A Novel Information Contents Based Similarity Metric for Comparing TFBS Motifs

Shaoqiang Zhang, Lifen Jiang
College of Computer and Information Engineering
Tianjin Normal University
Tianjin 300387, China
sqzhang@163.com

Chuanbin Du, Zhengchang Su
Department of Bioinformatics and Genomics
the University of North Carolina at Charlotte
Charlotte, NC 28223, USA
zcsu@uncc.edu

Abstract—**Identifying binding sites recognized by transcription factors (TFs) is one of major challenges to decipher complex genetic regulatory networks encoded in a genome. A set of binding sites recognized by the same TF, called a motif, can be accurately represented by a position frequency matrix (PFM) or a position-specific scoring matrix (PSSM). Very often, we need to compare motifs when searching for similar motifs in a motif database for a query motif, or clustering motifs possibly recognized by the same TF. In this paper, we have designed a novel metric, called SPIC (Similarity between Positions with Information Contents), for quantifying the similarity between two motifs using their PFMs, PSSMs, and column information contents, and demonstrated that this metric outperforms the other state-of-the-art methods for clustering motifs of the same TF and differentiating motifs of different TFs.**

Keywords: transcription factor binding sites (TFBS); information contents; motifs, regulatory networks; similarity metric

I. INTRODUCTION

Deciphering complex genetic regulatory networks encoded in a genome is a highly challenging problem in the post-genomic era [1]. Identifying all *cis*-regulatory binding sites (BSs) recognized by the transcription factors (TFs) in a genome is the first step towards this goal [2]. A TF binds to a 5-25 bp (base pairs) DNA sequence called a BS, and changes the rate of transcription of a nearby target gene. The BSs of the same TF show some level of conservation but can be rather degenerate. A set of similar *cis*-regulatory BSs recognized by the same TF is called a *motif*, which can only be predicted by comparing multiple sequences that potentially contain the BSs. A motif-finding algorithm identify BSs based on the assumption that the BSs are usually more conserved than their flanking non-functional sequences [3]. A motif can be represented by a $4\times n$ position frequency matrix (PFM), which consists of nucleotide frequencies at each position of the motif, or a $4\times n$ position-specific scoring matrix (PSSM), whose elements are the log-odds ratio of nucleotides at each position of the motif over a background model [4]. The PFM of a motif is derived from the alignment of its BSs; it largely reflects the corresponding TF's binding preference at each position. Thus, given the PFM of a TF, we can predict some new BSs by scanning it against the potential sequences in a genome.

Very often after we obtain some new putative motifs, for example using motif-finding algorithms, we hope to either compare them with known TFs' motifs in a database to infer their cognate TFs, or cluster them to remove redundancies and form unique motifs [5, 6]. In another case, we intend to merge motifs of different TFs of a structurally related class to form a familial binding profile (FBP) in order to find motifs for a particular TF family [7]. In all these applications, we need a metric to measure the similarity between any two motifs. The majority of current methods for comparing two motifs typically contain two components: a similarity metric for column-to-column comparisons between the PFMs of the two motifs, and an algorithm to find the optimal column-to-column alignment between the two motifs based on the metric. The column similarity metrics used in current methods include Pearson's correlation coefficient (PCC) [8], average Kullback-Leibler (AKL), average log-likelihood ratio (ALLR) [9], *p*-value of Chi-square (pCS) [10], and sum of squared distances (SSD), etc. [7, 11]. Either the Smith-Waterman [12] or the Needleman-Wunsch algorithm [13] is usually employed to find the optimal alignment. Mahony *et al.* [14] have evaluated these column similarity metrics along with the two alignment algorithms and implement them as a web tool STAMP [15]. Additionally, Mosta proposed by Pape *et al.* [16], and KFV by Xu and Su [17], are two alignment-free motif comparison methods. Xu and Su showed that their KFV method outperforms the Mosta and the methods in the STAMP [17].

