



Profile Hidden Markov Model Construction for Regulatory Motif Identification in DNA Sequence Families

¹J. Jeslin, ²A. Radhika*, ³K. Geetha

^{1,3}Research Scholar, Department of Statistics, Periyar University, Salem, India

²Assistant Professor, Department of Statistics, Periyar University, Salem, India*

Abstract: Biological motifs consist of short conserved sequence patterns that serve essential regulatory and structural functions throughout DNA/RNA/protein molecules. The growing genomic data requires developers to solve the essential problem of motif identification and classification in computational genomics. The detection of position-specific conservation between multiple sequences proves difficult for traditional alignment based methods. The Profile Hidden Markov Model (PHMM) framework provides improved motif conservation representation through position dependent emission and transition probabilities to address the limitations. Current PHMM based studies use frequency calculations with weak sequence weighting methods that decrease model accuracy when motif sequences have different lengths. We develop a statistical PHMM model to analyze eukaryotic regulatory DNA motifs using a multiple sequence alignment of a few canonical motifs. The method uses Henikoff & Henikoff (1994) position-based sequence weighting together with Laplace pseudocounts to enhance probability estimation accuracy. To assess motif similarity, the Viterbi algorithm enables log-odds scoring. The results show that PHMMs successfully identify regulatory elements and provide statistical evidence for motif detection.

Index Terms - Profile Hidden Markov Model, DNA motifs, multiple sequence alignment, sequence weighting, Laplace pseudocounts, Viterbi algorithm, computational genomics.

I. INTRODUCTION

The classification of biological sequences and grouping them into families remains a major challenge in genomics. The essential information within these sequences consists of motifs which represent short recurring nucleotide patterns that hold structural value. The DNA/ RNA/protein sequences contain motifs which serve vital biological functions by acting as transcription factor binding sites, splice junctions, promoter elements and protein-protein interaction interfaces. Over the last few decades, genomic databases have expanded, creating a need for efficient and accurate methods to discover biological motifs (Satam et al., 2023). The traditional pairwise alignment method succeeds in sequence comparison but fails when multiple sequences contain position-specific conserved patterns. The analysis of high-yield sequencing data has led to the development of multiple computational methods which enhance motif detection accuracy and processing speed. The Codon-based scoring system developed by Baridam in 2014 demonstrates reliable performance for detecting conserved regions that span different motif lengths (Baridam, 2014). The 2019 research by Yu et al. introduced two methods which divided large datasets into smaller parts to enhance motif detection of rare patterns and improve motif refinement (Yu et al., 2019). The RL-MD framework operates through Machine Learning (ML) to identify motifs from unlabelled data by applying relative scoring methods for motif discovery (Wang et al., 2022). Genetic algorithms have also been applied by Muttakin & Huq (2017) to unaligned sequences, improving transcription start site detection using evolutionary search methods (Muttakin and Huq, 2017). Identification on optimal methods which achieve fast computation and accurate motif detection in various biological systems continues to be a significant obstacle according to Joseph (2023) (Joseph, 2023).

Tools developed for the classification work need to apply established structural patterns from specific families and perform sequence alignment against suitable family models to detect these patterns and establish statistical significance for accurate family member identification. The fundamental principles of Profile Hidden Markov Models (PHMMs) match the operational mechanisms of these models which serve as essential tools for studying conserved sequence motifs. The models extract vital statistical patterns which exist in biological sequences.

The research by Dai & Cheng used PHMMs to identify conserved DNA sequences which advanced both gene structure prediction and database search capabilities (Dai and Cheng, 2008). The application of PHMM by Horan et al., (2010) in protein family modeling enabled the identification of essential functional areas through sub-HMMs generated from Pfam database information (Horan et al., 2010). The pattern-based filtering method from Sun & Buhler (2007; 2009) demonstrates how to enhance search speed without compromising sensitivity (Sun and Buhler, 2007). In our earlier work, we used Markov models to classify DNA sequences highlighting how powerful probabilistic methods can be for understanding sequence patterns (Jeslin et al., 2025). However, the majority of PHMM-based research continues to use basic frequency calculations while disregarding advanced sequence weighting methods which produce biased results when motifs appear infrequently or with balanced distribution. While HMMs are widely used

in protein family classification, fewer research have adapted these structural conservation methods to classify eukaryotic DNA regulatory motifs. Existing motif detection methods from EM models to genetic algorithms and ML techniques often prioritize either accuracy or speed, but rarely manage interpretability, efficiency, and biological reliability. These gaps highlight the need for a statistical PHMM framework.

