The ISC Int'l Journal of Information Security

November 2020, Volume 12, Number 3 (pp. 29–36)

http://www.isecure-journal.org

Selected Paper at the ICCMIT'20 in Athens, Greece

CEMD: A Cluster-based Ensemble Motif Discovery Tool**

Sumayia Al-Anazi *,[1], Isra Al-Turaiki [1], and Najwa Altwaijry [1]

[1] *Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia.*

ARTICLE INFO.

Keywords:
Clustering, DNA Motif, Transcription Factor Binding Site

Type: Research Article

doi: 10.22042/isecure.2021.271073.623

ABSTRACT

Motif discovery is a challenging problem in bioinformatics. It is an essential step towards understanding gene regulation. Although numerous algorithms and tools have been proposed in the literature, the accuracy of motif finding is still low. In this paper, we tackle the motif discovery problem using ensemble methods. A review and classification of current ensemble motif discovery tools is presented. We then propose our cluster-based ensemble motif discovery tool (CEMD) which is based on k-medoids clustering of state-of-art stand-alone motif finding tools. We evaluate the performance of CEMD on benchmark datasets, and compare the results to both stand-alone and similar ensemble tools. Experimental results indicate that CEMD has better sensitivity than state-of-art stand-alone tools when dealing with human datasets. CEMD also obtains better values of sensitivity when motifs are implanted in real promoter sequences. As for the comparison of CEMD with ensemble motif discovery tools, results indicate that CEMD achieves better results than MEME-ChIP on all evaluation measures. CEMD shows comparable performance to RSAT peak-motifs and MODSIDE.

1 Introduction

Bioinformatics is a multi-disciplinary field where computer science, information technology, statistics, and engineering coincide, in order to organize and understand biological data [1]. The field emerged as a result of the continuing growth of available biological data produced by high-throughput technologies. Thus, there is a need to solutions in order to analyze, understand, and facilitate discovery of hidden knowledge in this huge amount of data. Sequential patterns are one important kind of hidden knowledge buried in biological sequences (DNA,RNA,orprotein). Such patterns, also *motifs*, are responsible for critical biological functions in the cell. The problem of finding recurring patterns is called the *motif discovery problem.*

Motif discovery plays an important role in understanding gene regulation, disease detection, and drug discovery [2]. However, it remains one of the most challenging problems in bioinformatics. Discovering motifs in the wet lab is a difficult and time-consuming task. Thus, computational approaches have been developed to solve this problem. Today, many stand-alone tools are available for the discovery of motifs

* Corresponding author.

**The ICCMIT'20 program committee effort is highly acknowledged for reviewing this paper.

Email addresses: missing@missing.mis, ialturaiki@ksu.edu.sa, ntwaijry@ksu.edu.sa

ISSN: 2008-2045

in sequential data. However, research shows that the performance of these tools in terms of accuracy is still limited [3]. This is because motifs are very short segments ranging in size from 8 to 20 bases [4]. They are hidden within enormous amounts of DNA data. In addition, motifs are not exact patterns; they allow some variability while preserving the same biological function. Motif variations are not well understood, so this leads to a very large search space. In fact, it has been shown that the motif discovery problem is inherently NP-hard [5].

In order to enhance prediction accuracy, studies have started investigating the potential of using ensemble methods, also called *pipelines or meta-servers*. The term ensemble is borrowed from machine learning literature. An ensemble combines the results of multiple weak classifiers to enhance the accuracy of predictions. Recent years have witnessed an increasing interest in using ensemble methods in bioinformatics [6–8]. The main idea is for multiple stand-alone tools to cooperate by combining their individually obtained results [9]. Ensemble methods have unique advantages in dealing with smaller sample sizes, high-dimensionality, and complex data structures [10]. Although several ensemble tools have been proposed for motif finding over the years, the accuracy of these tools still needs improvement.

The *motif* discovery problem can be formulated as follows: Given a set of N promoter regions $S = s_1, s_2, \ldots, s_N$, corresponding to co-regulated genes, each $s_i \in S$ is defined over the alphabet $\sum$. Given two integers $\{e, q | e \geq 0, 2 \leq q \leq N\}$, find all repeated patterns that are present in at least q promoters such that any motif occurs with at most e mismatches.

