Detecting Sequence Motifs Associated with Continuous Biological Traits

Practical statistics using R, 88895 - Final Assignment

Ariel Bernstein 316072685

Introduction

Short sequence motifs—recurring nucleotide patterns within genomic or transcriptomic sequences—are known to influence a wide range of biological processes. These include transcriptional regulation [1], RNA stability [2], alternative splicing, mRNA localization, and translation efficiency, among others. Identifying such motifs is crucial to understanding gene regulation, interpreting high-throughput sequencing data, and discovering novel functional elements in the genome.

Over the past decades, several computational tools have been developed to identify overrepresented motifs in biological sequences [3]. One of the most widely used is the MEME Suite, which includes tools such as MEME [4] and DREME [5]. These approaches have proven highly effective in cases where the data can be separated into two distinct groups—typically referred to as "positive" and "negative" sets.

However, many biological traits are not naturally binary but rather are measured on a continuous scale. In such cases, researchers often convert continuous measurements into binary labels (for example, "high" vs. "low") to enable the use of existing tools. This process, while convenient, involves arbitrary thresholding and may lead to the loss of important information and statistical power.

In this project, I explore initial directions toward a computational framework for detecting motifs associated with continuous biological traits. The work focuses on several methodological challenges: the choice of an appropriate statistical test (KS versus Mann–Whitney), maintaining statistical power under multiple testing, preventing cases where statistical significance reflects biologically negligible effects, and developing strategies to cluster and visualize the large number of candidate motifs in a biologically interpretable way.

Methods

Cliff’s delta

Cliff’s delta is a non-parametric effect size statistic defined as $\delta = P(X > Y) - P(X < Y)$ for two independent random variables X and Y [6]. Its values range between -1 and +1, with $\delta = 0$ indicating no difference between groups. Because it depends only on the relative ordering of values, δ is unaffected by linear rescaling of the data and does not rely on distributional assumptions. In this study, it was used both as an effect size criterion and as a parameter controlling group separation in the simulations.

Simulation 1 – Power analysis of KS versus Mann–Whitney

Data were simulated from two independent normal distributions. The baseline group was sampled from $\mathrm{Y} \sim \mathcal{N}(\mu_0, \sigma_0^2)$ with parameters $\mu_0 = 5$ and $\sigma_0 = 1$. The affected group was sampled from $\mathrm{X} \sim \mathcal{N}(\mu_1, \sigma_1^2)$. The total sample size was fixed at $\mathrm{N} = 1000$, with $\mathrm{n}_{\text{affected}} = 50$ samples drawn from the affected group and $\mathrm{n}_{\text{baseline}} = 950$ samples from the baseline group. For each parameter configuration 1000 Monte Carlo replicates were performed with significance level $\alpha = 0.05$.

Two scenarios were examined. In the variance-shift scenario, the means were held equal ($\mu_1 = \mu_0 = 5$), ensuring that effect size was zero. while the standard deviation of the affected group varied in the range $\sigma_1 = 1.0, 1.1, \ldots, 2.5$. In the location-shift scenario, the variances were fixed ($\sigma_1 = \sigma_0 = 1$) and Cliff’s delta was varied between $\delta = 0.00, 0.05, \ldots, 0.60$. For each value of δ, the mean of the affected group was calculated using $\mu_1 = \mu_0 + \Phi^{-1}\left(\frac{\delta+1}{2}\right)\sqrt{\sigma_0^2 + \sigma_1^2}$ where Φ^{-1} is the quantile function of the standard normal distribution. Test power was defined as $\mathrm{Power} = \frac{\#\{p<0.05\}}{1000}$ and was computed for both the Kolmogorov–Smirnov and Mann–Whitney tests.

