Supplementary to "Improved similarity scores for comparing motifs"

Emi Tanaka, Timothy Bailey, Charles E. Grant, William Stafford Noble, and Uri Keich

1 Asymptotic analysis of the BLiC score

Claim 1. Let $\{T_N\}_N$ be a sequence of (target) motif columns where for each N, T_N has N sites. Assume that as $N \to \infty$, $\hat{P}^{T_N} \to \psi$ where $\psi \neq \pi$ (the background distribution). Then $S_{\mathrm{BLiC}}(Q, T_N) \geq N\varepsilon$ for some $\varepsilon > 0$ and all sufficiently large N.

Proof. We first show that $|S_1|$ is bounded. Let $n_a := N_{T_a}$, $m_a := N_{Q_a}$, $N := N_Q$ and $M := N_T$. Then, ignoring the issue of pseudo counts due to the Dirichlet prior (merely to simplify the presentation), we have

$$\begin{aligned} S_1 &= \sum_a (n_a + m_a) \log \frac{n_a + m_a}{N + M} - n_a \log \frac{n_a}{N} - m_a \log \frac{m_a}{M} \\ &= \underbrace{\sum_a n_a \log \left(\frac{n_a + m_a}{n_a} \frac{N}{N+M} \right)}_{S_{11}} + \underbrace{\sum_a m_a \log \left(\frac{n_a + m_a}{N+M} \frac{M}{m_a} \right)}_{S_{12}}. \end{aligned}$$

It is clear that $|S_{12}|$ is bounded (recall that Q and therefore M and m_a are all fixed). Assuming $\psi_a \neq 0$ (otherwise we can ignore a) $n_a \to \infty$ and therefore

$$x_N := \frac{n_a + m_a}{n_a} \frac{N}{N+M} - 1 = \frac{1}{n_a} \left[n_a \left(\frac{N}{N+M} - 1 \right) + m_a \frac{N}{N+M} \right] = \frac{1}{n_a} \cdot O(1) \quad \text{as } N \to \infty.$$

Therefore,

$$S_{11} = \sum_{a \in \mathcal{A}} n_a \left[x_N + O(X_N^2) \right] = \sum_a O(1) + \frac{1}{n_a} \cdot O(1) = O(1) \qquad \text{as } N \to \infty.$$

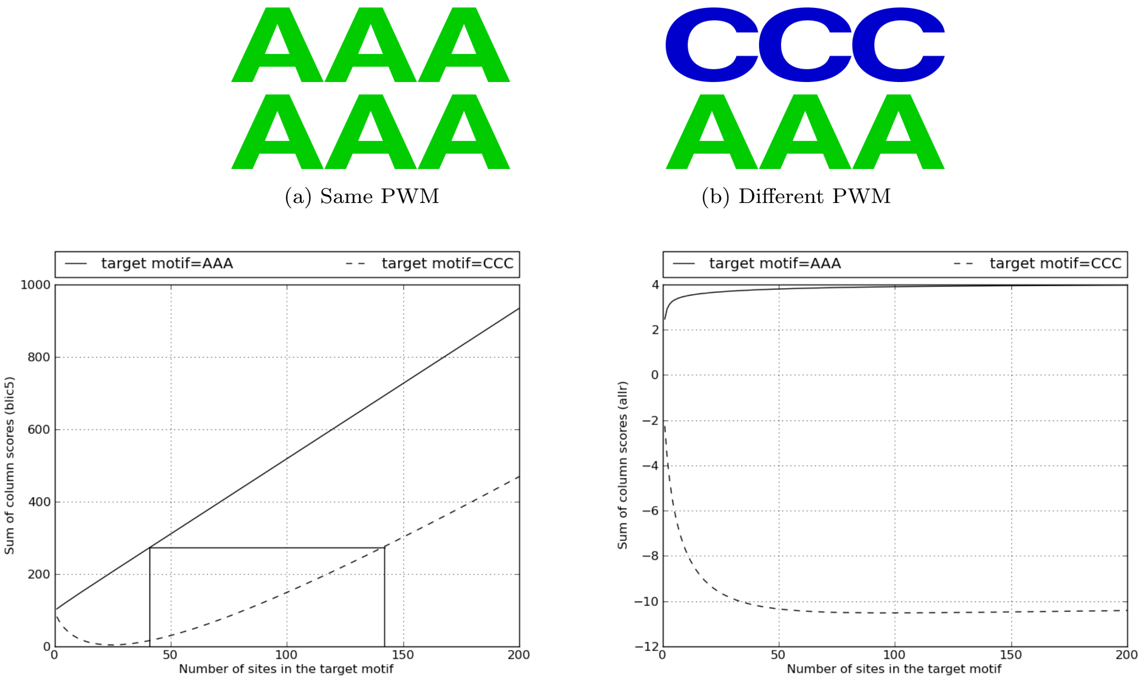
Hence $S_1 = S_{11} + S_{12}$ is bounded from above and below as $N \to \infty$. As for S_2, we have

