Markov Modelling for Mucoviscidosis using Genomic Data

Jeniffer David Packia Muthu, and Senthamarai Kannan Kaliyaperumal

***Abstract* — The CFTR gene, which encodes a protein present in the cell membranes of epithelial tissues, has an effect on a number of organ systems in the human body. Mutations in the CFTR gene lead to incorrect regulation of cell electrolytes and water levels. The importance of this gene for typical human development has been clearly stated through studies on the CFTR mutation. A new born that inherits one mutant copy of the CFTR gene from each parent may have cystic fibrosis, which is an autosomal recessive disease. This paper establishes a model for mutant genes that will assist in determining whether or not the embedded gene is mutated. Early detection enables the possibility of slightly earlier disease risk reduction. Modelling mutant genes and correspondingly matching the new gene with them would be a crucial and cost-effective method of preventing various chronic diseases and treatment resistance.**

***Keywords* — Cystic Fibrosis, Embedded Markov Chain, Gene Mutation, Markov Chain, Multiple Sequence Alignment.**

I. Introduction

There are billions of genes in the human body, each serving a different role. Cells can offer the structure necessary for the body to absorb nutrients from meals, transform these nutrients into energy, carry out certain tasks, etc. They are made of DNA, the genetic substance. The DNA contains the genetic information needed by an organism to control gene regulation. The primary functioning component of the human body is the gene, as each of us is aware. The DNA strands and the Gene are linked. It functions as heredity's essential physical unit as well. Each gene performs a unique job. They either supply the instructions for proteins to be made or they really shouldn't.

The Human Genome Project study states that the human body may have between 20,000 and 25, 000 genes. Chromosomes contain DNA sequences known as genes that encode specific proteins. These genes include the nucleotides Adenine (A), Thymine (T), Guanine (G), and Cytosine (C). Every gene exists in two copies in every human organism. Each gene is given a distinct name using a mix of letters and digits. For instance, CFTR gene stands for CF transmembrane conductance regulator refers to a gene on chromosome 7 that insists on producing protein and has been linked to the disease Cystic Fibrosis (OMIM No. 219700).

The hereditary disease known as Cystic Fibrosis causes severe injury to the lungs, digestive system, and other body parts. Cystic fibrosis affects the cells that produce mucus, sweat, and digestive secretions. These fluids are created are frequently slick and thin. But people with CF who have a defective gene get thick, gooey discharges. The fluids don't lubricate the tubes, ducts, and passages in the lungs or pancreas; rather, they clog them up.

Cystic fibrosis is an autosomal recessive condition, which means that a kid must inherit one mutant copy of the CFTR gene from each parent to develop the illness. Individuals nonetheless express enough healthy copies of the gene even if they only have one damaged copy of the gene. People with CF who have two mutant CFTR genes are unable to control their electrolytes adequately. Without prenatal or genetic screening, the mutant CFTR gene may be undiscovered since carriers of the disorder seem to be in good health [1], [2].

II. Review of Literature

In this study, we examined the mutated DNA sequences of thirty-five people with cystic fibrosis and created a model to detect and diagnose the disorder in a new one.

Instead of using voice recognition, [3] modified the Hidden Markov Models approach for use on biological sequences. In the paper, both theoretical and practical issues are explored. The aforementioned

Published on December 29, 2022.
J. D. P. Muthu, Manonmaniam Sundaranar University, Tirunelveli, India.
(corresponding e-mail: jenifferdavid1996@gmail.com)
S. K. Kaliyaperumal, Manonmaniam Sundaranar University, Tirunelveli, India.
(e-mail: senkannan2002@gmail.com)

example was shown by [4] who also suggested using HMM in the area of computational biology. In addition to the assumptions, the application of HMM in multiple sequence alignment and homology detection is also covered.

