

Motto: Representing Motifs in Consensus Sequences with Minimum Information Loss

Mengchi Wang,^{*1} David Wang,^{†1} Kai Zhang,^{*} Vu Ngo,^{*} Shicai Fan,^{†,*} and Wei Wang^{*,†,§,2}

^{*}Bioinformatics and Systems Biology, [†]Department of Chemistry and Biochemistry, and [§]Department of Cellular and Molecular Medicine, University of California at San Diego, La Jolla, California 92093 and [‡]School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China 610054

ORCID ID: 0000-0003-3454-7357 (K.Z.)

ABSTRACT Sequence analysis frequently requires intuitive understanding and convenient representation of motifs. Typically, motifs are represented as position weight matrices (PWMs) and visualized using sequence logos. However, in many scenarios, in order to interpret the motif information or search for motif matches, it is compact and sufficient to represent motifs by wildcard-style consensus sequences (such as [GC][AT]GATAAG[GAC]). Based on mutual information theory and Jensen-Shannon divergence, we propose a mathematical framework to minimize the information loss in converting PWMs to consensus sequences. We name this representation as sequence Motto and have implemented an efficient algorithm with flexible options for converting motif PWMs into Motto from nucleotides, amino acids, and customized characters. We show that this representation provides a simple and efficient way to identify the binding sites of 1156 common transcription factors (TFs) in the human genome. The effectiveness of the method was benchmarked by comparing sequence matches found by Motto with PWM scanning results found by FIMO. On average, our method achieves a 0.81 area under the precision-recall curve, significantly (P -value < 0.01) outperforming all existing methods, including maximal positional weight, Cavener's method, and minimal mean square error. We believe this representation provides a distilled summary of a motif, as well as the statistical justification.

KEYWORDS consensus; information theory; motif; sequence logo; transcription factor binding

MOTIF analysis is crucial for uncovering sequence patterns, such as transcription factor (TF) binding sites (Thompson *et al.* 2003), splicing sites (Murray *et al.* 2008), DNA methylation patterns (Wang *et al.* 2019), and histone modifications (Ngo *et al.* 2019b). A motif is typically represented as a position weight matrix (PWM), in which each entry shows the occurrence frequency of a certain type of nucleic acid at each position of the motif. PWMs are often visualized by the sequence logo (Schneider and Stephens 1990), which requires a graphical interface. However, when in a textual interface, representing PWMs requires an n by k

matrix, where n is the number of characters (such as A, C, G, T, for nucleotides), and k is the length of the motif. Recently, several studies have shown the usefulness of representing motifs using kmers (Fletez-Brant *et al.* 2013; Ghandi *et al.* 2014; Zeng *et al.* 2016; Guo *et al.* 2018); despite the power of this representation in machine learning models, it is cumbersome to have a set of kmers to characterize a single motif. In many scenarios, motifs can be sufficiently represented by regular expressions of the consensus sequences, such as [GC][AT]GATAAG[GAC] for the GATA2 motif. This representation is the most compact and intuitive way to delineate a motif. In the GATA2 motif example, the GATAAG consensus in the center is the most prominent pattern that would be read off the PWM or sequence logo. For this reason, consensus sequences are still widely used by the scientific community. Consensus sequences in regular expression form are the only supported textual format to highlight motif occurrence in popular genome browsers such as UCSC (Kent *et al.* 2002) and IGV (Robinson *et al.* 2011). Consensus sequences are assigned to *de novo* motifs and sequences for informative

Copyright © 2020 Wang *et al.*

doi: <https://doi.org/10.1534/genetics.120.303597>

Manuscript received May 31, 2020; accepted for publication August 17, 2020; published Early Online August 19, 2020.

Available freely online through the author-supported open access option.

This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

¹These authors contributed equally to this work.

²Corresponding author: University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093. E-mail: wei-wang@ucsd.edu

denotations (Bailey *et al.* 2009; Heinz *et al.* 2010; Whitaker *et al.* 2015; Wang *et al.* 2019). Wildcard-like sequence patterns are also supported in DNA oligo libraries synthesis by major vendors including Invitrogen, Sigma-Aldrich, and Thermo-Fisher.