$$S_2 = (N+M) \sum_{a \in \mathcal{A}} \frac{n_a + m_a}{N+M} \log \left(\frac{n_a + m_a}{N+M} \Big/ \pi_a \right) = (N+M) KL(\hat{P}^{Q,T_N} || \pi),$$

where $KL(P||P')$ stands for the (asymmetric) Kullback-Leibler divergence or the relative entropy of P with respect to P'. Since M is fixed and $N \to \infty$ it is clear that $\hat{P}^{Q,T_N} \to \psi$ as $N \to \infty$, and since we assumed that $\psi \neq \pi$ it follows that there exists $\varepsilon > 0$ such that $KL(\hat{P}^{Q,T_N} || \pi) > \varepsilon$ for all N sufficiently large. It follows that for all such N, $S_2 \geq (N+M)\varepsilon$ and since $|S_1|$ is bounded the claim follows (possibly with a different ε). □

It is obvious from the proof that $\hat{P}^{T_N}$ need not converge for the corollary to hold. Indeed, it suffices that $\hat{P}^{T_N}$ remains a certain minimal distance away from π for all sufficiently large N.

Supplementary Figure 1 below provides a visual confirmation of the asymptotic bias proved in the last claim.



(a) Same PWM

(b) Different PWM

(c) BLiC

(d) ALLR

Figure 1: **Plot of motif score vs number of sites**
Plots of the motif score as a functions of the number of sites in the target motif. The query motif is a fixed three all-A columns comprised of 25 sites. The BLiC similarity score of the query motif and an identical PWM comprised of 40 sites is roughly 280. This is also roughly the BLiC similarity score of the same query and an *unrelated* PWM comprised of 140 sites. With a higher number of sites the BLiC score would prefer the unrelated target PWM over the PWM which is identical to the target! The BLiC similarity score is defined here using the mixture of Dirichlet priors as described in [2].

2 Generating a null target motif under HMM

To generate a null target motif under our HMM, we generate a sequence of $|\boldsymbol{T}|+2$ states starting at the silent state I and ending at the silent state T. At each step we sample the next state according to its conditional probability given the current state and given that we have to be in state T at step $|\boldsymbol{T}|+2$. This conditional probability can be determined as follows.

$$\begin{aligned}
&P(S_{i+1} = s_{i+1} \big| S_{1:i} = s_{1:i}, S_{|\boldsymbol{T}|+2} = T) \\
&\quad = \frac{P(S_{i+1} = s_{i+1}, S_{1:i} = s_{1:i}, S_{|\boldsymbol{T}|+2} = T)}{P(S_{1:i} = s_{1:i}, S_{|\boldsymbol{T}|+2} = T)} \\
&\quad = \frac{P(S_{i+1} = s_{i+1}, S_{i} = s_{i}, S_{|\boldsymbol{T}|+2} = T)}{P(S_{i} = s_{i}, S_{|\boldsymbol{T}|+1} = T)} \\
&\quad = \frac{P(S_{|\boldsymbol{T}|+2} = T \big| S_{i+1} = s_{i+1}) P(S_{i+1} = s_{i+1} \big| S_i = s_i) P(S_i = s_i)}{P(S_{|\boldsymbol{T}|+2} = T \big| S_i = s_i) P(S_i = s_i)} \\
&\quad = \frac{P(S_{|\boldsymbol{T}|+2} = T \big| S_{i+1} = s_{i+1}) P(S_{i+1} = s_{i+1} \big| S_i = s_i)}{P(S_{|\boldsymbol{T}|+2} = T \big| S_i = s_i)}
\end{aligned}$$

where S_i is the state of the i-th column and $S_{1:i}$ is the states of the first i columns.

3 P-value Accuracy Assessment

Using MC sampling to calculate the overall p-value guarantees that the p-values will be fairly accurate for sufficiently large samples. However, the independent alignments assumption that we make when using the DP approach to calculate the overall p-value might compromise the accuracy of these p-values. We therefore need to verify that the p-values assigned to target motifs sampled according to the null iid model do indeed look like a sample from a uniform-$(0, 1)$ distribution.