This study constructs a PHMM from a curated multiple sequence alignment of five regulatory DNA motifs (the TATA box, GC-box, telomeric repeat, octamer motif and E-box). By combining position based sequence weights using the Henikoff and Henikoff (1994) method and utilising Laplace pseudocounts for estimation of transition and emission probabilities, the study aims to evaluate motif similarity using the log-odds Viterbi algorithm (Borodovsky and Ekisheva, 2006). Through this approach, the study demonstrates how PHMMs can be effectively applied to model DNA motifs.

II. PROPOSED FRAMEWORK

2.1. Motif selection and Multiple Sequence Alignment

Sequence alignment arranges DNA, RNA or protein to identify similarity regions that reflect evolutionary relationships (Rashed et al., 2021). While Pairwise Alignment considers two sequences, Multiple Sequence Alignment (MSA) aligns multiple related sequences at once. In this study, five conserved DNA motifs were chosen (listed in Table 1) and used as training sequences for PHMM construction.

Table 1. List of regulatory DNA Motifs

Motif Name	Sequence Pattern							
TATA box	-	-	T	A	T	A	A	A
GC-box	-	-	G	G	G	C	G	G
Telomeric repeat	-	-	T	T	A	G	G	G
Octamer motif	A	T	G	C	A	A	A	T
E-box	-	-	C	A	C	G	T	G

The sequences were then aligned to identify conserved columns. Columns with <50% gaps were set as match states, while others were set as insert states. Then the alignment is reduced to the ungapped block used for further model construction.

2.2. Henikoff-Henikoff Sequence weighting

Different scoring methods are used in the sequence alignment to know the level of similarity (Petescia et al., 2025; Tawfeek et al., 2025; Vedula et al., 2025). In this study we used, the Henikoff and Henikoff weighting scheme (Henikoff and Henikoff, 1994). The weights of the sequences in the rows of block are derived as a linear combination of the column-specific weights. The weight α_{ij} for sequence i per column j is calculated by,

$$\alpha_{ji} = \frac{1}{m_j k_{j(s_i^j)}} \quad \dots (2)$$

where m_j is the number of letter types in column j , s_i^j is the letter at column j of sequence i , and $k_{j(s_i^j)}$ is the number of letters of s_i^j in column j . For example, the column-specific weights for the first column are: $\alpha_{11} = \frac{1}{6}, \alpha_{12} = \frac{1}{6}, \alpha_{13} = \frac{1}{6}, \alpha_{14} = \frac{1}{6}, \alpha_{15} = \frac{1}{3}$. Similarly, the weights for other columns were derived. Summing and normalizing the column specific weights for each sequence produced the final sequence weights: $\alpha_1 = \frac{29}{144}, \alpha_2 = \frac{23}{108}, \alpha_3 = \frac{71}{432}, \alpha_4 = \frac{29}{144}, \alpha_5 = \frac{95}{432}$

2.3. Weighted Emission and Transition counts

To determine the number of match states if the PHMM, we applied the heuristic rule that columns with less than 50% are marked as match state; otherwise it is marked as an insert state. Based on this criterion, six match states $M_1 \dots M_6$ were identified, corresponding to columns 3 – 8 of the MSA (1). For parameter estimation, weighted emission counts $E_k(s)$ and weighted transition counts A_{kl} were computed. For any given state k , the weighted emission (or transition) count is obtained by summing the weights of all training sequences that emit a particular symbol from that state. For example, in match state M_6 , the weighted emission count for nucleotide G is calculated by summing the weights of those sequences that contain G in that column (sequences 2, 3, and 4).

$$E_6^a(G) = \alpha_2 + \alpha_3 + \alpha_5 = \frac{23}{108} + \frac{71}{432} + \frac{95}{432} = \frac{43}{72}$$

Table 2. Weighted Observations of Nucleotide Emissions and Transitions between Hidden States

		0	1	2	3	4	5	6
Emission counts for match states	A	-	0	91/216	79/216	29/72	79/216	29/144
	C	-	95/432	29/144	95/432	23/108	95/432	0
	G	-	179/432	23/108	23/108	83/216	23/108	43/72
	T	-	79/216	71/432	29/144	0	29/144	29/144
Emission counts for insert states	A	29/144	0	0	0	0	0	0
	C	0	0	0	0	0	0	0
	G	0	0	0	0	0	0	0
	T	29/144	0	0	0	0	0	0
Counts of transitions between hidden states	M-M	-	1	1	1	1	1	1
	M-D	0	0	0	0	0	0	-
	M-I	0	0	0	0	0	0	0
	I-M	29/144	0	0	0	0	0	0
	I-D	0	0	0	0	0	0	-
	I-I	29/144	0	0	0	0	0	0
	D-M	-	0	0	0	0	0	0
	D-D	-	0	0	0	0	0	-
	D-I	-	0	0	0	0	0	0

