in the motif M, and then it chooses a new motif instance in the sequence S based on the detected dependencies. Clearly, this model can find a new motif instance better than the CPWM model. However, this model is not practical in the de novo motif finding problem because there is no dependency information about an unknown motif. Zare's model is applied in the motif search and de novo motif finding problems, but this model considers dependency between any two positions and ignores conservation in each position.

In this study, a new MR model is constructed in which it is applicable in the both motif discovery problems by using the improvement of Zare's model. This new MR model is called EPWM. At first, 107 motifs from the JASPAR database (Sandelin et al., 2004) and 25 motifs from the TRANSFAC database (Wingender et al., 1996), based on the Model-Real benchmark (Sandve et al., 2007), are extracted and analyzed. Our analytical results can be summarized as follows: 1) the information between any two adjacent positions is more than the information between any two positions; 2) the information of any two adjacent positions is more than one position and three adjacent positions; 3) a motif is a cycle chain in which the first position may depends on the last position; 4) both conserved positions and dependent positions are crucial and cannot be disregarded. Therefore, these four features are the key observations that should be considered in an MR model.

Based on these features, we design a new model called EPWM model. The accuracy of the EPWM model on the motif search problem and de novo motif finding problem are evaluated by the jackknife test and Gibbs sampling method, respectively. Gibbs sampling method and jackknife test are performed on the real extracted motifs from the JASPAR and TRANSFAC databases. Then, the Gibbs sampling method and jackknife test with CPWM, Tomovic’s PWM (Tomovic & Oakeley, 2007) and Zare’s PWM (Zare-Mirakabad et al., 2009) are implemented and performed on the JASPAR and TRANSFAC databases. The results show that the EPWM model increases the efficiency of the Gibbs sampling method and jackknife test for de novo motif finding and motif search problems, respectively. Our result show that, the MR models can be simply replaced with the EPWM model to increase the accuracy of the methods.

2. METHODS

In this section, some MR models are introduced. In following, the new MR model is described.

2.1. Motif representation models

Common position weight matrix, CPWM, is an MR model with independent positions in the motif (Hertzberg et al., 2005). Biological experiments show that the structure and the sequence of the motif are important for the TFs. The CPWM shows only the sequence information of the motif. The structural information is included in an MR model by considering both dependent and conserved positions. Tomovic (Tomovic & Oakeley, 2007) proposed a method to find position dependencies in the motif, and then constructed a PWM named TPWM based on the detected dependencies. This model is applicable in the motif search problem. In this problem, a known motif is given as an input, and then based on the motif, the dependency between positions is computed, and afterwards TPWM model is constructed. Finally, TPWM model searches an unknown motif instance in the given sequence. However this model cannot be used in de novo motif finding problem, because there is no motif as an input to obtain the dependency information about it. Zare et al. (Zare-Mirakabad et al., 2009) defined another PWM (ZPWM) in which any two positions are dependent based on the joint information content and mutual information. This model is applicable in both motif discovery problems. However, this model only considers dependency information on any two positions and it loses independency information (conservation) on each position.

2.2. EPWM model

A sequence is a string on a given alphabet Γ where $\Gamma = \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}$. A substring $\mathsf{s}[\mathsf{j}]\dots\mathsf{s}[\mathsf{j}+\ell-1]$ with the length ℓ of a given string s is referred as a subsequence of the length ℓ. Let the set S be defined on t sequences, $\mathsf{S} = \{\mathsf{s}_1,\dots\mathsf{s}_\mathsf{t}\}$, where s_i is the i^{th} sequence with the length n_i. Assume that the motif M with the length ℓ occurs in the sample set S and it is represented by $\mathsf{M} = \{\mathsf{I}_1,\mathsf{I}_2,\dots,\mathsf{I}_\mathsf{t}\}$, where I_i is a subsequence (motif instance) with the length ℓ in the i^{th} sequence.

Some known motifs are extracted and analyzed from the JASPAR and TRANSFAC databases. We find four common features among them. These features are analyzed in the results section. In the following, we introduce a new MR model based on these features called EPWM.

Assume that a set M as a known motif is given, and then the matrix $\mathsf{W}^{\mathsf{M}}_{4\times4\times\ell}$ is constructed as an EPWM on the motif as follows:

1. The probability matrix $\mathsf{F}'^{\mathsf{M}}_{4\times\ell}$ is constructed where $\mathsf{F}'^{\mathsf{M}}_{[\alpha,\mathsf{j}]}$ shows the probability of the nucleotide $\alpha \in \Gamma$ in the j^{th} position of the motif M.
2. The probability matrix $\mathsf{F}''^{\mathsf{M}}_{4\times4\times\ell}$ is made where $\mathsf{F}''^{\mathsf{M}}_{[\alpha_1,\alpha_2,\mathsf{j}]}$ shows the probability of the two nucleotides $\alpha_1,\alpha_2 \in \Gamma$ in the positions j and $\mathsf{j}+1$ of the motif M.
3. The array $\mathsf{B}'^{\mathsf{S}}_{[\alpha]}$ shows the probability of nucleotide $\alpha \in \Gamma$ in the set S as a background.
4. The array $\mathsf{B}''^{\mathsf{S}}_{[\alpha_1,\alpha_2]}$ identifies the probability of two adjacent nucleotides $\alpha_1,\alpha_2 \in \Gamma$ in the set S as a background.
5. The matrix $\mathsf{W}'^{\mathsf{M}}_{4\times\ell}$ is obtained as:
$$\mathsf{W}'^{\mathsf{M}}_{[\alpha,\mathsf{j}]} = \log_2(\mathsf{F}'^{\mathsf{M}}_{[\alpha,\mathsf{j}]}/\mathsf{B}'^{\mathsf{S}}_{[\alpha]}).$$
6. The matrix $\mathsf{W}''^{\mathsf{M}}_{4\times4\times\ell}$ is made as:
$$\mathsf{W}''^{\mathsf{M}}_{[\alpha_1,\alpha_2,\mathsf{j}]} = \log_2(\mathsf{F}''^{\mathsf{M}}_{[\alpha_1,\alpha_2,\mathsf{j}]}/\mathsf{B}''^{\mathsf{S}}_{[\alpha_1,\alpha_2]}).$$
7. The new position weight matrix (EPWM), $\mathsf{W}^{\mathsf{M}}_{4\times4\times\ell}$, is constructed as:
$$\mathsf{W}^{\mathsf{M}}_{[\alpha_1,\alpha_2,\mathsf{j}]} = \omega^{\mathsf{M}}[\mathsf{j}] \times \mathsf{W}''^{\mathsf{M}}_{[\alpha_1,\alpha_2,\mathsf{j}]} + (1-\omega^{\mathsf{M}}[\mathsf{j}]) \times \mathsf{W}'^{\mathsf{M}}_{[\alpha_1,\mathsf{j}]}.$$
where $\omega^{\mathsf{M}}[\mathsf{j}]$, $1 \le \mathsf{j} \le \ell$, determines the probability of the dependency between two adjacent positions j and $\mathsf{j}+1$. This weight is obtained as follows:
    i) For each position j, Joint Information Content (JIC) and Mutual Information (MI) on the motif M is obtained as follows (Cover & Thomas, 2005):
$$\mathsf{JIC}^{\mathsf{M}}[\mathsf{j}] = \sum_{\alpha_1\in\Gamma}\sum_{\alpha_2\in\Gamma} \mathsf{F}''^{\mathsf{M}}_{[\alpha_1,\alpha_2,\mathsf{j}]} \times \mathsf{W}''^{\mathsf{M}}_{[\alpha_1,\alpha_2,\mathsf{j}]},$$
$$\mathsf{MI}^{\mathsf{M}}[\mathsf{j}] = \sum_{\alpha_1\in\Gamma}\sum_{\alpha_2\in\Gamma}(\mathsf{F}''^{\mathsf{M}}_{[\alpha_1,\alpha_2,\mathsf{j}]} \times \log_2(\mathsf{F}''^{\mathsf{M}}_{[\alpha_1,\alpha_2,\mathsf{j}]}/\left(\mathsf{F}'^{\mathsf{M}}_{[\alpha_1,\mathsf{j}]} \times \mathsf{F}'^{\mathsf{M}}_{[\alpha_2,\mathsf{j}+1]}\right))).$$
    ii) For each position j, $\omega^{\mathsf{M}}[\mathsf{j}]$ is computed as follows:
$$\omega^{\mathsf{M}}[\mathsf{j}] = (\mathsf{MI}^{\mathsf{M}}[\mathsf{j}] + \mathsf{JIC}^{\mathsf{M}}[\mathsf{j}])/4.$$

Let s' be a subsequence with the length ℓ. This subsequence is scored based on W^{M} as follows:
$$\mathsf{Score}(\mathsf{s}',\mathsf{W}^{\mathsf{M}}) = \textstyle\sum_{\mathsf{j}=1}^{\ell} \mathsf{W}^{\mathsf{M}}_{[\mathsf{s}'[\mathsf{j}],\mathsf{s}'[\mathsf{j}+1],\mathsf{j}]}.$$

3. RESULTS AND DISCUSSION

In this section, the accuracy of four features which are common in the most motifs are discussed and then the new model, EPWM, is made based on these features. Afterwards, this MR model is planted in the Gibbs sampling method and jackknife test to demonstrate that the new MR model is an appropriate model for both motif discovery problems. Then, for comparing the EPWM model to some well-known (the CPWM, ZPWM and TPWM) MR models, these MR models are planted in the Gibbs sampling method and jackknife test. The Gibbs sampling method and jackknife test with each MR model is run on some extracted motifs from the JASPAR and TRANSFAC databases. Finally, the results of them are compared based on the normalized nCC, nSp and nSn.

3.1. Databases

Some TFs are extracted from two public databases, JASPAR (Sandelin et al., 2004) and TRANSFAC (Wingender et al., 1996). We select 107 TFs from the JASPAR database and implant TFBSs of these TFs in some random sequences which are generated similar to (Tomovic & Oakeley, 2007). For extracting motifs from the TRANSFAC database, we use the generated benchmark by Sandve et al. (Sandve et al., 2007). They have made a benchmark from the TRANSFAC database based on the collections of binding site fragments that are ranked according to the optimal level of discrimination. This benchmark contains three parts, Markov-algorithm', 'Real-algorithm', and 'Model-Real'. In this paper, we use 'Model-Real' because there are at least 10 motif instances for each motif in this benchmark.