However, current methods that convert PWMs to consensus sequences are often heuristic. One simple approach is taking the nucleotide with maximal frequency at each position to define the consensus sequence (*e.g.*, GGTC AAGTCAC for ESRRB). Unsurprisingly, this could misrepresent positions with similar frequencies (*e.g.*, 0.26, 0.25, 0.25, 0.24, which should have been assigned as N). Alternatively, in 1987, Cavener proposed to follow a set of rules: use the single nucleotide with the highest frequency when it exceeds 0.50 and two times the second-highest frequency; else, use the top two dinucleotides when their total frequencies exceed 0.75; else, use N (Cavener 1987). However, these rules are arbitrary, inflexible, and lack a mathematical framework.

Here, we present Motto, a sequence consensus representation of motifs based on information theory, which ensures minimal information loss when converted from a PWM (Figure 1). We provide a standardized solution that determines the optimal motif consensus sequence. We have also implemented a lightweight and easy-to-use Python package with versatile options for biologists.

Materials and Methods

A positional weight matrix (PWM) defines $P(i, j)$, the probability of the j th character (out of n characters) at the i th position (out of k positions), where $\sum P(i, j) = 1$, $i \in [1, k]$, and $j \in [1, n]$. For a given position i , let $M(i) = \{S_{i,1}, S_{i,2}, \dots, S_{i,m}\}$ denote a output consensus character set, where m is the number of characters to be presented, $m \in [1, n]$, and let $C(i, m)$ denote the perceived frequencies for a combination of m characters, defined by equal frequencies shared among included characters:

$$C(i, 1) = C(i, 2) = \dots = C(i, m) = 1/m$$

For example, a $M(i) = \{"A", "C", "T"\}$ is a case for $m = 3$ and $C(i, m) = 0.333$ with frequencies of [0.333, 0.333, 0, 0.333] for [A, C, G, T], respectively. Thus, we consider the optimal consensus sequence as a series of combination of characters $M(i)$ that has the most similarity between $C(i, m)$ and $P(i)$ for each position $i \in [1, k]$.

For convenience, in the following discussion we will omit the index i when possible, as we note that optimal $M(i)$ is independent of the position $i \in I$ under consideration. To further simplify the discussion, we use the second ($i = 2$) position of the human TF P73 (Figure 1), where the $P("A") = 0.077$, $P("C") = 0$, $P("G") = 0.726$, and $P("T") = 0.197$, for demonstration below.

Motto method using minimal Jensen-Shannon divergence

Here, we propose to use Jensen-Shannon divergence (JSD) to measure the similarity between $C(m)$ and P . JSD has been widely used in information theory to characterize the

difference between distributions (Lin 1991). Using this metric, the combination of nucleotides with the least JSD from $C(m)$ to P will have the minimal "information loss," and is thus considered as the optimal consensus nucleotide.

To efficiently compare JSD between all possible nucleotide combinations, we propose the following algorithm (Figure 1). First, we sort the nucleotides of the PWM in descending order, so that:

$$P(j_1) \geq P(j_2) \dots \geq P(j_n).$$

For example, at the second position of the human TF P73 (Figure 1), the nucleotides are sorted by occurrence frequencies so that:

$$\begin{aligned} P("G") &= 0.726 \geq P("T") = 0.197 \geq P("A") \\ &= 0.077 \geq P("C") = 0. \end{aligned}$$

Next, we reasoned that if a nucleotide with probability $P(S_j)$ is included in the output consensus sequence set, then all nucleotides with frequencies larger than $P(S_j)$ must also be included. Therefore, the optimal consensus character set M (denoted as M^*) is given by the optimal m (denoted as m^*), where:

$$M^* = (S_1, S_2, \dots, S_{m^*})$$

$$P(S_1) \geq P(S_2) \dots \geq P(S_{m^*}).$$

For example, if $m^* = 2$, then the optimal output character set will be $M^* = \{"G", "T"\}$, where $P("G") = 0.726 \geq P("T") = 0.197$.

The closer this distribution is to the original distribution of nucleotide frequencies, the better approximation of the consensus motif is to the original PWM. Thus, m^* can be determined by minimizing the JSD between the two distributions:

$$m^* = \operatorname{argmin}_m (JSD(C(m), P) + q^2 \cdot m)$$

$$JSD(A, B) = \frac{1}{2} KLD(A, M) + \frac{1}{2} KLD(B, M)$$

$$M = \frac{1}{2}(A + B)$$

$$KLD(A, B) = \sum_{j=1}^n \ln \left(\frac{A(j)}{B(j)} \right).$$

Here, $q \in [0, 1]$ is the ambiguity penalty, a parameter input from the user to penalize a larger m , in case a more definite output is preferred. When $q = 0$ (the default value), the optimal m^* marks the canonical minimal JSD, which we deem to have retained the most information about the original PWM. When $q = 1$, m^* is guaranteed to be 1, thus the output consensus nucleotide is $M^* = \{S_1\}$, equivalent to using nucleotides with the maximal frequency.