The process for determining Hidden Markov Models for computational biology was clarified by [5]. The mathematical algorithms and extensions of the HMM are studied and shown using the SH2 domain. In the study work of [6] the three Markov Models-conventional Markov Model, Hidden Markov Model, and Profile Hidden Markov Model-are classified; differences and drawbacks are also covered. By taking into account the three fundamental tenets of models created to analyse biological sequences, Profile HMM is provided. Software packages, libraries, and the profile HMM for apperception have been described by [7].

Gene-HMM, which is used to model pre-mRNA and predict its gene, has been presented as a way of enhancing the already available resources. Reference [8] discuss the HMMER and SAM programmes, the parametric technique, and the consensus profile of the amino acids.

To analyse any sequence, multiple sequence alignment is a prerequisite. Multiple Sequence Alignment was reported by [9]. Multiple sequence alignment is the main goal of any sequence analysis. A rigid representation known as an MSA averages outmatched residues that might not be consistently matched over the lengths of the sequences to give a coherent understanding of a sequence property. The authors offer a thorough examination of MSA methods and approaches. In their article, the authors talked about finding homologies and homogenous traits in biological sequences. The main ideas to be looked at are HMM notations, analysis, HMM training, MSA, gene disclosure, and genetic mapping. For the profile-profile alignment, several sequence alignments may be conducted. To do so, the HMMER and SAM modules were used. The profile-profile approach is applied in the profiler comparer software package. Additionally, studies on multiple sequence alignment have been conducted [4] and [10].

Reference [11] took a quick look into how HMM modifications are seen visually. The conundrums are studied using profile-csHMM, pairHMM, csHMM, and profile HMM. Machine learning is a branch of computing that is expanding quickly. It may also be used to computational biology. Semi-supervised learning is used to explain biological sequence analysis utilising Hidden Markov Models. Based on the above justifications and recommended techniques, several research projects can be finished. Numerous researchers have adapted and utilised the aforementioned works to cure a variety of ailments, and the writers have added their phrenic notions.

By following the path of the study and researchers, several adjustments and implementations are produced in the years that follow. Reference [12] made an effort to estimate the HMM's parameters by assuming that the route was unknown. The Ant Colony Optimization method for estimate was modified by the authors. An effective technique for analysing biological sequences and illness prognosis has recently emerged in bioinformatics [12]. The HMM for random sequences was created and trained by [13]. When deciding on the optimum path, the probability analysis path method was taken into account. The article's major emphasis was on the mathematical foundations of the intron and exon hidden states. In 2021,[14] published a systematic review with Hidden Markov Models as its focus.

Reference [15] calculated the Transition Probability Matrix using the TP53 gene sequence alignment procedure. The profile HMM is created and trained using the aligned sequences. The EM method was applied for the sequence alignment in this work, and a unique tool was suggested [15]. According to [16] JalView is a tool for viewing aligned sequences. Reference [17] investigated the selection of SAR-CoV zones using an entropy-based biological sequence analysis. Reference [18] utilised the BaMM service to detect motifs after doing the regulatory sequence analysis on the biological sequences. Reference [19] improved HMM training by modifying the Baum-Welch method.

Reference [20] employed the PHMM models to detect malware in ProDroid Android apps with an accuracy rate of 94.6 percent. Reference [21] compared the homology in Hidden Publishing Services using the PHMM characteristics. Reference [22] constructed the Markov model, Hidden Markov Model, Profile Hidden Markov Model, and Artificial Neural Network while modelling the gene sequences of cancer patients. The performance accuracy was also assessed. In this study, the achievements from past research will be used to do illness prediction.

III. Data Description and Methodology

The HGMD databank was utilised to provide the study's data. Thirty-five Cystic Fibrosis patients were involved in the study and their mutant gene sequences were taken into consideration for prediction. Combinations of the four nucleotides adenine (A), guanine (G), cytosine (C), and thymine (T) make up the data sequences (T). The following are the research techniques used:

Before Constructing the Markov Chain for the Data, the Data has to be preprocessed. For the preprocessing, Disparity Index, Agglomerative Clustering cum Multiple Sequence Alignment, Generating the Consensus Sequence will be applied.

