RESEARCH PAPER

A Hidden Markov–Entropy Dynamical Framework for Characterizing Mutation Hotspots in β-Thalassemia HBB Gene Sequences

S. D. Jeniffer[1], P. Maheswari[2]
[1,2]Department of Mathematics, Mepco Schlenk Engineering College, Sivakasi – 626 005.
Email: jenifferdavid1996@gmail.com, maheswari23061993@gmail.com
Received: 25th May, 2026; **Revised:** 6th June, 2026; **Accepted:** 8th June, 2026; **Available Online:** 16th June, 2026

ABSTRACT
β-Thalassemia is a hereditary blood disorder caused by mutations in the HBB gene that impair β-globin synthesis. This study proposes a Hidden Markov–Entropy framework to analyze mutation dynamics in 130 aligned HBB gene sequences of length 851 nucleotides. Sequence positions were classified into Match, Mutation, and Deletion states, and a transition probability matrix was constructed to model stochastic state transitions. Stationary distribution, mutation hotspot profiling, and Shannon entropy analysis were employed to investigate long-term state behaviour and sequence variability. The results revealed strong persistence of Match and Deletion states, while entropy and mutation hotspot analyses identified highly variable regions within the gene. A clear association between mutation burden and entropy was observed, indicating increased sequence complexity in mutation-prone regions. The proposed framework provides a mathematical and information-theoretic approach for characterizing mutation patterns in β-thalassemia and can be extended to the study of other genetic disorders.
Keywords: β-Thalassemia, HBB Gene, Hidden Markov Model, Stochastic Modelling, Mutation Hotspots, Shannon Entropy, Bioinformatics.
How to cite this article: Jeniffer SD, Maheswari P. A Hidden Markov–Entropy Dynamical Framework for Characterizing Mutation Hotspots in β-Thalassemia HBB Gene Sequences. Int J Drug Deliv Technol. 2026;16(60s): 1133-1139. DOI: 10.25258/ijddt.16.60s.124
Source of support: Nil.
Conflict of interest: None.

1. Introduction :

β-Thalassemia is a hereditary haemoglobin disorder caused by mutations in the HBB gene on chromosome 11, resulting in reduced or absent synthesis of the β-globin chain. The disease exhibits considerable genetic heterogeneity, with numerous mutations contributing to varying clinical manifestations ranging from mild anaemia to severe transfusion-dependent conditions. Consequently, the identification and characterization of mutation patterns within the HBB gene remain important areas of research in molecular genetics and bioinformatics.

Recent advances in sequencing technologies have generated extensive genomic datasets, enabling the application of mathematical and computational approaches to investigate sequence variability. Traditional mutation analysis methods primarily identify individual nucleotide changes but often provide limited information regarding the stochastic behaviour of mutations across biological sequences. Hidden Markov Models (HMMs) have emerged as effective probabilistic tools for modelling sequential data and have been successfully applied to gene prediction, sequence alignment, and mutation analysis. Similarly, Shannon entropy offers an information-theoretic measure for quantifying sequence variability and identifying conserved and mutation-prone regions.

HMMs and entropy measures have been widely employed independently, studies integrating stochastic state-transition analysis with entropy-based characterization of HBB gene mutations remain limited. Furthermore, the long-term behaviour of mutation states through stationary distribution analysis has received comparatively less attention in β-thalassemia research. Addressing these gaps may provide deeper insights into mutation dynamics and sequence complexity.

This study proposes a Hidden Markov–Entropy Dynamical Framework for analyzing aligned β-thalassemia HBB gene sequences. The framework employs Match, Mutation, and Deletion states to model sequence evolution, estimates transition probabilities and stationary distributions, identifies mutation hotspots, and quantifies sequence variability using Shannon entropy. By integrating stochastic modelling and information-theoretic analysis, the proposed approach provides a mathematical framework for investigating mutation dynamics in β-thalassemia gene sequences.