Thus, the optimal consensus nucleotide set at the i th position is:

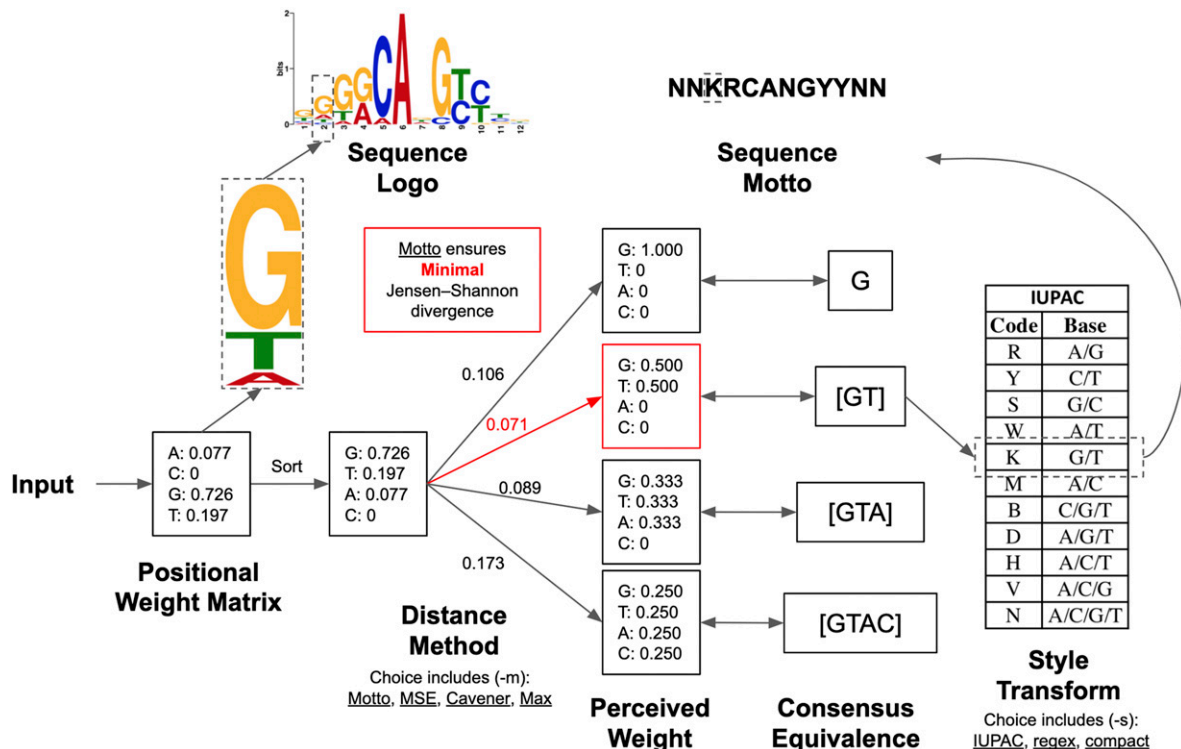


Figure 1 Overview of sequence Motto and comparison with sequence logo. Given a motif PWM as the input, Motto outputs a consensus that minimizes information loss. Here we show how the sequence Motto of the human transcription factor P73 is determined.

$$M^* = \{S_{1,1}, S_{1,2}, \dots, S_{1,m^*}\}.$$

Repeat this procedure for every position $i \in [1, k]$, the final optimal consensus sequence is given by:

$$\{S_{1,1}, S_{1,2}, \dots, S_{1,m^*}\}, \{S_{2,1}, S_{2,2}, \dots, S_{2,m^*}\}, \dots, \{S_{k,1}, S_{k,2}, \dots, S_{k,m^*}\}.$$