3.2. Methods for Comparison

Some comparisons are needed to evaluate the new MR model. The comparisons are preceded in nucleotide level, with respect to the position of motifs in the main sequences. For this reason, we first introduce the following criteria for comparison (Tompa et al., 2005), (Burset & Guigo, 1996), (Lenhard & Wasserman, 2002).

1. nTP: the number of nucleotide positions in both known sites and the predicted sites,
2. nFP: the number of nucleotide positions not in the known sites but in the predicted sites,
3. nFN: the number of nucleotide positions in the known sites but not in the predicted sites,
4. nTN: the number of nucleotide positions neither in the known sites nor in the predicted sites.

Based on above criteria, three different measurements for the evaluation of the algorithm are introduced.

1. Nucleotide Correlation Coefficient (nCC) is defined as:
   $$nCC = (nTP \times nTN - nFN \times nFP)/\sqrt{(nTP + nFN)(nTN + nFP)(nTP + nFP)(nTN + nFN)}$$
   The value of nCC varies from -1 (indicating perfect anti-correlation between two known sites the predicted sites) to $+1$ (indicating the perfect correlation and match).
2. Nucleotide Specificity (nSp) is a statistical measure for the correctness prediction of positions of a non-motif sequence which is equal to
   $$nSp = nTN/(nTN + nFP).$$
3. Nucleotide Sensitivity (nSn) is the fraction of the known site nucleotides that are predicted as motifs and is defined by
   $$nSn = nTP/(nTP + nFN).$$

None of the above measurements can provide the correlation of the motif discovery methods perfectly by themselves. Therefore, in any case, we need a way of summarizing the performance of a given motif finding method over all samples. For each method, each measurement i (one of the above three measurements), over all samples, is obtained and the performance of each method on all samples are compared by normalization. For each motif, the measurement i is normalized by subtracting the mean and dividing by the standard deviation over all the methods on that motif, and the average of these normalized scores over all motifs is obtained. Application of this method allows us to put easy and hard motifs on the same scale.

3.3. The evaluation of the new motif representation model

Some extracted motifs from the JASPAR and TRANSFAC databases are analyzed and four following features are found:

1. Any two adjacent positions give more information than each two positions in the motifs.
2. The information of two adjacent positions is more than one position and three adjacent positions.
3. A motif is a cycle chain.
4. The information on the conserved positions and dependent positions is crucial and important in motifs.

In following, these features are evaluated based on the comparison and statistical analysis.

3.3.1. Evaluation and comparison

These four features are studied on each motif from JASPAR and TRANSFAC databases based on various measurements of the information theory.

For the first feature, we would like to compare two assumptions, any two adjacent positions and any two positions, in the motifs. For each motif M, the first and the second assumptions are evaluated based on the normalized joint information content plus mutual information (NJICMI^{M}). For the first assumption, any two adjacent positions, NJICMI^{M} is computed as follows:

$$\mathrm{NJICMI}^{M} = (1/4\ell) \times \sum_{j=1}^{\ell}(\mathrm{JIC}^{M}[j] + \mathrm{MI}^{M}[j]),$$

where $\mathrm{MI}^{M}[j]$ and $\mathrm{JIC}^{M}[j]$ shows the mutual information and joint information content between two positions j and $j + 1$, respectively in the motif M. For the second assumption, any two positons, NJICMI^{M} is computed as follows:

$$\mathrm{NJICMI}^{M} = (1/(4\ell(\ell-1)/2)) \times \sum_{i=1}^{\ell-1}\sum_{j=i+1}^{\ell}(\mathrm{JIC}^{M}[i,j] + \mathrm{MI}^{M}[i,j]),$$

where $\mathrm{MI}^{M}[i,j]$ and $\mathrm{JIC}^{M}[i,j]$ shows the mutual information and joint information content between two positions i and j in the motif M, respectively.

The values of NJICMI^{M} based on two assumptions are computed on each motif M, $1 \le M \le 107$, from the JASPAR database and they are shown in Figure 1. For each motif M, $1 \le M \le 25$, from the TRANSFAC database (Model-Real benchmark), the values of NJICMI^{M} are illustrated in Figure 2. In these figures, it can be seen that the information of any two adjacent positions is more than any two positions because in the folding of macromolecules, local interaction among residues play more important role than non-local or distant residues.

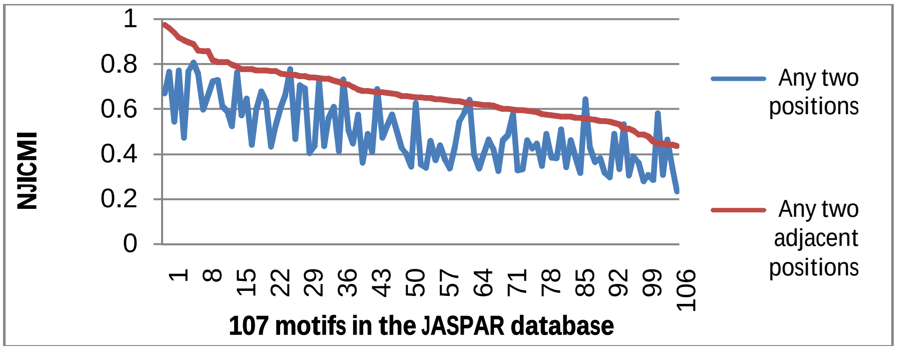

Figure 1. Results obtained from 107 motifs based on any two adjacent positions and any two positions. These motifs are selected from the JASPAR database. For each of them, the normalized **JIC** plus **MI** (**NJICMI**) based on any two adjacent positions and any two positions are computed