2. Review of Literature :

Hidden Markov Models (HMMs) have been widely applied in biological sequence analysis since their introduction to computational biology by Rabiner (1989) and subsequent extensions by Eddy

et al. (1995, 1998) for sequence alignment, homology detection, and profile modelling. Further advancements by Hughey et al. (1996), Stanke et al. (2006), and Schuster et al. (2007) established Profile HMMs as effective tools for analyzing genomic sequences. In parallel, multiple sequence alignment techniques developed by Simossis et al. (2003) and other researchers became fundamental for identifying conserved and variable regions in biological sequences.

Recent studies have focused on enhancing HMM training, parameter estimation, and disease-related applications. Emdadi et al. (2019) improved HMM parameter estimation using optimization techniques, while Karuppusamy et al. (2021) and Li et al. (2021) proposed advancements in sequence modelling and Baum–Welch training procedures. Jeniffer et al. (2021) employed Profile HMMs for TP53 gene analysis and transition probability estimation, demonstrating the usefulness of stochastic models in mutation studies. More recently, Jeniffer (2026) introduced a hybrid framework integrating higher-order Markov chains and fuzzy automata for predictive analysis of haemoglobin beta gene mutations. By combining sequence alignment, entropy measures, Hidden Markov Models, and co-occurrence networks, the study achieved high predictive performance and highlighted the effectiveness of stochastic and fuzzy-based methods for modelling mutation dynamics in β-thalassemia sequences. These developments emphasize the growing importance of probabilistic and information-theoretic approaches in genomic sequence analysis and disease prediction.

3. Materials and Methods :

3.1 Dataset Description :

The dataset used in this study consisted of 130 aligned β-thalassemia HBB gene sequences obtained in FASTA alignment format. Following multiple sequence alignment, each sequence possessed a uniform length of 851 nucleotide positions. The aligned sequences formed the basis for stochastic state modelling, mutation hotspot analysis, and entropy-based characterization of sequence variability.

The aligned dataset can be represented as

$$X = \{X_1, X_2, \dots, X_N\}$$

where $N = 130$ denotes the total number of sequences. Each sequence is expressed as $X_i = (x_{i1}, x_{i2}, \dots, x_{iL})$ where $L = 851$ represents the alignment length and x_{ij} in $\{A, C, G, T, -\}$ corresponds to the nucleotide or gap symbol at position j of sequence i.

3.2 Consensus Sequence Construction :

A consensus sequence was generated from the aligned dataset by selecting the most frequently occurring nucleotide at each alignment position. Let f_{kj} denote the frequency of nucleotide k at position j. The consensus nucleotide at position j is defined as $c_j = max_k\, f_{kj}$ for $k\ in\{A, C, G, T\}$

The resulting consensus sequence is given by $C = (c_1, c_2, \dots, c_L)$ and serves as the reference sequence for identifying mutations and deletions.

3.3 Hidden State Representation :

Each nucleotide position was assigned to one of three hidden states: M = Match, U = Mutation, D = Deletion. The state assignment was performed by comparing each aligned sequence with the consensus sequence.

For position j, $S_j = f(x) =$

$$\begin{cases} M,\ x_j = c_j \\ U,\ x_j \neq c_j\ and\ x_j \neq - \\ D,\ x_j = - \end{cases}$$

Thus, every sequence was transformed into a state sequence $S = (s_1, s_2, \dots, s_L)$ where $s_j\ in\ \{M, U, D\}$ This state representation forms the stochastic process used throughout the analysis.

3.4 Transition Probability Matrix :

The transition probability matrix quantifies the probability of moving from one hidden state to another. Let N_{ij} denote the number of observed transitions from state i to state j.

The transition probability is estimated as, $p_{ij} = \frac{N_{ij}}{\sum_j N_{ij}}$. The transition probability matrix is therefore $P = \begin{bmatrix} p_{MM} & p_{MU} & p_{MD} \\ p_{UM} & p_{UU} & p_{UD} \\ p_{DM} & p_{DU} & p_{DD} \end{bmatrix}$ where $\sum_j p_{ij} = 1$ for every state i. This matrix characterizes the mutation dynamics across the aligned HBB sequences.

