Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 23s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Sequence Entropy And Markov Modelling Of Mutated *CFTR* Genes: Insights From Multiple Sequence Alignment

S. D. Jeniffer[1]

[1]Department of Mathematics, Mepco Schlenk Engineering College, Sivakasi, India - 626005.
Email: jenifferdavid1996@gmail.com

Cite this paper as: S. D. Jeniffer, (2025) Sequence Entropy And Markov Modelling Of Mutated *CFTR* Genes: Insights From Multiple Sequence Alignment. *Journal of Neonatal Surgery*, 14 (23s), 1043-1050.

ABSTRACT
Mutations in the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene are central to the development of cystic fibrosis, a severe genetic disorder affecting epithelial function. This study analyses 35 mutated CFTR gene sequences to explore underlying sequence variation and structural patterns. Multiple sequence alignment was performed to organize the sequences, followed by Shannon entropy analysis to identify regions of high variability and conservation. Hierarchical clustering provided insights into relationships among the mutated sequences, while sequence logo plots visually highlighted nucleotide distribution at each alignment position. To model the sequence behaviour statistically, a first-order Markov chain was constructed, capturing transition probabilities between nucleotides across the aligned sequences. Together, these methods offer a comprehensive view of the mutational landscape within the CFTR gene. The findings enhance our understanding of sequence-level mutation dynamics and provide a foundation for further computational modelling and genotype-phenotype correlation studies in cystic fibrosis research.

Keywords: *Cystic Fibrosis, CFTR Gene, Gene Mutation, Multiple Sequence Alignment, Shannon Entropy, Markov Chain Model, Genotype-Phenotype Correlation*

1. INTRODUCTION

The Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene encodes a membrane protein responsible for regulating chloride and bicarbonate transport across epithelial cells. Mutations in CFTR are the primary cause of cystic fibrosis (CF), a genetic disorder that affects the respiratory, digestive, and reproductive systems. To date, over 2,000 mutations have been identified in CFTR, with varying effects on gene function and disease severity. Understanding the structural and sequence-level implications of these mutations is critical for advancing both diagnostic and therapeutic strategies.

Computational approaches such as multiple sequence alignment (MSA), entropy analysis, and probabilistic modelling have emerged as powerful tools for characterizing genetic variation. Sequence alignment allows for the identification of conserved and divergent regions across multiple variants, while Shannon entropy quantitatively captures the variability at each alignment position. These metrics can provide valuable insights into mutation hotspots and functionally important domains.

Beyond descriptive analysis, probabilistic frameworks like Markov chains enable the modelling of nucleotide transitions within aligned sequences. By capturing the likelihood of one nucleotide following another, such models can reveal patterns in mutation behaviour and potentially inform evolutionary or structural constraints on the gene.

In this study, 35 mutated CFTR sequences were analysed using a combination of information- theoretic and statistical techniques. The goals were to identify conserved and variable regions through entropy profiling, group sequences based on similarity using hierarchical clustering, visualize nucleotide distributions with sequence logos, and model sequence behaviour using Markov chains. This integrative approach aims to deepen our understanding of how mutations shape the CFTR gene's sequence architecture and to lay the groundwork for future computational analyses of gene variants.

2. REVIEW OF LITERATURE

The Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene encodes a chloride channel essential for regulating fluid transport across epithelial surfaces. Mutations in this gene disrupt ion balance, resulting in cystic fibrosis (CF), a prevalent autosomal recessive disorder, particularly in the Caucasian population. More than 2,000 CFTR mutations have been identified, including missense, nonsense, splicing, and regulatory variants (Sosnay et al., 2013). These mutations differ in their functional impact, influencing both the severity and manifestation of disease phenotypes.

Beyond classic CF, CFTR mutations have been linked to non-classical disorders, such as congenital bilateral absence of the vas deferens and idiopathic pancreatitis. Emerging studies have also implicated CFTR variants in forms of monogenic diabetes, including neonatal diabetes, suggesting a potential role in pancreatic β-cell dysfunction (Hasan et al., 2022; Choi et al., 2021).

