Techincal Report, LTU-22-06-2013

pp. **1-7**

B-MISCORE: A NEW SIMILARITY METRIC FOR SELF-ORGANIZATION OF DNA k-MERS

Dianhui Wang

Department of Computer Science and Computer Engineering
La Trobe University, Melbourne, Vic 3086, Australia

Email: dh.wang@latrobe.edu.au

Abstract. *This technical report presents a new similarity metric to characterize the closeness between a k-mer and a background, which can be a whole genomic DNA sequence or a collection of non-coding regions of relevant species. Such a similarity metric plays an important role in modeling DNA regulatory motifs, and also it will provide us with a sensible and interpretable measure for data assignments in self-organization learning schemes. This original contribution can be regarded as a counterpoint of our proposed MISCORE concept. Some primary results on signal separability and monotonic property of mixed models are reported.*

Keywords: Background modeling, DNA motifs, self-organization learning, similarty.

1. **Introduction.** The SOM-based clustering techniques in DNA motif discovery mostly use traditional similarity metrics to form the motif (signal) and noise clusters of k-mers in the datasets, which often fails to offer an explicable interpretation to the noise-dominated clusters. Specifically, the same similarity metric is used to generate opposite clusters of k-mers, despite their well-known distinctive characteristics and opposite statistical properties in the datasets. Usually, the employed similarity metric has a special interpretation and design for functional discrimination of motif instances from the random (noise) k-mers through the use of the embedded motif properties in the clusters. This causes an inconsistency (and an inexplicability) to occur, when these similarity metrics are used for k-mer assignment to the clusters with no considerable presence of motif characteristics. Consequently, this allows a reasonable interpretation only for the clusters with significant degree of motif properties, and practically limits a consistent interpretation for the noise-dominated clusters that occupy the largest portion of the maps.

In [1], we presented MISCORE as a new similarity metric for effective discrimination of functional or putative motif models through the characterization of the functional motif properties. Recently, MISCORE has been employed in a fuzzy SOM-based motif discovery algorithm and its value can be confirmed by the promising results [2]. In this report, our previously reported MISCORE is extended to a so-called background MISCORE (B-MISCORE), which is a new metric to compute the similarity between a k-mer and the background. The objective of this development is to obtain a regularized similarity measure for an explicable self-organization of k-mers in the clustering process, where the MISCORE and B-MISCORE will be combined through an adaptive weight in a regularized similarity metric. This new metric will assign weights between MISCORE and B-MISCORE in accordance with the discrete mixture of signal and noise in each cluster, which will offer an improved explicability of the clustering process, since the practical fuzziness in terms of mixture between signals and embedded noise in each cluster will have an appropriate treatment.

The remainder of this report is organized as follows. Section 2 provides relevant preliminaries including a brief introduction of MISCORE. Section 3 describes the proposed B-MISCORE. Section 4 then presents some observations on the B-MISCORE using several DNA datasets, and also discusses the functions of both MISCORE and B-MISCORE in motif-finding. Section 5 concludes this report with some remarks.

2. **Preliminaries.**

2.1. **Model representation.** In this report, Positional Frequency Matrix (PFM) is employed as the motif model [3]. The PFM-based motif model, denoted by M, is a matrix, i.e., $M = [f(b_i, i)]_{4\times k}$, where $b_i \in \chi = \{A, C, G, T\}$ and $i = 1, \ldots, k$, and each entry $f(b_i, i)$ represents the probability of nucleotide b_i at position i. Similarly, a k-mer $K_s = q_1 q_2 \ldots q_k$ is encoded as a binary matrix $K = [k(b_i, i)]_{4\times k}$ with $k(q_i, i) = 1$ and $k(b_i, i) = 0$ for $b_i \neq q_i$. For example, a k-mer K_s=$AGCGTGT$ can be encoded as,

$$K = encode(K_s) = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}_{4\times k}.$$

For a given binary encoded set of k-mers, $S = \{K_1, K_2, ..., K_P\}$, the motif PFM model M_S can be computed by $M_S = \frac{1}{P}\sum_{i=1}^{P} K_i$.

2.2. **MISCORE.** MISCORE is a new scoring function for modeling motif signals that uses a combined characterization on the model conservation, the background rareness and the compositional complexity of functional motifs. It quantifies a similarity between a k-mer K and a putative model M with respect to the background reference model M_{ref}, that is,

$$r(K, M) = \frac{d(K, M)}{d(K, M_{ref}) + c(K)}, \tag{1}$$

where $d(K, M)$ is defined as a generalized Hamming distance, expressed as,

$$d(K, M) = 1 - \frac{1}{k}\sum_{i=1}^{k}\sum_{\forall b_i \in \chi} f(b_i, i)k(b_i, i), \tag{2}$$

where $f(b_i, i)$ and $k(b_i, i)$ are the observed frequencies of base b_i at position i in M and K, respectively.

