Motif finding in DNA sequences

Adam Sambor, Jakub Żuk

June 2025

1 Introduction

The main goal of this project is to find motifs in DNA sequences. Such sequences consist of four nitrogenous bases A, C, G, T, occurring in certain 'random' configurations.

For simplicity, we assume that the probability of occurrence of $a \in \{A, C, G, T\}$ as the i'th element of each gene sequence $\boldsymbol{x}_i = (x_{i1}, ..., x_{iw})$ is independent due to the earlier and later occurring bases and it is equal to $\theta_{a,i}$ or θ_a^b (with probabilities α and $1 - \alpha$ accordingly for some $\alpha \in [0, 1]$ known).

It means that each gene sequence $\boldsymbol{x}_i$ of length w is an independent realization of a random variable $X = (X_1, ..., X_w)$, where each of X_i follows the distribution $\alpha\boldsymbol{\theta}_i + (1 - \alpha)\boldsymbol{\theta}^b$, where $\boldsymbol{\theta}_i = (\theta_{A,i}, \theta_{C,i}, \theta_{G,i}, \theta_{T,i})$, and $\boldsymbol{\theta}^b = (\theta_A^b, \theta_C^b, \theta_G^b, \theta_T^b)$.

Let us note $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_w)$. Given a random sample $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_w)$, we want to reproduce $\Theta = (\boldsymbol{\theta}, \boldsymbol{\theta}^b)$. To accomplish this task, we will use the EM algorithm.

2 EM algorithm

The EM algorithm is an iterative method used to find maximum likelihood estimates of unknown parameters in statistical models. The algorithm consists of two main parts, in which we alternately determine expectation of missing parameters (based on sample) and modify them to maximize log-likelihood function.

In our case, we get k samples from some unknown distribution. Starting from randomly chosen distribution of nitrogenous bases, each step of EM algorithm tries to correct initial distribution and maximize log-likehood.

Algorithm 1 General EM Algorithm

Require: Observed data X, latent variables Z, initial parameters $\theta^{(0)}$ (that give us initial distribution $d^{(0)}$), max iterations T, tolerance ε

Ensure: Estimated parameters $\widehat{\theta}$

1: Set $t \leftarrow 0$, $\mathrm{LL}^{(0)} \leftarrow \log L(X \mid \theta^{(0)})$ (where L is the likelihood function)

2: **repeat**

3: **E-step:** compute

$$Q(\theta \mid \theta^{(t)}) \;=\; \mathbb{E}_{Z \sim d^{(t)}}\big[\log L(X, Z \mid \theta)\big]$$

4: **M-step:** update

$$\theta^{(t+1)} \;=\; \arg\max_{\theta}\; Q(\theta \mid \theta^{(t)})$$

5: Compute log-likelihood

$$\mathrm{LL}^{(t+1)} \;=\; \log L(X \mid \theta^{(t+1)})$$

6: $t \leftarrow t + 1$

7: **until** $|\mathrm{LL}^{(t+1)} - \mathrm{LL}^{(t)}| < \varepsilon$ **or** $t \geq T$

8: **return** $\theta^{(t)}$

2.1 EM algorithm in finding DNA motifs

In our case, X is a $k \times w$ matrix of sample DNA sequences where w is the length of sequence and k is the number of samples. Therefore, the initial parameters $\theta^{(0)}$ is also represented as a matrix $(w \times 4)$ with entries (θ_{ij}) representing the probability of finding the j-th nitrogenous base at the i-th position. Z is a binary random variable deciding whether we sample from motif $(Z = 1)$ or background $(Z = 0)$.

- In each "E-step" we compute the expectation of log-likelihood assuming that our samples come from the probability distribution defined in the previous step.
- Then in "M-steps" we overwrite previous parameters of distribution with new ones maximizing the expectation of log-likelihood function computed earlier, so in fact we compute something similar to the maximum likelihood estimator (MLE) in every step. Instead of computing the MLE for original complete data (which we do not know), we find it for expected complete data.

We hope that repeating this procedure will give us a quite good estimate of the true parameters (parameters of distribution from which we sampled the DNA sequences). That "hope" is based on the fact (that can be strictly proven [1]), that every iteration of EM algorithm **will not** reduce the likelihood.