In this paper we designed a new column similarity metric called *Similarity between Positions with Information Contents* (SPIC). The metric is inspired by our early work used in the pipelines GLECLUBS and eGLECLUBS for genome-wide binding site prediction in prokaryotes [5, 6]. More specifically, to measure the similarity between the columns X and Y of two motifs M_1 and M_2, respectively, SPIC first uses the column X and its information contents (IC) of M_1's PSSM to match the column Y of M_2's PFM and then uses the column Y and its IC of M_2's PSSM to match the column X of M_1's PFM. In the following sections, we will describe the SPIC metric in details, and evaluate its performance using some datasets from STAMP [17] and GLECLUBS [5, 6].

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

II. Methods and Materials

A. Previous metrics

We compared our metric with the following ones for their ability to either retrieve motifs from a database or cluster relevant motifs. In the definitions of these metrics, X_b is the probability of base $b \in \{A,C,G,T\}$ in a column X of the position frequency matrix of a motif.

1) Pearson correlation coefficient (PCC).

The PCC was first introduced by Pietrokovski [8] for computing the similarity of two columns *X* and *Y* of two motifs, and is defined as,

$$PCC(X,Y) = \frac{\sum_{b\in\{A,C,G,T\}} (X_b - \overline{X})(Y_b - \overline{Y})}{\sqrt{\sum_{b\in\{A,C,G,T\}} (X_b - \overline{X})^2 \sum_{b\in\{A,C,G,T\}} (Y_b - \overline{Y})^2}} \quad (1),$$

where $\overline{X}$ and $\overline{Y}$ are the averages of X_b and Y_b, respectively.

2) Average Kullback-Leibler (AKL, or relative entropy).

For two columns X and Y,

$$AKL(X,Y) = 10 - \frac{\sum_{b\in\{A,C,G,T\}} X_b \log\frac{X_b}{Y_b} + \sum_{b\in\{A,C,G,T\}} Y_b \log\frac{Y_b}{X_b}}{2} \quad (2)$$

3) Average log-likelihood ratio (ALLR).

The ALLR formula was proposed by Wang and Stormo [9]. For two columns *X* and *Y*,

$$ALLR(X,Y) = \frac{\sum_b N_{X_b}\left(\frac{Y_b}{q_b}\right) + \sum_b N_{Y_b}\left(\frac{X_b}{q_b}\right)}{\sum_b (N_{X_b} + N_{Y_b})} \quad (3),$$

where N_{X_b} and N_{Y_b} are the counts of base $b \in \{A,C,G,T\}$ in column *X* and *Y*, respectively, and q_b is the background probability of $b \in \{A,C,G,T\}$.

4) $1-$p-value of Chi-square (pCS).

The pCS was proposed by Schones *et al.* [10]. For two columns X and Y, we calculate $1-$p-value of

$$\chi_3^2(X,Y) = \sum_{b\in\{A,C,G,T\}} \frac{(N_{X_b} - N_{X_b}^e)^2}{N_{X_b}^e} + \sum_{b\in\{A,C,G,T\}} \frac{(N_{Y_b} - N_{Y_b}^e)^2}{N_{Y_b}^e} \quad (4).$$

Note that $N_{X_b}^e = (N_X \cdot N_b)/N$, where N_X is the total number of counts in column *X*, N_b is the total number of counts for base *b* in the columns *X* and *Y*, and *N* is the total number of counts for all bases in the columns *X* and *Y*. $N_{Y_b}^e$ is similarly defined.

5) Sum of squared distances (SSD).

The SSD formula is a variant of Euclidean distance defined as

$$SSD(X,Y) = 2 - \left(\sum_{b\in\{A,C,G,T\}} (X_b - Y_b)^2 \right) \quad (5).$$

6) Asymptotic covariance (AC).

The AC formula was recently proposed by Pape *et al.* [16] based on the asymptotic covariance between the frequency matrixes of two motifs *A* containing a set of binding sites $\{a\}$ and *B*= $\{b\}$. Let the number of counts of site a in a background sequence of length m is $N_a(m)$ and the sum of $N_a(m)$ is $N_A(m) = \sum_{a\in A} N_a(m)$. Let A' and B' be the reverse complementary sequence sets of *A* and *B*, respectively, the similarity between motifs *A* and *B* is defined as:

$$AC(A,B) = \lim_{m\to\infty} m^{-1} \operatorname{cov}(N_A(m) + N_{A'}(m), N_B(m) + N_{B'}(m)) \quad (6).$$

We used the software package Mosta downloaded from http://mosta.molgen.mpg.de to calculate the AC scores.