A. Lemma 1 (Disparity Index)

The widely used analytical tool for evaluating the homogeneity of taken sequences is the disparity index. Let X and Y and be the respective sequences in the pair. Let x_i represent the number of times nucleotide 'i' appears in the first sequence, and y_i represent the number of times nucleotide 'i' appears in sequence two. I might here be any nucleotide between A, C, G, and T.

Let D_C be the composition difference within the sequences under consideration and expressed as,

$$D_C = \frac{1}{2}(x_i - y_i)^2, \text{ where i = A, C, G or T.}$$

By determining the expected value for D_C, the following are derived.

$$E(D_C) = \frac{1}{2} E\left(\left(\sum_{k=1}^{L} \delta_i{}^k\right)^2\right) \quad (1)$$

$$\text{Where, } \delta_i{}^k = \begin{cases} +1 & \text{for } a_k = i \text{ and } b_k \neq i \\ -1 & \text{for } a_k = i \text{ and } b_k \neq i \\ 0 & \text{otherwise} \end{cases}$$

To determine the expected value, the homogeneous condition will be applied. The formula will become,

$$I_D = \frac{1}{2}\sum_i \left(x_i - y_i\right)^2 - N_d, \text{ where } N_d \text{ is used as an estimator of } D_C.$$

If the valued of $E(I_D)$ is equals zero, it can be assumed that the homogeneity requirement if satisfied [23].

B. Lemma 2 (Hierarchical Clustering cum Multiple Sequence Alignment)

All of the sequences must be aligned in order to analyze them. Determine the degrees of homology among individuals in a group of globally related sequences using multiple sequence alignment. In a multiple sequence alignment, homologous residues from a group of sequences are arranged in columns and are considered to be similar both structurally and evolvable.

There are also other heuristic techniques for multiple alignment. The fundamental principle behind progressive alignment techniques is to look at all pairwise alignments and combine them to create a multiple alignment.

All pairwise alignment scores are computed for the Multiple Sequence Alignment. The degree of connection between the sequences might be calculated using these alignment scores. The clustering may be done using the alignment scores that were previously determined.

Agglomerative In order to integrate units sequentially, hierarchical clustering is employed. Clustering has been employed in the sequence analysis to align the sequence in terms of similarity.

Let (x_i, x_j) represent the distance between the two sequences x_i and x_j, and let (X,Y) represent the distance between the sequences X and Y. The sequence dissimilarity measurements may thus be described as,

$$\delta(X,Y) = \max_{x_i \in X, x_j \in Y} d(x_i, x_j) \quad (2)$$

If there are p sequences to merge, the first two closest sequences are combined. The only sequences left are p-1 sequences. The nearest two sequences can be combined once the distance has been recalculated. The procedure will then keep going until there is just one sequence remaining.

Sequence : sequence, profile: profile, profile : profile alignment is all that is involved in this procedure. The dendrogram, a type of graph, may be used to display the alignment.

C. Lemma 3 (Consensus Sequence)

A single sequence that reflects the "best fit" for all of its constituent sequences after alignment of all of them. If the same amino acid does not appear in all of the constituent sequences at a certain location, the amino acid that does is chosen using a voting process or some other selection method.

The Disparity Index was used to determine if the gene sequences used were homogeneous. The Agglomerative Clustering was used to combine the homogeneous gene sequences, and the progressive Alignment was used to align them. In order to build the Markov Chain, the aligned sequences were then inferred into consensus sequences.

D. Markov Chain

A stochastic process $\{X_n\}$ with a finite or denumerable state space S is said to have the Markov property if for any positive integers k, n, $k \leq n$ and for any choice of states $i_0, \ldots, i_{n+1}$ in S.

We have, $Pr\{X_{n+1} = i_{n+1} \mid X_n = i_n, \ldots\ldots, X_{k+1} = i_{k+1}, X_k = i_k\} = Pr\{X_{n+1} = i_{n+1} \mid X_n = i_n\}$ ...(3)

Whenever the involved conditional probabilities are defined, that is whenever the conditioning events have positive probabilities. If $\{X_n\}$ has the Markov Property, we say that $\{X_n\}$ is a Markov chain with state space S.