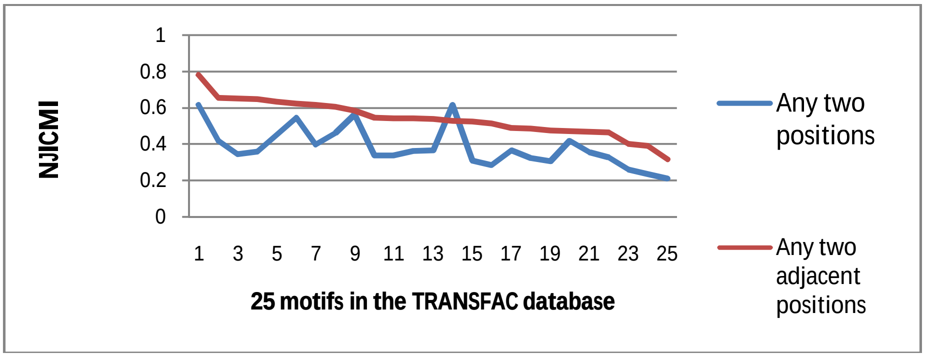

Figure 2. Results obtained from 25 motifs based on any two adjacent positions and any two positions. These motifs are selected from TRANSFAC database (Model-Real benchmark). For each of them, the normalized JIC plus MI (NJICMI) based on any two adjacent positions and any two positions are computed.

The second feature describes that how many adjacent positions may be dependent in a motif. For each motif, we study three assumptions, one position, two adjacent positions and three adjacent positions based on the normalized entropy. For each motif M, entropy on k adjacent positions is defined as follows:

$$\mathrm{E}_{\mathrm{k}}^{\mathrm{M}} = -\sum_{\alpha_1 \in \Gamma} \cdots \sum_{\alpha_k \in \Gamma} \sum_{j=1}^{\ell} \left(\mathrm{F}_{[\alpha_1,\ldots,\alpha_k,j]}^{\mathrm{M}} \log_2 \mathrm{F}_{[\alpha_1,\ldots,\alpha_k,j]}^{\mathrm{M}}\right),$$

where $\mathrm{F}_{[\alpha_1,\ldots,\alpha_k,j]}^{\mathrm{M}}$ shows the probability of the nucleotides $\alpha_1, \ldots, \alpha_k \in \Gamma$ in the positions $\mathrm{j}, \ldots, \mathrm{j} + \mathrm{k} - 1$ of the motif M.The normalized entropy is obtained as follows:

$$\mathrm{NE}_{\mathrm{k}}^{\mathrm{M}} = \mathrm{E}_{\mathrm{k}}^{\mathrm{M}} / (\min\{\mathrm{t}, 2\ell\} \times \ell) \ ,$$

where t, ℓ are the number of motif instances in the set M and the length of motif, respectively. For each motif M, the normalized entropy for $\mathrm{k} = 1, 2$ and 3 is computed. In the most motifs from the JASPAR and TRANSFAC (Mode-Real benchmark) databases, Figure 3 and Figure 4 show that the value of $\mathrm{NE}_2^{\mathrm{M}}$ is less than $\mathrm{NE}_1^{\mathrm{M}}$ and $\mathrm{NE}_3^{\mathrm{M}}$.

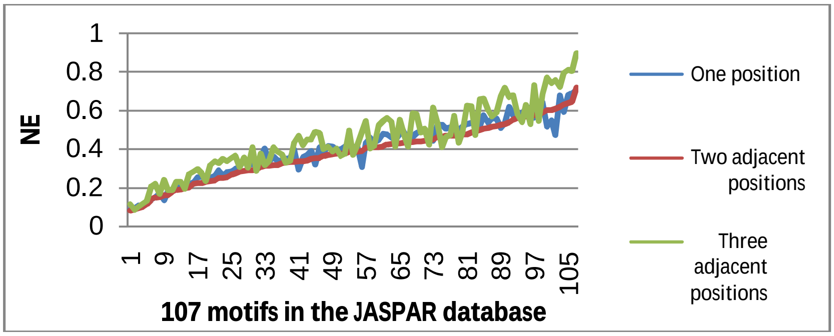

Figure 3. Results obtained from 107 motifs on one, two and three adjacent positions based on the normalized entropy. These motifs are selected from JASPAR database. For each of them, the normalized entropy NE based on one position, two and three adjacent positions are computed.