To test the uniformity of null p-values, we adopt a similar setup to the one used by Gupta et al. [1]. We select a random motif from the mouse PBM database and then query it against a random shuffle of the columns of the remaining 385 motif in the database, while keeping the distribution of database motif lengths unchanged. Repeating this process 1000 times produces a sample of 385,000 null DP-computed p-values. The resulting quantile-quantile plot can be viewed in Supplementary Figure 2, where the vertical axis corresponds to ordered computed motif p-values and the horizontal axis corresponds to the rank p-values.

4 Performance of the complete score for different quantiles

The complete scores defined in (3) assign the median, or the 0.5 quantile of the null target column similarity score to each unaligned column. We varied this quantile to gauge the sensitivity of our approach in terms of both retrieval accuracy and filtering of uninformative alignment.

Supplementary Table 1 confirms that the mean AUC is fairly robust to variations in the quantile within a fairly large range of 0.3-0.7. It is evident from Supplementary Figure 3 that there is, as expected, increased filtering with a decreasing quantile. However, for all quantiles in the studied range of 0.3-0.7 the filtering is significantly more effective than that offered by any of the raw or incomplete scores in Supplementary Figure 4: up to FPR of 0.1 there are very little to no uninformative alignments using any of the considered quantiles.

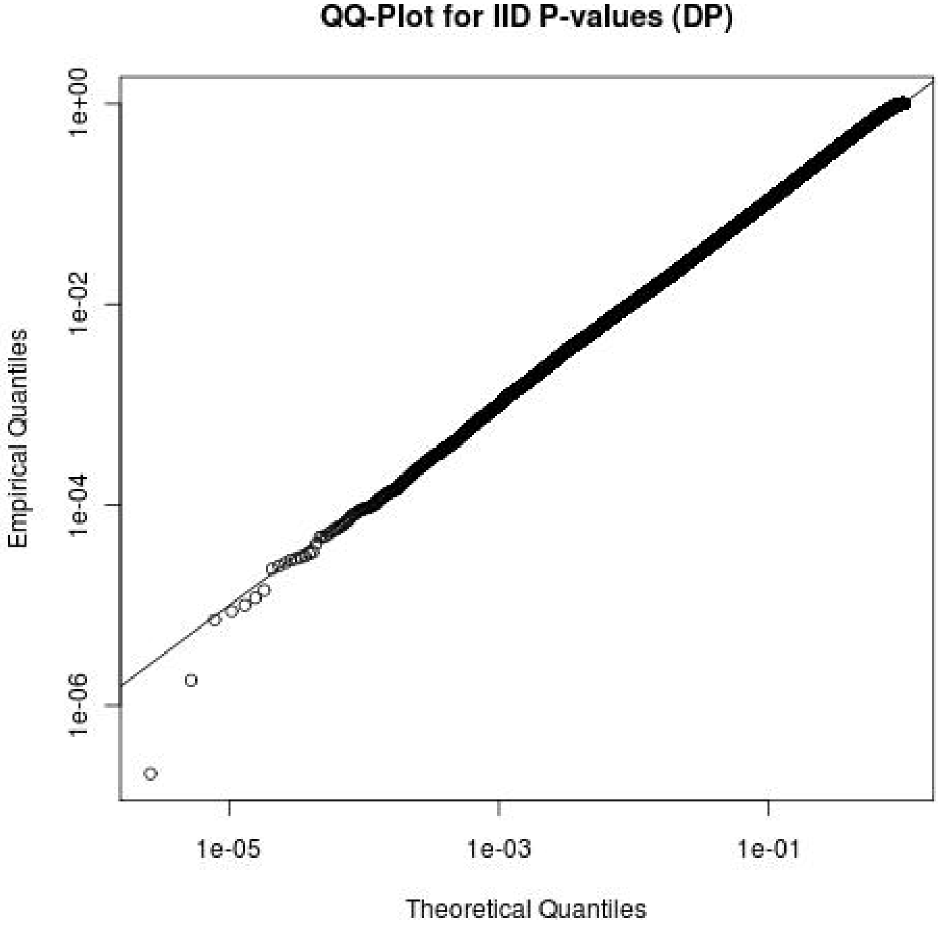
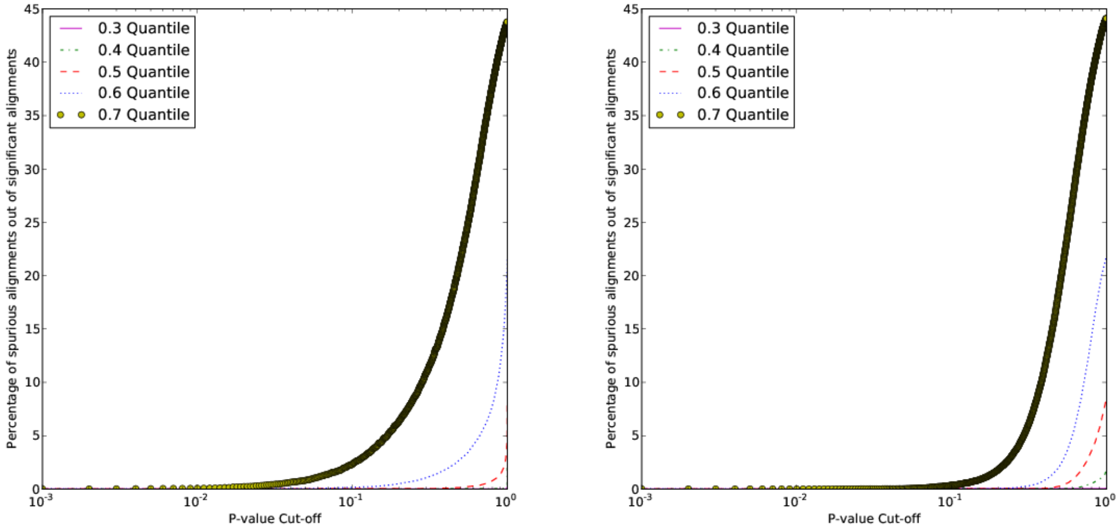