7) KFV (k-mer frequency vector).

The KFV metric was more recently designed by Xu and Su [17]. Each PFM is first converted into a 4^k -dimensional composition vector called a KFV with each element representing the likelihood score for a particular short k-mer sequence fitting the PFM model, and then the similarity between two motifs is calculated by a distance measure between their corresponding KFVs.

The first five column metrics have been surveyed by Gupta *et al.* [18] and Mahony *et al.* [14]. The last two alignment-free methods have been compared by Xu and Su [17].

B. Our Metric SPIC

For a motif M_x containing n_x sequences with length L_x, let $F_x = (f_x(b,X))_{4\times L_x}$ be its position frequency matrix (PFM) and P_x be its position-specific scoring matrix (PSSM) defined as

$$P_x = (P_x(b,X))_{4\times L_x} = \left(\log\frac{p_x(b,X)}{q_x(b)} \right)_{4\times L_x} \quad (7),$$

where $f_x(b,Y)$ and $p_x(b,Y)$ are the count and probability of base $b \in \{A,C,G,T\}$ appearing at position Y of M_x (i.e., column *Y* of P_x), respectively, and $q_x(b)$ is the probability of base *b* appearing in the background sequences. A pseudo-count is added when computing these probabilities. The information content (IC) of column X of the PSSM P_x is defined as

$$I(X,P_x) = \sum_{b\in\{A,C,G,T\}} p_x(b,X) P_x(b,X) = \sum_{b\in\{A,C,G,T\}} p_x(b,X) \log\frac{p_x(b,X)}{q_x(b)} \quad (8).$$

For two motifs M_1 and M_2 with PSSMs P_1 and P_2, and PFMs F_1 and F_2, respectively, the similarity score between position *X* of M_1 and position *Y* of M_2 is defined as

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

$$\mathrm{Sim}\big(M_1(X), M_2(Y)\big) =$$

$$\min\left\{1, \frac{\max\{\mathrm{score}\big(P_1(X), F_2(Y)\big), \mathrm{score}\big(P_2(Y), F_1(X)\big)\}}{\max\{\mathrm{score}\big(P_1(X), F_1(X)\big), \mathrm{score}\big(P_2(Y), F_2(Y)\big)\}}\right\} \quad (9),$$

where $\mathrm{score}(P_x(A), F_y(B))$

$$= I(A, P_x)\sum_b \left(f_y(b,B) \cdot \log \frac{p_x(b,A)}{q_x(b)} \right),$$

$$x, y \in \{1,2\}, A, B \in \{X, Y\}. \qquad (10)$$

In the function (10), we use the information content of each column to attenuate the influence of the low information parts, and to enhance the effect of the high information parts of the PSSM on the similarity score. Notably, the function (10) reflects the likelihood for $P_x(A)$ to generate $F_y(B)$.

C. Pairwise motif alignment and calculation of empirical p-values

We use Needleman-Wunsch (NW) global alignment [13] and Smith-Waterman (SW) local alignment [12] to evaluate the metrics. Both methods allow for affine gap penalties. For this study, the gap-extension penalty is set to be half the value of the gap-opening penalty. We also implemented an ungapped, extended Smith-Waterman alignment method as described by Mahony *et al.* [14]. In order to determine the likelihood of any score given the lengths of aligned matrices, Sandelin and Wasserman [7] assigned empirical *p*-values to the alignment scores. Following Mahony *et al.* [14], we also used the method of Sandelin and Wasserman for the calculation of empirical *p*-values based on simulated PSSMs [7].

D. Datasets of motifs

1) Dataset-1

Dataset-1 was originally created by Mahony *et al.* [14] from JASPAR, containing PFMs and PSSMs of 96 motifs with known TF structural classes, 25 of them belong to the Zinc-Finger(ZF) families. The dataset was also used for evaluating the KFV metric by Xu and Su [17].