In this paper, we address the motif discovery problem using ensemble methods. We present the following contributions. First, we review the ensemble motif discovery literature. Then, we classify ensemble motif discovery tools into a number of categories: clustering-based ensembles, machine learning ensembles, complementary search space ensembles, and motif ranking ensembles. Next, we propose our *cluster-based ensemble motif discovery tool* (CEMD), which is based on k-medoids clustering of state-of-art stand-alone motif finding tools. We evaluate the prediction performance of CEMD using benchmark datasets and compare the obtained results with similar approaches in the literature.

2 Related Work

In ensemble motif discovery, complementary algorithms are combined to improve the accuracy of motif prediction. As shown in Figure 1, an ensemble motif discovery tool is composed of three main components: weak classifiers, learning rule, and an output reporting module. Several stand-alone motif discovery tools are used to process the same input data sequences. Then, a learning rule is used to aggregate the results together and rank the final motif predictions using an appropriate scoring function. Finally, results of the final filtered outcomes are reported by selecting top-ranking motifs in different formats such as text files and visual graphs. Efforts in ensemble motif discovery started in 2005. Since then, many tools have been proposed in the literature with the goal of improving accuracy of motif prediction. Here, we classify and discuss the currently available ensemble tools for DNA motif finding. Based on the available literature, we propose the following classification: clustering-based ensembles, machine learning ensembles, complementary search space ensembles, and motif ranking ensembles.

Clustering is the process of placing similar objects into groups (clusters). In motif discovery, motifs predicted by different tools are pooled together and divided into groups based on similarity. The main idea is that similar motifs predicted by different tools are more likely to represent real motifs. RGSMiner [11] is a clustering-based ensemble that combines three popular motif discovery programs: Gibbs Sampler [12], MEME [13], and AlignACE [14].

MultiFinder [15] is one of the early ensemble algorithms for motif discovery. It includes four motif discovery programs: MDscan [16], BioProspector [17], AlignACE [14, 18] and MEME [13]. The results from these tools are merged and ranked using hierarchical clustering. Other examples include: EMD [19], WebMotifs [20], MEMOFinder [21], MotifVoter [22], Complete MOTIFs [23], and Gimme Motifs [24]. In the machine learning literature, ensemble methods refer to combining the results of weak classifiers to enhance the accuracy of predictions. Machine learning ensemble techniques include: Bayesian averaging, error-correcting output coding, Bagging, and boosting [25]. Examples of machine learning motif discovery tools include: MotifBooster [26], SVMotifPWM [27], and W-ChIPMotifs [28].

In the motif discovery literature, some tools that are based on combining stand-alone tools that target different search spaces. For example: SCOPE [29, 30] and MEME-ChIP [31, 32]. Other tools such as BEST [33], Promzea [34], MODSIDE [35] do not actually combine the results of individual tools, but work on optimizing and enhancing the predictions of each component tool. We found that the majority of ensemble tools are clustering-based. They incorporate the most accurate component modules such as: MEME, Bioprospector, Weeder, AlignACE and MDscan. Some of the ensemble tools, such as EMD, MotifVoter, and CE3 [36], have the ability to be extended with new component modules. There are many challenges facing current ensemble motif discovery tools. Some tools are offered as stand-alone software that take a long time

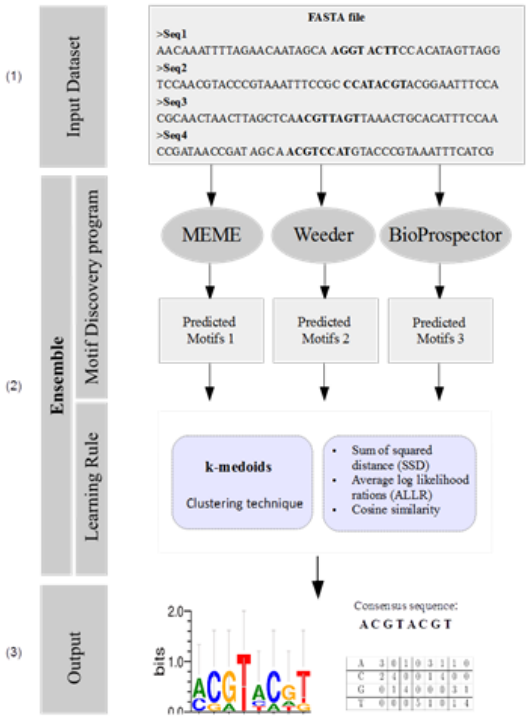

Figure 2. CEMD architecture

to download and install. Sometimes libraries and packages are required for proper installation. As for online web-based tools, maintenance and support teams are needed to keep the tool alive. Some online tools are no longer available, such as RGSMiner [11], SCOPE [29, 30], and BEST [33]. Motif discovery tools based on machine learning techniques need to have a training dataset with known labels for the classifiers, which is difficult to obtain.