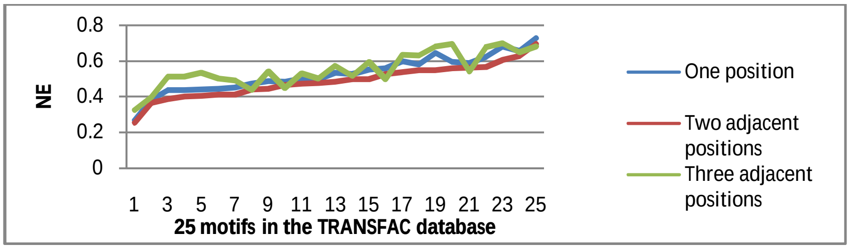

Figure 4. Results obtained from 25 motifs on one, two and three adjacent positions based on the normalized entropy. These motifs are selected from TRANSFAC database (Model-Real benchmark). For each of them, the normalized entropy (NE) are computed based on one position, two and three adjacent positions.

The third feature identifies that a motif is a cycle or a linear chain. These assumptions, cycle or linear chain, are evaluated based on the normalized mutual information in each motif. Normalized mutual information on the cycle chain assumption is computed as follows:

$$\mathrm{NMI}^{\mathrm{M}} = (1/4\ell) \times \textstyle\sum_{\mathrm{j}=1}^{\ell} \mathrm{MI}^{\mathrm{M}}[\mathrm{j}],$$

where $\mathrm{MI}^{\mathrm{M}}[\mathrm{j}]$, $1 \leq \mathrm{j} \leq \ell - 1$, shows the mutual information between two positions j and $j + 1$ and for $\mathrm{MI}^{\mathrm{M}}[\ell]$, is equal to the mutual information between two positions ℓ and 1. Normalized mutual information on the linear chain assumption is computed as follows:

$$\mathrm{NMI}^{\mathrm{M}} = 1/(4(\ell - 1)) \times \textstyle\sum_{\mathrm{j}=1}^{\ell-1} \mathrm{MI}^{\mathrm{M}}[\mathrm{j}],$$

where $\mathrm{MI}^{\mathrm{M}}[\mathrm{j}]$ shows the mutual information between two positions j and $j + 1$. For the most motifs from

JASPAR database and TRANSFAC database (Model-Real benchmark), Figure 5 and Figure 6 show that NMI on the cycle chain is equal or larger than the linear chain. It can be guessed that when two strands of DNA are separated to transcript, the first and the last positions of the single strand come near each other for making a structure.

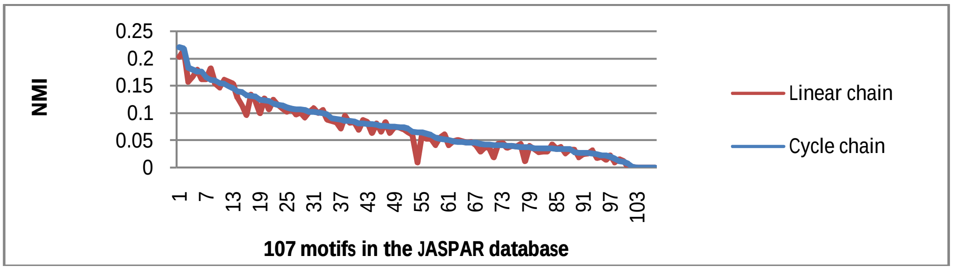

Figure 5. Results obtained from 107 motifs on the linear and cycle chain based on the normalized mutual information.These motifs are selected from JASPAR database. For each of them, the normalized mutual in formation (NMI) is computed based on the linear and cycle chain.

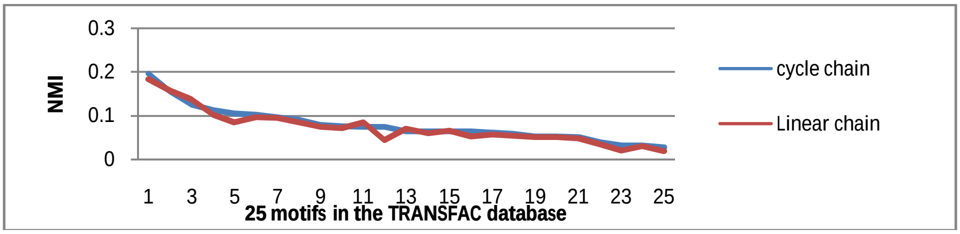

Figure 6. Results obtained from 25 motifs on the linear and cycle chain based on the normalized mutual information. These motifs are selected from TRANSFAC database (Model-Real benchmark). For each of them, the normalized mutual information (NMI) is computed based on the linear and cycle chain.

In the last feature, the assumption is that some positions are conserved and independent of the other positions. To evaluate this feature, we select MA0053 motif from the JASPAR database and analyze single and two adjacent positions in the motif. In Figure 7, it can be seen that the fourth position is exactly conserved and the fifth position is non-conserved. If we consider only position dependencies, the value of the conservation in some positions are ignored. In this motif, mutual information plus joint information content between the fourth and the fifth positions is almost equal to $0 + 2.63$ and Information Content (IC) for the fourth position is equal to 2. So, a multiplier of value of conservation of each position should be added to the value of dependency.