Minimal mean squared error method

For comparison purposes, we have also implemented a minimal mean squared error (MSE) method, which is another metric used widely to measure distribution discrepancy (Lele 1993). The rest of the implementation is unchanged, except that the optimal m (m^*) is now determined by minimizing the MSE between the two distributions:

$$m^* = \operatorname{argmin}_m (MSE(C(m), P) + q^2 \cdot m)$$

$$MSE(A, B) = \frac{1}{n} \sum_{j=1}^n (A(j) - B(j))^2.$$

Evaluating motif occurrence sites

We have collected 1156 common TFs from human and mouse from the databases of Transfac (Matys *et al.* 2006), Jaspar (Portales-Casamar *et al.* 2010), Uniprobe (Robasky and Bulyk 2011), hPDI (Xie *et al.* 2010), and HOCOMOCO (Kulakovskiy *et al.* 2018). Each PWM is converted into consensus sequences,

using default options of the four discussed methods: JSD (described above), MSE (described above), Cavener (Cavener 1987), and the naive approach of using the maximal frequency. Motif occurrence sites are determined in the human genome (hg19), matched by their regular expressions. The ground truth of the occurrence sites is determined by scanning the original PWMs with FIMO (Grant *et al.* 2011) using a $1e-5$ P -value cutoff. The resulting P -values are converted into a significance score $[-\log(P\text{-value})]$ and assigned to the matched motif occurrence sites from sequence Mottos. Thus, the area under the precision-recall curves (Davis and Goadrich 2006) (auPRC) is calculated by comparing the motif occurrence sites and their significance scores. Resulting auPRCs are averaged and a paired (by each motif) t -test is conducted to determine performance. Comparisons with significance (P -value < 0.01) are shown (Figure 3).

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article. Motto is freely available at <https://github.com/MichaelMW/motto>. Motto representation of all 1156 common TFs, as well as their sequence logo used in this study can be found at <http://wanglab.ucsd.edu/star/motto>.

Results and Discussion

Motto takes the MEME format of PWM as the input because of its popularity. The MEME format is supported by the majority