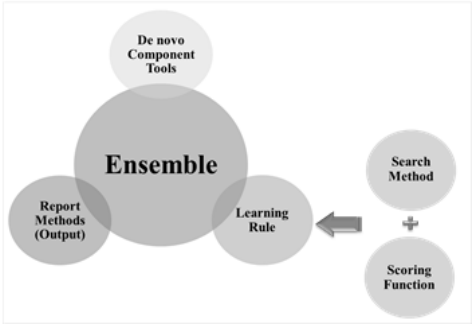

Figure 1. Basic components of an ensemble tool

3 CEMD Architecture

In this section, we describe the components of CEMD, our proposed cluster-based ensemble motif discovery tool. Figure 2 shows the architecture of CEMD, which is composed of the following modules: motif discovery programs, learning rule, and output. The motif discovery programs produce the initial set of predicted motifs. Then, the learning rule, which is a clustering ensemble in our work, predicts motifs using the k-mediods algorithm. Finally, motifs are scored and presented to the user.

3.1 Motif Discovery Programs

In order to develop CEMD, we selected three of the best motif discovery tools as our component modules: [7] MEME, [37] Weeder, and [13] BioProspector. MEME and Weeder were found to be the best motif discovery tools in the study of Tompa *et al.* [3]. In addition, we integrate the results of BioProspector, which is considered a fast, accurate and complementary searching tool [34]. Next, we provide a brief description of each tool.

- MEME [37] is an expectation maximization based algorithm for motif discovery. It optimizes the expected value of a score based on the information content. From each l-mer, the algorithm iterates between the E-step and M-step collecting candidates and updating the model with the new sites that lead to higher expected values. MEME searches for motifs of length 6 to 18 base pairs. It then returns the top five discovered motifs.
- Weeder [38] is a suffix-tree based word enumeration algorithm. It compares the observed occurrences of a motif to the expected frequencies in the promoter regions of the same organism. Weeder searches for motifs of widths: 6, 8, 10, 12 and 14. It allows for mismatches between 1 and 4. The top ten discovered motifs are returned to the user.
- BioProspector [39] is based on Gibbs sampling. A zero to third-order Markov model is used to model the background. This strategy enhances the accuracy of the reported motifs. BioProspector allows input sequences to contain zero to multiple copies of the motif. The top five discovered motifs will be returned.

3.2 Learning Rule Module

The learning rule in our proposed ensemble consists of a clustering algorithm, which is used to group similar motifs together. The clustering algorithm employs similarity measures that measure motif similarity.

3.2.1 Clustering Algorithm

CEMD is based on K-medoids, a well-known clustering algorithm. First, CEMD runs the three motif discovery programs to find motifs. CEMD then clusters the motifs and outputs the most representative one. K-

medoids allows motifs to move between clusters at any point in the clustering process. K-medoids performs the following steps:

1. Specify the number of clusters, K.
2. Select a total of K initial cluster representative objects, called medoids, mi.
3. Assign the remaining objects to the cluster with the closest medoid.
4. Randomly select another object, o, which is a non-representative object.
5. Compute the distance between o and other objects in each cluster, if the total distance $D < 0$, then swap m with o to form a new set of medoids.
6. Repeat steps 4 and 5 until the cluster quality and convergence distance is satisfied.

A demonstration of D-medoids algorithm, where $D = 2$ is shown in Figure 3.

3.2.2 Similarity Measures

The similarity measures or scoring functions are used in the clustering process to determine similarity. In CEMD, we choose three distance metrics for motif alignments: [7] Sum of squared distance, [37] Average log likelihood rations, and [13] Cosine similarity with range. Based on previous research, these measures were found to produce promising results in motif discovery [40, 41].