Finding $\arg\max_{\theta}\; Q(\theta \mid \theta^{(t)})$ is not trivial, so let us see how we approach this problem. We denote

$Q_i^{(t)}(0), Q_i^{(t)}(1)$ as posterior probabilities that the sequence was generated by background or motif, respectively. We have

$$\begin{aligned}
Q_i^{(t)}(0) &= P(z_i = 0 | \boldsymbol{x}_i, \Theta^{(t)}) \\
&= \frac{P(z_i = 0, \boldsymbol{x}_i | \Theta^{(t)})}{P(\boldsymbol{x}_i | \Theta^{(t)})} \quad \text{(by definition of conditional probability)} \\
&= \frac{P(\boldsymbol{x}_i | z_i = 0, \Theta^{(t)}) \cdot P(z_i = 0 | \Theta^{(t)})}{P(\boldsymbol{x}_i | \Theta^{(t)})} \quad \text{(by the chain rule of probability)} \\
&= \frac{\left(\prod_{j=1}^{w} \theta_{x_{ij}}^{b,(t)} \right) \cdot (1 - \alpha)}{P(\boldsymbol{x}_i | \Theta^{(t)})}
\end{aligned}$$

In the final step, we substituted $P(\boldsymbol{x}_i | z_i = 0, \Theta^{(t)}) = \prod_{j=1}^{w} \theta_{x_{ij}}^{b,(t)}$ like in the project description and the probability of the background class, $P(z_i = 0 | \Theta^{(t)}) = 1 - \alpha$. Similarly we obtain

$$Q_i^{(t)}(1) = P(z_i = 1 | \boldsymbol{x}_i, \Theta^{(t)}) = \frac{\alpha \cdot \prod_{j=1}^{w} \theta_{x_{ij},j}^{(t)}}{P(\boldsymbol{x}_i | \Theta^{(t)})}.$$

Maximizing Q is difficult, so we split it into two independent ones:

$$\begin{aligned}
Q(\Theta, \Theta^{(t)}) &= \sum_{i=1}^{k} \sum_{j=0}^{1} Q_i^{(t)}(j) \log P(\mathbf{x}_i, z_i = j | \Theta) \\
&= \sum_{i=1}^{k} Q_i^{(t)}(0) \log P(\mathbf{x}_i, z_i = 0 | \Theta) + \sum_{i=1}^{k} Q_i^{(t)}(1) \log P(\mathbf{x}_i, z_i = 1 | \Theta) \\
&= \underbrace{\sum_{i=1}^{k} Q_i^{(t)}(0) \log \left((1 - \alpha) \prod_{j=1}^{w} \theta_{x_{ij}}^{b} \right)}_{\substack{\text{dependent only on background parameters } \theta^b \\ Q_1(\theta^b, \alpha)}} + \underbrace{\sum_{i=1}^{k} Q_i^{(t)}(1) \log \left(\alpha \prod_{j=1}^{w} \theta_{x_{ij},j} \right)}_{\substack{\text{dependent only on motif parameters } \theta \\ Q_2(\theta, \alpha)}}
\end{aligned}$$

so we can write

$$Q(\Theta, \Theta^{(t)}) = Q_1(\theta^b, \alpha) + Q_2(\theta, \alpha)$$

maximize each summand separately. We will do this using Lagrange multipliers.

2.2 Maximization of likelihood expectation

Background model

Let the constraint function be:

$$g(\boldsymbol{\theta}^b) = \theta_1^b + \theta_2^b + \theta_3^b + \theta_4^b - 1$$

The Lagrangian is defined as:

$$\mathcal{L}(\boldsymbol{\theta}^b) = Q_1(\boldsymbol{\theta}^b) - \lambda \cdot g(\boldsymbol{\theta}^b)$$

We compute the partial derivative with respect to θ_r^b for $r \in \{1, 2, 3, 4\}$.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \theta_r^b} &= \frac{\partial}{\partial \theta_r^b}\left(\sum_{i=1}^{k} Q_i^{(t)}(0)\left(\log(1-\alpha) + \sum_{j=1}^{w} \log \theta_{x_{ij}}^b\right) - \lambda\left(\theta_1^b + \theta_2^b + \theta_3^b + \theta_4^b - 1\right)\right) \\ &= \sum_{i=1}^{k}\left(Q_i^{(t)}(0)\frac{1}{\theta_r^b} \cdot |\{j : x_{ij} = r\}|\right) - \lambda \quad \text{since } \frac{\partial \log \theta_{x_{ij}}^b}{\partial \theta_r^b} = \begin{cases} 1/\theta_r^b & \text{if } x_{ij} = r \\ 0 & \text{if } x_{ij} \neq r \end{cases}\end{aligned}$$

where $|\{j : x_{ij} = r\}|$ denotes the number of occurrences of base r in sequence $\boldsymbol{x}_i$.