Advancements in computational biology have enhanced the ability to analyze such genetic variations. Multiple Sequence Alignment (MSA) remains a foundational tool in bioinformatics for comparing homologous sequences and identifying conserved or variable regions. Shannon entropy- based techniques are frequently used to quantify sequence variability; high entropy often highlights mutational hotspots, while low entropy may indicate functionally important conserved regions (Capra & Singh, 2007). Complementary to entropy, nucleotide diversity (π) provides insights into the polymorphic burden of a gene, helping to distinguish pathogenic from benign variants.

To explore relationships among sequences, hierarchical clustering has been applied to group similar mutation profiles, aiding in evolutionary and functional classification (Shannon et al., 2003). Probabilistic models, particularly first-order Markov chains, have been employed to model the statistical dependencies between nucleotides, revealing context-dependent mutation patterns (Durbin et al., 1998).

Building on these foundations, several studies have explored Hidden Markov Models (HMMs) and other computational tools to enhance disease gene analysis. HMM parameter estimation has been refined using techniques such as Ant Colony Optimization when the hidden path is unknown (Emdadi et al., 2019). Comparative gene alignment tools have also been used to analyse disease genes. Sequence modelling approaches have evolved to model intron and exon structures based on mathematical formulations of hidden states (Karuppusamy et al., 2021).

Feature extraction from biological sequences has been improved through mathematical descriptors using tools like MathFeature (Bonidia et al., 2021). Efforts to align SARS-CoV-2 gene sequences (Kumari et al., 2021) and TP53 sequences (Jeniffer et al., 2021) led to enhanced training of profile HMMs using EM algorithms. Systematic reviews have examined HMM applications in bioinformatics, further supporting their validity for gene prediction and sequence analysis (Mor et al., 2021).

Advanced entropy-based techniques have been used to identify key regions of viral genomes (Sarkar, 2021), while motif detection has been conducted using probabilistic models like BaMM (Roth et al., 2021). The Baum–Welch algorithm has been optimized to improve HMM training in biological data (Li et al., 2021), and JalView continues to be a valuable platform for aligned sequence visualization (Proctor et al., 2021).

Applications of HMM and PHMM extend beyond biology. For instance, HMMs have achieved high accuracy in malware detection (Sasidharan et al., 2021) and have been used to assess homology in publishing platforms (Meng et al., 2022). In the context of CF, HMM-based approaches using CFTR sequences have shown potential for early disease detection (Muthu et al., 2022). Additionally, hybrid computational approaches combining HMM, Markov models, and artificial neural networks have been employed to study malignancies at the sequence level (Senthamarai Kannan et al., 2022).

Despite these advances, the role of CFTR mutations in emerging phenotypes such as neonatal diabetes remains underexplored. The application of entropy analysis, sequence diversity metrics, and Markov-based modeling offers a promising framework to unravel the functional and clinical impact of CFTR sequence variation.

3. MATERIALS AND METHODS

3.1 Data Collection

A total of 35 mutated *CFTR* gene sequences were retrieved from publicly available genetic databases such as GenBank (NCBI) and UniProt. Sequences were selected based on the presence of known point mutations or deletions affecting the CFTR coding region. All sequences were downloaded in FASTA format for downstream analysis.

3.2 Multiple Sequence Alignment (MSA)

Multiple sequence alignment was performed using Clustal Omega, a widely used tool for accurate alignment of nucleotide or protein sequences. The aligned output was saved in both FASTA and Clustal formats. This alignment served as the foundation for entropy analysis, clustering, and modeling.

3.3 Shannon Entropy Calculation

To quantify sequence variability at each aligned position, Shannon entropy was computed using the formula:

$$H(x) = \sum_{i=1}^{n} p_i \log_2(p_i)$$

where p_i represents the probability of nucleotide i at a given position. Positions with higher entropy

values indicate greater variability. A custom Python script utilizing the Biopython library was used for this computation.