3.5 Stationary Distribution Analysis:

The long-term behaviour of the stochastic process was investigated through stationary distribution analysis. A stationary probability vector $\pi = [\pi_M, \pi_U, \pi_D]$ satisfies $\pi P = \pi$ subject to $\pi_M + \pi_U + \pi_D = 1$. The stationary distribution represents the equilibrium probabilities of occupying the Match, Mutation, and Deletion states after a large number of transitions.

3.6 Mutation Hotspot Analysis :

Mutation hotspots were identified by calculating the mutation frequency at each nucleotide position. Let m_j denote the number of sequences that differ from the consensus sequence at position j. The mutation frequency at position j is given by $F_j = \frac{m_j}{N}$ where $0 \leq F_j \leq 1$. Positions with larger values of F_j correspond to highly variable regions and are considered potential mutation hotspots. The mutation hotspot profile is expressed as, $F = [F_1, F_2, \dots, F_L]$ across the entire alignment.

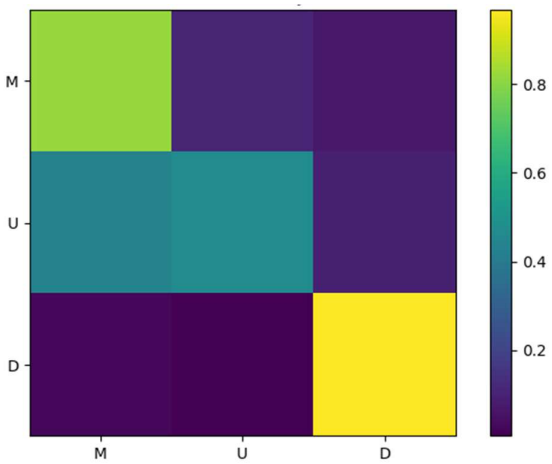
3.7 Shannon Entropy Analysis :

Sequence variability was quantified using Shannon entropy. For position j, let $p_k(j)$ represent the probability of observing nucleotide k.

The entropy at position j is defined as,

$$H_{j=} - \sum_k p_k(j)\, log_2 p_k(j) \text{ where } k \in \{A, C, G, T, -\}$$

Entropy values close to zero indicate highly conserved positions, whereas larger entropy values correspond to increased sequence variability and uncertainty.



The entropy profile is represented by, $H = [H_1, H_2, \ldots, H_L]$ and provides an information-theoretic measure of genetic diversity across the HBB gene alignment.

3.8 Computational Environment

All computations were performed using Python. Sequence processing and alignment analysis were carried out using the Biopython library, while numerical computations were performed using NumPy and SciPy. Data visualization and graphical representations were generated using Matplotlib. The transition probability matrix, stationary distribution, mutation hotspot profile, and entropy measures were computed directly from the aligned β-thalassemia HBB gene sequences.

4. Results and Discussion

4.1 Mutation Score Statistics :

The aligned β-thalassemia dataset consisted of 130 HBB gene sequences with a uniform alignment length of 851 nucleotides. Mutation scores were calculated by comparing each sequence with the generated consensus sequence. The mutation score represents the proportion of positions that differ from the consensus sequence. The statistical summary of mutation scores is presented in Table 1.

Table 1. Summary of Mutation Scores

Statistic	***Value***
Mean	0.0593
Maximum	0.2409
Minimum	0.0000

The mean mutation score of 0.0593 indicates that approximately 5.93% of nucleotide positions differ from the consensus sequence across the analyzed dataset. The maximum mutation score of 0.2409 suggests the presence of highly divergent sequences containing substantial nucleotide variation. Conversely, several sequences exhibited mutation scores close to zero, indicating strong similarity to the consensus profile.

The observed variation in mutation scores reflects the genetic heterogeneity associated with β-thalassemia and demonstrates the suitability of stochastic modelling techniques for characterizing sequence variability.

4.2 Transition Probability Matrix Analysis :

The transition probability matrix estimated from the hidden-state sequences is given by

$$P = \begin{bmatrix} 0.8190 & 0.1092 & 0.0718 \\ 0.4377 & 0.4688 & 0.0935 \\ 0.0266 & 0.0067 & 0.9667 \end{bmatrix}$$

where M, U, and D denote Match, Mutation, and Deletion states respectively. The corresponding heatmap is shown in Figure 1.