Simulation 2 – KS test power under multiple testing (Šidák correction)

The same framework was used to evaluate the performance of the KS test under large-scale multiple testing. The baseline distribution was defined by $\mu_0 = 5$ and $\sigma_0 = 1$, while the affected distribution had $\sigma_1 = 1$ and μ_1 computed from the desired effect size δ using the same formula as above. The total sample size was fixed at $\mathrm{N} = 7000$. Two setups were examined. In **Setup A**, small effect sizes were tested with $\delta \in 0.10, 0.20$ and the number of affected samples varied as $\mathrm{n}_{\text{affected}} = 50, 100, \ldots, 2000$. In **Setup B**, moderate effect sizes were tested with $\delta \in 0.30, 0.40, 0.50$ and the number of affected samples varied as $\mathrm{n}_{\text{affected}} = 5, 10, \ldots, 200$. In both setups, 1000 Monte Carlo replicates were performed with $\alpha = 0.05$. To account for multiple testing, a Šidák correction was applied with K = 21760. Empirical power was defined as $\mathrm{Power} = \frac{\#\mathrm{p}<\alpha_{\text{eff}}}{1000}$.

Simulated datasets

For the analyses in this study, I used data generated by a controlled simulation framework. Artificial gene-like sequences were created with realistic nucleotide composition and dinucleotide transition probabilities derived from a reference genome. Motifs were embedded into a predefined fraction of the sequences, and a continuous biological trait was simulated such that the presence of a motif shifted the trait value relative to a baseline distribution. The simulation output included both the sequences (FASTA format) and their associated continuous trait values.

For the analysis of Cliff's delta, I used a simple dataset in which a single motif was embedded into the simulated sequences. For the evaluation of the full pipeline, I used a dataset containing six distinct motifs, two with positive effect on the trait and four with negative effect. The positive-effect motifs were GCACTT and TATTTA, while the negative-effect motifs were TTTTTT, TTCGC, AATTG, and GGAAGG.

Initial pipeline

The analyses in this work started from an existing workflow developed in the lab by PhD student Leor Fishman, designed to connect k-mer composition with continuous biological traits. For each dataset, a count matrix was first generated, recording the occurrences of all k-mers of a predefined length in every sequence. In the statistical analysis, only k-mer presence or absence was considered by dividing the sequences into two groups: those containing the k-mer and those not containing it. Trait distributions between the two groups were then compared using a two-sample Kolmogorov–Smirnov (KS) test. To account for the large number of k-mers tested, p-values were adjusted with the Benjamini–Hochberg procedure. In addition, an effect size was calculated as the fold change between the mean trait values of the two groups, with the threshold chosen manually based on inspection of the results. Finally, the combined significance and effect size criteria were visualized in a volcano plot.

RESULTS

Initial framework and research questions

As described in the Methods, I started from an existing pipeline developed in the lab by, designed to connect k-mer composition with continuous biological traits through Kolmogorov–Smirnov testing and volcano plot visualization. In this project, I focused on four methodological questions that arose from this framework: (i) whether the KS test is the most appropriate choice compared to alternatives such as Mann–Whitney, (ii) how statistical power behaves under large-scale multiple testing correction, (iii) how to introduce an effect size criterion to ensure that significant results also reflect meaningful biological differences, and (iv) how to cluster and summarize large sets of significant k-mers into interpretable motifs**.**