3.4 Hierarchical Clustering

Pairwise sequence distances were calculated using Hamming distance, based on nucleotide mismatches across the aligned sequences. A distance matrix was generated and used to perform hierarchical clustering with average linkage (UPGMA method). The clustering results were visualized as a dendrogram using the SciPy and Matplotlib libraries in Python.

3.5 Sequence Logo Generation

To visualize nucleotide conservation and variation, a sequence logo was generated using WebLogo, an online tool that creates graphical representations of aligned sequences. The height of each letter in the logo corresponds to its frequency at that position, reflecting the information content derived from entropy.

3.6 Markov Chain Construction

A first-order Markov chain model was developed to capture the transition probabilities between nucleotides across the aligned sequences. Each nucleotide (A, T, G, C) was treated as a discrete state. Transition probabilities were computed by counting adjacent nucleotide pairs throughout the alignment and normalizing by the total transitions from each state. The resulting transition matrix provides insights into common mutation paths or conserved nucleotide flows.

3.7 Tools and Libraries

All computational analyses were performed using Python (v3.10) with the following libraries:

- Biopython for sequence parsing and manipulation
- NumPy and Pandas for data handling
- Matplotlib and Seaborn for visualization
- SciPy for clustering and statistical operations

4. RESULTS

4.1 Multiple Sequence Alignment Overview

The multiple sequence alignment of 35 mutated *CFTR* gene sequences revealed well- conserved regions interspersed with highly variable segments. These patterns provided a basis for downstream entropy profiling, diversity analysis, and probabilistic modelling.

4.2 Shannon Entropy and Annotated Regions

Entropy profiling across the alignment showed extensive variability at several nucleotide positions. As illustrated in *Figure 1*, many alignment positions exceed the high entropy threshold of

1.5 (red dashed line), indicating mutational hotspots. Regions such as R9 (positions 35–38, mean entropy ≈ 1.80) were particularly variable, while early segments like R1 (positions 1–3, mean entropy

≈ 0.21) were highly conserved.

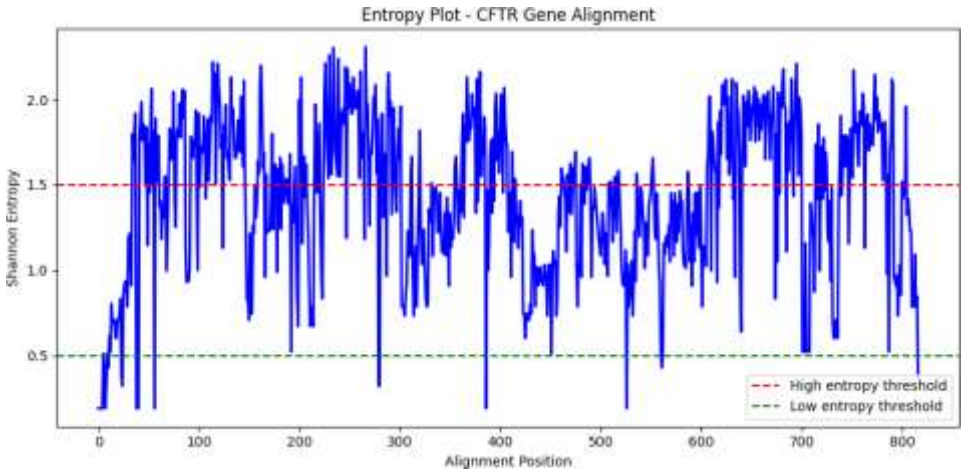

Figure 1: Shannon entropy across alignment positions.

The red line indicates high entropy (1.5), and the green line indicates low entropy (0.5). A total of 102 annotated regions were identified based on entropy peaks and valleys, supporting localized sequence variability critical for understanding mutation behaviour in CFTR.

4.3 Nucleotide Diversity (π) :

The nucleotide diversity across the sequences was calculated to be $\pi = 0.556$, indicating a moderate degree of polymorphism. This supports the entropy findings and reflects a mix of conserved and diverse regions across the mutated sequences.