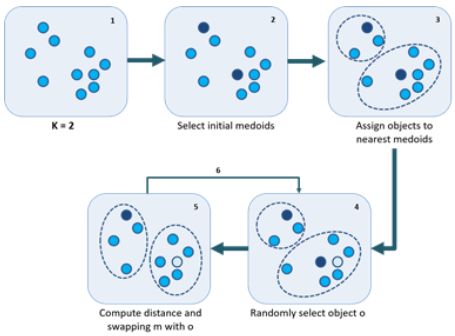

Figure 3. K-medoids algorithm process

(1) Sum of squared distance (SSD):

$$SSD(x,y) = 2 - \sum_{b=a}^{T} (f_x(b) - f_y(b))^2 \qquad (1)$$

(2) Average log likelihood rations (ALLR):

$$ALLR_{(x,y)} = \frac{\sum_{b=A}^{T} n_x(b) \log \frac{f_y(b)}{P_{ref}(b)} + \sum_{b=A}^{T} n_y(b) \log \frac{f_x(b)}{P_{ref}(b)}}{\sum_{b=A}^{T} (n_x(b) + n_y(b))}$$

(3) Cosine similarity:

$$d_{cos}(a,b) = 1 - \frac{a \cdot b}{\|a\| \cdot \|b\|} \qquad (2)$$

CEMD users are allowed to select one measure from these three distance metrics. Since the range of the results from each scoring function is different, a scaling function (see Equation 3) is used as a unified scale for output results from all three scoring functions in the CEMD tool, with $Min = 0$ meaning more similar, and $Max = 1$ meaning there is a difference.

$$MinMax - Scaling = \frac{x - \min}{\max - \min} \qquad (3)$$

3.3 Motif Representation Module

As discussed in Section 3.1, the CEMD tool contains three motif discovery programs that produce different output formats such as text files, XML files, and HTML files. The motif representation module is responsible for converting the output motifs to a unified representation (i.e. all motifs can be represented as a consensus sequence). Each motif discovery program will process the input file separately. The CEMD tool will pool and integrate the output files to improve the accuracy, and extract the motifs' consensus sequences.

3.4 Output Module

Each motif that was extracted by the CEMD tool will be represented by the following: consensus, instances, component module, sequence logo, location, and position weight matrix. The final results are provided on the CEMD webpage, and can be downloaded as a text file, or sent to an email address as a text file with all motifs' graphical logos.

4 Experimental Results

Here, we evaluate the performance of CEMD and report the prediction results. The performance of CEMD is compared to other similar tools. CEMD was implemented and tested on an IntelR CoreTM i7 Duo CPU 2.40 GHz machine, with 16GB RAM.

4.1 Dataset and Evaluation Measures

We use the benchmark dataset of Tompa *et al.*, [3]. This dataset consists of 56 datasets representing four different species: human, mouse, fly, and yeast. The average number of sequences per dataset is 7. Each sequence has an average length of 1000 (approximately) nucleotides. The benchmark dataset is divided into three parts, each representing different types of background sequences. In the dataset, real transcription factor binding sites are placed at their original positions. There are three types of background sequence: real promoters, randomly chosen promoters from the same genome, and sequences generated by a Markov

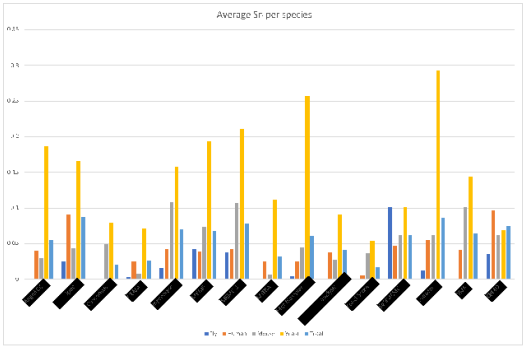

Figure 4. The average sensitivity per species obtained for CEMD and stand-alone motif finding tools

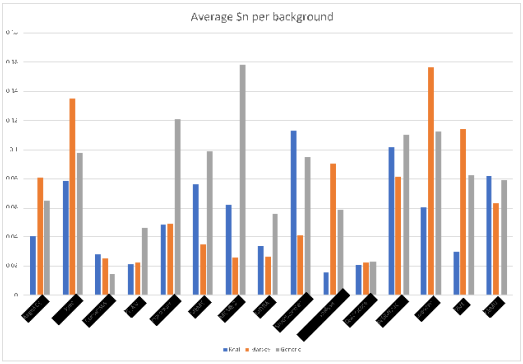

Figure 5. The average sensitivity per background type obtained for CEMD and stand-alone motif finding tools

chain of order 3. The performance is measured using the following measures calculated at the nucleotide level:

- *Sensitivity* (Sn): the proportion of known nucleotide positions that are correctly predicted.

$$S_n = \frac{TP}{TP + FN} \tag{4}$$

- *Positive Predictive Value* (PPV): the proportion of predicted nucleotide positions that are known.

$$PPV = \frac{TP}{TP + FP} \tag{5}$$

- *Specificity* (Sp): the proportion of background nucleotides that are correctly identified as background:

$$S_p = \frac{TN}{TN + FP} \tag{6}$$

Where TP refers to true positives, FN to false negatives, TN to True Negatives, and FP to false positives.