Figure 2: **Accuracy of DP computed p-values.** A qq-plot of the computed p-value against the rank p-value which is the rank of the p-value in the list of all p-values divided by the number of p-values or 385,000. The DP computed p-values using complete-ED were generated from comparison of random queries from the mouse PBM database against the shuffled target databases. The straight line corresponds to $y = x$.

Table 1: **Mean AUC using different quantiles.** The mean AUC computed using different quantiles for the complete-ED with HMM p-values and DP computed iid p-values.

Quantile	HMM P-value (MC)	iid P-value (DP)
0.3	0.9874	0.9993
0.4	0.9871	0.9994
0.5	0.9869	0.9994
0.6	0.9870	0.9995
0.7	0.9877	0.9995



(a) iid p-value (DP)

(b) HMM p-value (MC)

Figure 3: **Percentage of uninformative alignments using different quantiles.** Comparison of the effects of variation in the quantile on removing uninformative alignments for quantile in the range of 0.3 to 0.7. All test were done using complete-ED and with overall p-values computed as indicated.

The design of the experiments whose results are summarized below is the same as described in Sections 5.1 and 5.2 (main paper).

5 Retrieval Accuracy Results with TRANSFAC

To show that our results in Section 5.2 (main paper) are not unique to the mouse PBM database we conducted a similar experiment with the TRANSFAC database. Again, we randomly select 80 template query motifs only this time from the TRANSFAC database and proceed exactly as described in Section 5.2 except that the same number of sites for a motif is fixed for all randomized versions of the template by randomly choosing one of 5, 10 or 20 sites. This process yields a total of 560 motifs which we query against the TRANSFAC database. The AUC is calculated from the ROC curve for each query motif and the averaged AUC is reported in Supplementary Table 2.

Consistently with Table 2 (main paper), all the target functions which compute the p-values using MC sampling are ranked lowest, although by a smaller margin here. The results for the other target functions are also consistent with the ALLR producing the lowest mean AUC and little to no variation among other target functions.

6 Retrieval Accuracy Results for Compromised Target Motifs

In general we believe that database motifs are less likely to omit key motif columns. Nevertheless, it is important to test the effects such compromised motifs have on the retrieval accuracy of the complete scores. In order to simulate motifs that are missing key columns without generating artificially short target motifs we randomly extended query motifs by adding columns sampled from the target database.