Based on the weighted emission and transition counts in Table 2, PHMM parameters were obtained using weighted maximum-likelihood estimates of the emission probabilities $e_k(s)$ and transition probabilities a_{kl} . For each state k and each permitted transition $k \rightarrow l$, the transition probabilities are computed as

$$a_{kl} = \frac{A_{kl}^{(\alpha)}}{\sum_{l'} A_{kl'}^{(\alpha)}} = \frac{\sum_{i=1}^N \varepsilon_{kl}(i) \alpha_i}{\sum_{l'} \sum_{i=1}^N \varepsilon_{kl'}(i) \alpha_i}, \quad \dots (3)$$

and for every match or insert state k emitting nucleotide s , the emission probabilities are computed as

$$e_k(s) = \frac{E_k^\alpha(s)}{\sum_{s'} E_k^\alpha(s')} = \frac{\sum_{i=1}^N \delta_{ks}(i) \alpha_i}{\sum_{s'} \sum_{i=1}^N \delta_{ks'}(i) \alpha_i} \quad \dots (4)$$

Here α_i is the weight of sequence i ; $\varepsilon_{kl}(i)$ is the number of observed transitions from state k to l in sequence i ; and $\delta_{ks}(i)$ is the number of emissions of nucleotide s from state k in sequence i . Applying these equations to the counts in Table 2 yields the estimated emission and transition probabilities presented in Table 3.

Table 3. Parameters of the PHMM obtained using weighted sequence alignment

		0	1	2	3	4	5	6
Emission probabilities for match states	A	-	0.000	0.421	0.366	0.403	0.403	0.201
	C	-	0.220	0.201	0.220	0.213	0.220	0.000
	G	-	0.414	0.213	0.213	0.384	0.377	0.597
	T	-	0.366	0.164	0.201	0.000	0.201	0.201
Emission probabilities for insert states	A	0.500	0.000	0.000	0.000	0.000	0.000	0.000
	C	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	G	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	T	0.500	0.000	0.000	0.000	0.000	0.000	0.000
Transition probabilities between hidden states	M-M	-	1.000	1.000	1.000	1.000	1.000	1.000
	M-D	0.000	0.000	0.000	0.000	0.000	0.000	-
	M-I	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	I-M	0.500	0.000	0.000	0.000	0.000	0.000	0.000
	I-D	0.000	0.000	0.000	0.000	0.000	0.000	-
	I-I	0.500	0.000	0.000	0.000	0.000	0.000	0.000
	D-M	-	0.000	0.000	0.000	0.000	0.000	0.000
	D-D	-	0.000	0.000	0.000	0.000	0.000	-
	D-I	-	0.000	0.000	0.000	0.000	0.000	0.000

2.4. Laplace Pseudocount Regularization

Laplace pseudocounts were applied to prevent zero probability estimates in small MSAs. A normalized pseudocount of $1/5$ (or $1/N$ for N training sequences) was added to each weighted count in Table 2, excluding entries corresponding to states from which emissions or transitions are not permitted under the standard PHMM structure. The pseudocount-adjusted values obtained through this procedure are presented in Table 4.

Table 4. Observed nucleotide emissions and transitions between hidden states calculated using Laplace's pseudocounts

		0	1	2	3	4	5	6
Emission counts for match states	A	-	1/5	611/1080	611/1080	217/360	611/1080	289/720
	C	-	907/2160	289/720	907/2160	223/540	1/5	1/5
	G	-	1327/2160	223/540	223/540	631/1080	1247/2160	287/360
	T	-	611/1080	787/2160	289/720	1/5	907/2160	289/720
Emission counts for insert states	A	289/720	1/5	1/5	1/5	1/5	1/5	1/5
	C	1/5	1/5	1/5	1/5	1/5	1/5	1/5
	G	1/5	1/5	1/5	1/5	1/5	1/5	1/5
	T	289/720	1/5	1/5	1/5	1/5	1/5	1/5
Counts of transitions between hidden states	M-M	-	6/5	6/5	6/5	6/5	6/5	6/5
	M-D	1/5	1/5	1/5	1/5	1/5	1/5	-
	M-I	1/5	1/5	1/5	1/5	1/5	1/5	1/5
	I-M	289/720	1/5	1/5	1/5	1/5	1/5	1/5
	I-D	1/5	1/5	1/5	1/5	1/5	1/5	-
	I-I	289/720	1/5	1/5	1/5	1/5	1/5	1/5
	D-M	-	1/5	1/5	1/5	1/5	1/5	1/5
	D-D	-	1/5	1/5	1/5	1/5	1/5	-
	D-I	-	1/5	1/5	1/5	1/5	1/5	1/5