For each n in the $\{X_n\}$ of finite state space, Let $\mathbf{P}(n) := [(\mathbf{P}(n))^i_j]_{i,j} \in S$ be the matrix.

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ p_{31} & p_{32} & p_{33} & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

We call the stochastic matrix $\mathbf{P}(n)$ be the matrix of transition probabilities or the Transition Probability Matrix of the process $\{X_n\}$ from n^{th} step to the $(n+1)^{th}$ step. For every $n \geq 0$, the matrix $\mathbf{P}(n+1)$ links the distribution of the random variables X_n and X_{n+1}. [24], [25].

IV. Results and Discussion

All additional sequences are associated with each individual sequence. The first of the matched sequences is known as X, while the last one will be Y. The terms "x_A" and "y_A" denote how many adenine nucleotides are present in the sequences X and Y, respectively. All further sequence metrics are established in this manner.

For the first set of biological sequences, the values are as follows: $x_A = 246$; $y_A = 226$; $x_C = 235$; $y_C = 275$; $x_G = 235$; $y_G = 237$; $x_T = 209$; $y_T = 217$; $D_C = 1574$. Disparity index I_D then becomes 537.18. Since I_D is greater than 0, the sequences under consideration satisfy the homogeneity requirement. This method is used to verify the remaining sequence pairings.

First, we used the Clustal Omega alignment tool to align the DNA sequences that were obtained. The technique used the distance matrix as the foundation for its multiple sequence alignment. This is arranged similarly as a guiding tree. Below is a phylogenetic tree that was used to align the selected DNA sequences. The homogenous biological sequences' dissimilarity measurements are established. The measurements are used to generate the phylogenetic tree. The determined phylogenetic guide tree will be used to create the hierarchical clustering. According to the clusters, multiple sequence alignment may be determined. The alignment is carried out using the seq-seq alignment and seq-profile alignment methods. The following is a section of the phylogenetic tree with the dissimilarity measure.

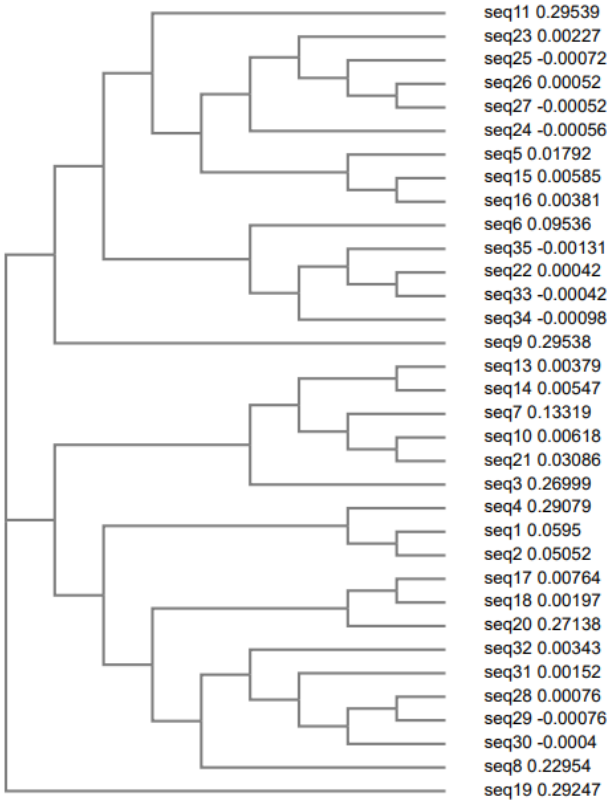

Fig. 1. Phylogenetic tree for the alignment of five DNA sequences.