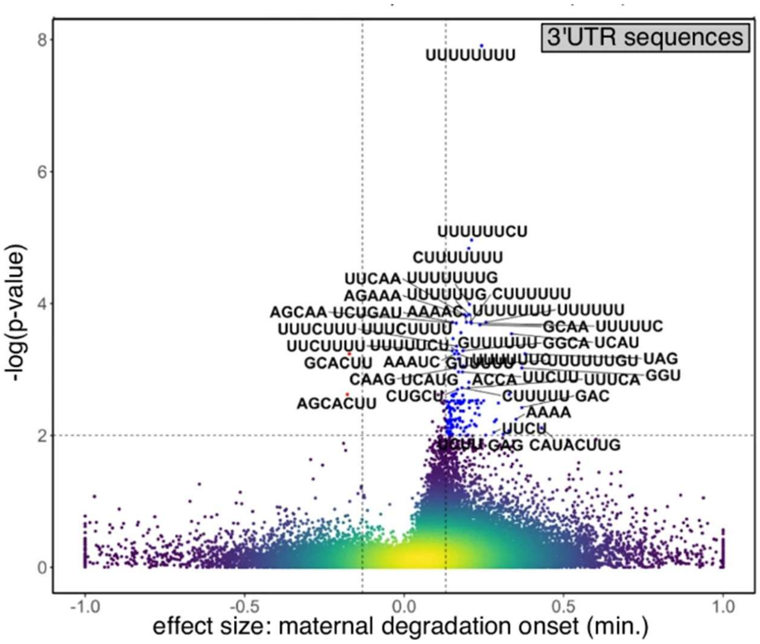

Figure 1. Example volcano plot of k-mer results (adapted from Fishman et al., *Nat Commun*, 2024).
Each point represents a k-mer, with effect size on the x-axis and adjusted p-value on the y-axis. Vertical dashed lines mark the effect-size threshold (chosen manually for this dataset), and the horizontal dashed line marks the significance cutoff. While informative for a global overview, the dense overlap of points makes it difficult to infer underlying motifs.

Choice of test: KS versus Mann–Whitney

To evaluate the relative performance of the Kolmogorov–Smirnov (KS) and Mann–Whitney (MW) tests, we performed power analyses under two conditions (see Methods). In the first, a mean shift was introduced while variance was held constant (Figure 2.A). In the second, the standard deviation was altered while the mean remained fixed, modeling a change in distributional shape without an effect on central tendency (Figure 2.B).

When the difference between groups was a mean shift, both tests achieved similar power, with MW performing slightly better. By contrast, when only the variance changed, KS retained high power whereas MW failed to detect the effect. Taken together, these results indicate that while MW may be marginally more powerful for detecting location shifts, KS is more versatile, capturing distributional changes beyond differences in the mean.

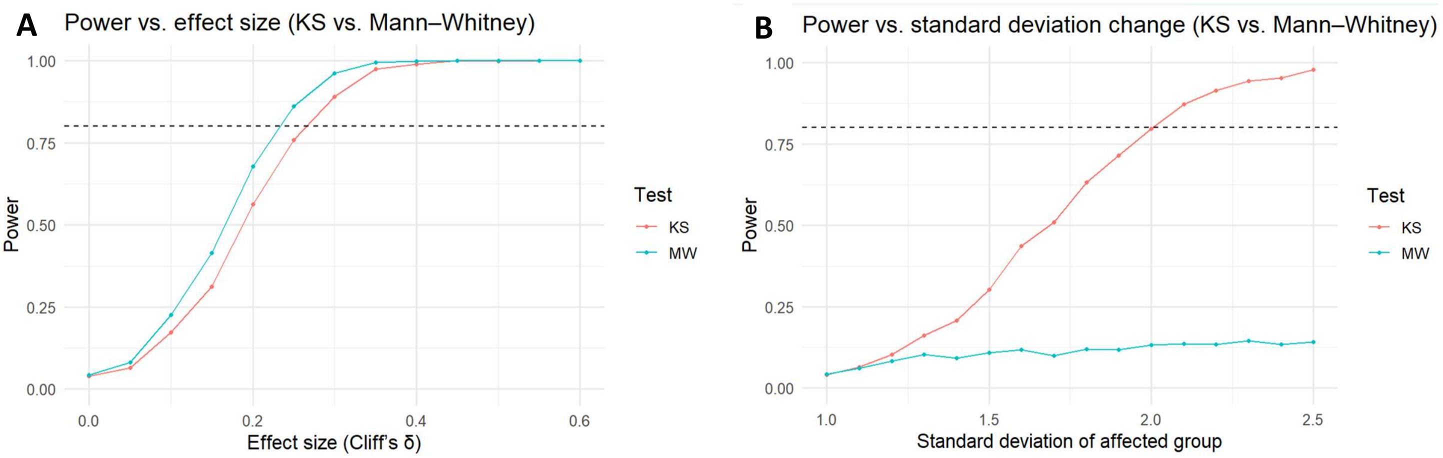

Figure 2. Power analysis comparing Kolmogorov–Smirnov (KS) and Mann–Whitney (MW) tests.
(A) Power as a function of effect size, quantified by Cliff's δ. Both KS and MW show increasing power with larger effects, with MW performing slightly better for location shifts.
(B) Power as a function of variance change while holding the mean constant. KS retains high power to detect distributional shifts, whereas MW remains largely insensitive. The dashed line marks the 80% power threshold.