Then, setting the derivative to zero, $\frac{\partial \mathcal{L}}{\partial \theta_r^b} = 0$, gives:

$$\theta_r^b = \frac{1}{\lambda} \cdot \sum_{i=1}^{k} Q_i^{(t)}(0)|\{j : x_{ij} = r\}|$$

But we know that $\sum_{r=1}^{4} \theta_r^b = 1$, which implies:

$$\lambda = w \cdot \sum_{i=1}^{k} Q_i^{(t)}(0)$$

Therefore, we get:

$$\theta_r^b = \frac{\sum_{i=1}^{k} Q_i^{(t)}(0)|\{j : x_{ij} = r\}|}{w \cdot \sum_{i=1}^{k} Q_i^{(t)}(0)}$$

where we know the formula for $Q_i^{(t)}(0)$.

Motif model

In very similar way we derive the formula for θ_r:

$$\theta_{r,j} = \frac{\sum_{i=1}^{k} Q_i^{(t)}(1)\mathbb{1}(x_{ij} = r)}{\sum_{i=1}^{k} Q_i^{(t)}(1)}$$

3 Total variation distance

To assess the performance of our EM algorithm, we must measure how closely our estimated parameters match the true parameters used to generate the data. For this purpose, we will use the total variation distance, a standard metric for comparing probability distributions.

For two discrete probability distributions $\boldsymbol{p} = (p_1, ..., p_n)$ and $\boldsymbol{q} = (q_1, ..., q_n)$ on $\{1, ..., n\}$, we define the total variation distance as:

$$d_{tv}(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{2} \sum_{i=1}^{n} |p_i - q_i|$$

Note that $0 \leq d_{tv}(\boldsymbol{p}, \boldsymbol{q}) \leq 1$

Let the original parameters (used to generate the data) be:

$$\boldsymbol{\theta}^{b,orig} = (\theta_1^{b,orig}, \theta_2^{b,orig}, \theta_3^{b,orig}, \theta_4^{b,orig}), \;\; \boldsymbol{\theta}^{orig} = (\boldsymbol{\theta}_1^{orig}, ..., \boldsymbol{\theta}_w^{orig})$$

where each $\boldsymbol{\theta}_i^{orig}$, for $i = 1, ..., w$, is a column vector of size 4.

Similarly, let the estimated parameters be:

$$\boldsymbol{\theta}^{b,estim} = (\theta_1^{b,estim}, \theta_2^{b,estim}, \theta_3^{b,estim}, \theta_4^{b,estim}), \;\; \boldsymbol{\theta}^{estim} = (\boldsymbol{\theta}_1^{estim}, ..., \boldsymbol{\theta}_w^{estim})$$

As a final performance measure, we will compute the average total variation distance across all distributions:

$$d_{tv} = \frac{1}{w+1} \left[d_{tv}(\boldsymbol{\theta}^{b,orig}, \boldsymbol{\theta}^{b,estim}) + \sum_{i=1}^{w} d_{tv}(\boldsymbol{\theta}_i^{orig}, \boldsymbol{\theta}_i^{estim}) \right]$$

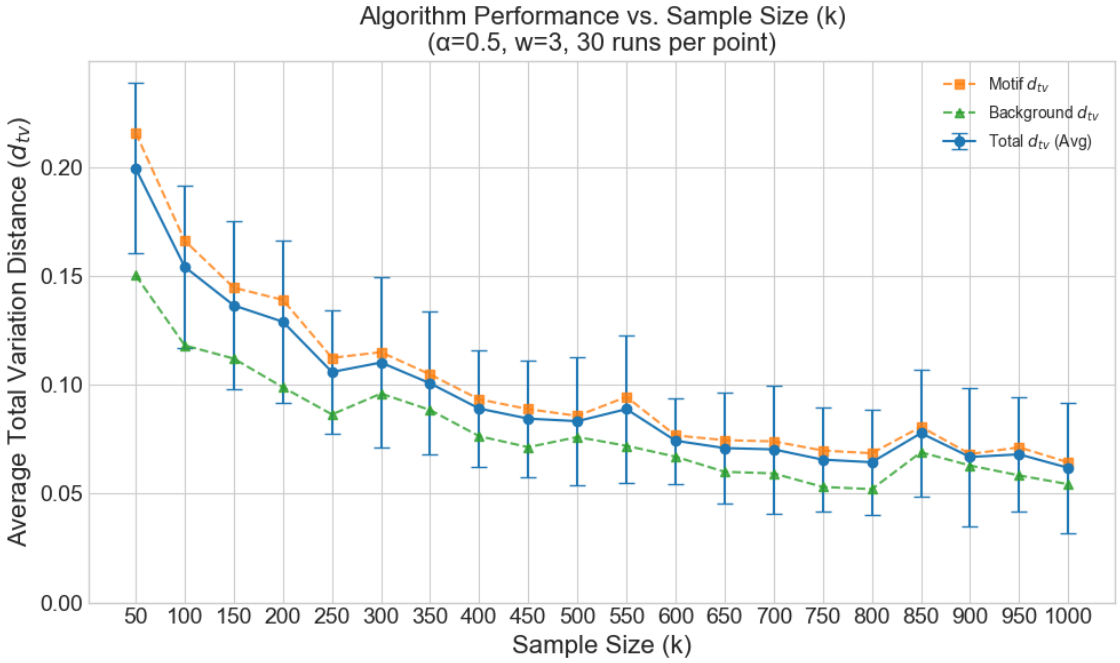
This final d_{tv} score provides a single, comprehensive measure of our model's accuracy. We will use it throughout the following sections to evaluate the performance of our estimates under various conditions.