4.4 Hierarchical Clustering of Mutated Sequences

Hierarchical clustering grouped the sequences into three main clusters based on nucleotide similarity (Hamming distance). The dendrogram in *Figure 2* reveals patterns of shared mutation profiles. For instance, sequences indexed 0–8 formed a tightly connected sub-cluster, indicating similar mutational characteristics.

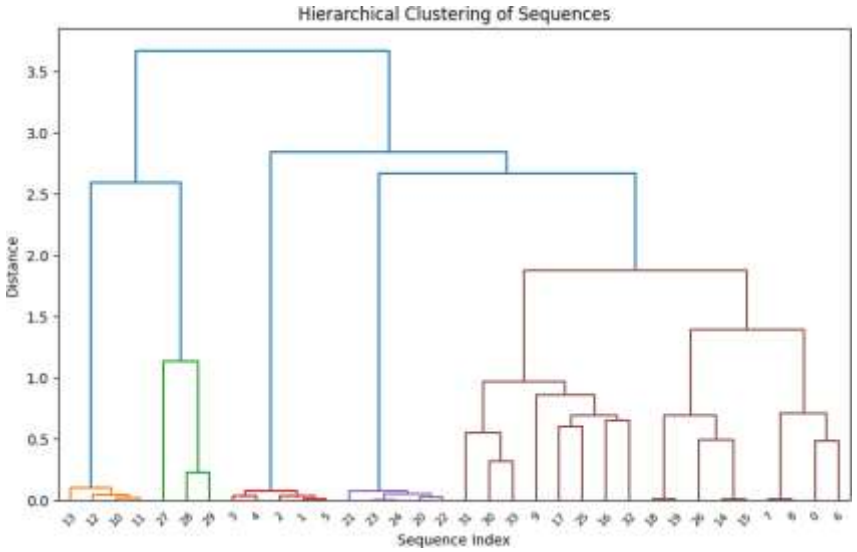

Figure 2: Dendrogram showing hierarchical clustering of the 35 mutated CFTR sequences.

These clusters may reflect shared evolutionary origins or mutation mechanisms affecting specific regions of the gene.

4.5 Sequence Logo Visualization

The sequence logo, showed in *Figure 3*, reflected the entropy results: conserved positions appeared as dominant bases, while high-entropy regions showed a stacked mix of nucleotides. This visually validated the entropy-based annotation.

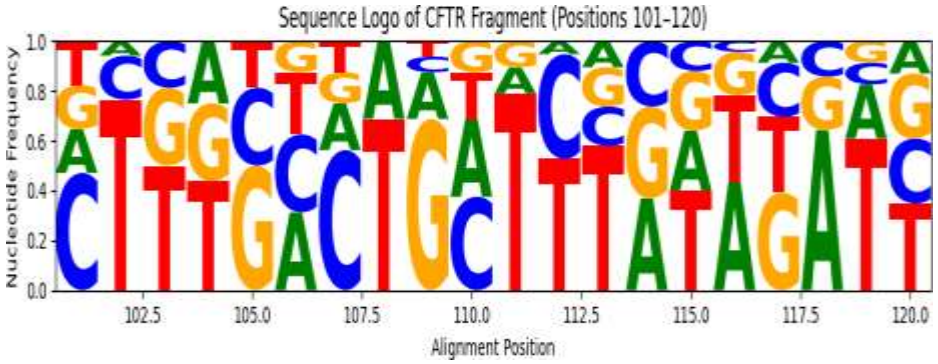

Figure 3: Fragment of Sequence logo

4.6 First-Order Markov Chain Modelling

Each nucleotides inter and intra transitions are documented, and the counts are determined to be the transition frequencies. Each transition would be divided by the sum of the associated row to obtain the transition probabilities, as per the accepted

definition of a Markov chain. Below is a digraph of transition probabilities as well as a matrix of transition probabilities (from A, C, G T to A, C, G, T) for each nucleotide.

$$P = \begin{pmatrix} 0.2857 & 0.1774 & 0.2684 & 0.2684 \\ 0.3750 & 0.1667 & 0.0278 & 0.4306 \\ 0.3392 & 0.1754 & 0.1813 & 0.3041 \\ 0.1963 & 0.1845 & 0.2741 & 0.3481 \end{pmatrix}$$