Statistical power under multiple testing

Because directly incorporating Benjamini–Hochberg correction into the simulation would require modeling the joint distribution of thousands of p-values—a task that is computationally demanding and difficult to approximate realistically—we instead applied the Šidák adjustment. This conservative choice means that the reported power values should be interpreted as lower bounds.

As shown in Figure 3.A–B, the KS test retained high sensitivity even under correction for tens of thousands of hypotheses. For moderate to large effect sizes (Cliff's δ = 0.3–0.5), 80% power was achieved with only a few dozen affected sequences, and the required sample size decreased further as δ increased. Even at smaller effect sizes (δ ≈ 0.2), 80% power was reached with well under 500 affected sequences. By contrast, for very small effects (δ ≈ 0.1), reaching 80% power required close to 2000 affected sequences, which is less realistic in practice given the typical heterogeneity and variant forms of motifs in real biological data.

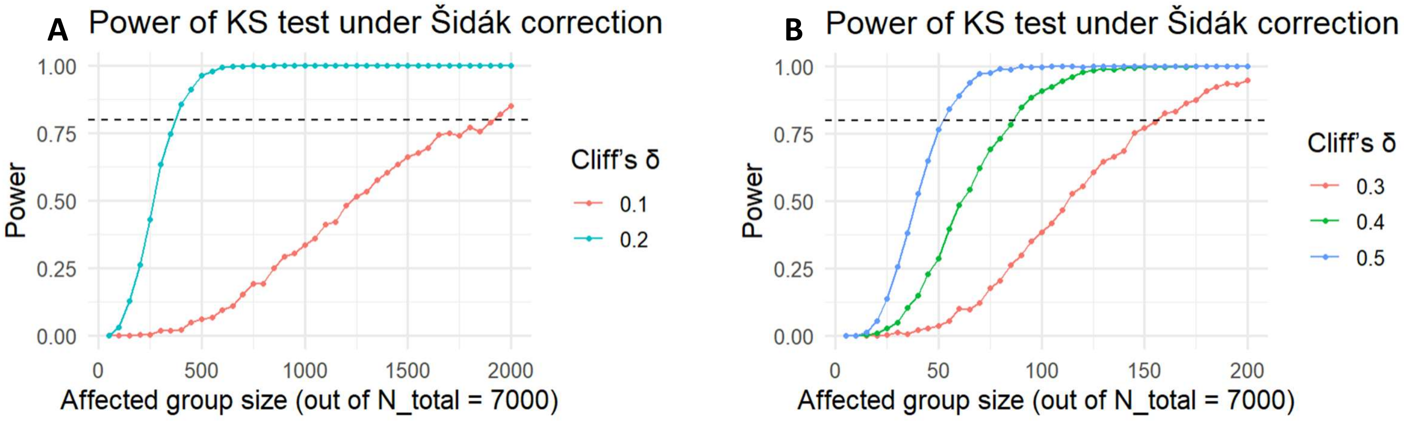

Figure 3. Power analysis of the KS test under Šidák correction.
(A) Power curves for small effect sizes (Cliff's δ = 0.1–0.2). Achieving 80% power (dashed line) required nearly 2000 affected sequences for δ = 0.1, but fewer than 500 for δ = 0.2.
(B) Power curves for moderate to large effect sizes (Cliff's δ = 0.3–0.5). In these cases, 80% power was reached with only a few dozen affected sequences, and the required sample size decreased further as δ increased.
Sample size refers to the number of affected sequences out of a total of 7000.

Cliff's delta as a robust effect size

Because our statistical test is highly powered in large samples, it may flag many k-mers as significant even when the biological effect is negligible. Variants or shifted instances of motifs can also appear significant, although they are far from the true motif of interest. Both issues hinder motif discovery by inflating runtime and obscuring convergence to meaningful patterns. To mitigate this, we filtered results using Cliff's delta (δ), a non-parametric and scale-free effect size. As shown in Figure 4.A, applying a δ threshold separates weak from strong signals, while Figure 4.B demonstrates that k-mers above the threshold are typically closer to the true motif, whereas those below are often distant variants.