2.5. PHMM Parameter Estimation

The parameters of the PHMM are estimated by substituting the counts from Table 4 into the equations 3 and 4. The resulting emission and transition probability estimates are presented in Table 5. As shown, the inclusion of pseudocounts ensures that all generic transitions and emissions obtain strictly positive probabilities, establishing a complete and numerically stable parameterization for the PHMM(Nishida et al., 2009). Consequently, the model is capable of accommodating any nucleotide sequence, while assigning higher likelihoods to those sequences that exhibit greater similarity to the original training alignment.

Table 5. Final estimates of PHMM parameters

		0	1	2	3	4	5	6
Emission probabilities for match states	A	-	0.11	0.32	0.31	0.33	0.32	0.22
	C	-	0.23	0.23	0.23	0.23	0.11	0.11
	G	-	0.34	0.24	0.23	0.32	0.33	0.44
	T	-	0.31	0.21	0.22	0.11	0.24	0.22
Emission probabilities for insert states	A	0.33	0.25	0.25	0.25	0.25	0.25	0.25
	C	0.17	0.25	0.25	0.25	0.25	0.25	0.25
	G	0.17	0.25	0.25	0.25	0.25	0.25	0.25
	T	0.33	0.25	0.25	0.25	0.25	0.25	0.25
Transition probabilities between hidden states	M-M	-	0.75	0.75	0.75	0.75	0.75	0.86
	M-D	0.50	0.13	0.13	0.13	0.13	0.13	-
	M-I	0.50	0.13	0.13	0.13	0.13	0.13	0.14
	I-M	0.40	0.33	0.33	0.33	0.33	0.33	0.50
	I-D	0.20	0.33	0.33	0.33	0.33	0.33	-
	I-I	0.40	0.33	0.33	0.33	0.33	0.33	0.50
	D-M	-	0.33	0.33	0.33	0.33	0.33	0.50
	D-D	-	0.33	0.33	0.33	0.33	0.33	-
	D-I	-	0.33	0.33	0.33	0.33	0.33	0.50

III. RESULTS AND DISCUSSION

The PHMM achieved its final emission and transition probabilities (Table 5) through MSA analysis and two-stage estimation method. The emission distributions of match states (M1–M6) show the nucleotide composition from the aligned sequences while M1, M4 and M5 states demonstrate higher preference for specific nucleotides. The insert states showed uniform emission patterns at all positions because of pseudocount smoothing which indicates that inserted regions have lower positional specificity.

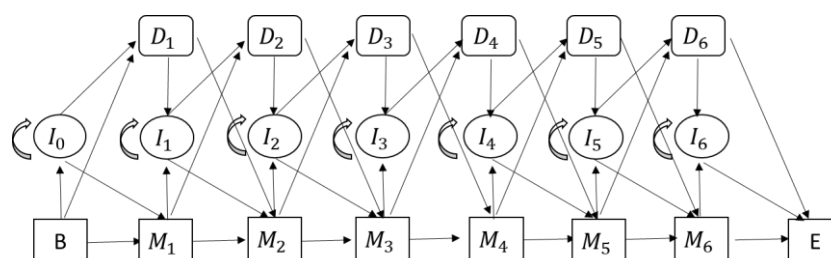


Fig. 1 PHMM state diagram with parameters estimated using pseudocounts

The transition probability estimates show that sequences follow the core motif structure through match-to-match (M–M) transitions between columns with high probabilities (0.75 for M1–M5 and 0.86 for M6). The alignment positions allow gaps because the transition probabilities from match to deletion (M–D $\approx$ 0.13) and insertion (M–I $\approx$ 0.13–0.14) states remain significant, although lower than match-to-match transitions. The insert states maintain equal transition probabilities between I–M, I–D and I–I states at 0.33. The delete states in all columns except the last one show equal transition probabilities to M, D and I states at 0.33. The PHMM parameters show that the model detects a moderately conserved motif at the beginning of the alignment while allowing insertions and deletions.

IV. ACKNOWLEDGMENT

The authors would like to thank the Department of Statistics, Periyar University, Salem, for providing the basic infrastructure and instrumentation facility for this research.