2) Dataset-2

Dataset-2 containing about 10^5 putative motifs were predicted in a total of 2,313 orthologous inter-operonic sequence sets from 55 closely-related γ -proteobacterial genomes including *E. coli* K12 using phylogenetic foot-printing during the development of our GLECLUBS pipeline for genome-wide prediction of TF binding sites in prokaryotic genomes [5]. These 105 putative motifs contain 1,411 known binding sites belonging to 122 TFs (true motifs) in *E. coli* K12 according to the RegulonDB v6.0 database [19]. The dataset is available at: http://motifclick.uncc.edu.

E. Performance evaluation by accuracy and ROC analysis

We used the ROC (Receiver Operating Characteristic) analysis to compare these methods for their ability to identify the TFBS motifs of structural and/or evolutionarily related TFs in the Dataset-1. The performance "accuracy" is measured as the percent of motifs whose structural class are correctly recovered via the best hit in database searches. The ROC curves were plotted based on the following criteria. Given a dataset containing n motifs with known TF structural classes, n(n+1)/2 pair-wise comparisons (including self-comparisons) were conducted and pair-wise similarity scores were computed using our algorithm or the other compared methods. We consider a pair of motifs as a match (positive) if "1-similarity score" between the two motifs within a threshold, or a mismatch (negative), otherwise. We consider a positive as a true positive if the two associated TFs come from the same structural class, and a negative as a true negative if the associated two TFs are from different structural classes. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR), computed for different thresholds of pair-wise scores.

III. Results

A. Performance evaluation for motif retrieval

One of the major uses of motif similarity metric is in motif database search, where a query motif is compared to all motifs in the database by an alignment method and metric, and the motifs in database that are similar enough to the query motif are returned as the hits. However, it is well-known that structurally and/or evolutionarily related TFs tend to bind similar motifs, and because of the highly degenerate nature of binding sites, it is often not easy to precisely differentiate these similar motifs in database searches. We compared SPIC with the PCC, SSD and KFV metrics for their abilities to find the same structural class in the database as that of the query motif using Dataset-1. We choose these three exiting metrics for the comparison, because it has been shown by Mahony *et al.* [14] that PCC and SSD have the best overall performance among the five metric evaluated when combined with an appropriate alignment algorithm, and that KFV even outperforms the prior metrics, both alignment-based or alignment-free ones [17]. As in Mahony *et al.* [14], we computed the accuracy of a metric as the percentage of query motifs whose structural classes are correctly recovered by the metric as the best hit.

For the PCC and SSD metrics, we combine them with the best alignment algorithm and parameters according to Mahony *et al.* [14], i.e., for PCC, we used the ungapped Smith-Waterman algorithm (PCC/SWU), and for SSD, we adopted the gapped Smith-Waterman algorithm with gap open =1 and gap extension=0.5 (SSD/SW). For the alignment free metric KFV, we choose the *k*-mer length *k*=4 and cosine angle for vector comparison for the best overall performance according to the authors [17]. We tested our metric using both SW and NW alignment algorithms with different gap open and extension penalties. As shown in Table 1, our metric combined with the SW alignment algorithm with open = 1 achieves the best results, and it outperforms PCC/SWU and SSD/SW implemented in STAMP and KFV with their best parameter settings on Dataset-1.

TABLE I. COMPARISON OF SPIC WITH DIFFERENT PARAMETERS WITH THE EXISTING METHODS FOR MOTIF RETRIEVAL USING DATASET-1.