Figure 1 Heatmap of the Transition Probability Matrix

The transition probability from Match to Match was found to be 0.8190, indicating a strong tendency for conserved nucleotide positions to remain conserved. This behaviour is expected in functional genomic regions where sequence preservation is biologically important. The Mutation state exhibited a self-transition probability of 0.4688. This indicates that once a mutation occurs, neighbouring positions frequently remain within the mutation state, suggesting the existence of mutation clusters within specific regions of the HBB gene. The Deletion state displayed the highest self-transition probability of 0.9667, indicating the persistence of deletion blocks throughout the aligned sequences. This observation suggests that deletion-associated regions contribute significantly to sequence variability within the dataset. The transition probability from Mutation to Match (0.4377) exceeded the transition probability from Mutation to Deletion (0.0935), implying that many

mutation events remain localized rather than expanding into larger deletion regions.

4.3 Stationary Distribution Analysis :

The stationary distribution associated with the transition matrix was obtained by solving $\pi P = \pi$. The resulting stationary probability vector is,

$$\pi = [0.2455 \quad 0.0593 \quad 0.6952]$$

The stationary probabilities indicate that, in the long term, approximately 24.55% of positions are associated with the Match state, 5.93% with the Mutation state, and 69.52% with the Deletion state. The relatively low equilibrium probability of the Mutation state suggests that mutation events are localized and occur intermittently throughout the alignment. In contrast, the dominant stationary probability of the Deletion state indicates that deletion-related patterns exert substantial influence on the overall sequence dynamics.

From a stochastic perspective, the stationary distribution reveals the equilibrium structure of the aligned HBB sequences and highlights the importance of structural variations in shaping long-term sequence behaviour.

4.4 Mutation Hotspot Identification :

Mutation hotspot analysis was performed by computing mutation frequencies across all 851 alignment positions. The resulting mutation hotspot profile is illustrated in Figure 2.

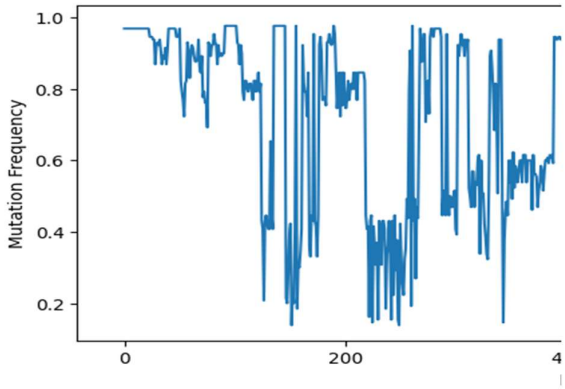


Figure 2. Mutation Hotspot Profile of the aligned sequences

The profile reveals substantial variation in mutation frequencies throughout the alignment. Several positions exhibited mutation frequencies approaching unity, indicating that a large proportion of sequences differed from the consensus sequence at those locations.

Three categories of regions can be identified:

- Conserved regions exhibiting low mutation frequencies.
- Moderately variable regions characterized by intermediate mutation frequencies.
- Mutation hotspots displaying exceptionally high mutation frequencies.

The hotspot regions likely correspond to locations within the HBB gene that experience elevated mutational activity and contribute significantly to the genetic diversity associated with β-thalassemia. The non-uniform distribution of mutation frequencies suggests that mutation accumulation is concentrated within specific genomic segments rather than occurring randomly across the entire gene.

4.5 Entropy-Based Variability Analysis :

The entropy profile of the aligned HBB sequences is presented in Figure 3. Shannon entropy was used to quantify sequence uncertainty at each nucleotide position.