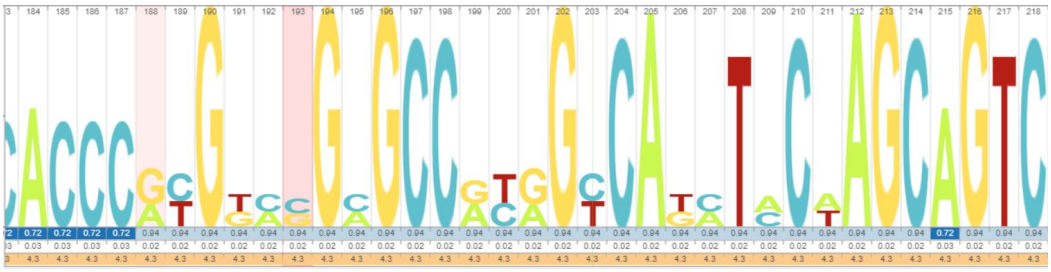

Fig. 2. Consensus sequence Logo for Multiple Sequence Alignment.

The consensus logos, developed by Schneider and Stephen, can be used to illustrate the aligned sequences. To produce a logo, there are several tools at your disposal. To depict DNA sequences visually, we utilized the internet application skylign. This logos may be used to study how each residue will predominate at each place and how much it will predominate. The logo shown below anticipates a lot of information. Consensus sequences can be created by recognizing frequently occurring residues. Additionally, we may create a % identity matrix for the alignment of DNA sequences. It shows how comparable the DNA sequences that were chosen are.