4 Tests

4.1 Impact of Data Characteristics on Performance

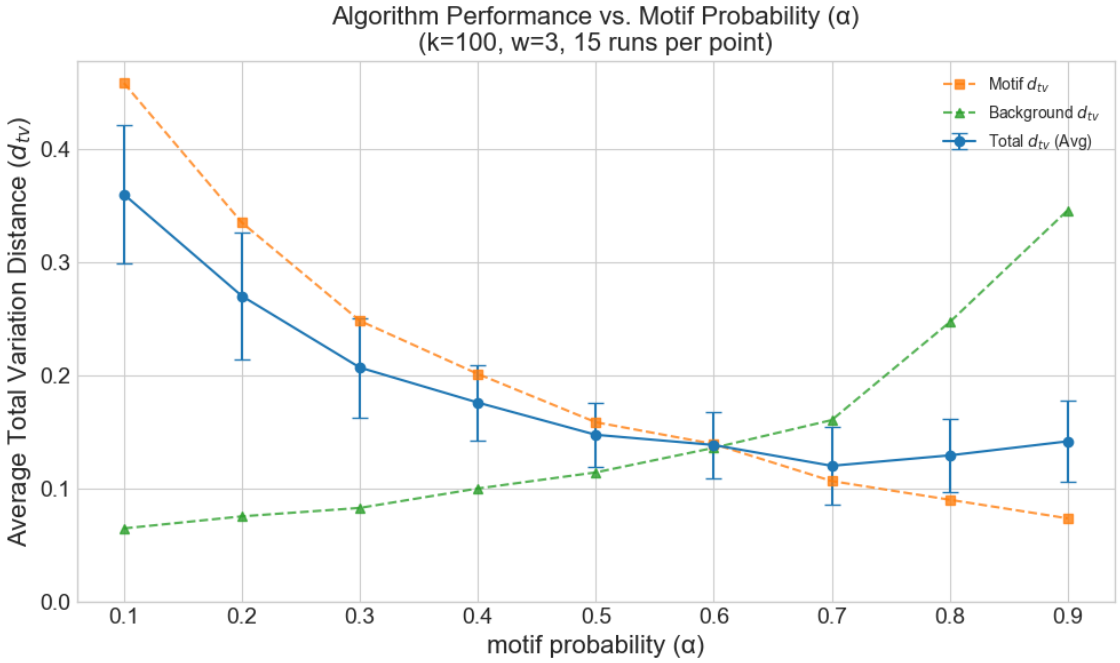
In this section, we will explore how various parameters of our initial data (such as sample size k or the motif probability α) impact the performance of the EM algorithm. Since most parameters have a monotonic influence on the quality of estimation, all visualizations are one-dimensional, with the hope that other dependencies do not occur. We set the tolerance of 10^{-10} as boundary value up to which the algorithm performs. It is more than sufficient for our goals.

We begin our analysis by examining the influence of sample size k on algorithm performance, as measured by the average Total Variation Distance d_{tv}.