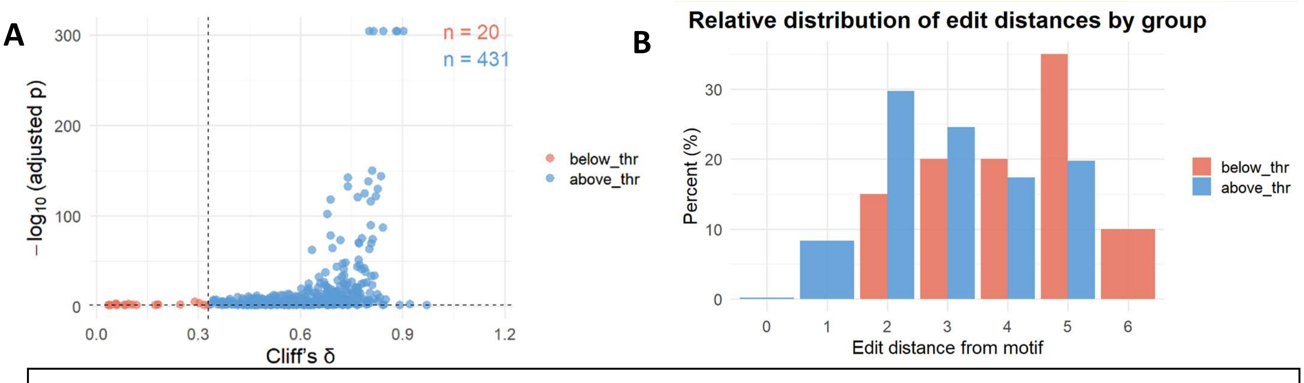

Figure 4. Filtering by Cliff's delta (δ).
(A) Cliff's δ versus adjusted p-values for all k-mers passing the significance threshold (adjusted p < 0.01). A δ cutoff of 0.33 (vertical line) separates weaker (red) from stronger (blue) signals.
(B) Distribution of edit distances from the true motif by group. K-mers below the δ cutoff (red) are typically more distant, whereas those above (blue) cluster closer.

From k-mers into interpretable consensus motifs

Following statistical filtering, the output still consisted of a large and redundant set of significant k-mers. Many of these overlapped substantially, reflected shifted versions of the same underlying motif, or represented minor sequence variants. As a result, the raw list was difficult to interpret in biological terms. To address this, we implemented a multi-stage procedure that merges overlapping hits into extended windows, computes distance between these windows and ultimately consolidates them into clusters from which consensus motifs can be derived.

From k-mers to genomic regions

After statistical testing identified significant k-mers, the next step was to map them back onto the original genomic sequences. All genomic occurrences of each significant k-mer were recorded, and overlapping or adjacent hits were merged into extended windows. These windows typically contained multiple related k-mers — including shifted instances and sequence variants — thereby capturing the broader sequence context of the motif. This step ensured that downstream analyses were grounded in biologically meaningful genomic intervals, rather than in redundant lists of isolated k-mers (Figure 5).

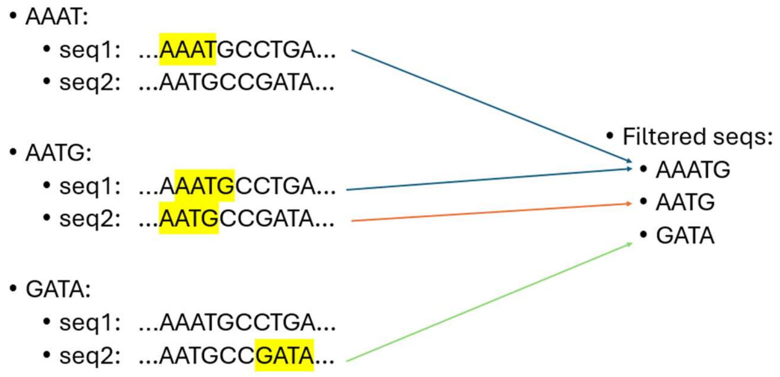

Figure 5. Illustration of the transition from raw k-mers to merged genomic windows. Significant k-mers (highlighted in yellow) are identified across sequences (seq1, seq2). Overlapping or shifted k-mers are then consolidated, producing extended windows (right) that capture the broader sequence context of the motif.