4.2 Prediction Performance

In this section, we discuss the performance of CEMD and compare it to stand-alone motif finding tools [3], including: MEME and Weeder. We examine the prediction performance species-wise and also with respect to the type of background sequences in the dataset (i.e. real, generic, or Markovian). In addition, we compare the performance of CEMD to other ensemble motif discovery tools [35].
Figure 4 shows the average sensitivity values of CEMD as compared with fourteen stand-alone motif finding tools. We observe that the obtained sensitivity values for CEMD are greater than the values of all the tools when dealing with human datasets. CEMD also achieved better results for the fly and mouse datasets. In general, CEMD achieved better results than eleven tools across all species. We now look at the average sensitivity per background type. As shown in Figure 5, CEMD performed similarly on the three types of background sequences. For the real datasets, CEMD obtained better values of sensitivity than most stand-alone tools. Also, CEMD was able to achieve better results compared with some tools when dealing with generic and Markovian datasets. As for specificity, Figure 6 and Figure 7 show that CEMD is comparable to other tools.
In terms of positive predictive value, Figure 8 and Figure 9 show the performance of CEMD and other tools per species and per background type, respectively. CEMD again performs better on fly and human datasets. Figure 9 shows that CEMD performs better than eleven tools on the real dataset.
We now compare CEMD to recently published experimental results using a subset of 16 datasets from Tompa's benchmark dataset. The datasets represent three species: human, mouse, and yeast. The background sequences are either generic or Markovian. We compare the performance of CEMD to the recently published results presented with MODSIDE[35]. Figure 10 shows the performance of CEMD compared to the stand-alone motif finding tools ChIPMunk [42], MEME [43], Weeder [38], XXMotif [44], and MODSIDE [35]. CEMD outperforms MEME and XXMotif on all evaluation measures. CEMD has better sensitivity than ChIPMunk, but comparable values for positive predictive value and specificity. Tran and Huang [35] presented the performance of MEME-ChIP [32], RSAT peak-motifs [45], and MODSIDE [35]. Figure 11 shows the performance of CEMD and ensemble motif finding tools using 16 selected datasets. The results indicate that CEMD achieves better results than MEME-ChIP in terms of all evaluation measures. However, RSAT peak-motifs and MODSIDE perform better than CEMD.

5 Conclusion

In this paper, we present CEMD, a clustering based ensemble for motif discovery. CEMD incorporates three stateof-the-art stand-alone motif finding tools. We used K-medoids in the learning rule as a clustering algorithm. We also used three distance measures: sum

of squared distance, average log likelihood rations, and cosine similarity. The performance of CEMD was evaluated using the gold standard benchmark dataset developed by Tompa *et al.*, [3]. CEMD showed better performance than stand-alone tools when dealing with human datasets and motifs planted in real promoter sequences. Our tool has comparable performance to RSAT peak-motifs and MODSIDE. CEMD may be further enhanced by making it an extensible tool to allow for the integration of advanced stand-alone motif finding tools.

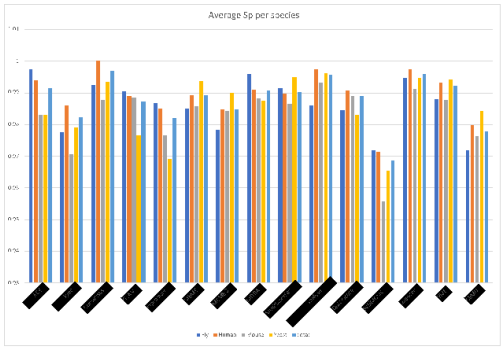

Figure 6. The average specificity per species obtained for CEMD and stand-alone motif finding tools

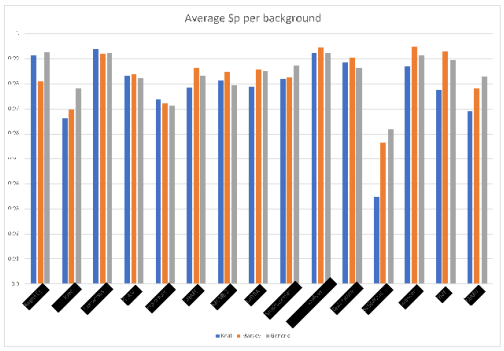

Figure 7. The average specificity per background type obtained for CEMD and stand-alone motif finding tools