As expected, the estimation error decreases as the sample size (k) increases. For large values of k with α fixed at 0.5, both the motif $\boldsymbol{\theta}$ and background $\boldsymbol{\theta^b}$ parameters are estimated with comparably high accuracy. However, at smaller sample sizes, the difference is more noticeable, and the estimation of the background distribution is significantly more accurate than that of the motif. This is likely attributable to two factors: the simplicity of the uniform background distribution $\boldsymbol{\theta^b} = (0.25, 0.25, 0.25, 0.25)$ and the short motif length ($w = 3$), which provides limited information within each sequence.

It is important to notice that each point on the graph represents the mean d_{tv} calculated across 30 independent runs. Furthermore, the standard deviation error bars illustrate the algorithm's sensitivity to two sources of randomness: the specific data sample generated in each run and, critically, the random initialization of the parameters $\Theta^{(0)}$. The critical role of initialization is underscored by the observation that while the standard deviation does decrease for some increments in sample size, this reduction is not monotonic. This inconsistency highlights that the variability introduced by initialization can be significant regardless of the amount of data.

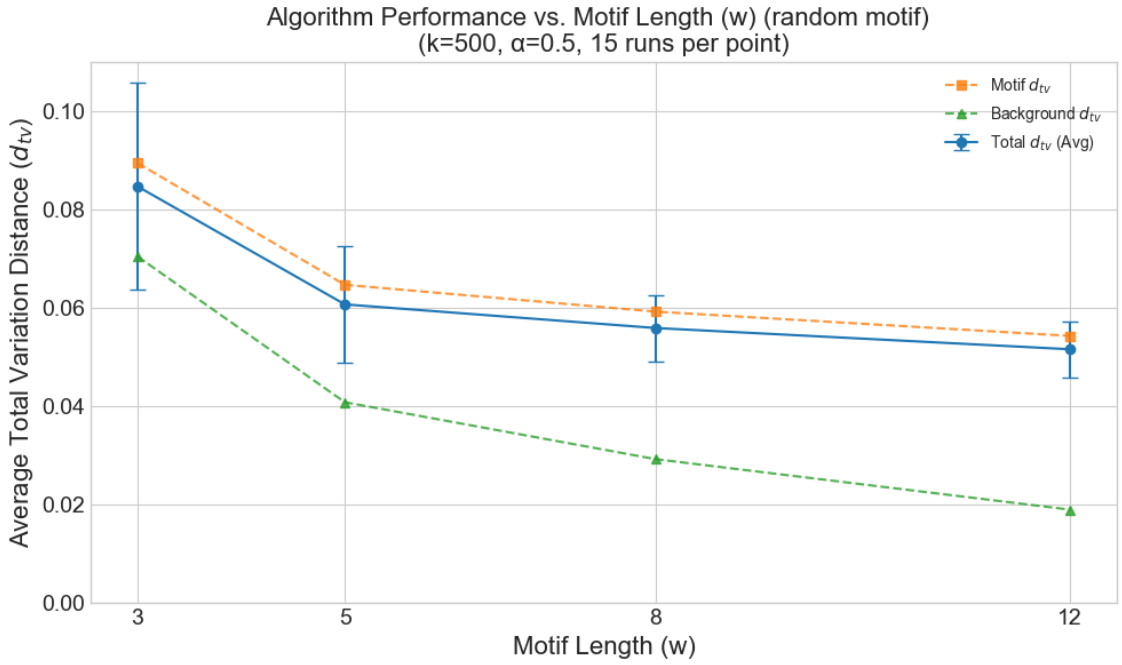


It seems that raising the motif probability reduces the error of finding the motif distribution and increases the error of finding the background motif. Naturally, a higher chance of taking a sample from one model makes estimations easier, so that observation is reasonable. It is worth noting that motif d_{tv} has a greater impact on total average total variation distance than the background. Let us recall how we defined the total d_{tv}. The part coming from the motif is w times more important than the part coming from the background.

The interesting feature of this plot is the increasing average total variation distance for boundary values of motif probability. We can find the minimum of variation distance around $\alpha \approx 0.7$. For higher values of w we expect that minimum to move right for stronger motif, or left for weak motif. High probability of sampling from motif or from the background results in very poor classification of the less probable samples and therefore a sudden increase in error.