1: seq11	100.00	35.14	35.62	35.60	35.22	35.58	41.40	42.70	43.24	40.72	38.11	37.08	37.50	37.10	25.97	26.62	32.47	31.62	30.77	31.47	31.49	31.49	31.49	32.44	32.44	34.13	31.91	32.35	29.31	33.89	33.33	30.65	28.95	16.67
2: seq23	35.14	100.00	99.70	99.71	99.71	99.71	47.71	42.86	43.43	35.03	38.39	38.46	38.36	38.12	28.29	28.29	25.98	27.21	25.68	25.68	26.99	26.99	26.99	27.19	27.19	30.00	28.89	29.45	29.33	30.00	33.33	31.67	32.31	50.00
3: seq24	35.62	99.70	100.00	100.00	100.00	100.00	48.16	43.45	44.05	36.47	38.39	38.46	38.36	38.12	28.29	28.29	25.98	27.91	25.00	25.00	27.40	27.40	27.40	27.65	27.65	30.53	29.63	30.09	30.35	30.72	33.33	31.67	32.52	50.00
4: seq25	35.60	99.71	100.00	100.00	100.00	100.00	47.95	43.01	43.55	35.64	38.39	38.46	38.36	38.12	28.29	28.29	25.98	27.66	25.00	25.00	27.63	27.63	27.63	27.65	27.65	30.53	29.63	30.75	29.19	29.81	33.33	31.67	32.06	50.00
5: seq26	35.22	99.71	100.00	100.00	100.00	100.00	47.27	41.67	42.22	34.62	38.39	38.46	38.36	38.12	28.29	28.29	25.98	26.62	25.00	25.00	27.31	27.31	27.31	27.65	27.65	30.53	29.63	29.79	29.19	29.81	33.33	31.67	32.06	50.00
6: seq27	35.58	99.71	100.00	100.00	100.00	100.00	47.54	42.53	43.10	35.23	38.39	38.46	38.36	38.12	28.29	28.29	25.98	27.41	25.00	25.00	27.11	27.11	27.11	27.65	27.65	30.53	29.63	29.85	29.47	30.19	33.33	31.67	32.56	50.00
7: seq5	41.40	47.71	48.16	47.95	47.27	47.54	100.00	96.62	96.62	33.82	32.29	32.73	31.98	32.21	25.90	26.62	28.26	24.81	26.83	26.83	31.98	32.29	31.98	31.75	31.75	25.00	26.13	33.13	30.48	30.00	33.33	34.91	31.47	16.67
8: seq15	42.70	42.86	43.45	43.01	41.67	42.53	96.62	100.00	99.03	34.00	32.58	33.72	32.58	33.72	18.18	18.18	23.17	26.96	22.92	22.92	34.48	34.48	34.48	34.33	34.33	24.32	37.93	36.95	31.47	31.25	33.33	33.33	32.17	33.33
9: seq16	43.24	43.43	44.05	43.55	42.22	43.10	96.62	99.03	100.00	35.29	33.71	34.88	33.71	34.88	18.18	18.18	25.61	26.96	22.92	22.92	33.79	33.79	33.79	33.58	33.58	25.23	37.93	36.45	30.77	30.56	33.33	33.33	32.17	33.33
10: seq6	40.72	35.03	36.47	35.64	34.62	35.23	33.82	34.80	35.29	100.00	79.66	79.31	79.73	79.09	39.66	39.66	34.62	41.10	38.03	39.44	32.48	32.48	32.48	32.74	32.74	33.33	43.90	39.39	33.33	35.91	33.33	37.59	35.84	34.35
11: seq35	38.11	38.39	38.39	38.39	38.39	38.39	32.29	32.58	33.71	79.66	100.00	100.00	100.00	100.00	29.58	28.87	39.34	39.74	29.27	30.89	30.68	30.68	30.68	30.68	30.68	36.78	29.20	36.36	37.08	38.32	39.88	38.80	40.17	35.29
12: seq34	37.08	38.46	38.46	38.46	38.46	38.46	32.73	33.72	34.88	79.31	100.00	100.00	100.00	100.00	29.58	28.87	39.34	39.74	29.27	30.89	30.89	30.92	30.65	30.65	30.65	36.71	29.20	36.61	37.56	38.60	39.88	38.80	41.07	35.29
13: seq22	37.50	38.36	38.36	38.36	38.36	38.36	31.98	32.58	33.71	79.73	100.00	100.00	100.00	100.00	29.10	28.36	39.34	39.74	28.45	30.17	30.04	30.04	30.04	30.04	30.04	36.44	28.46	36.11	36.99	38.51	39.88	38.59	40.00	35.29
14: seq33	37.10	38.12	38.12	38.12	38.12	38.12	32.21	33.72	34.88	79.09	100.00	100.00	100.00	100.00	28.93	28.10	39.34	39.74	30.39	32.35	29.86	29.86	29.86	29.86	29.86	36.44	27.59	36.43	37.37	38.65	39.88	39.79	40.72	35.29
15: seq13	25.97	28.29	28.29	28.29	28.29	28.29	25.90	18.18	18.18	39.66	29.58	29.58	29.10	28.93	100.00	99.07	44.19	36.11	28.06	27.34	41.00	41.00	41.00	41.00	41.00	35.76	36.88	47.57	39.52	38.16	46.15	43.28	52.00	64.29