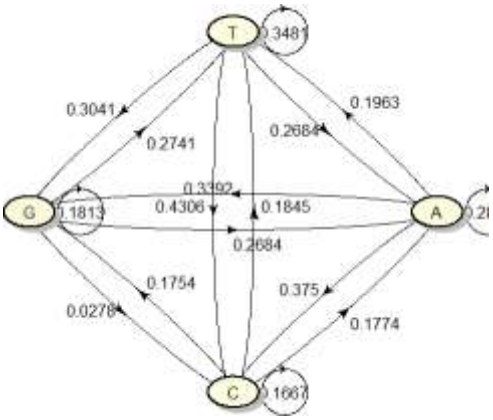

Figure 4: Transition digraph of Markov Chain

Transitions such as **C → A (43%)** and **T → T (34%)** were more common, indicating preferences toward purine transitions and thymine conservation. Rare transitions like **C → G (2.7%)** suggest sequence constraints or selective pressures in those contexts.

The comprehensive analysis of 35 mutated CFTR gene sequences using entropy profiling, nucleotide diversity metrics, hierarchical clustering, and Markov modelling sheds light on the underlying mutational dynamics and sequence variability of this clinically significant gene. Each methodological layer contributed unique insights into the functional and structural implications of the mutations observed.

2. Discussion

4.7 Entropy and Evolutionary Implications

Shannon entropy served as a proxy for nucleotide variability at each aligned position. High- entropy regions (above 1.5, as visualized in Figure 1) signify positions with frequent nucleotide substitutions across the sequences. These regions may reflect sites under relaxed purifying selection, allowing more mutational flexibility without immediate functional loss. For example, regions such as R9 (entropy ≈ 1.8) may lie in loop or linker domains of the *CFTR* protein, where variability is structurally or functionally tolerated. In contrast, low-entropy regions (e.g., R1, R3) likely encode conserved functional motifs or structural cores essential for proper protein folding or channel activity. Importantly, the entropy-based region annotation provides a means to prioritize positions for functional assays or deep mutational scanning in future work. It also supports the use of entropy as a

feature in machine learning models predicting variant pathogenicity.

4.8 Nucleotide Diversity and Sequence Constraint

The nucleotide diversity ($\pi = 0.556$) observed in this dataset is consistent with moderate-to- high levels of mutational heterogeneity. Such a value suggests that, despite the presence of conserved motifs, the CFTR gene accommodates a significant spectrum of variations, possibly due to its large size and multiple regulatory domains. It is notable that this value reflects both synonymous and nonsynonymous changes and thus captures the total variation landscape.

This degree of polymorphism aligns with clinical observations of CFTR mutation diversity in cystic fibrosis patients, where over 2,000 unique variants have been documented—many of which are rare and population-specific. The entropy and diversity analyses together suggest that certain hotspots are more prone to harbouring rare or novel mutations.

4.9 Hierarchical Clustering and Sequence Grouping

The hierarchical clustering dendrogram (Figure 2) uncovered three distinct clusters of sequences based on pairwise similarity.

These clusters may reflect subgroups of mutations with shared origins (e.g., recombination events or founder mutations) or functional consequences. For instance, the leftmost cluster comprising sequences 13 to 0 (in reverse index order) may share common mutations in high-entropy regions, whereas other clusters may involve dispersed changes or affect more conserved segments.

In applied contexts, such clustering could inform genotype-phenotype mapping, especially if the clustered sequences are linked with specific disease severities, ethnic backgrounds, or drug response profiles. Integrating these clusters with metadata such as patient outcomes could yield new biomarkers or mutation classifications.

4.10 Markov Modelling and Mutational Dynamics

The first-order Markov chain transition matrix captured the probabilistic structure of base-to- base transitions across the aligned sequences. Transitions such as **C → A (43%)** and **T → T (34%)** suggest a bias towards specific substitution types, notably **purine transitions (A↔G)**, which are known to occur more frequently due to the chemical similarity and DNA repair bias.