	Accuracy		
Method	***Non-ZF PFMs(71)***	***ZF PFMs(25)***	***Total(96)***
SPIC(gap open=1.00, SW)	**0.921**	**0.620**	**0.841**
SPIC(gap open=0.75, SW)	0.918	0.613	0.837
SPIC(gap open=0.50, SW)	0.916	0.614	0.837
SPIC(gap open=1.50, SW)	0.916	0.605	0.835
SPIC(gap open=0.25, SW)	0.915	0.606	0.835
SPIC(ungapped, SW)	0.916	0.610	0.836
SPIC(gap open=1.0, NW)	0.792	0.584	0.721
SPIC(gap open=1000, NW)	0.801	0.592	0.730
KFV(k=4, cosine)	0.915	0.600	0.833
STAMP(PCC/SWU)	0.887	0.600	0.813
STAMP(SSD/SW)	0.859	0.560	0.781

The results are shown separately for the zinc-finger (ZF) and non-ZF families. The values in bold indicate the highest accuracy achieved for each category. The results of STAMP (PCC/SWU and SSD/SW) are taken from [14],. The results of KFV are taken from [17].

To further compare our algorithm with STAMP and KFV for retrieving motifs in a database, we conducted Receiver Operating Characteristic (ROC) analysis of the performance of the three algorithms on Dataset-1 using their respective best parameter settings. We chose PCC/SWU in STAMP for the comparison as it outperforms SSD/SW on the dataset (Table 1). As shown in Figure 1, SPIC largely outperforms both the STAMP(PCC/SWU) and KFV algorithms for motif retrieval.

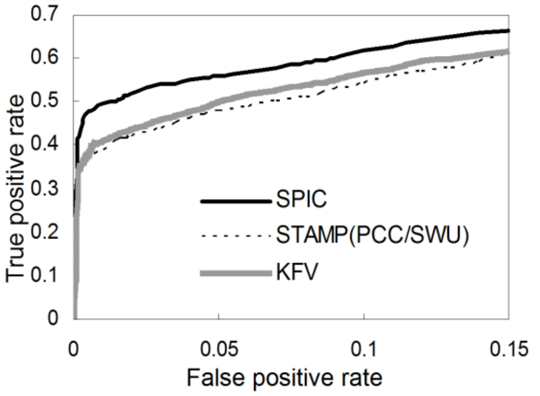

Figure 1. Evaluation of the three motif comparison algorithms using ROC analysis on Dataset-1.

B. Separation of true motifs from spurious ones

In genome-scale TF binding site motifs prediction applications, redundant and sub-motifs of the same TFs are often returned by motif finding programs, and they are required to be clustered to form unique motifs [9]. In order to facilitate the separation of true motifs from spurious ones in dataset-2, we need a motif similarity metric that not only accurately measures the similarity between each pair of true motifs, but also can be efficiently computed. Specifically, we need a motif similarity metric that gives a high score for two relevant motifs, i.e., two sub-motifs of the motif of a TF, but a low score for two irrelevant motifs, i.e., two motifs for evolutionarily unrelated TFs or two spurious motifs. To this end, we compare our metric SPIC with seven existing metrics for their capability of differentiating between relevant motifs and irrelevant ones.

In phylogenetic foot-printing based genome-scale TFBS prediction algorithms such as PhyloNet [7, 11] and GLECLUBS [5, 6], redundant and sub-motifs of the same TFs need to be clustered to form unique motifs, and at the same time spuriously identified motifs need to be removed [5, 6, 9]. To achieve such a goal, a motif similarity metric is needed that not only accurately measures the similarity between each pair of a large number of predicted-motifs by a phylogenetic foot-printing procedure, but also can be efficiently computed. Specifically, we need a motif similarity metric that gives a high score for two relevant motifs, i.e., two sub-motifs of the motif of a TF, but a low score for two irrelevant motifs, i.e., two motifs for evolutionarily unrelated TFs or two spurious motifs (as the chance for two spurious motifs to be very similar to each other is usually very low). To evaluate our metric SPIC in such applications, we compare it with seven existing metrics for their capability of differentiating between relevant motifs and irrelevant ones using Dataset-2.