Defining distances between windows

The next challenge was to quantify similarity between sequence windows, which often differed in length and contained subtle variations. A standard approach such as edit distance would penalize length differences. To address this, we employed a sliding q-gram distance with q=3 (Figure 6). In this method, each sequence is decomposed into all possible 3-letter substrings (q-grams). For two sequences $s_1$and s_2 , the distance is defined as:

$$d(s_1, s_2) = 1 - \max_{\text{offset}} \frac{|Q(s_1) \cap Q_{\text{offset}}(s_2)|}{\min(|Q(s_1)|,|Q(s_2)|)}$$

where $Q(s)$ denotes the set of 3-grams from sequence s, and the maximum is taken over all possible relative offsets of s_1 against s_2. This measure is bounded between 0 (identical) and 1 (completely dissimilar). Crucially, it captures local sequence composition while remaining robust to differences in length, making it particularly well suited for grouping windows that represent variants or shifted instances of the same underlying motif.

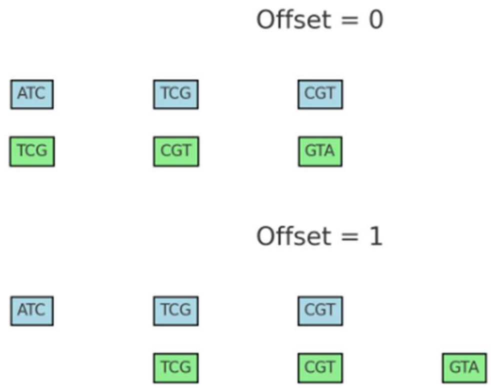

Figure 6. Illustration of sliding q-gram distance.
Each sequence is decomposed into all possible 3-grams (boxes). At offset = 0 (top), q-grams are compared in their original alignment. At offset = 1 (bottom), one sequence is shifted relative to the other, creating new potential overlaps. The similarity score is based on the maximum fraction of shared q-grams across all offsets.

Determining the number of clusters

After statistical filtering, sequences were first divided into positive and negative groups according to the direction of their effect. Within each group, clustering was performed on the distance matrix using the Partitioning Around Medoids (PAM) algorithm, which—unlike k-means—can operate directly on a distance matrix.

The number of clusters, k, was chosen using the silhouette width. For each sequence i, let $a(i)$ be its average distance to other sequences in the same cluster, and let $b(i)$ be the minimum, over all other clusters, of the average distance from iii to the sequences in that cluster. The silhouette value $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$ ranges from –1 to 1: values close to 1 indicate strong separation, values near 0 suggest overlap, As shown in Figure 7.A and 7.B, and negative values reflect possible misclassification. the optimal number of clusters was *k = 2* for the positive group and *k = 4* for the negative group, in line with the simulated ground truth. Figures 8.A and 8.B show a two-dimensional t-SNE projection based on the distance matrix. While this visualization is only an approximation, since t-SNE reduces high-dimensional relationships into two dimensions, it provides an intuitive indication that the partitions are relatively well separated, supporting the validity of the clustering procedure.