The following figure was made using random $\boldsymbol{\theta}$ (with average $\boldsymbol{\theta}_i = (0.25, 0.25, 0.25, 0.25)$) and

$$\boldsymbol{\theta}^b = (0.25, 0.25, 0.25, 0.25)^T$$



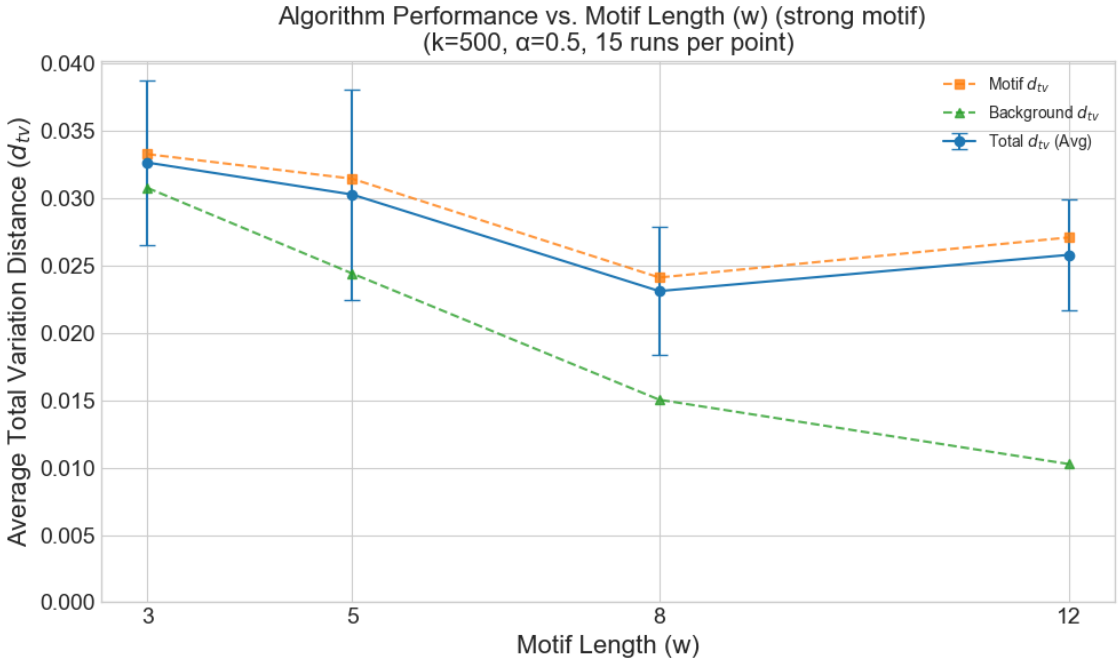
As the plot illustrates, when the motif's position weight matrix $\boldsymbol{\theta}$ is random, the estimation error d_{tv} decreases as the sequence length w increases.

For such $\boldsymbol{\theta}$, each column represents a weak signal, being only slightly different from the uniform background distribution. Based on a single position, it is nearly impossible for the algorithm to distinguish a motif sequence from a background one. However, as w grows, these small, independent differences accumulate. The combined likelihood of a true motif sequence across many positions becomes significantly more distinct from the background, allowing the EM algorithm to perform a more accurate classification and, consequently, a more precise parameter estimation.

Let us now consider a different scenario. Suppose

$$\boldsymbol{\theta} = \begin{pmatrix} 0.85 & 0.05 & 0.05 & 0.05 & 0.85 & \cdots \\ 0.05 & 0.85 & 0.05 & 0.05 & 0.05 & \cdots \\ 0.05 & 0.05 & 0.85 & 0.05 & 0.05 & \cdots \\ 0.05 & 0.05 & 0.05 & 0.85 & 0.05 & \cdots \end{pmatrix}$$

and $\boldsymbol{\theta}^b$ as before.