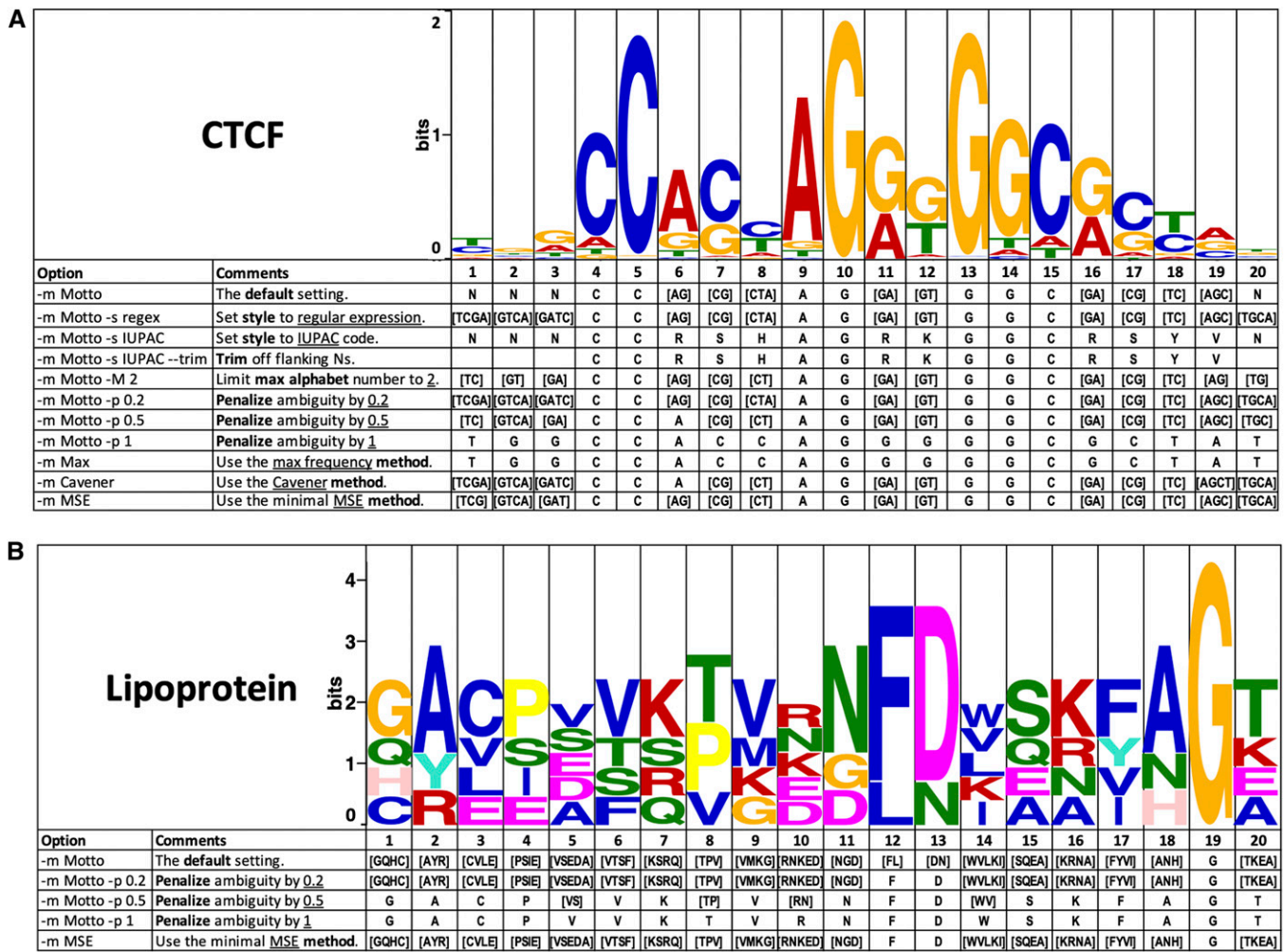


Figure 2 Example usage using human CTCF (upper panel) and lipoprotein binding sites from Bailey and Elkan (1994) (lower panel). The original PWM is shown in a sequence logo. Different Motto options resulted in various consensus sequence output at each position. In particular, “-m/-method” specifies the method: Motto (default), MSE (minimal mean square error), Cavener (Cavener 1987), or Max (using maximal frequency at each position); “-s/-style” specifies the output style: IUPAC (Johnson 2010) (single character for nucleotide combinations), regex (regular expression), or compact (convert [ACGT] to N in regex); “-t/-trim” is an option for trimming off the flanking Ns; “-p/-penalty” specifies a weight between 0 and 1 that penalizes ambiguity at each position (for details see *Materials and Methods*).

of the motif databases (Kulakovskiy *et al.* 2018), and the MEME suite provides packages for integrative analysis and conversion from other motifs formats (Bailey *et al.* 2009). The recently proposed kmer-based motif models also support conversion to MEME format (Fletez-Brant *et al.* 2013; Ghandi *et al.* 2014; Zeng *et al.* 2016; Guo *et al.* 2018). Our package is lightweight and open-source. The algorithm is efficiently implemented in Python and the conversion for 1000 motif sequences typically takes <2 sec. In addition, perhaps expectedly, downstream analysis like matching motif occurrences using Motto is much faster (~5 sec for a common PWM on a chromosome, implemented inhouse with Python) than a conventional PWM scanning (about 1 min, scanned with FIMO (Grant *et al.* 2011)). By default, the Motto package takes a motif in the MEME format, parses the header to get the nucleotide, computes the optimal consensus sequences based on the Motto method, and then outputs the

sequence in a compact format (Figure 1). Motto provides flexibility at each step along this process. Input can be from a file, or from standard input, and Motto can consider nucleotides, amino acids, and customized characters such as CpG/non-CpG methylation (Ngo *et al.* 2019a) and protein phosphorylation (Amanchy *et al.* 2011). We have provided four methods for comparison: maximal probability (Max), heuristic Cavener’s method (hereafter referred to as Cavener), minimal MSE method, and our proposed Motto method using the minimal JSD (Motto) (see *Materials and Methods*). Three output styles are provided: (1) IUPAC uses a single character to represent the combination of nucleotides (*e.g.*, S for [CG]) and is the most compact form, but requires reference to the nomenclature (Johnson 2010); (2) regular expression (“regex”) enumerate all output consensus nucleotide ranked by occurrences and is recommended for downstream analysis, such as motif occurrence and oligo designs;