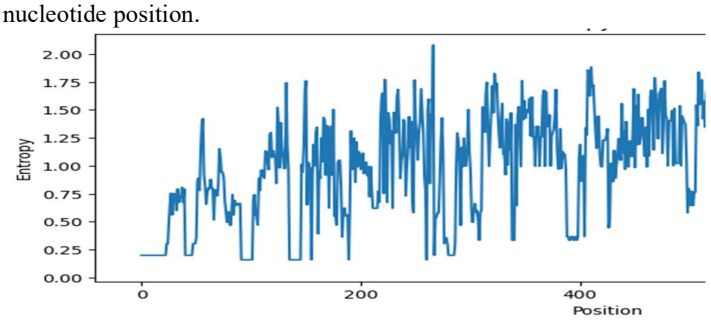


Figure 3 Entropy Profile of the aligned sequences

The entropy values varied considerably across the alignment, ranging from near-zero values in highly conserved regions to values exceeding 2.0 in highly variable regions. Low entropy values indicate positions where a single nucleotide dominates the alignment, reflecting strong evolutionary conservation. Conversely, high entropy values correspond to positions containing multiple nucleotide variants and therefore greater uncertainty. The central portion of the alignment exhibited elevated entropy levels, suggesting increased sequence diversity within these regions. Several entropy peaks were observed throughout the alignment, indicating positions characterized by substantial nucleotide variability.

4.6 Mutation Accumulation Dynamics :

To investigate the cumulative behaviour of mutations across the HBB gene alignment, the mutation accumulation function $M(t)$ was computed. This function represents the average mutation frequency observed up to position ttt and provides a quantitative measure of mutation burden along the sequence. The mutation accumulation profile is shown in Figure 4.

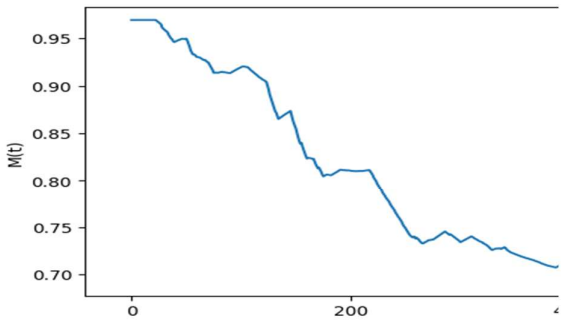


Figure 4 Mutation Accumulation Function of the aligned sequences

The curve initially exhibits high values approaching 0.97, indicating that the early alignment positions contain a substantial proportion of mutations relative to the consensus sequence. As the sequence position increases, the mutation burden gradually decreases and stabilizes around 0.70. The decreasing trend suggests that mutation events are not uniformly distributed throughout the alignment. Instead, the early regions contribute disproportionately to the cumulative mutation burden, whereas later regions contain a mixture of conserved and variable positions.

After approximately position 550, the mutation accumulation function begins to increase slightly and approaches a value of approximately 0.75 near the terminal region. This behaviour indicates the presence of additional mutation hotspots within the latter portion of the HBB alignment. From a stochastic perspective, the mutation accumulation function provides a global measure of sequence variability and demonstrates that mutation burden evolves dynamically across genomic positions rather than remaining constant.

4.7 Dynamical Mutation Burden Model :

To describe the evolution of mutation burden mathematically, a dynamical system was introduced based on the mutation accumulation function. The mutation burden acts as an external driving force influencing the system dynamics.

The governing equation is $\frac{dD}{dt} = rD\left(1 - \frac{D}{K}\right) + \beta M(t)$

where,

- $D(t)$ denotes the Mutation severity index,
- r represents the intrinsic growth parameter,
- K is the carrying capacity,
- $M(t)$ is the mutation accumulation function,
- β quantifies the influence of mutation burden.

The numerical solution of the model is illustrated in Figure 5.

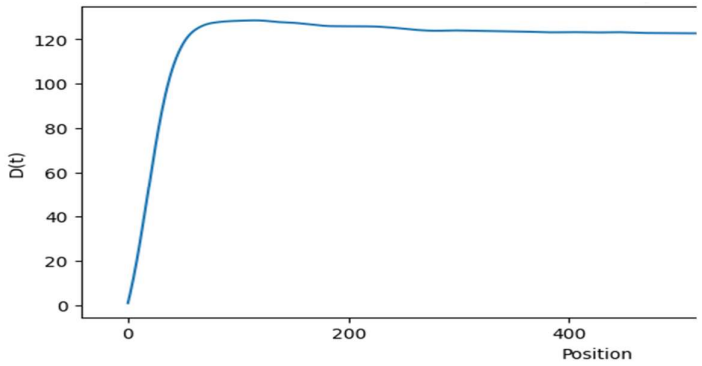


Figure 6 Mutation Severity Index Derived from the Dynamical Mutation Burden Model