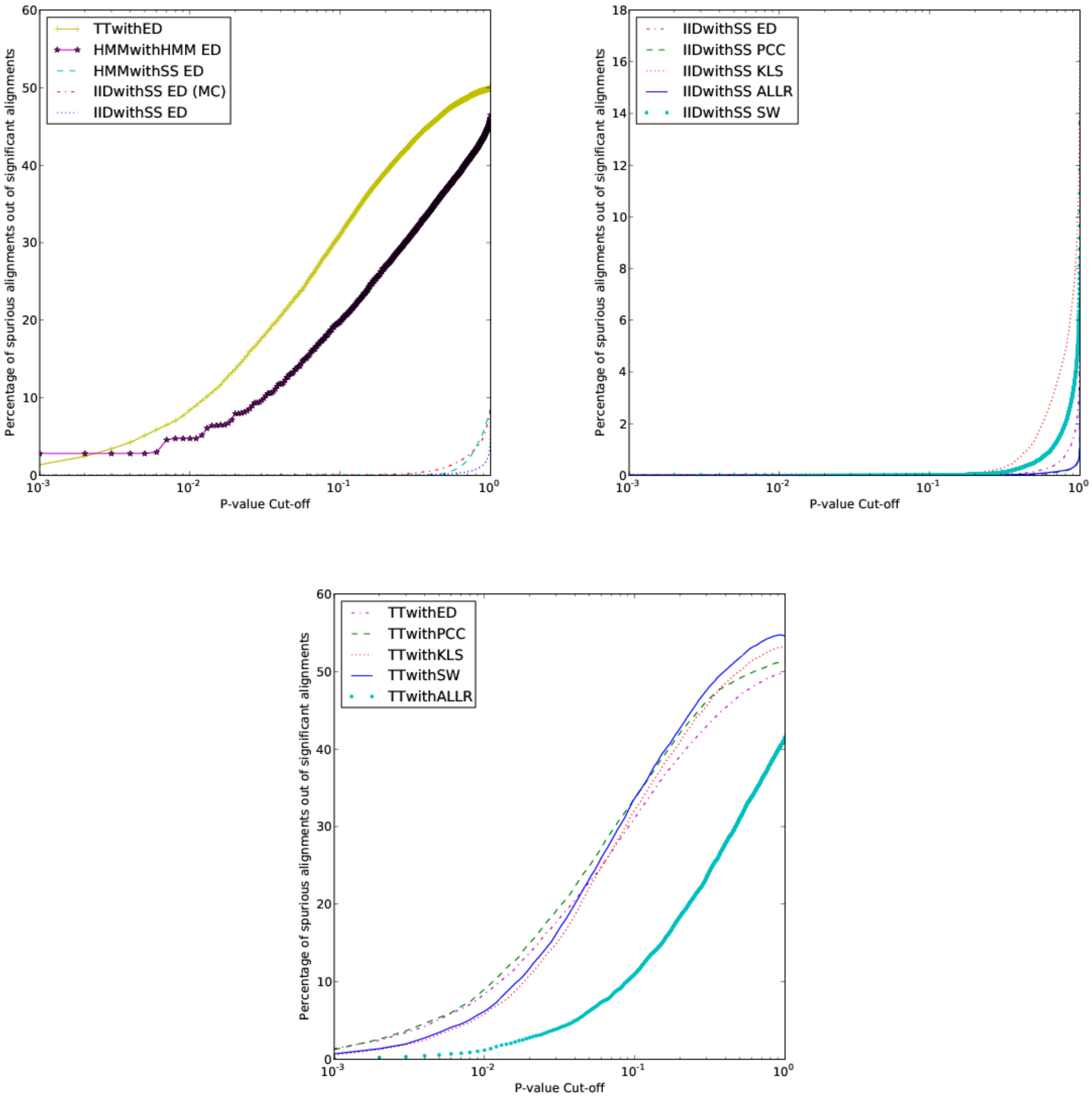


Figure 4: **Percentage of uninformative alignments vs p-value cut-off** Plots of the percentage of significant alignments that are uninformative for several combinations of target functions. All plots are generated by querying the mouse PBM database against itself using the specified target function (column similarity and an overall p-value estimation). TTwithX corresponds to results reported by Tomtom using similarity score X. HMMwithHMM ED uses the ED score with the HMM to evaluate the offset p-values as well as to assign an overall p-value. HMMwithSS ED uses the complete-ED with the HMM estimated overall p-values. IIDwithSS ED (MC) and IIDwithSS ED use the complete-ED with p-values computed by MC sampling and the DP approach, respectively. Similar notations apply to IIDwithSS for PCC, KLS, ALLR, and SW.