Converted Sequences Recapitulate Motif Occurrences

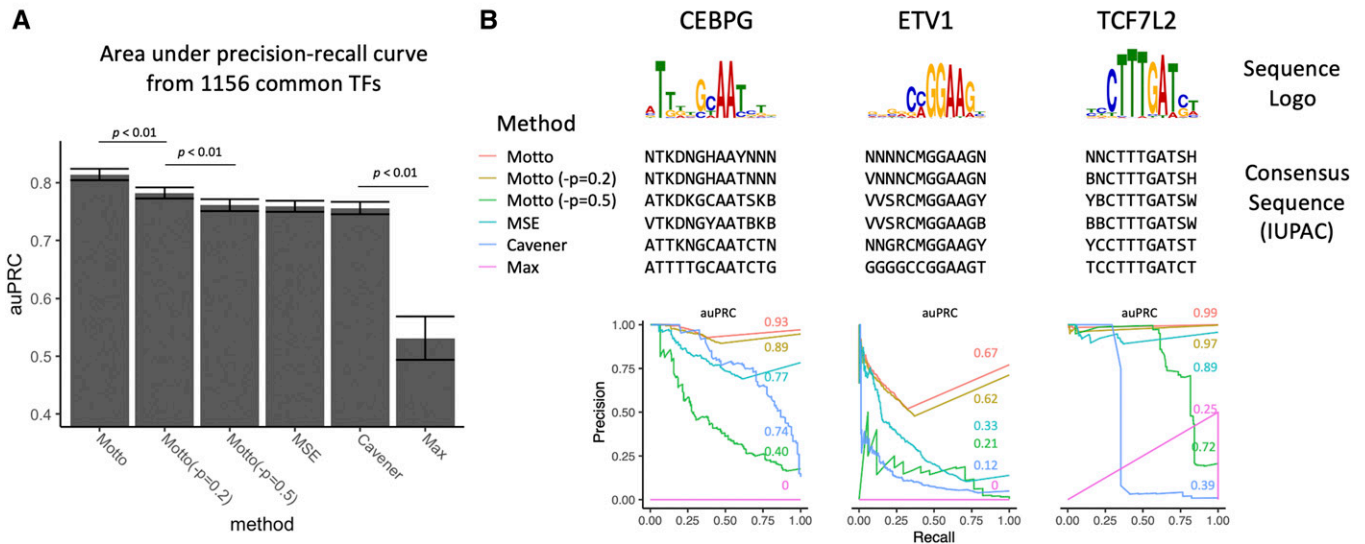


Figure 3 Converted sequence Mottos recapitulate motif occurrence sites of 1156 common human and mouse transcription factors (TFs) in the human genome (hg19). (A) The averaged area under the precision-recall curve (auPRC) using Motto (default method with minimal JSD, ambiguity penalty at $-P = 0.2$, and at $-P = 0.5$) compared with existing alternative methods. P -value determined by paired t -test. (B) Comparison in three examples TFs showing the differences of consensus sequences [shown in IUPAC (Johnson 2010) coding for better alignment] and performances.

(3) “compact” (the default) is the same as “regex,” except that it replaces [ACGT] with N. To trim off Ns ([ACGT]s) at both ends of the output sequences, an optional flag “-trim” is provided. If the users prefer consensus with more certainty (e.g., prefer [AC] to [ACG]), they can use either “-maxCharacter” as a hard limit to the number of characters allowed, or use “-penalty” to penalize ambiguity (see *Materials and Methods*).

Effects of these options are shown using an example of human TF CTCF (Figure 2, upper panel). Unsurprisingly, MSE, Motto, and Cavener are more representative than the naive maximal probability methods. For example, positions 1, 2, 3 and 20 with low information content (<0.2) in CTCF, are justifiably called as “N” by Motto and Cavener, which is an improvement over strictly calling the top nucleotide. MSE considers [TCG] and [GAT] more representative at the first and the third position but agrees with Motto and Cavener at the 2nd and 20th. Similarly, Motto, MSE, and Cavener successfully capture strong double-consensus patterns at indices 7, 11, 12 and 16, which maximal probability fails to capture. The advantage of Motto over Cavener is noticeable at index 6, where the logo of CTCF shows a dominating AG consensus. While Motto finds this co-consensus, Cavener disregards G that barely misses the cutoff. In addition, at index 19, the logo of CTCF shows a strong three-way split among A, G and C, but Cavener, by its rules (as described previously), ignores all such triple patterns. In addition, among the four methods, only the Motto and MSE are capable of generating consensus sequences for amino acid motifs (Bailey and Elkan 1994) (Figure 2, lower panel). Due to its arbitrary nature, heuristic methods like Cavener have difficulties defining decision boundaries for motifs of more than four nucleotides or customized character sets found in motifs containing methylated