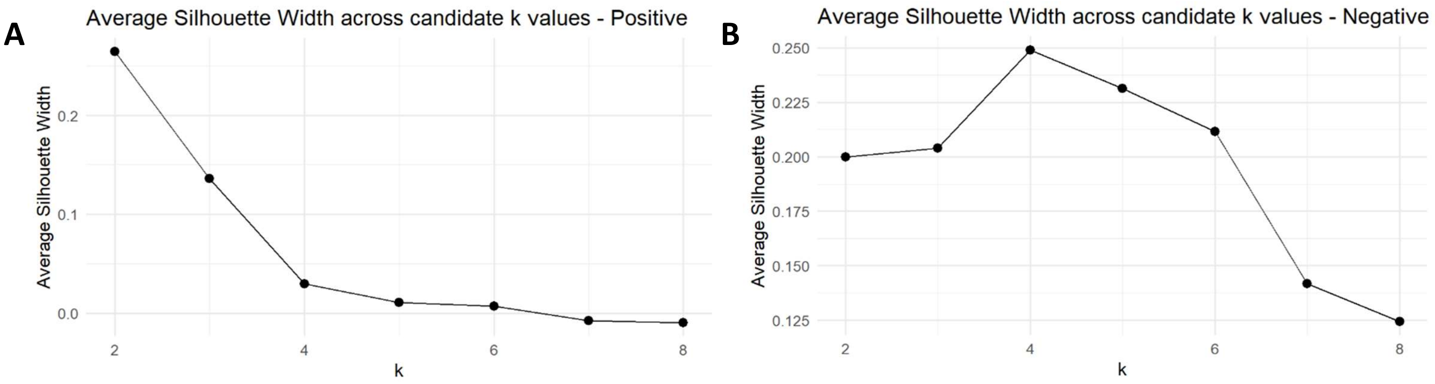

Figure 7. Silhouette analysis for cluster selection.
(A) Positive group: average silhouette width across candidate *k* values, with the optimal *k* = 2.
(B) Negative group: average silhouette width across candidate *k* values, with the optimal *k* = 4.

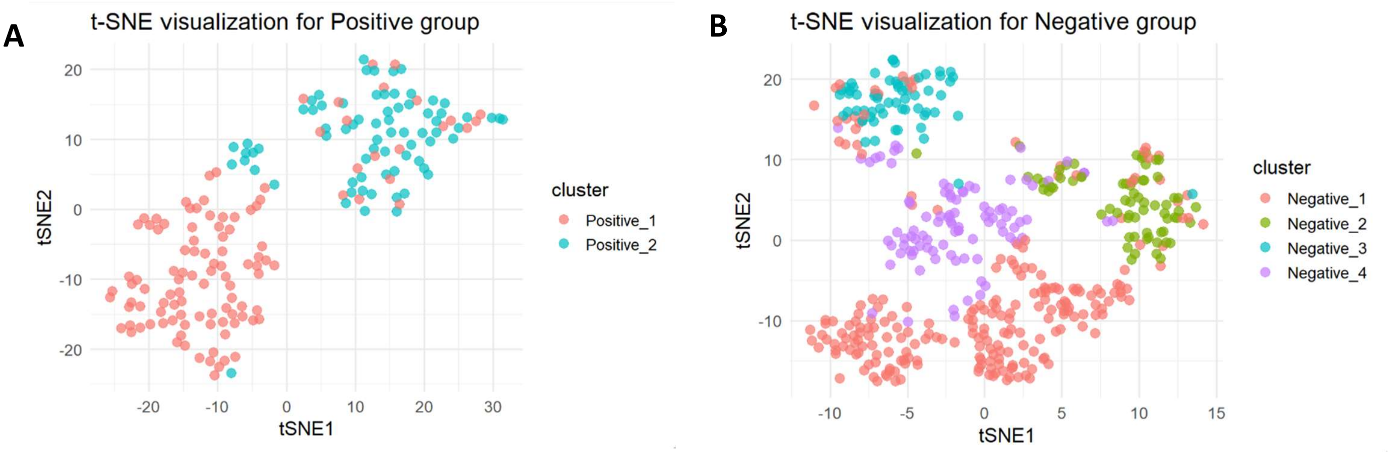

Figure 8. Visualization of clustering results using t-SNE projection.
(A) Positive group (k = 2).
(B) Negative group (k = 4).

Consensus motif recovery

Finally, for each cluster, I extracted all associated windows and subjected them to multiple sequence alignment. Alignment positions with excessive gaps were removed to emphasize the conserved regions. The aligned sequences were then visualized as sequence logos, providing a concise graphical summary of the nucleotide preferences at each position.

As shown in Figure 9, I successfully recovered all simulated motifs. The consensus logos closely matched the ground-truth motifs embedded during simulation.