Column Score	Optimal Alignment Selection	Overall p-value	Mean AUC
ED	iid offset p-value	ind alignments (TT)	0.9997
PCC	iid offset p-value	ind alignments (TT)	0.9993
ALLR	iid offset p-value	ind alignments (TT)	0.9972
KLS	iid offset p-value	ind alignments (TT)	0.9997
SW	iid offset p-value	ind alignments (TT)	0.9996
ED	HMM offset p-value	HMM (MC)	0.9914
Complete-ED	Motif Score	HMM (MC)	0.9969
Complete-ED	Motif Score	iid (MC)	0.9962
Complete-ED	Motif Score	ind alignments (DP)	0.9998
Complete-PCC	Motif score	ind alignments (DP)	0.9996
Complete-ALLR	Motif score	ind alignments (DP)	0.9983
Complete-KLS	Motif score	ind alignments (DP)	0.9997
Complete-SW	Motif score	ind alignments (DP)	0.9997

Table 2: **Retrieval accuracy**. See Table 2 (main paper) for the description of the methods. The above results are mean AUC with sample query database composed from the TRANSFAC database. For comparison, using the BLiC score gives an AUC of 0.6967 which is consistent with the fact that TRANSFAC has some deep motifs.

We applied our test procedure to three different motif databases: MacIsaac's set of yeast motifs (referred to as the Yeast database), the mouse PBM database and TRANSFAC. A key feature of these experiments was the extension of the queries with an increasing number of sampled database columns. In addition, to broaden the scope of our test, we generated queries here by sampling motif sites rather than motif columns as we did in the previously described experiments[1]. The precise protocol for generating the sites follows.

For the Yeast database we select all motifs that have more than 5 sites. For each of these motifs we sample with replacement 5 or 10 sites. We then change the coverage for all motifs in a similar fashion described in Section 5.2 (main paper) by removing up to 30% of the columns from one end. At the other end we add columns sampled without replacement from the Yeast database (including reverse complements). The number of added columns is determined by drawing uniformly from 1 to $x\%$ of the original length of the motifs, where x is a parameter that we varied from 10 to 100. This yields a total of 178 motifs[2] to query against the Yeast database.

The experimental setup for mouse PBM database and TRANSFAC are similar except we randomly select 200 motifs from the database and the columns are determined by sampling 5 or 10 nucleotides from a multinomial distribution with weights given by the column's frequencies. This gives a total of 400 motifs which we query against the database that contains the originally selected motif.

Supplementary Table 3 reports the mean AUC results, calculated as described in Section 5.2 (main paper), for these three datasets with different values of x.

7 Column Scores

The following is the list of column scores used in this paper. Q_a is the frequency of the letter $a \in \mathcal{A}$ in column Q (similarly for column T) while N_{Q_a} is the count of a in the column.

[1] Of course, the extension of the queries was still done by sampling columns.

[2] Doubling the number of query motifs gave similar results.