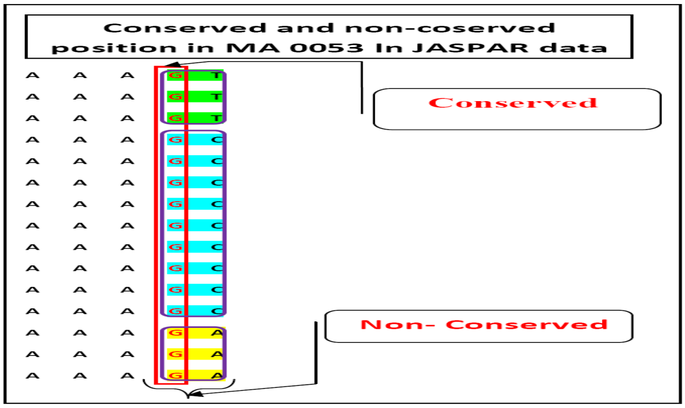

Figure 7. The motif MA0053 is selected from the JASPAR database. The fourth position is exactly conserved. If we consider the fourth and fifth positions together, then the value of the fourth position conservation is lost.

3.3.2. Statistical analysis

Let $P_{i,j,k}$ be a set of values of measurement i on database j and assumption k. The average of $P_{i,j,k}$ is called $\mu_{i,j,k}$. We would like to compare $\mu_{i,j,k}$ and $\mu_{i,j,k'}$. At first, we investigate normal distribution on each set $P_{i,j,k}$. This is done by using Kolmogorov-Smirnov test. Now, if the pvalue is more than 0.05 then the null hypothesis (H_0), stating the data come from normal distribution, is accepted. For comparing $\mu_{i,j,k}$ and $\mu_{i,j,k'}$, we use the paired t-test, if $P_{i,j,k}$ and $P_{i,j,k'}$ have normal distribution otherwise Wilcoxon signed-rank test is applied. Now, if the pvalue is more than 0.05 thus the null hypothesis (H_0), stating $\mu_{i,j,k} \geq \mu_{i,j,k'}$, is accepted. In Table 1, the statistical results on the features can be seen. For each measurement i, database j, assumptions k and k′, the null hypothesis $\mu_{i,j,k} \geq \mu_{i,j,k'}$ is accepted because pvalue is more than *0.05*. So, all features are acceptable in the most motifs.

Table 1. For each measurement i, database j, assumptions k and k′, pvalue is computed on null sis (H_0).

i	j	K	k′	Type of test	H0: $\mu_{i,j,k} \geq \mu_{i,j,k'}$ pvalue
NJICMI	JASPAR	two adjacent positions	two positions	Wilcoxon	1
NJICMI	TRANSFAC	two adjacent positions	two positions	Wilcoxon	1
NE	JASPAR	one position	two adjacent positions	T – test	1
NE	JASPAR	three adjacent positions	two adjacent positions	T – test	1
NE	TRANSFAC	one position	two adjacent positions	T – test	1
NE	TRANSFAC	three adjacent positions	two adjacent positions	T – test	1
NMI	JASPAR	cycle chain	linear chain	Wilcoxon	0.9998
NMI	TRANSFAC	cycle chain	linear chain	T – test	0.9915

3.4. Motif search problem

To evaluate the EPWM model for the motif search problem, we implement a jackknife test with each MR model (EPWM, CPWM, ZPWM and TPWM model). The jackknife test with one of the MR models is implemented as follows. Assume that a motif M has t known TFBSs (motif instances) with the length ℓ which are implanted in t sequences. The k^{th} motif instance is ignored in the set M. The MR model is made from $t-1$ remaining motif instances. In the k^{th} sequence where $1 \leq k \leq t$, each subsequence is scored by the MR model. Finally, a subsequence with maximum score is selected as a predicted motif instance in the k^{th} sequence. In the next steps, the above steps are repeated t times for all motif instances. The jackknife test with each MR model (EPWM, CPWM, ZPWM and TPWM) is performed on the motifs from the JASPAR database and Model-Real benchmark from the TRANSFAC database (we cannot perform the jackknife test with the TPWM model on the Model-Real benchmark because we do not have dependent positions based on Tomovic method (Tomovic & Oakeley, 2007)). Figures 8 and 9 show the normalized nCC, nSp and nSn on the motifs from the JASPAR and TRANSFAC databases, respectively. These figures illustrate that the EPWM model has the best statistical accuracy in the jackknife test.

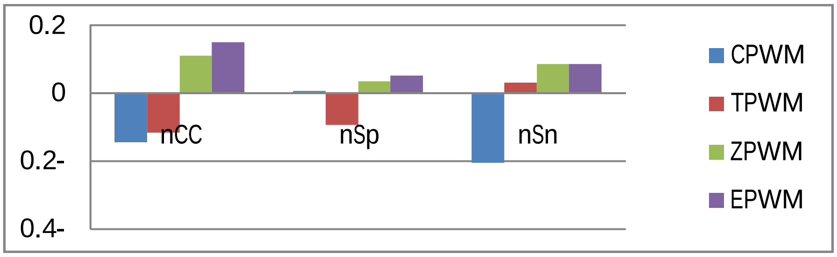

Figure 8. The jackknife test results obtained on 107 motifs from the JASPAR database with each MR model (CPWM, TPWM, ZPWM and EPWM model).The normalized nCC, nSp and nSn values show that the EPWM model is more successful than the other MR models in the motif search problem.