Motivated by the well-known Gini index to quantify impurity of data clusters, we define $c(K)$ in (1) to compute the compositional complexity of K as follows:

$$c(K) = \frac{4}{3}\left[1 - \frac{1}{k^2}\sum_{\forall b_i \in \chi}\left(\sum_{i=1}^{k} k(b_i, i)\right)^2\right], \tag{3}$$

where the complexity is scored according to the distribution of bases (A, C, G, T) in the K. An equal distribution gives the maximum score of 1 and a dominant distribution, i.e., a nucleotide appears at all positions of the K, gives the minimum complexity of 0.

The complexity measure given in (1) helps in automatically eliminating the low-complex motifs from the top rank. In this way, an empirical threshold-based filtering [4] for filtering the low-complex candidate motifs can be avoided.

Binding sites are evolutionarily constrained with limited mutations, hence a K can be a putative motif instance if $d(K, M) < d(K, M_{ref})$ holds, which implies a smaller mismatch to a true motif model M than the background reference model M_{ref}. Note that the

M_{ref} is a PFM that can be constructed by all k-mers from the background sequences. For a large sized background, each column of the M_{ref} approximates the nucleotides background frequency. Thus, the M_{ref} can be conveniently composed of the nucleotides pre-computable background frequency in each column. Large sequence-portions that have a minimal chance of having the true binding sites can be taken as the backgrounds, e.g., random chunks of large genomic portions or a large collection of upstream regions from the relevant species. Note that a smaller $r(K, M)$ score characterizes a higher similarity of that K to M in respect to its dissimilarity to M_{ref} and a better nucleotide complexity in K, which implies a combined characterization on the model conservation, the background rareness and the compositional complexity.

The mathematical expectation of the MISCORE values of a set of k-mers can be viewed as a metric to characterize the candidate motifs. Given a set of k-mers S and its PFM model M_S, a MISCORE-based Motif Score (MMS), denoted as $R(S)$, can be evaluated by,

$$R(S) = \frac{1}{|S|} \sum_{\forall K \in S} r(K, M_S), \tag{4}$$

where $|*|$ is the set cardinality and $r(*,*)$ is the MISCORE given in (1).

3. **B-MISCORE.** This new similarity metric uses a large random sampling of the backgrounds to offer a more effective way to model the background, where only the first order of Markov chain (MC) model is used. Concretely, a large collection of random sets are generated from the background sequences to represent the random sampling of the background, where each random set is composed of a group of randomly selected k-mers from the background. It aims to offer an effective and computationally efficient background representation that can be used to evaluate a group of k-mers in computational search and evaluation of DNA motifs.

Firstly, a large collection of random sets, denoted as $\zeta = \{G_1, G_2, G_3 \ldots, \}$, $|\zeta| \geq 1000$, are generated where each random set G_l has a set of randomly grouped k-mers from the background, i.e., $G_l = \{K_1, K_2, K_3, \ldots\}$, $25 \leq |G_l| \leq 50$. Then, the background probability with denotation $P(K|M_B)$ of every $K \in G_l$ is computed using the first order MC transition matrix $\beta = [\pi(a, a')]_{4\times 4}$ built from the background, where $\forall a, a' \in \chi$, as,

$$P(K|M_B) = p(b_1) \prod_{\forall (a,a')} \pi(a, a')^{k(a,a')}, \tag{5}$$

where $k(a, a')$ gives the *count* of di-nucleotide aa' in K, and $p(b_1)$ is the independent background probability of the nucleotide appearing at the first position in K.

Let $P_n(K|M_B)$ denote a normalized background probability of the k-mer K, and defined as

$$P_n(K|M_B) = \frac{P(K|M_B) - \min\limits_{\forall K \in \Gamma}\{P(K|M_B)\}}{\max\limits_{\forall K \in \Gamma}\{P(K|M_B)\} - \min\limits_{\forall K \in \Gamma}\{P(K|M_B)\}}, \tag{6}$$

where Γ represents the k-mer dataset produced from the ζ.