DNA (Ngo *et al.* 2019a) and phosphorylated amino acids (Amanchy *et al.* 2011). In such cases, Motto and MSE provide more mathematically rigorous information than Cavener and oversimplified maximal consensus methods. With increased penalty level at 0, 0.2, 0.5, and 1 respectively, the consensus sequence smoothly progresses toward single nucleotide consensus (Figure 2). Such flexibility gives an advantage to users that are biased toward more defined consensus results.

To quantify how well these four methods summarize the information in the original PWMs, we converted 1156 common human and mouse TFs to consensus sequences and compared their matched occurrences (by regular expression) in the human genome (hg19) with conventional motif sites scanned by FIMO (Grant *et al.* 2011) with PWMs, which is how conventionally motif sites are determined (see *Materials and Methods*). We observe that using the Motto method has resulted in the best (0.81 ± 0.01) area under the Precision-Recall curve (auPRC), significantly (P -value < 0.01) better than existing alternative methods, including MSE (0.76 ± 0.01), Cavener (0.76 ± 0.01), and maximal frequency (0.53 ± 0.04) (Figure 3). In addition, we observed Motto performs better with lower ambiguity penalty, where the default setting with minimal ambiguity penalty ($-P = 0$) performs significantly (P -value < 0.01) better than setting penalty at 0.2 (0.78 ± 0.01) or 0.5 (0.76 ± 0.01) (Figure 3). This is consistent with the finding that setting the ambiguity penalty to the maximal value of 1 ($-P = 1$) is equivalent to using the max frequency method, resulting in the worst performance. These results confirm that Motto conversion minimizes information discrepancy (per JSD) from the original PWM, while setting larger penalties will

result in more determined sequences at the cost of accurately recapitulating the original PWM.

In summary, Motto provides a mathematical framework and a set of convenient features to textualize PWMs in a compact, intuitive and accurate manner.

Acknowledgments

This work was partially supported by National Institutes of Health (NIH) (U54HG006997 R01HG009626) and California Institute for Regenerative Medicine (CIRM) (RB5-07012). The authors declare no conflict of interest.

Author contributions: M.W. conceived the idea, implemented the package, performed the analyses, and wrote the manuscript; D.W. implemented key aspects of the package, contributed to the data analysis and manuscript preparation; K.Z., V.N., and S.F. contributed to data analysis and manuscript preparation; W.W. supervised the analyses of the project and contributed to manuscript preparation.