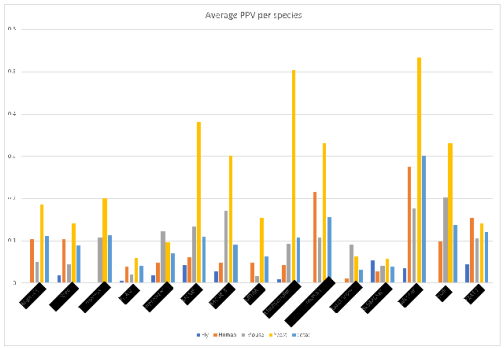

Figure 8. The average positive predictive value per species obtained for CEMD and stand-alone motif finding tools

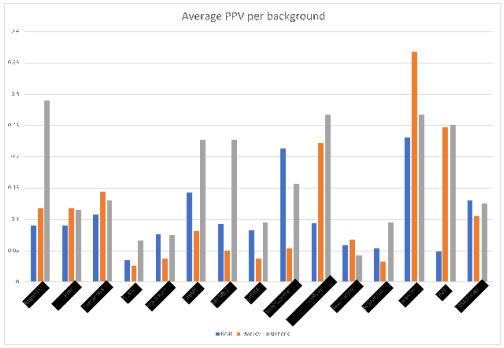

Figure 9. The average positive predictive value per background type obtained for CEMD and stand-alone motif finding tools

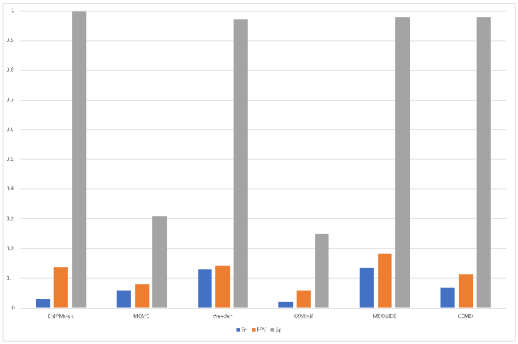

Figure 10. Evaluation measures for CEMD and stand-alone motif finding tools using sixteen selected datasets

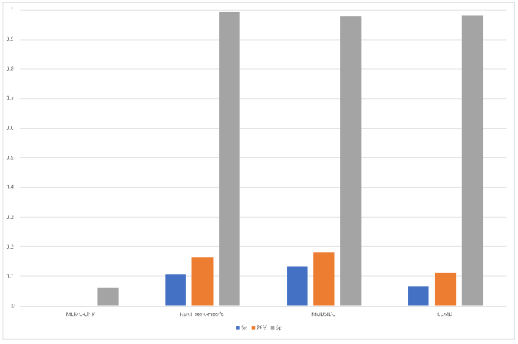

Figure 11. Evaluation measures for CEMD and ensemble motif finding tools using sixteen selected datasets.

References

[1] Nicholas M Luscombe, Dov Greenbaum, and Mark Gerstein. What is bioinformatics? a proposed definition and overview of the field. *Methods of information in medicine*, 40(04):346–358, 2001.

[2] Lachlan Coff, Jeffrey Chan, Paul A Ramsland, and Andrew J Guy. Identifying glycan motifs using a novel subtree mining approach. *BMC bioinformatics*, 21(1):42, 2020.

[3] Martin Tompa, Nan Li, Timothy L Bailey, George M Church, Bart De Moor, Eleazar Eskin, Alexander V Favorov, Martin C Frith, Yu-

tao Fu, W James Kent, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, 23(1): 137–144, 2005.

[4] Federico Zambelli, Graziano Pesole, and Giulio Pavesi. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in bioinformatics*, 14 (2):225–237, 2013.

[5] Jaime Davila, Sudha Balla, and Sanguthevar Rajasekaran. Fast and practical algorithms for planted (l, d) motif search. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(4):544–552, 2007.

[6] Yanju Zhang, Sha Yu, Ruopeng Xie, Jiahui Li, André Leier, Tatiana T Marquez-Lago, Tatsuya Akutsu, A Ian Smith, Zongyuan Ge, Jiawei Wang, et al. Pengaroo, a combined gradient boosting and ensemble learning framework for predicting non-classical secreted proteins. *Bioinformatics*, 36(3):704–712, 2020.