To this end, we first generated a series of sub-motifs for each of the known 122 TF motifs *E. coli* K12 as follows. For each motif containing n known binding sites (we only consider the motifs that have at least 3 known binding sites), we randomly selected ($n-k+1$) sub-sets (sub-motifs) of size k with replacement from the n binding sites, $k=1,\ldots,n$. Therefore, there are $n(n+1)/2$ sub-motifs for each known motif. We then used each of the metrics with their best alignment methods and parameters for motif clustering and/or retrieval (Table 1) [14, 17] to compute pair-wise similarity scores among the sub-motifs of the same motif as well as the pair-wise similarity scores of the $\sim 10^5$ putative motifs in Dataset-2. Figure 2 shows the distribution of the normalized pair-wise motif similarity scores among the motifs in Dataset-2 (labeled by "all pairs") and that of the normalized scores among the sub-motifs of a known motif in RegulonDB (labeled by "known inner"), computed by each of these metrics. Since the majority of the motifs in Dataset-2 are irrelevant to one another, a good metric should well-separate the bulk of the distribution of the similarity scores among the motifs and that of the similarity scores among the sub-motifs of a known motif. As shown in Figure 2, of all the metrics examined, our metric resulted in the smallest overlap between the distribution of the similarity scores among all motifs in Dataset-2 and that of the similarity scores among the sub-motifs of each known motif, suggesting that our metric outperforms these existing metrics in separating the relevant motifs from irrelevant ones.

IV. CONCLUSION

In this paper, we proposed a new column similarity metric SPIC, when combined with the Smith-Waterman alignment algorithm, it outperforms the existing state-of-the-art metrics in both retrieving motifs in database search and clustering motifs, or separating true motifs from spurious ones. Particularly, the SPIC metric can be used in some phylogenetic footprinting based genome-wide TF binding sites prediction algorithms, such as PhyloNet [7, 11] and GLECLUBS [5, 6].

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

ACKNOWLEDGMENTS

This work was partially supported by a grant (61103073, SZ) from National Science Foundation of China, a grant (11JCYBJC26600, SZ) from Natural Science Funds of Tianjin, a grant from Doctoral Funds of Tianjin Normal University (52X09013, LJ), and two grants (EF0849615 and CCF1048261, ZS) from National Science Foundation of USA.

REFERENCES

[1] H. D. Kim, T. Shay, E. K. O'Shea, and A. Regev, "Transcriptional regulatory circuits: predicting numbers from alphabets," Science, vol. 325, pp. 429-32, Jul 24 2009.

[2] J. L. Reed, I. Famili, I. Thiele, and B. O. Palsson, "Towards multidimensional genome annotation," Nat Rev Genet, vol. 7, pp. 130-41, Feb 2006.

[3] D. GuhaThakurta, "Computational identification of transcriptional regulatory elements in DNA sequence," Nucleic Acids Res, vol. 34, pp. 3585-98, 2006.

[4] V. R. Akmaev, S. T. Kelley, and G. D. Stormo, "Phylogenetically enhanced statistical tools for RNA structure prediction," Bioinformatics, vol. 16, pp. 501-512, 2000/06// 2000.

[5] S. Zhang, M. Xu, S. Li, and Z. Su, "Genome-wide de novo prediction of cis-regulatory binding sites in prokaryotes," Nucleic Acids Res, vol. 37, p. e72, Jun 2009.

[6] S. Zhang, S. Li, P. T. Pham, and Z. Su, "Simultaneous prediction of transcription factor binding sites in a group of prokaryotic genomes," BMC Bioinformatics, vol. 11, p. 397, 2010.

[7] A. Sandelin and W. W. Wasserman, "Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics," J Mol Biol, vol. 338, pp. 207-15, Apr 23 2004.

[8] S. Pietrokovski, "Searching databases of conserved sequence regions by aligning protein multiple-alignments," Nucleic Acids Res, vol. 24, pp. 3836-45, Oct 1 1996.

[9] T. Wang and G. D. Stormo, "Combining phylogenetic data with co-regulated genes to identify regulatory motifs," Bioinformatics, vol. 19, pp. 2369-80, Dec 12 2003.

[10] D. E. Schones, P. Sumazin, and M. Q. Zhang, "Similarity of position frequency matrices for transcription factor binding sites," Bioinformatics, vol. 21, pp. 307-13, Feb 1 2005.