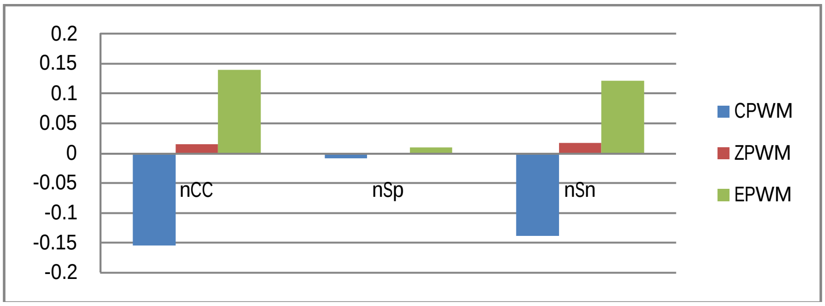

Figure 9. The jackknife test results obtained on 25 motifs from the TRANSFAC database with each MR model (CPWM, TPWM, ZPWM and EPWM model). The normalized nCC, nSp and nSn values show that the EPWM model is more successful than the other models in the motif search problem.

3.5. De novo motif finding problem

In the motif finding problem, the set S is given as an input in which it contains t sequences and the length of the sequence $s_k \in S$ is denoted by $|s_k| = n_k$. The goal is to find a motif with length ℓ. The Gibbs sampling method is a method for de novo motif finding. In the first step, this algorithm selects one random motif instance from each sequence. These random motif instances are shown as a set $M$and then a motif instance I_k is randomly selected from the set M and is removed. Now, a PWM is made from the set M and each subsequence with the length ℓ is scored in the k^{th} sequence. Afterwards, a subsequence with maximum score is added to the set M as a predicted motif instance. Randomly deleting step from the set M is repeated twice the number of sequences. Finally, the updated set M is introduced as a predicted motif. For each motif M, the Gibbs sampling method is performed on $max_{j=1}^{t_M}|n_j|/\ell_M$ random motif instance sets where t_M, n_j and ℓ_M are the number of the sequences, the length of sequence j and the length of an unknown motif, respectively. The Gibbs sampling method with each MR model (EPWM, ZPWM, CPWM and TPWM) is performed on the motifs from the JASPAR and TRANSFAC databases (the Gibbs sampling method with TPWM is not performed on the TRANSFAC database). Figures 10 and 11 illustrate the results of the normalized value of nCC, nSp and nSn on the JASPAR database and TRANSFAC, respectively. These figures show that the EPWM model has better statistical accuracy than the other models in the Gibbs sampling method.

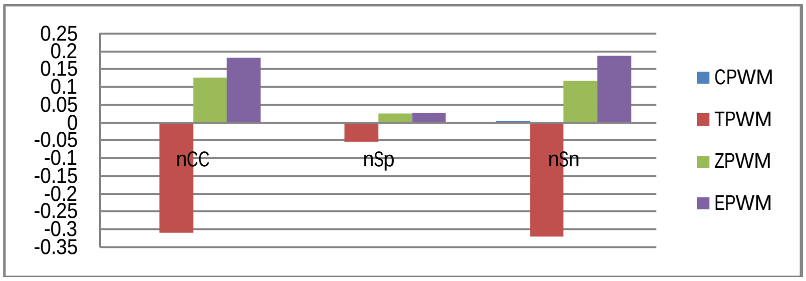

Figure 10. The Gibbs sampling results obtained on 107 motifs from the JASPAR database with each MR model (CPWM, TPWM, ZPWM and EPWM model). The normalized nCC, nSp and nSn values show that the EPWM model is more successful than the other MR models in the de novo motif finding problem.

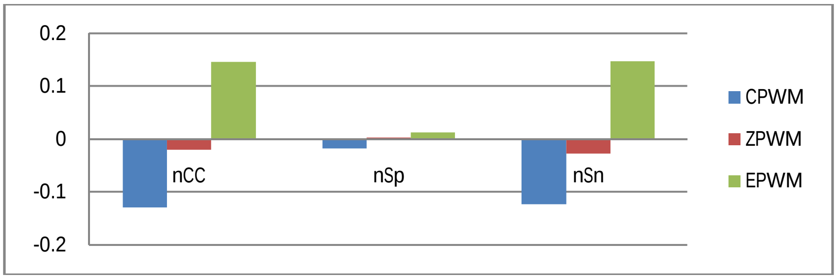

Figure. 11. The Gibbs sampling results obtained on 25 motifs from the TRANSFAC database with each MR model (CPWM, TPWM, ZPWM and EPWM model). The normalized nCC, nSp and nSn values show that the EPWM model is more successful than the other MR models in the de novo motif finding problem.

4. Conclusions

In the present study, a new MR model based on the dependency between adjacent positions along with conserved positions in the binding sites is proposed. We used our MR model in the jackknife test and Gibbs sampling method for motif search and de novo motif finding problems, respectively. In this MR model (EPWM), joint information content and mutual information were used as a measure of dependency between adjacent positions and information content was utilized as a measure of conserving for each position in TFBSs. Finally, the results of CPWM, TPWM and ZPWM on the JASPAR (Sandelin et al., 2004) and TRANSFAC (Wingender et al., 1996) (and Model-Real benchmark (Sandve et al., 2007)) databases were compared to our MR model. Our results show that the new MR model has better specificity (Sp) and sensitivity (Sn) than the other MR models.