[7] Mehmet Eren Ahsen, Robert Vogel, and Gustavo A Stolovitzky. R/py-summa: An r/python package for unsupervised ensemble learning for binary classification problems in bioinformatics. *Journal of Computational Biology*, 27(9):1337–1340, 2020.

[8] Kanica Sachdev and Manoj K Gupta. Predicting drug target interactions using dimensionality reduction with ensemble learning. In *Proceedings of ICRIC 2019*, pages 79–89. Springer, 2020.

[9] Juho Kim, Seunghak Yu, and Sungroh Yoon. Ensemble algorithms for dna motif finding. In *2014 International Conference on Electronics, Information and Communications (ICEIC)*, pages 1–2. IEEE, 2014.

[10] Pengyi Yang, Yee Hwa Yang, Bing B Zhou, and Albert Y Zomaya. A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5(4): 296–308, 2010.

[11] Hsien-Da Huang, Jorng-Tzong Horng, Yi-Ming Sun, Ann-Ping Tsou, and Shir-Ly Huang. Identifying transcriptional regulatory sites in the human genome using an integrated system. *Nucleic acids research*, 32(6):1948–1956, 2004.

[12] Charles E Lawrence, Stephen F Altschul, Mark S Boguski, Jun S Liu, Andrew F Neuwald, and John C Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *science*, 262(5131):208–214, 1993.

[13] Timothy L Bailey, Charles Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.

[14] Jason D Hughes, Preston W Estep, Saeed Tavazoie, and George M Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae. *Journal of molecular biology*, 296(5):1205–1214, 2000.

[15] Bertrand R Huber and Martha L Bulyk. Meta-analysis discovery of tissue-specific dna sequence motifs from mammalian gene expression data. *BMC bioinformatics*, 7(1):1–25, 2006.

[16] X Shirley Liu, Douglas L Brutlag, and Jun S Liu. An algorithm for finding protein–dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature biotechnology*, 20(8):835–839, 2002.

[17] Xiaole Liu, Douglas L Brutlag, and Jun S Liu. Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. In *Biocomputing 2001*, pages 127–138. World Scientific, 2000.

[18] Frederick P Roth, Jason D Hughes, Preston W Estep, and George M Church. Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nature biotechnology*, 16(10):939–945, 1998.

[19] Jianjun Hu, Yifeng D Yang, and Daisuke Kihara. Emd: an ensemble algorithm for discovering regulatory motifs in dna sequences. *BMC bioinformatics*, 7(1):1–13, 2006.

[20] Katherine A Romer, Guy-Richard Kayombya, and Ernest Fraenkel. Webmotifs: automated discovery, filtering and scoring of dna sequence motifs using multiple programs and bayesian approaches. *Nucleic acids research*, 35(suppl_2): W217–W220, 2007.

[21] Bartek Wilczynski, Milosz Darzynkiewicz, and Jerzy Tiuryn. Memofinder: combining de novo motif prediction methods with a database of known motifs. *Nature Precedings*, pages 1–1, 2008.

[22] Edward Wijaya, Siu-Ming Yiu, Ngo Thanh Son, Rajaraman Kanagasabai, and Wing-Kin Sung. Motifvoter: a novel ensemble method for fine-grained integration of generic motif finders. *Bioinformatics*, 24(20):2288–2295, 2008.

[23] Lakshmi Kuttippurathu, Michael Hsing, Yongchao Liu, Bertil Schmidt, Douglas L Maskell, Kyungjoon Lee, Aibin He, William T Pu, and Sek Won Kong. Completemotifs: Dna motif discovery platform for transcription factor binding experiments. *Bioinformatics*, 27(5): 715–717, 2011.

[24] Simon J van Heeringen and Gert Jan C Veenstra. Gimmemotifs: a de novo motif prediction pipeline for chip-sequencing experiments. *Bioinformatics*, 27(2):270–271, 2011.

[25] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on*

multiple classifier systems, pages 1–15. Springer, 2000.

[26] Pengyu Hong, X Shirley Liu, Qing Zhou, Xin Lu, Jun S Liu, and Wing H Wong. A boosting approach for motif modeling using chip-chip data. *Bioinformatics*, 21(11):2636–2643, 2005.