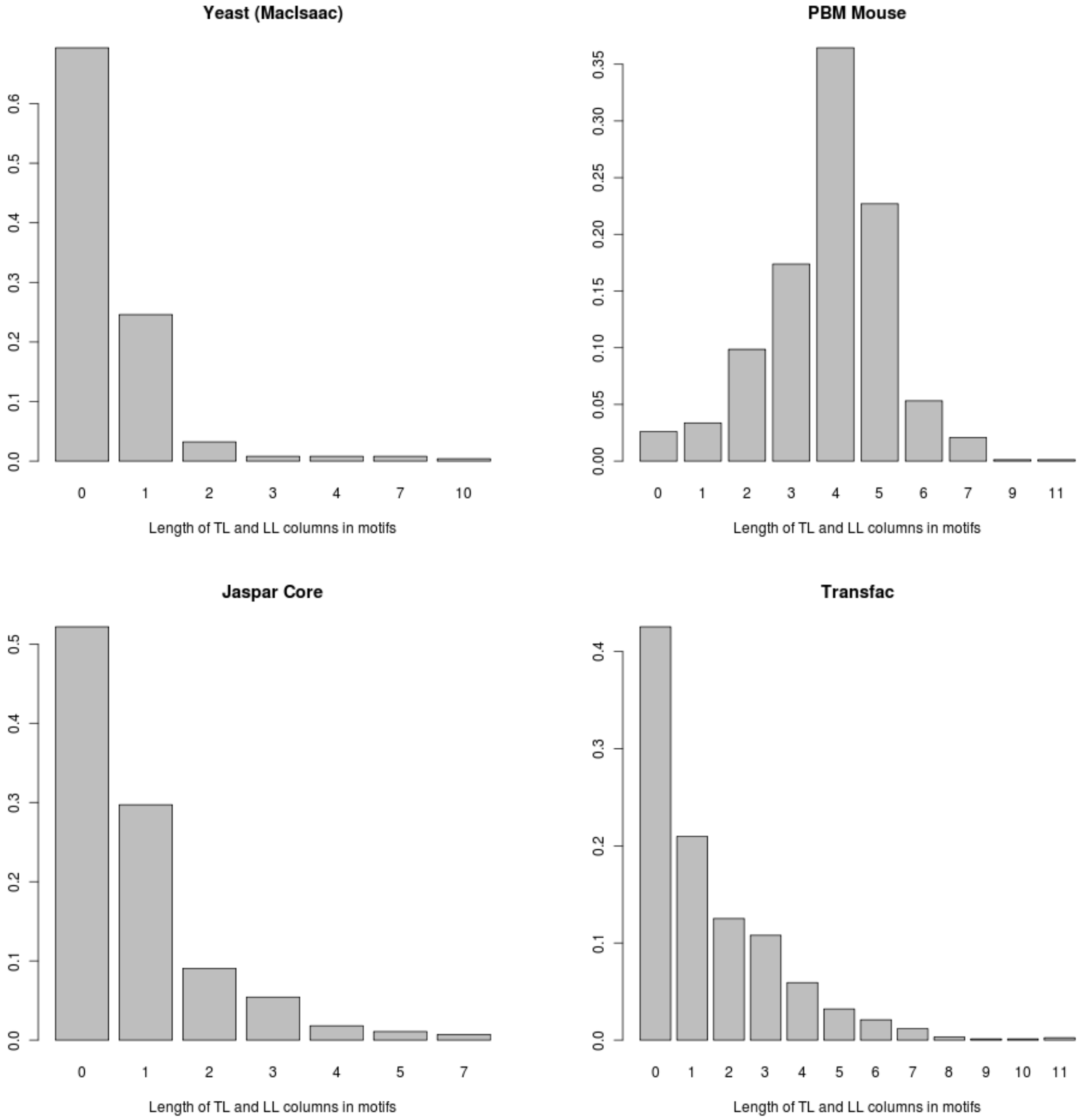


Figure 5: **The distribution of the consecutive number of uninformative columns found at the ends of database motifs.**

The bar plots indicate the proportion of each length of leading or trailing stretches of uninformative columns (information content lower than 0.5). These histograms demonstrate the potential problem of uninformative alignments in commonly used motif databases. While the mouse PBM database is particularly exposed to this kind of spurious alignments, all databases have some motifs containing uninformative ends which could result in alignments such as in Figure 2 (main paper).

Table 3: **Retrieval accuracy**. See Table 2 for the description of the methods. Below x is the maximum percentage of columns added on one end of the query motifs out of the length of the motifs that served as the template.

$x\%$	Yeast		mouse PBM		TRANSFAC	
	ED	complete-ED	ED	complete-ED	ED	complete-ED
10	0.9990	0.9985	0.9990	0.9989	0.9981	0.9982
20	0.9991	0.9982	0.9987	0.9986	0.9994	0.9990
30	0.9991	0.9975	0.9986	0.9986	0.9990	0.9988
40	0.9989	0.9964	0.9985	0.9985	0.9994	0.9989
50	0.9988	0.9961	0.9978	0.9976	0.9991	0.9987
60	0.9990	0.9964	0.9985	0.9979	0.9996	0.9989
70	0.9971	0.9914	0.9989	0.9987	0.9996	0.9979
80	0.9977	0.9903	0.9985	0.9980	0.9991	0.9978
90	0.9965	0.9873	0.9980	0.9966	0.9994	0.9983
100	0.9968	0.9873	0.9989	0.9985	0.9980	0.9976

7.1 Euclidean Distance

$$S_{\mathrm{ED}}(Q,T) = -\sqrt{\sum_{a\in\mathcal{A}}(Q_a - T_a)^2}$$

7.2 Pearson Correlation Coefficient

$$S_{\mathrm{PCC}}(Q,T) = \frac{\sum_{a\in\mathcal{A}}(Q_a-\bar{Q})(T_a-\bar{T})}{\sqrt{\sum_{a\in\mathcal{A}}(Q_a-\bar{Q})^2\sum_{a\in\mathcal{A}}(T_a-\bar{T})^2}}$$

where

$$\bar{Q} = \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} Q_a$$

$$\bar{T} = \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} T_a$$

Since we work with DNA motifs, $|\mathcal{A}| = 4$. Also $\sum_{a\in\mathcal{A}} Q_a = 1$ and $\sum_{a\in\mathcal{A}} T_a = 1$ so $\bar{Q} = \bar{T} = \frac{1}{4}$.