We have to emphasize that we have not designed or implemented a new algorithm or software for motif finding. We only claim that this model of motif representation will make an existing algorithm more efficient. In this study we used the Gibbs sampling algorithm as an example to show this, so if we want to test this with the other program such as MEME, we replace the motif representation model in MEME (CPWM) with our new model (EPWM). But in order to do so, we need to have source code of MEME or we have to implement it again. For a future work one can improve some of well known motif finding methods with this new MR model and compare them to the original methods.

REFERENCES

[1] Soleimani-damaneh, M., (2011). An optimization modelling for string selection in molecular biology using Pareto optimality. *Applied Mathematical Modelling*, 35, 3887-3892.

[2] Zörnig, P., (2011). Improved optimization modelling for the closest string and related problems. *Applied Mathematical Modelling*, 35, 5609-5617.

[3] Bailey, T., Elkan, C., (1995). The value of priori knowledge in discovering motifs with MEME. *In Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology AAAI Press, Menlo Park,CA*, 21-29.

[4] Hughes, J., Estep, P., Tavazoie, S., Church, G., (2000). Computational identification of cisregulatory elements associated with functionally coherent groups of genes in Saccharomyces Cerevisiae. *JMol Biology* 296(5), 1205-1214.

[5] Lawrence, C., Altschul, S., Bogusky, M., Liu, J., Neuwald, A., Wootton, J., (1993) .Detecting subtle sequence signals: Gibbs sampling strategy for multiple alignment. *Science*, 262, (5131), 208-214.

[6] Luo, J., & Wang, T., (2010). Motif discovery using an immune genetic algorithm. *Journal of Theoretical Biology*, 264 (2), 319-325.

[7] Sinha, S., Tompa, M., (2003). YMF, A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* 31(13), 3586-3588.

[8] Zare-Mirakabad, F., Ahrabian, H., Sadeghi, M., Hashemifar, S., Nowzari-Dalini A., Goliaei, B., (2009). Genetic algorithm for dyad patterns finding DNA sequences. *Genes Genet. syst.* 84 (1), 81-93.

[9] Kim, J. T., Martinetz, T., Polani, D., (2003). Bioinformatic principles underlying the information content of transcription factor binding site, *Journal of Theoretical Biology*, 220 (4), 529–544.

[10] Tomovic, A., & Oakeley, E., (2007). Position dependencies in transcription factor binding sites. *Bioinformatics*, 23 (8), 933-941.

[11] Benos, P., Bulyk, M., Stormo, G., (2002). Additivity in Protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, 30 (20), 4442-4451.

[12] Bulyk, M., Johnson, P., Church, G., (2002). Nucleotides of transcription factor binding site exert independent effects on the binding affinities of transcription factors. *Nucleic Acids Res.,* 30 (5), 1255-1261.

[13] Barash, Y., Elidan, G., Friedman, N., Kaplan, T., (2003). Modeling dependencies in protein- DNA binding sites., *In Proceedings of the seventh annual international conference on Research in computational molecular biology Berlin, Germany: ACM, New York, NY*, 28-37.

[14] Ellrott, K., Yang, C., Sladek, F., Jiang, T., (2002). Identifying transcription factor binding sites through Markov chain optimization. *Bioinformatics*, 18, Suppl 2:S100- S109.

[15] King, O., & Roth, F., (2003). A non-parametric model for transcription factor binding sites. *Nucleic Acids Res.*, 31 (19), e116.

[16] Marinescu, V., Kohane, I., Riva, A., (2005). MAPPER: A search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics*, 6(79), 79.

[17] Zhao, X., Huang, H., Speed, T., (2005). Finding short DNA motifs using permuted Markov models. *J. Comput. Biol.*, 12, 894-906.

[18] Zhou, Q., Liu, J., (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics,* 20 (6), 909-916.

[19] Zare-Mirakabad, F., Ahrabian, H., Sadeghi, M., Nowzari-Dalini A., Goliaei, B., (2009). New Scoring Schema for Finding Motifs in DNA Sequences. *BMC Bioinformatics*, 10, 93.

[20] Sandelin, A., Alkema, W., Engstrom, G. P., Wasserman, W. W., Lenhard, B., (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32, 91-94.

[21] Wingender, E., Dietze, P., Karas, H., Knuppel, R., (1996). TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, 24 (1), 238-241.

[22] Sandve, G., Abul, O., Walseng, V., Drablos, F., (2007). Improved benchmarks for computational motif discovery. *BMC Bioinformatics,* 8, 193.

[23] Hertzberg, L., Zuk, O., Getz, G., Domany, E., (2005). Finding Motifs in Promoter Regions. *J. Comput. Biol.*, 12 (3), 314-330.

[24] Cover, T. M., Thomas, J. A., (2005). Elements of Information Theory.

[25] Tompa, M., Li, N., Bailey, T., Church, G., De Moor, B., Eskin, E., Favorov, A., Frith, M., Fu, Y., Kent, W., Makeev, V., Mironov, A., Noble, W., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, W., Weng, Z., Workman, C., Ye, C., Zhu, Z., (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23, 137-144.

[26] Burset, M., Guigo, R., (1996). Evaluation of gene structure prediction programs. *Genomics*, 34 (3), 353-367.

[27] Lenhard, B., & Wasserman, W., (2002). TFBS: Computational framework for transcription factor binding site analysis. *Bioinformatics*, 18 (8), 1135-1136.