Interestingly, the relatively low transition probability from **C → G (2.7%)** indicates evolutionary or biochemical constraints against such changes, possibly due to their disruptive nature in protein-coding regions. These insights reinforce previous findings in molecular evolution that transitions are generally more tolerated than transversions.

This Markovian framework can be expanded in future studies to higher-order models (e.g., 2nd or 3rd order), capturing more complex dependencies and perhaps incorporating codon context or neighbouring base effects. Additionally, comparing Markov chains between disease-causing and benign variants could help identify signature transition profiles linked to pathogenicity.

4.11 Clinical Relevance: CFTR Mutations and Neonatal Diabetes

While *CFTR* mutations are classically associated with cystic fibrosis, emerging reports highlight a broader disease spectrum, including neonatal diabetes. Neonatal diabetes is a rare monogenic disorder characterized by early-onset insulin deficiency. Recent studies have implicated CFTR variants in disrupted pancreatic function, potentially affecting insulin production or secretion. The high-entropy regions identified in this study could intersect with exonic or splicing elements expressed in pancreatic tissue, particularly in early development. Such overlap may contribute to β-cell dysfunction or developmental anomalies, thereby linking CFTR mutations to the neonatal diabetes phenotype. Furthermore, the clustering results may capture sequence patterns shared by individuals with atypical CFTR presentations, including diabetes without full-blown cystic

fibrosis.

Although clinical validation is necessary, this computational insight suggests a testable hypothesis: certain mutation clusters or entropy-rich segments in CFTR may serve as molecular markers for neonatal diabetes predisposition.

4.12 Integrative Perspective

Taken together, the combined use of information-theoretic and probabilistic methods enables a layered interpretation of sequence variation. Entropy identifies where variation occurs, diversity measures how extensive it is, clustering organizes similar mutational profiles, and Markov models describe the direction and frequency of mutational changes.

This multi-faceted approach is particularly valuable for complex genes like CFTR, where mutations span coding regions, regulatory motifs, and structural elements, and where single mutations can have diverse clinical implications. The tools and insights presented here can inform not only basic mutation biology but also clinical diagnostics, variant classification, and therapeutic targeting.

4.13 Broader Implications and Future Directions

The integrative framework applied here demonstrates how information theory and probabilistic models can illuminate mutational behaviour in medically relevant genes. Beyond CFTR, this approach is generalizable to other genetic disorders where distinguishing pathogenic variants is challenging.

Future extensions could include integrating structural modelling, transcriptomic data, or functional assays to enhance biological interpretation. Moreover, applying this methodology to larger and clinically annotated datasets could refine genotype-phenotype correlations, paving the way for personalized diagnostics and therapeutics.

5. CONCLUSION

This study offers an in-depth and integrative exploration of mutated CFTR gene sequences using a combination of Shannon entropy, nucleotide diversity (π), hierarchical clustering, and first- order Markov modelling. Through entropy analysis, regions of high mutational activity—referred to as "hotspots"—as well as conserved, functionally important domains were identified. These patterns shed light on the structural and evolutionary importance of specific nucleotide positions. Nucleotide diversity provided an additional layer by quantifying genetic variability across sequences, supporting the entropy-based

findings.

The use of hierarchical clustering grouped sequences with similar mutation patterns, suggesting the existence of subtypes that may have evolutionary or clinical significance. Meanwhile, the construction of a first-order Markov chain allowed for the estimation of nucleotide transition probabilities, revealing context-specific mutational tendencies and directional trends in sequence evolution. This probabilistic approach helps to explain how certain mutation patterns might emerge and persist over time.

Interestingly, the study also points to a possible link between distinct CFTR mutation profiles and the development of neonatal diabetes, a less common manifestation of CFTR-related dysfunction. This aligns with growing evidence that CFTR mutations can contribute to conditions beyond classical cystic fibrosis, including disorders involving pancreatic β-cell dysfunction.