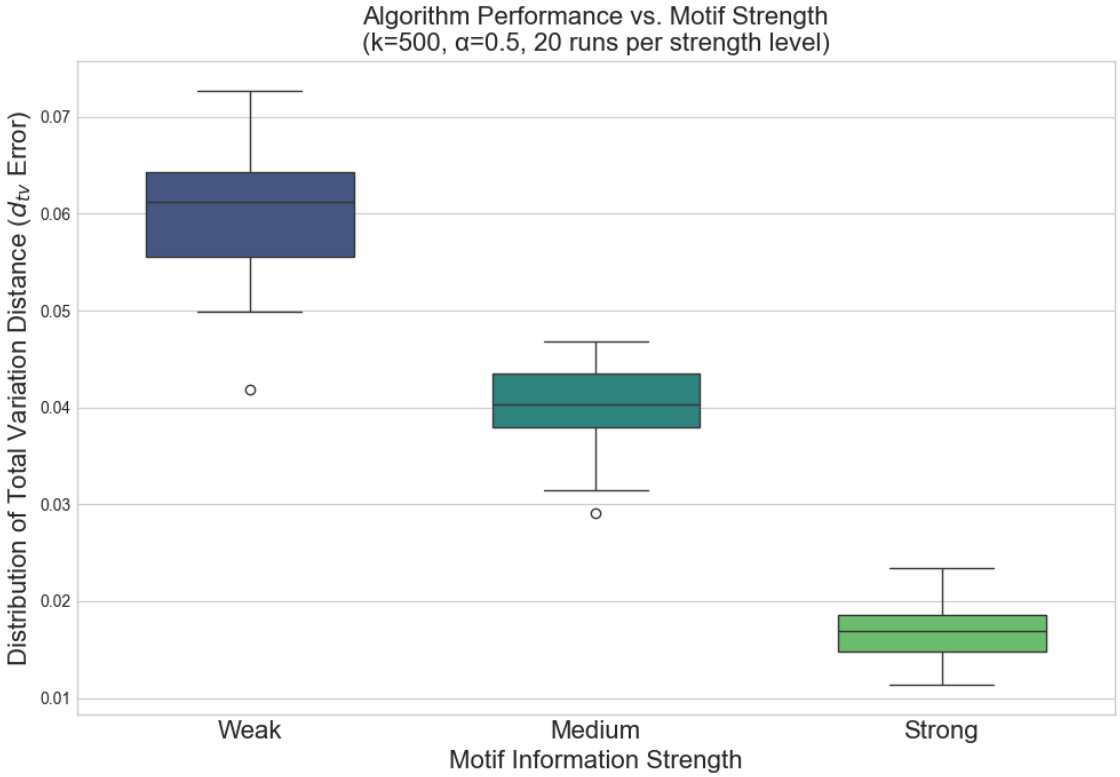


The results from this experiment reveal a fundamentally different behavior compared to the random motif case.

First, as expected, the overall estimation error d_{tv} is much lower across all values of w. The strong signal provided by this motif allows the algorithm to perform a much more accurate estimation.

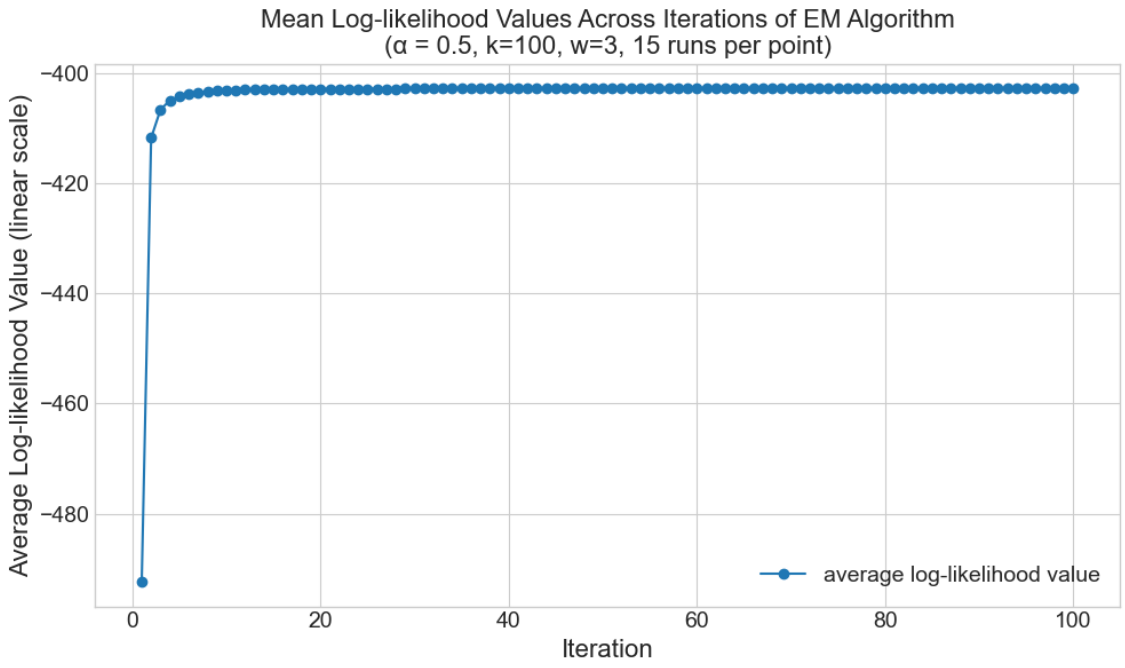
Second, and more importantly, the relationship between w and d_{tv} is no longer monotonic. In fact, the error slightly increases for $w = 12$ compared to $w = 8$. With such a strong signal, the algorithm can reliably distinguish motif sequences from background sequences even for smaller values of w, so there is limited benefit from a greater motif length. Furthermore, as we increase the motif length, we also increase the number of parameters in $\boldsymbol{\theta}$ we need to estimate. With a fixed amount of data, estimating more parameters can lead to the model fitting not just the true signal, but also the random noise specific to our dataset, thus slightly degrading its generalization performance and increasing the d_{tv} error.