REFERENCES

- [1] Ali, A. 2001. Macroeconomic variables as common pervasive risk factors and the empirical content of the Arbitrage Pricing Theory. *Journal of Empirical finance*, 5(3): 221–240.
- [2] Basu, S. 1997. The Investment Performance of Common Stocks in Relation to their Price to Earnings Ratio: A Test of the Efficient Markets Hypothesis. *Journal of Finance*, 33(3): 663–682.
- [3] Bhatti, U. and Hanif, M. 2010. Validity of Capital Assets Pricing Model. Evidence from KSE-Pakistan. *European Journal of Economics, Finance and Administrative Science*, 3 (20).
- [1] Baridam, B., 2014. Application of the Codon-based Scoring Method in Motif Detection. *International Journal of Computer Applications* 92, 1–4.
- [2] Borodovsky, M., Ekisheva, S., 2006. Problems and Solutions in Biological Sequence Analysis. Cambridge University Press.
- Dai, J., Cheng, J., 2008. HMMEditor: a visual editing tool for profile hidden Markov model. *BMC Genomics* 9 Suppl 1, S8. <https://doi.org/10.1186/1471-2164-9-S1-S8>
- [3] Henikoff, S., Henikoff, J.G., 1994. Position-based sequence weights. *Journal of Molecular Biology* 243, 574–578. [https://doi.org/10.1016/0022-2836\(94\)90032-9](https://doi.org/10.1016/0022-2836(94)90032-9)
- [4] Horan, K., Shelton, C.R., Girke, T., 2010. Predicting conserved protein motifs with Sub-HMMs. *BMC Bioinformatics* 11, 205. <https://doi.org/10.1186/1471-2105-11-205>
- [5] Jeslin, J., Radhika, A., HariPriya, M., 2025. A Comparative Analysis of Probabilistic and Machine Learning Models for Biological Sequence Classification. *JP Journal of Biostatistics* 25, 273–294. <https://doi.org/10.17654/0973514325014>
- [6] Joseph, P., 2023. Comparative analysis of computational based motif refinement methods. *International Journal of Research in Engineering and Science (IJRES)* 11, 13–16.
- [7] Muttakin, A., Huq, M.R., 2017. Motif discovery in unaligned DNA sequences using genetic algorithm, in: 2017 4th International Conference on Advances in Electrical Engineering (ICAEE). Presented at the 2017 4th International Conference on Advances in Electrical Engineering (ICAEE), pp. 725–730. <https://doi.org/10.1109/ICAEE.2017.8255450>
- [8] Nishida, K., Frith, M.C., Nakai, K., 2009. Pseudocounts for transcription factor binding sites. *Nucleic Acids Res* 37, 939–944. <https://doi.org/10.1093/nar/gkn1019>
- [9] Petescia, A., Denti, L., Gafurov, A., Hodorová, V., Nosek, J., Brejová, B., Vinař, T., 2025. Alignment-free detection of differences between sequencing datasets. *iScience* 28, 113828. <https://doi.org/10.1016/j.isci.2025.113828>
- [10] Rashed, A.E.E.-D., Obaya, M., Moustafa, H.E., 2021. Accelerating DNA pairwise sequence alignment using FPGA and a customized convolutional neural network. *Computers & Electrical Engineering* 92, 107112. <https://doi.org/10.1016/j.compeleceng.2021.107112>
- [11] Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R.P., Banday, S., Mishra, A.K., Das, G., Malonia, S.K., 2023. Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology (Basel)* 12, 997. <https://doi.org/10.3390/biology12070997>
- [12] Sun, Y., Buhler, J., 2007. Designing patterns for profile HMM search. *Bioinformatics* 23, e36–e43. <https://doi.org/10.1093/bioinformatics/btl323>
- [13] Tawfeek, M.A., Alrashdi, I., Alruwaili, M., Elhady, G.F., 2025. A multi-objective fuzzy model based on enhanced artificial fish Swarm for multiple RNA sequences alignment. *Egyptian Informatics Journal* 29, 100627. <https://doi.org/10.1016/j.eij.2025.100627>
- [14] Vedula, S., Bronstein, A.M., Marx, A., 2025. Improving prediction accuracy in chimeric proteins with windowed multiple sequence alignment. *Computational and Structural Biotechnology Journal* 27, 3292–3298. <https://doi.org/10.1016/j.csbj.2025.07.039>
- [15] Wang, W., Wang, J., Si, S., Huang, Z., Xiao, J., 2022. RL-MD: A Novel Reinforcement Learning Approach for DNA Motif Discovery. pp. 1–7. <https://doi.org/10.1109/DSAA54385.2022.10032340>
- [16] Yu, Q., Zhao, X., Huo, H., 2019. A new algorithm for DNA motif discovery using multiple sample sequence sets. *J Bioinform Comput Biol* 17, 1950021. <https://doi.org/10.1142/S0219720019500215>