The integration of statistical, evolutionary, and computational tools in this study provides a robust framework for analysing gene sequence variability within a disease context. These methods not only enhance our understanding of CFTR mutation behaviour but also present a scalable strategy for investigating other monogenic diseases where variant interpretation remains complex. The study underscores the potential of computational genomics to advance precision medicine and improve the clinical relevance of genetic findings.

Acknowledgement:

The author would like to express their gratitude to the editor and reviewers for their valuable comments and suggestions. There is no conflict of interest as declared by the authors.

REFERENCES

[1] Bonidia RP, Domingues DS, Sanches DS, de Carvalho AC. MathFeature: feature extraction package for DNA, RNA and protein sequences based on mathematical descriptors. *Briefings in bioinformatics*. 2022 Jan;23(1)

[2] Capra, J. A., & Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics*, *23*(15), 1875-1882.

[3] Donaldson, S. H., Samulski, T. D., LaFave, C., Zeman, K., Wu, J., Trimble, A., ... & Davis, S. D. (2020). A four week trial of hypertonic saline in children with mild cystic fibrosis lung disease: effect on mucociliary clearance and clinical outcomes. *Journal of Cystic Fibrosis*, *19*(6), 942-948.

[4] Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

[5] Emdadi A, Moughari FA, Meybodi FY, Eslahchi C. A novel algorithm for parameter estimation of Hidden Markov Model inspired by Ant Colony Optimization. *Heliyon.* 2019 Mar 1;5(3):e01299.

[6] Hasan, S., Soltman, S., Wood, C., & Blackman, S. M. (2022). The role of genetic modifiers, inflammation and CFTR in the pathogenesis of Cystic fibrosis related diabetes. *Journal of Clinical & Translational Endocrinology*, *27*, 100287.

[7] Jeniffer S D and Senthamarai Kannan K (2021) Stochastic modelling for identifying malignant diseases. *Advances and Applications in Mathematical Sciences*, 20(9) : 1923-1936.

[8] Kannan KS, Jeniffer SD. Hidden Markov Modelling for Biological Sequence. *In Proceedings of International Conference on Computational Intelligence: ICCI* (2022) Oct 4 (p. 383). Springer Nature.

[9] Karuppusamy T. Biological Gene Sequence Stucture Analysis Using Hidden Markov Model. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*. 2021 Apr 11;12(4):1652-66.

[10] Kumar S, Gadagkar SR. Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics.* 2001 Jul 1;158(3):1321-7.

[11] Li J, Lee JY, Liao L. A new algorithm to train hidden Markov models for biological sequences with partial labels. *BMC bioinformatics.* 2021 Dec;22(1):1-21.

[12] Meng Y, Fei J. Hidden service publishing flow homology comparison using profile-hidden markov model. *International Journal of Intelligent Systems.* 2022 Feb;37(2):1081-112.

[13] Mor B, Garhwal S, Kumar A. A systematic review of hidden Markov models and their applications. *Archives of computational methods in engineering*. 2021 May;28(3):1429-48.

[14] Muthu, J. D. P., & Kaliyaperumal, S. K. (2022). Markov Modelling for Mucoviscidosis using Genomic Data. *European Journal of Mathematics and Statistics*, *3*(6), 27-34.

[15] Roth C. *Statistical methods for biological sequence analysis for DNA binding motifs and protein contacts* (Doctoral dissertation, Georg-August-Universität Göttingen).

[16] Sarkar BK. Entropy Based Biological Sequence Study. In Entropy and Exergy in Renewable Energy 2021 Mar 29. *IntechOpen*.

[17] Sasidharan SK, Thomas C. ProDroid—An Android malware detection framework based on profile hidden Markov model. *Pervasive and Mobile Computing*. 2021 Apr 1;72:101336.

[18] Schuster-Böckler B, Bateman A. An introduction to hidden Markov models. *Current protocols in bioinformatics.* 2007 Jun;18(1):A-3A.

[19] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, *13*(11), 2498-2504.

[20] Sosnay, P. R., Siklosi, K. R., Van Goor, F., Kaniecki, K., Yu, H., Sharma, N., ... & Cutting, G. R. (2013). Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nature genetics*, *45*(10), 1160-1167.