Then, a local background score of the k-mer K (i.e., the similarity of the k-mer to the background) can be measured by using a random k-mers set from the background, namely G_l, that is,

$$d_B(K, G_l) = \frac{1}{|G_l|} \sum_{\forall K_p \in G_l} P_n(K_p|M_B) d(K, K_p), \tag{7}$$

where $|*|$ is the set cardinality, and $d(\cdot,\cdot)$ is the Hamming distance (in percentage) between two k-mers, defined as,

$$d(K_1, K_2) = 1 - \frac{1}{k}\sum_{i=1}^{k}\sum_{b_i \in \chi} f_1(b_i, i) f_2(b_i, i), \tag{8}$$

The $d_B(K, G_l)$ can be regarded as a weighted measure of K of being a background class element in respect to G_l, where $d(K, K_p)$ acts as the weight to the contribution of each K_p in G_l in evaluating the similarity of K to the background. A single random set G_l gives an insufficient sampling of the background. Hence, a large collection of random sets ζ is used to reliably obtain a discriminative background score, denoted as $b_f(K)$, that is,

$$r_b(K) = \min_{\forall G_l \in \zeta} \{d_B(K, G_l)\}. \tag{9}$$

Given a set of k-mers, namely S, a B-MISCORE-based Model Score (BMMS) can be evaluated by

$$R_b(S) = \frac{1}{|S|}\sum_{\forall K \in S} r_b(K), \tag{10}$$

where a *larger* $R_b(*)$ score represents a higher potential of the models to be functional, and vice versa.

4. **Primary Results.** In this section, we investigate the separability performance of B-MISCORE using several datasets. Firstly, we observe how well B-MISCORE can separate true motif models from a large collection of random ones. Secondly, we observe how the separability performance holds for true models and degenerated models which were artificially created by mixing true binding sites and random noises.

4.1. **Separability:** In this section, we investigate the separability performance of B-MISCORE using four real datasets, namely MEF2, SRF, CREB and E2F, where each dataset contains a known motif of the respective Transcription Factors (TFs). These datasets were previously used in [5]. The objective is to observe how well B-MISCORE can separate the true motif models from the random ones in terms of discriminative score-gaps. In our simulations, the true model (M_t) of each TF was generated by aligning the known binding sites using the ClustalW [6] tool. Then, 1000 random models, denoted as $M_{r_q}, q = 1, 2, \ldots, 1000$, were generated by randomly collecting k-mers the sequences, where each random model contained the same number of k-mers as the true model. Then, the BMMS score of the models are generated using (10) and presented in Figure 1. The projection of the scores show that B-MISCORE has useful ability of separating the true models from the random one in respect with the background.

4.2. **Monotonic Property:** In this section, we observe how the separability performance of B-MISCORE holds for the true models and the degenerated models using 10 real DNA datasets, collected from [5, 7]. The objective is to observe how well B-MISCORE can separate the true motif models (M_t) from the degenerated models, denoted as M'_t, in terms of the BMMS score. The M'_t models were generated by inserting specific amounts of noise in the true models. 5000 of such M'_t models were generated for each degree of noise insertion. Then, the BMMS score of the M_t and M'_t models were computed using (10), and expectation $E\{*\}$ and $std\{*\}$ of the M'_t models were presented in Table 1. The results show that the expectation of BMMS score demonstrates a monotonic decrease alone with the increase of noise level (from 10% to 50%) in the models. This clearly indicates the rationality and applicability of the B-MISCORE metric in finding putative motifs through self-organizing learning schemes.