To summarize and generalize our previous conclusion, we decided to investigate the direct relationship between the motif signal strength and the estimation error.

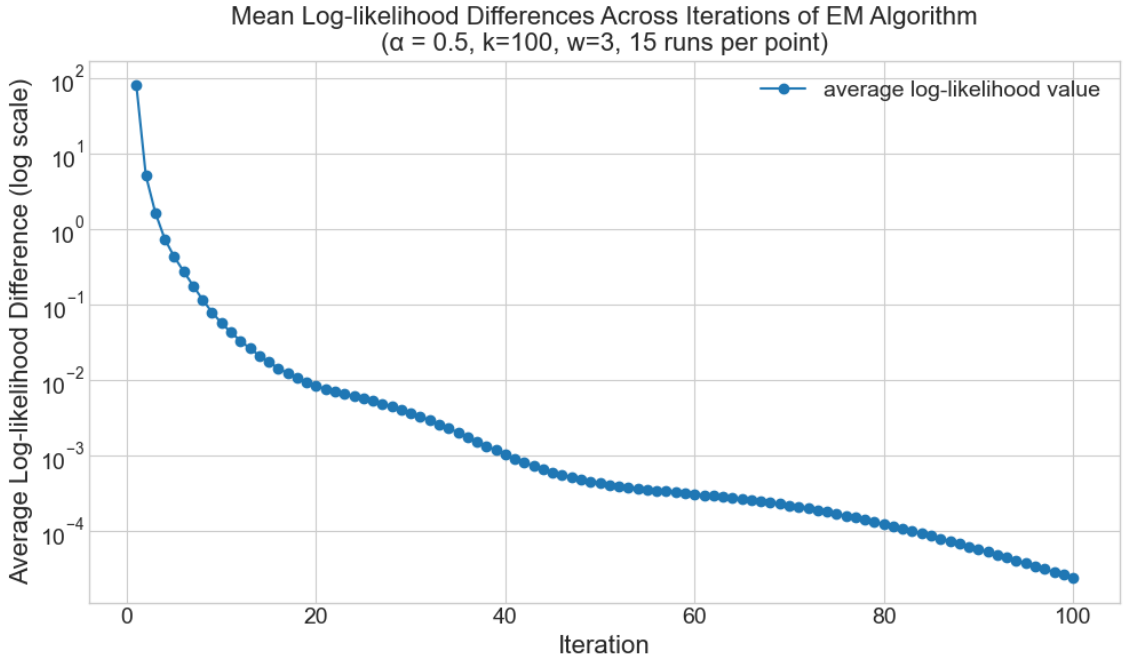


The graph clearly reveals a strong negative correlation - the greater the motif information strength, the lower the estimation error d_{tv}. This result is consistent with our earlier observations: the random motif (with weak information strength) led to much higher d_{tv} than the motif with a strong, distinct signal.

4.2 Analysis of Algorithm Behavior



The log-likelihood grows very rapidly in the first few steps to de facto stabilize on a certain level. Of course, the algorithm corrects the previous prediction every step, but when we are closer and closer to the true values, the differences are small. We will illustrate this on the next plot.



As we can see, algorithm does most of the error reduction in the first few steps (as it was shown in the plot above). Each iteration, the algorithm predicts final motif and background distributions better and better, so that the distributions are closer to their true values. However, because of more accurate predictions in every step, differences in log-likelihood between steps are, in fact, monotonically decreasing.

5 Summary

In this project, we faced the problem of finding DNA sequence motifs based only on samples coming from unknown distribution influenced by hidden latent variable. Depending on various parameters (motif probability, motif length, sample size, motif strength) the EM algorithm performed better or worse, but generally well. Concluding, EM algorithm may require some harder or easier analytical calculations, but it is a useful and powerful tool for these types of tasks.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". eng. In: *Journal of the Royal Statistical Society. Series B, Methodological* 39.1 (1977), pp. 1–38. ISSN: 0035-9246.