Initially, the severity index increases rapidly from its starting value due to the large mutation burden present in the early sequence positions. The growth subsequently slows and approaches a stable equilibrium near $D(t) \approx 125$. The stabilization of the curve indicates that the mutation burden reaches a dynamic balance with the nonlinear regulatory component of the logistic model. The equilibrium behaviour suggests that the cumulative effects of mutations eventually become bounded rather than increasing indefinitely. Although the model does not represent clinical disease progression directly, it provides a mathematically interpretable mutation-severity index derived solely from sequence information.

The logistic framework was adopted as a generic nonlinear growth model to represent bounded mutation severity. Since clinical measurements were unavailable, the parameters r, K, and β were selected empirically to demonstrate the influence of mutation accumulation on the dynamical behaviour of the system.

4.8 Relationship Between Mutation Burden and Entropy :

A comparison of the mutation hotspot profile and entropy profile reveals a strong correspondence between mutation frequency and sequence uncertainty. Regions exhibiting high mutation frequencies generally coincide with entropy peaks, indicating that mutation accumulation contributes directly to increased sequence diversity. Similarly, conserved regions characterized by low mutation frequencies correspond to low entropy values. This relationship confirms that mutation burden and entropy are closely linked descriptors of sequence variability. While mutation frequency quantifies the occurrence of sequence alterations, entropy measures the resulting uncertainty within nucleotide composition. The combined use of these metrics therefore provides a comprehensive representation of mutation dynamics within β-thalassemia HBB sequences.

4.9 Biological Implications :

The proposed Hidden Markov–Entropy Dynamical Framework successfully identified conserved regions, mutation hotspots, and long-term stochastic characteristics of β-thalassemia HBB sequences. The transition probability analysis revealed strong persistence within Match and Deletion states, while stationary distribution analysis highlighted the dominance of deletion-related patterns in long-term sequence behaviour. Mutation hotspot profiling identified regions of elevated mutational activity, and entropy analysis quantified the associated sequence uncertainty.

Together, these findings demonstrate that β-thalassemia-associated HBB sequences exhibit heterogeneous mutation patterns characterized by localized mutation clusters and variable levels of sequence complexity. The integration of Hidden Markov modelling and entropy analysis provides a mathematically interpretable framework for studying mutation dynamics and may be extended to other hereditary genetic disorders.

5. Conclusion:

This study presented a Hidden Markov–Entropy Dynamical Framework for characterizing mutation dynamics in β-thalassemia HBB gene sequences. A dataset consisting of 130 aligned HBB sequences with an alignment length of 851 nucleotides was analyzed using stochastic state-transition modelling, mutation hotspot profiling, entropy-based variability assessment, and dynamical mutation burden analysis.

The Hidden Markov representation successfully transformed aligned nucleotide sequences into Match, Mutation, and Deletion states, enabling the estimation of transition probabilities governing sequence evolution. The resulting transition probability matrix revealed strong persistence within Match and Deletion states, indicating the coexistence of highly conserved regions and structurally variable segments. Stationary distribution analysis further demonstrated the dominant influence of deletion-associated patterns on the long-term stochastic behaviour of the aligned sequences.

Mutation hotspot analysis identified substantial heterogeneity across the HBB gene, revealing regions characterized by elevated mutation frequencies. Shannon entropy analysis complemented these findings by quantifying positional uncertainty and sequence complexity. The observed correspondence between mutation hotspots and entropy peaks confirmed the close relationship between mutation accumulation and sequence variability.

To provide a dynamical perspective, a mutation accumulation function was introduced and incorporated into a nonlinear differential equation framework. The resulting mutation severity index illustrated how cumulative mutation burden evolves across genomic positions and eventually approaches a stable equilibrium. This mathematical representation offers an interpretable mechanism for studying mutation-driven sequence dynamics using genomic information alone. By providing a quantitative framework for analyzing sequence variability, the proposed approach contributes to the growing intersection of mathematical biology, bioinformatics, and stochastic modelling.