Literature Cited

- Amanchy, R., K. Kandasamy, S. Mathivanan, B. Periaswamy, R. Reddy *et al.*, 2011 Identification of novel phosphorylation motifs through an integrative computational and experimental analysis of the human phosphoproteome. *J. Proteomics Bioinform.* 4: 22–35. <https://doi.org/10.4172/jpb.1000163>
- Bailey T. L., and C. Elkan, 1994 Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *Proc Int Conf Intell Syst Mol Biol.* 2: 28–36
- Bailey, T. L., M. Boden, F. A. Buske, M. Frith, C. E. Grant *et al.*, 2009 MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37: W202–W208. <https://doi.org/10.1093/nar/gkp335>
- Cavener, D. R., 1987 Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Res.* 15: 1353–1361. <https://doi.org/10.1093/nar/15.4.1353>
- Davis, J., and M. Goadrich, 2006 The Relationship Between Precision-Recall and ROC Curves, pp. 233–240 in *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*. ACM, New York. 10.1145/1143844.1143874 <https://doi.org/10.1145/1143844.1143874>
- Fletez-Brant, C., D. Lee, A. S. McCallion, and M. A. Beer, 2013 kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic Acids Res.* 41: W544–W556. <https://doi.org/10.1093/nar/gkt519>
- Ghandi, M., D. Lee, M. Mohammad-Noori, and M. A. Beer, 2014 Enhanced regulatory sequence prediction using gapped k-mer features. *PLOS Comput. Biol.* 10: e1003711 (erratum: *PLOS Comput. Biol.* 10: e1004035). <https://doi.org/10.1371/journal.pcbi.1003711>
- Grant, C. E., T. L. Bailey, and W. S. Noble, 2011 FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27: 1017–1018. <https://doi.org/10.1093/bioinformatics/btr064>
- Guo, Y., K. Tian, H. Zeng, X. Guo, and D. K. Gifford, 2018 A novel k-mer set memory (KSM) motif representation improves regulatory variant prediction. *Genome Res.* 28: 891–900. <https://doi.org/10.1101/gr.226852.117>
- Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin *et al.*, 2010 Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38: 576–589. <https://doi.org/10.1016/j.molcel.2010.05.004>
- Johnson, A. D., 2010 An extended IUPAC nomenclature code for polymorphic nucleic acids. *Bioinformatics* 26: 1386–1389. <https://doi.org/10.1093/bioinformatics/btq098>
- Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle *et al.*, 2002 The human genome browser at UCSC. *Genome Res.* 12: 996–1006. <https://doi.org/10.1101/gr.229102>
- Kulakovskiy, I. V., I. E. Vorontsov, I. S. Yevshin, R. N. Sharipov, A. D. Fedorova *et al.*, 2018 HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* 46: D252–D259. <https://doi.org/10.1093/nar/gkx1106>
- Lele, S., 1993 Euclidean Distance Matrix Analysis (EDMA): estimation of mean form and mean form difference. *Math. Geol.* 25: 573–602. <https://doi.org/10.1007/BF00890247>
- Lin, J., 1991 Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* 37: 145–151. <https://doi.org/10.1109/18.61115>
- Matys, V., O. V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land *et al.*, 2006 TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 34: D108–D110. <https://doi.org/10.1093/nar/gkj143>
- Murray, J. I., R. B. Voelker, K. L. Henscheid, M. B. Warf, and J. A. Berglund, 2008 Identification of motifs that function in the splicing of non-canonical introns. *Genome Biol.* 9: R97. <https://doi.org/10.1186/gb-2008-9-6-r97>
- Ngo, V., Z. Chen, K. Zhang, J. W. Whitaker, M. Wang *et al.*, 2019b Epigenomic analysis reveals DNA motifs regulating histone modifications in human and mouse. *Proc. Natl. Acad. Sci. USA* 116: 3668–3677. <https://doi.org/10.1073/pnas.1813565116>
- Ngo, V., M. Wang, and W. Wang, 2019a Finding de novo methylated DNA motifs. *Bioinformatics* 35: 3287–3293. <https://doi.org/10.1093/bioinformatics/btz079>
- Portales-Casamar, E., S. Thongjuea, A. T. Kwon, D. Arenillas, X. Zhao *et al.*, 2010 JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 38: D105–D110. <https://doi.org/10.1093/nar/gkp950>
- Robasky, K., and M. L. Bulyk, 2011 UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic Acids Res.* 39: D124–D128. <https://doi.org/10.1093/nar/gkq992>
- Robinson, J. T., H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander *et al.*, 2011 Integrative genomics viewer. *Nat. Biotechnol.* 29: 24–26. <https://doi.org/10.1038/nbt.1754>
- Schneider, T. D., and R. M. Stephens, 1990 Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18: 6097–6100. <https://doi.org/10.1093/nar/18.20.6097>
- Thompson, W., E. C. Rouchka, and C. E. Lawrence, 2003 Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.* 31: 3580–3585. <https://doi.org/10.1093/nar/gkg608>
- Wang, M., K. Zhang, V. Ngo, C. Liu, S. Fan *et al.*, 2019 Identification of DNA motifs that regulate DNA methylation. *Nucleic Acids Res.* 47: 6753–6768.
- Whitaker J. W., Z. Chen, and W. Wang, 2015 Predicting the human epigenome from DNA motifs. *Nat. Methods* 12: 265–272, 7 p following 272. <https://doi.org/10.1038/nmeth.3065>
- Xie, Z., S. Hu, S. Blackshaw, H. Zhu, and J. Qian, 2010 hPDI: a database of experimental human protein–DNA interactions. *Bioinformatics* 26: 287–289. <https://doi.org/10.1093/bioinformatics/btp631>
- Zeng, H., T. Hashimoto, D. D. Kang, and D. K. Gifford, 2016 GERV: a statistical method for generative evaluation of regulatory variants for transcription factor binding. *Bioinformatics* 32: 490–496. <https://doi.org/10.1093/bioinformatics/btv565>

Communicating editor: D. Nielsen