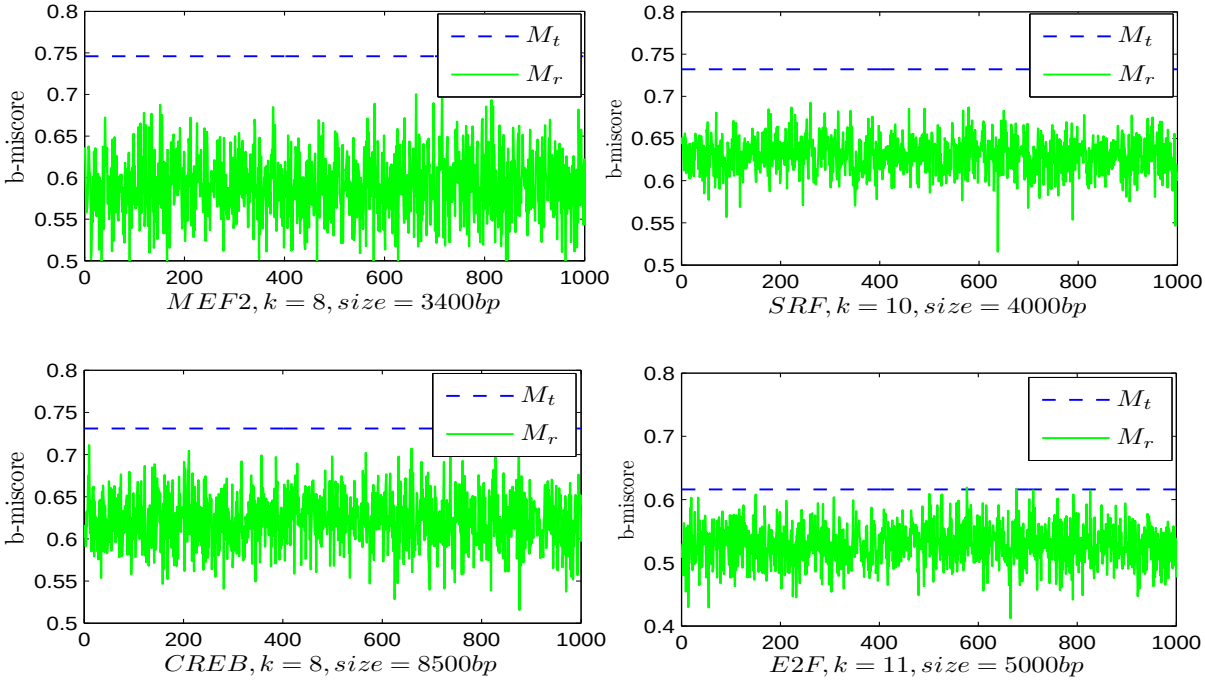

FIGURE 1. Separability demonstration of B-MISCORE

TABLE 1. $R_b(*)$ score comparison between M_t and noisy models (M'_t)

		$E\{R_b(M'_t)\} \pm std\{R_b(M'_t)\}$ scores on noisy models.				
TF	$R_b(M_t)$	10% noise	20% noise	25% noise	30% noise	50% noise
CREB	0.822	0.804±0.016	0.785±0.025	0.773±0.030	0.764 ±0.034	0.721 ±0.046
SRF	0.869	0.854± 0.011	0.839± 0.019	0.829± 0.023	0.822± 0.027	0.783± 0.040
MEF2	0.841	0.827± 0.015	0.813± 0.021	0.804± 0.025	0.796± 0.029	0.758± 0.040
MYOD	0.733	0.722± 0.012	0.710± 0.019	0.702± 0.022	0.697± 0.024	0.666± 0.036
ERE	0.796	0.783± 0.011	0.770± 0.018	0.762± 0.021	0.755± 0.024	0.721± 0.037
E2F	0.759	0.745± 0.012	0.732± 0.019	0.721± 0.023	0.715± 0.026	0.677± 0.037
CREB*	0.797	0.783± 0.013	0.767± 0.020	0.759± 0.023	0.751± 0.028	0.716± 0.040
SRF*	0.412	0.405± 0.008	0.399± 0.012	0.395± 0.014	0.391± 0.017	0.373± 0.024
MEF2*	0.464	0.455± 0.011	0.447± 0.014	0.440± 0.018	0.435± 0.021	0.410± 0.030
MYOD*	0.325	0.319± 0.007	0.314± 0.011	0.311± 0.013	0.308± 0.015	0.293± 0.020

Note: datasets with * are composed of promoters of 500bp length and the others have 200bp in length.

In order to obtain a statical evaluation, the z-score, also known as the standard score, can be computed on the BMMS score of the M_t and M'_t models, as:

$$Z_B(M_t, M'_t) = \frac{R_b(M_t) - E\{R_b(M'_t)\}}{std\{R_b(M'_t)\}}, \tag{11}$$

where $R_b(*)$ is the BMMS score given in (10). The higher the $Z_B(M_t, M'_t)$ score on the large collection of M'_t models can be obtained, the better statistical significance of the separability performance of the B-MISCORE would be. The $Z_B(M_t, M'_t)$ score observation on the separability performance of B-MISCORE is presented in Table 2.