[27] Yue Fan, Mark A Kon, and Charles DeLisi. Ensemble machine methods for dna binding. In *2008 Seventh International Conference on Machine Learning and Applications*, pages 709–716. IEEE, 2008.

[28] Victor X Jin, Jeff Apostolos, Naga Satya Venkateswara Ra Nagisetty, and Peggy J Farnham. W-chipmotifs: a web application tool for de novo motif discovery from chip-based high-throughput data. *Bioinformatics*, 25(23):3191–3193, 2009.

[29] Jonathan M Carlson, Arijit Chakravarty, Charles E DeZiel, and Robert H Gross. Scope: a web server for practical de novo motif discovery. *Nucleic acids research*, 35(suppl_2):W259–W264, 2007.

[30] A Chakravarty, JM Carlson, RS Khetani, and RH Gross. A parameter-free algorithm for improved de novo identification of transcription factor binding sites. *BMC Bioinformatics*, 8:29, 2007.

[31] Wenxiu Ma, William S Noble, and Timothy L Bailey. Motif-based analysis of large nucleotide data sets using meme-chip. *Nature protocols*, 9 (6):1428–1450, 2014.

[32] Philip Machanick and Timothy L Bailey. Meme-chip: motif analysis of large dna datasets. *Bioinformatics*, 27(12):1696–1697, 2011.

[33] Dongsheng Che, Shane Jensen, Liming Cai, and Jun S Liu. Best: binding-site estimation suite of tools. *Bioinformatics*, 21(12):2909–2911, 2005.

[34] Christophe Liseron-Monfils, Tim Lewis, Daniel Ashlock, Paul D McNicholas, François Fauteux, Martina Strömvik, and Manish N Raizada. Promzea: a pipeline for discovery of co-regulatory motifs in maize and other plant species and its application to the anthocyanin and phlobaphene biosynthetic pathways and the maize development atlas. *BMC plant biology*, 13(1):1–17, 2013.

[35] Ngoc Tam L Tran and Chun-Hsi Huang. Modside: a motif discovery pipeline and similarity detector. *BMC genomics*, 19(1):1–9, 2018.

[36] K Tillán, M Leoncini, and M Montangero. Ce 3: Customizable and easily extensible ensemble tool for motif discovery. 2012.

[37] Timothy L Bailey, Charles Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.

[38] Giulio Pavesi, Giancarlo Mauri, and Graziano Pesole. An algorithm for finding signals of unknown length in dna sequences. *Bioinformatics*, 17(suppl_1):S207–S214, 2001.

[39] Xiaole Liu, Douglas L Brutlag, and Jun S Liu. Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. In *Biocomputing 2001*, pages 127–138. World Scientific, 2000.

[40] Pilib Ó Broin, Terry J Smith, and Aaron AJ Golden. Alignment-free clustering of transcription factor binding motifs using a genetic-k-medoids approach. *BMC bioinformatics*, 16(1): 1–12, 2015.

[41] Shaun Mahony, Philip E Auron, and Panayiotis V Benos. Dna familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput Biol*, 3(3): e61, 2007.

[42] Ivan V Kulakovskiy, VA Boeva, Alexander V Favorov, and Vsevolod J Makeev. Deep and wide digging for binding motifs in chip-seq data. *Bioinformatics*, 26(20):2622–2623, 2010.

[43] Timothy L Bailey, Nadya Williams, Chris Misleh, and Wilfred W Li. Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic acids research*, 34(suppl_2):W369–W373, 2006.

[44] Sebastian Luehr, Holger Hartmann, and Johannes Söding. The xxmotif web server for exhaustive, weight matrix-based motif discovery in nucleotide sequences. *Nucleic acids research*, 40 (W1):W104–W109, 2012.

[45] Morgane Thomas-Chollier, Carl Herrmann, Matthieu Defrance, Olivier Sand, Denis Thieffry, and Jacques van Helden. Rsat peak-motifs: motif analysis in full-size chip-seq datasets. *Nucleic acids research*, 40(4):e31–e31, 2012.

Isra AL-Turaiki is an associate professor of computer science at King Saud University. She received her Ph.D. degree in 2014 from the college of computer sciences at King Saud University. Her research interests include data mining, machine learning, and bioinformatics.

Najwa Altwaijry is an assistant professor of computer science at King Saud University. She received her Ph.D. degree in 2014 from the college of computer sciences at King Saud University. Her research interests include machine learning, swarm intelligence, evolutionary computation, cybersecurity and bioinformatics.