Declaration :

The authors declared that there is no conflict of interest.

Funding :

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Reference:

1. Böer, J. (2016). Multiple alignment using hidden Markov models. *proteins*, *4*, 14.
2. Bonidia, R. P., Domingues, D. S., Sanches, D. S., & de Carvalho, A. C. (2021). MathFeature: feature extraction package for DNA, RNA and protein sequences based on mathematical descriptors. *Briefings in Bioinformatics*.
3. Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.
4. Eddy, S. R. (1995, July). Multiple alignment using hidden Markov models. In *Ismb* (Vol. 3, pp. 114-120).
5. Emdadi, A., Moughari, F. A., Meybodi, F. Y., & Eslahchi, C. (2019). A novel algorithm for parameter estimation of Hidden Markov Model inspired by Ant Colony Optimization. *Heliyon*, *5*(3), e01299.
6. Gagniuc, P. A. (2017). *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons.
7. Hughey, R., & Krogh, A. (1996). Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Bioinformatics*, *12*(2), 95-107.
8. Jeniffer, S. D. (2026). Integrating higher-order Markov modeling with fuzzy automata for predictive genomics of hemoglobin subunit beta mutations. *The European Physical Journal Plus*, *141*(5), 485.
9. Jeniffer, S.D., & Senthamarai Kannan, K., (2021). Stochastic modelling for identifying, *Advances and Applications of Mathematical Sciences*, *20*(9), 1923-1936.
10. Kannan, K. S., & Jeniffer, S. D. (2022, October). Hidden Markov Modelling for Biological Sequence. In *Proceedings of*

International Conference on Computational Intelligence: ICCI 2021 (p. 383). Springer Nature.
11. Karuppusamy, T. (2021). Biological gene sequence stucture analysis using hidden markov model. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(4), 1652-1666.
12. Kumar, S., & Gadagkar, S. R. (2001). Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics*, *158*(3), 1321-1327.
13. Li, J., Lee, J. Y., & Liao, L. (2021). A new algorithm to train hidden Markov models for biological sequences with partial labels. BMC bioinformatics, 22(1), 1- 21.
14. Meng, Y., & Fei, J. (2022). Hidden service publishing flow homology comparison using profile-hidden markov model. *International Journal of Intelligent Systems*.
15. Modica, G., & Poggiolini, L. (2012). Discrete Time Markov Chains. In *A First Course in Probability and Markov Chains* (pp. 168-240). John Wiley & Sons, Ltd.
16. Mor, B., Garhwal, S., & Kumar, A. (2021). A Systematic Review of Hidden Markov Models and Their Applications. Archives of computational methods in engineering, 28(3).
17. Procter, J. B., Carstairs, G. M., Soares, B., Mourão, K., Ofoegbu, T. C., Barton, D., & Barton, G. J. (2021). Alignment of biological sequences with Jalview. In Multiple Sequence Alignment (pp. 203-224). Humana, New York, NY.
18. Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257-286.
19. Roth, C. (2021). Statistical methods for biological sequence analysis for DNA binding motifs and protein contacts (Doctoral dissertation, Georg-August University).
20. Sarkar, B. K. (2021). Entropy Based Biological Sequence Study. In Entropy and Exergy in Renewable Energy. IntechOpen.
21. Sasidharan, S. K., & Thomas, C. (2021). ProDroid—An Android malware detection framework based on profile hidden Markov model. *Pervasive and Mobile Computing*, *72*, 101336.
22. Schuster-Böckler, B., & Bateman, A. (2007). An introduction to hidden Markov models. *Current protocols in bioinformatics*, *18*(1), A-3A.
23. Simossis, V., Kleinjung, J., & Heringa, J. (2003). An overview of multiple sequence alignment. *Current protocols in bioinformatics*, *3*(1), 3-7.
24. Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC bioinformatics*, *7*(1), 1-11.
25. Yoon, B. J. (2009). Hidden Markov models and their applications in biological sequence analysis. Current genomics, 10(6), 402-415.