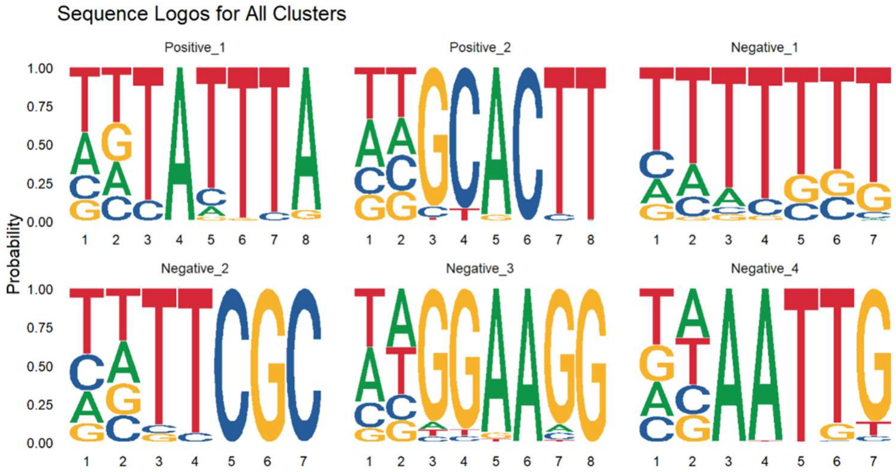

Figure 9. Sequence logos of recovered motifs from each cluster. The consensus motifs closely match the ground-truth sequences embedded in the simulated.

discussion

From the first simulation we observed that the Kolmogorov–Smirnov (KS) test maintains good sensitivity not only to shifts in central tendency but also to broader differences between distributions. This property makes KS the more suitable choice for the motif-discovery framework, where biological effects may not be limited to changes in mean values.

The second simulation demonstrated that, even under conservative multiple testing correction, statistical power is not a limiting factor. For effect sizes larger than 0.1, a sample size on the order of a few dozens to a few hundreds was sufficient to achieve power of 0.8.

Regarding Cliff's delta, while it is impossible to define in advance what magnitude of effect size should be considered biologically meaningful, the statistic provides a robust and interpretable measure of group separation. This allows the use of relatively permissive thresholds without the risk of losing potentially relevant motifs. The main drawback is computational: calculating Cliff's delta at scale is time-consuming and may become a bottleneck in large analyses.

Altogether, our results suggest that the central limitation of the framework is not the sensitivity of the statistical test but its specificity. This points to the possible benefit of applying more conservative multiple testing corrections than the Benjamini–Hochberg procedure in order to reduce false positives.

At the same time, the current simulations do not fully characterize the boundaries of the tool. Biological systems are highly complex, and the simplified simulation settings used here are insufficient to fully explore the conditions under which the framework succeeds or fails. To address this, I developed a richer simulation framework that allows systematic testing of such boundaries (and from which I generated the datasets used for the analyses in this work).

Finally, the most significant next step is to apply the pipeline to real biological data and compare its performance directly with existing motif-discovery tools. Such benchmarking is essential for establishing the practical value of the method.

References

1. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*. 2006;7:29-59. doi:10.1146/annurev.genom.7.080505.115623
2. Rabani M, Pieper L, Chew GL, Schier AF. A massively parallel reporter assay of 3′ UTR sequences identifies in vivo rules for mRNA degradation. Mol Cell. 2017;68(6):1083-94.e5. doi:10.1016/j.molcel
3. 3. Hashim FA, Mabrouk MS, Al-Atabany W. Review of different sequence motif finding algorithms. Avicenna J Med Biotechnol. 2019;11(2):130-48. PMID: 31057715; PMCID: PMC6490410.
4. Bailey TL, Elkan C. Unsupervised learning of multiple motifs in biopolymers using EM. Machine Learning. 1995;21(1–2):51-80.
5. Bailey TL. DREME: Motif discovery in transcription factor ChIP-seq data. Bioinformatics. 2011;27(12):1653-9.
6. Cliff N. Dominance statistics: Ordinal analyses to answer ordinal questions. Psychol Bull. 1993;114(3):494-509. doi:10.1037/0033-2909.114.3.494