[11] T. Wang and G. D. Stormo, "Identifying the conserved network of cis-regulatory sites of a eukaryotic genome," Proc Natl Acad Sci U S A, vol. 102, pp. 17400-5, Nov 29 2005.

[12] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," J Mol Biol, vol. 147, pp. 195-7, Mar 25 1981.

[13] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," J Mol Biol, vol. 48, pp. 443-53, Mar 1970.

[14] S. Mahony, P. E. Auron, and P. V. Benos, "DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies," PLoS Comput Biol, vol. 3, p. e61, Mar 30 2007.

[15] S. Mahony and P. V. Benos, "STAMP: a web tool for exploring DNA-binding motif similarities," Nucleic Acids Res, vol. 35, pp. W253-8, Jul 2007.

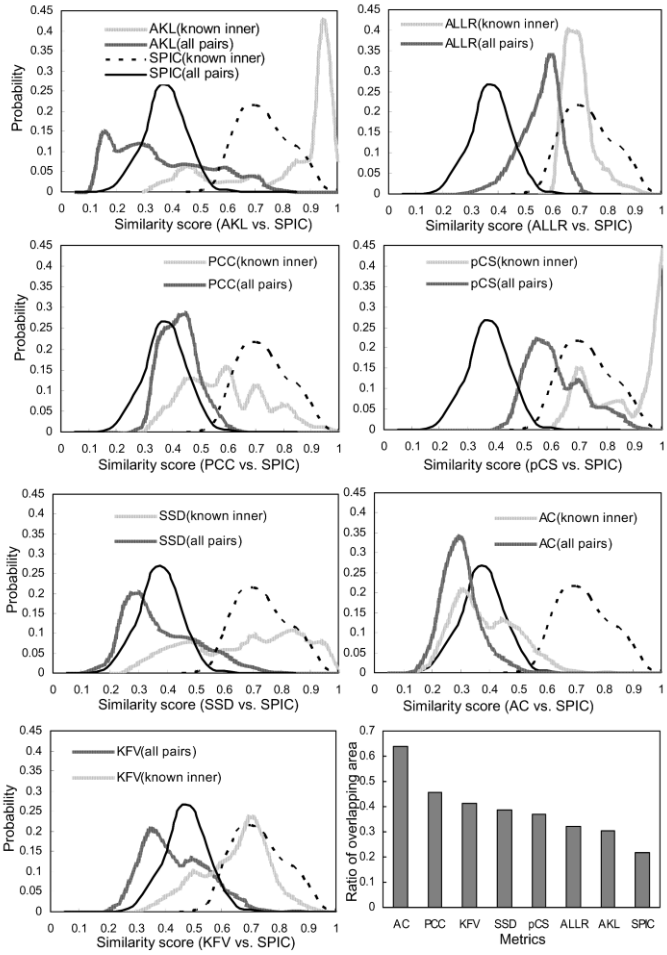

Figure 2. Comparison of the SPIC metric with seven existing methods for separation of true motifs from spurious ones on DataSet-2.

[16] U. J. Pape, S. Rahmann, and M. Vingron, "Natural similarity measures between position frequency matrices with an application to clustering," Bioinformatics, vol. 24, pp. 350-7, Feb 1 2008.

[17] M. Xu and Z. Su, "A novel alignment-free method for comparing transcription factor binding site motifs," PLoS One, vol. 5, p. e8797, 2010.

[18] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble, "Quantifying similarity between motifs," Genome Biol, vol. 8, p. R24, 2007.

[19] S. Gama-Castro, V. Jimenez-Jacinto, M. Peralta-Gil, A. Santos-Zavaleta, M. I. Penaloza-Spinola, B. Contreras-Moreira, J. Segura-Salazar, L. Muniz-Rascado, I. Martinez-Flores, H. Salgado, C. Bonavides-Martinez, C. Abreu-Goodger, C. Rodriguez-Penagos, J. Miranda-Rios, E. Morett, E. Merino, A. M. Huerta, L. Trevino-Quintanilla, and J. Collado-Vides, "RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation," Nucleic Acids Res, vol. 36, pp. D120-4, Jan 2008.