7.3 Average log-likelihood ratio [4]

$$S_{\mathrm{ALLR}}(Q,T) = \sum_{a\in\mathcal{A}}\left(\frac{N_{Q_a}\log\left(\frac{T_a}{\pi_a}\right) + N_{T_a}\log\left(\frac{Q_a}{\pi_a}\right)}{\sum_{a\in\mathcal{A}}(N_{Q_a}+N_{T_a})}\right)$$

7.4 Sandelin-Wasserman [3]

$$S_{\mathrm{SW}}(Q,T) = 2 - \sum_{a\in\mathcal{A}}(Q_a - T_a)^2$$

7.5 Symmetric Kullback-Leibler divergence

$$S_{\mathrm{KLS}}(Q,T) = \frac{1}{2}\left(\sum_{a\in\mathcal{A}} Q_a \log\left(\frac{Q_a}{T_a}\right) + \sum_{a\in\mathcal{A}} T_a \log\left(\frac{T_a}{Q_a}\right)\right)$$

References

[1] Gupta, S., J. Stamatoyannopoulos, T. Bailey, and W. Noble (2007, February). Quantifying similarity between motifs. *Genome Biology 8*(2).

[2] Habib, N., T. Kaplan, H. Margalit, and N. Friedman (2008, February). A novel Bayesian DNA motif comparison method for clustering and retrieval. *PLoS computational biology 4*(2).

[3] Sandelin, A. and W. W. Wasserman (2004, April). Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *Journal of molecular biology 338*(2), 207–15.

[4] Wang, T. and G. D. Stormo (2003). Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics 19*(18), 2369–2380.

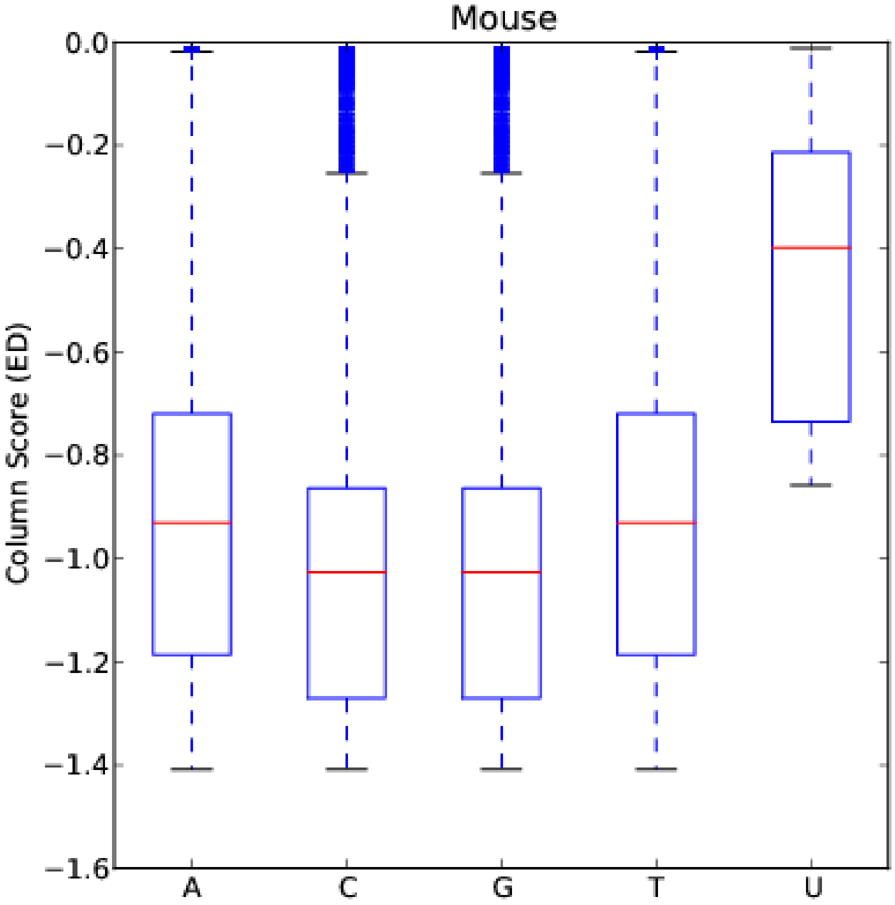


Figure 6: **Column score distribution.** Boxplots of the null alignment column scores of five different columns: all-A, all-C, all-G, all-T and a uniformly distributed column. The column similarity score is ED, and $\mathbb{T}$, the set of null database columns, is derived from from the mouse PBM database.