TABLE 2. Z-score computation between M_t and noisy models (M'_t)

$Z_B(M_t, M'_t)$ score evaluation for different amount of noise.					
TF	10% noise	20% noise	25% noise	30% noise	50% noise
CREB	1.087	1.445	1.636	1.708	2.207
SRF	1.349	1.542	1.691	1.747	2.152
MEF2	0.977	1.306	1.481	1.546	2.062
MYOD	0.949	1.223	1.405	1.473	1.837
ERE	1.162	1.450	1.600	1.668	2.028
E2F	1.140	1.462	1.663	1.732	2.193
CREB*	1.091	1.451	1.643	1.644	2.038
SRF*	0.867	1.093	1.222	1.248	1.633
MEF2*	0.894	1.200	1.350	1.400	1.832
MYOD*	0.801	0.984	1.096	1.095	1.544

Note: datasets with asterisk are composed of promoters with 500*bp*, while the others have 200*bp* in length.

4.3. **Discussion:** From the results reported in [1], we can see that a smaller MMS score indicates a better potential for a candidate motif to be functional. However, the MISCORE itself for each individual k-mer seems meaningless as the model M used in (1) is not a true motif model. Fortunately, for a given true motif model, such a MISCORE becomes a powerful feature of k-mers to characterize transcription factor binding sites.

A dual statement can be made using our proposed BMMS, i.e., a smaller BMMS score indicates a less potential for a candidate motif to be functional. Based on this understanding, we can have a brand new characterization on motifs using both MMS and BMMS. For instance, the following two scoring functions, $\frac{MMS}{BMMS}$ or $MMS + \frac{1}{BMMS}$, may result in better performances on signal separability and/or recognizability. Similarly, the B-MISCORE value itself for each individual k-mer seems not to be so meaningful. However, it is believed that such a score can be used as a numerical feature of k-mers in rule-based or learning-based prediction of transcription factor binding sites.

Finally, it should be pointed out that our main motivation behind this B-MISCORE concept is to get rid of the constraint on the model order uncertainty and computational burden using Markov chain method for background modeling.

5. **Conclusions.** While characterizing DNA motifs by our earlier proposed MISCORE-based model score, background modeling can be used as additional information to improve the discriminative power in motif-finding algorithms. From our observations in this report, we have seen that B-MISCORE-based model score can distinguish remarkably random models from true motifs. Statistical figures in Table 1 and 2 show a monotonic property of the noisy models alone with the amount of true binding sites contained in the models. This fact indeed implies rationality and usefulness of such a similarity metric for k-mer assignment in self-organizing learnings.

Recently, the B-MISCORE concept has been used in a fuzzy SOM-based motif discovery scheme [8]. It seems that a MISCORE-BMISCORE-based regularizing similarity metric works well, and most importantly, it makes the learning process understandable and the learning results interpretable.

Obviously, our proposed B-MISCORE is dependent of the random models collected from the background. Also, it is necessary to look into the robustness of the B-MISCORE with respect to the number of random models used in G_l. Further results on this aspect will be reported in our relevant publications.

6. **Acknowledgment.** I am grateful to Sarwar Tapan, one of my PhD students at La Trobe University, for his contributions to this technical report. Through personal discussions on campus and his extensive simulation studies, the concept and usefulness of the proposed B-MISCORE have been reasonably evaluated.

REFERENCES

[1] Wang DH, Sarwar T: **MISCORE: a new scoring function for characterizing DNA regulatory motifs in promoter sequences.** *BMC Systems Biology* 2012, **6**: S4.

[2] Wang DH, Sarwar T: **A robust elicitation algorithm for discovering DNA motifs using fuzzy self-organizing maps.** *IEEE Transactions on Neural Networks and Learning Systems* 2013, (conditional acceptance) .

[3] Stormo GD, Fields DS: **Specificity, free energy and information content in protein-DNA interactions**. *Trends in Biochemical Sciences* 1998, **23**(3):109–113.

[4] Mahony S, Hendrix D, Golden A, Smith TJ, Rokhsar DS: **Transcription factor binding site identification using the self-organizing map**. *Bioinformatics* 2005, **21**(9):1807–1814.

[5] Wei Z, Jensen ST: **GAME: detecting cis-regulatory elements using a genetic algorithm**. *Bioinformatics* 2006, **22**(13):1577–1584.

[6] Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Research* 1994, **22**(22):4673–4680.

[7] Blanco E, Farre D, Alba MM, Messeguer X, Guigo R: **ABS: a database of Annotated regulatory Binding Sites from orthologous promoters**. *Nucleic Acids Research* 2006, **34**(suppl1):D63–67.

[8] Sarwar T, Wang DH: **A further study on fuzzy SOM-based motif discovery.** *IEEE Transactions on Neural Networks and Learning Systems* 2013, (to be submitted).