16: seq14	26.62	28.29	28.29	28.29	28.29	28.29	26.62	18.18	18.18	39.66	28.87	28.87	28.36	28.10	99.07	100.00	41.86	38.89	28.06	27.34	41.00	41.00	41.00	41.00	41.00	35.10	35.46	47.03	39.52	38.16	46.15	42.54	52.00	64.29
17: seq4	32.47	25.98	25.98	25.98	25.98	25.98	28.26	23.17	25.61	34.62	39.34	39.34	39.34	39.34	44.19	41.86	100.00	38.42	36.36	37.88	41.26	41.26	41.26	41.26	41.26	38.27	37.21	35.56	45.27	46.20	41.21	47.15	32.99	32.45
18: seq9	31.62	27.21	27.91	27.66	26.62	27.41	24.81	26.96	26.96	41.10	39.74	39.74	39.74	39.74	36.11	38.89	38.42	100.00	34.43	36.07	32.64	32.64	32.64	33.58	33.58	39.87	40.00	41.38	41.78	41.78	46.97	45.45	31.18	22.22
19: seq17	30.77	25.68	25.00	25.00	25.00	25.00	26.83	22.92	22.92	38.03	29.27	29.27	28.45	30.39	28.06	28.06	36.36	34.43	100.00	99.04	33.08	33.08	33.08	33.08	33.08	42.75	38.40	40.40	39.60	40.00	41.67	42.74	38.30	41.67
20: seq18	31.47	25.68	25.00	25.00	25.00	25.00	26.83	22.92	22.92	39.44	30.89	30.89	30.17	32.35	27.34	27.34	37.88	36.07	99.04	100.00	33.08	33.08	33.08	33.08	33.08	42.75	39.20	40.40	40.59	42.00	44.44	43.55	38.30	41.67
21: seq32	31.49	26.99	27.40	27.63	27.31	27.11	31.98	34.48	33.79	32.48	30.68	30.89	30.04	29.86	41.00	41.00	41.26	32.64	33.08	33.08	100.00	99.46	99.46	99.44	99.44	40.28	37.97	45.61	36.22	35.71	50.00	43.75	37.67	31.82
22: seq31	31.49	26.99	27.40	27.63	27.31	27.11	32.29	34.48	33.79	32.48	30.68	30.92	30.04	29.86	41.00	41.00	41.26	32.64	33.08	33.08	99.46	100.00	99.73	99.72	99.72	40.28	37.97	46.81	36.19	36.68	50.00	43.75	37.67	31.82
23: seq28	31.49	26.99	27.40	27.63	27.31	27.11	31.98	34.48	33.79	32.48	30.68	30.65	30.04	29.86	41.00	41.00	41.26	32.64	33.08	33.08	99.46	99.73	100.00	100.00	100.00	40.28	37.97	45.92	36.22	35.71	50.00	43.75	37.67	31.82
24: seq29	32.44	27.19	27.65	27.65	27.65	27.65	31.75	34.33	33.58	32.74	30.68	30.65	30.04	29.86	41.00	41.00	41.26	33.58	33.08	33.08	99.44	99.72	100.00	100.00	100.00	40.28	37.97	45.06	37.30	36.92	50.00	43.75	38.50	31.82
25: seq30	32.44	27.19	27.65	27.65	27.65	27.65	31.75	34.33	33.58	32.74	30.68	30.65	30.04	29.86	41.00	41.00	41.26	33.58	33.08	33.08	99.44	99.72	100.00	100.00	100.00	40.28	37.97	45.06	37.30	36.92	50.00	43.75	38.50	31.82
26: seq20	34.13	30.00	30.53	30.53	30.53	30.53	25.00	24.32	25.23	33.33	36.78	36.71	36.44	36.44	35.76	35.10	38.27	39.87	42.75	42.75	40.28	40.28	40.28	40.28	40.28	100.00	38.71	45.40	39.81	43.35	47.17	49.01	36.19	39.13
27: seq19	31.91	28.89	29.63	29.63	29.63	29.63	26.13	37.93	37.93	43.90	29.20	29.20	28.46	27.59	36.88	35.46	37.21	40.00	38.40	39.20	37.97	37.97	37.97	37.97	37.97	38.71	100.00	41.79	43.07	49.37	41.67	45.39	44.12	44.83
28: seq3	32.35	29.45	30.09	30.75	29.79	29.85	33.13	36.95	36.45	39.39	36.36	36.61	36.11	36.43	47.57	47.03	35.56	41.38	40.40	40.40	45.61	46.81	45.92	45.06	45.06	45.40	41.79	100.00	46.12	49.31	28.57	42.00	38.53	41.88
29: seq1	29.31	29.33	30.35	29.19	29.19	29.47	30.48	31.47	30.77	33.33	37.08	37.56	36.99	37.37	39.52	39.52	45.27	41.78	39.60	40.59	36.22	36.19	36.22	37.30	37.30	39.81	43.07	46.12	100.00	89.00	48.02	48.90	34.82	31.16
30: seq2	33.89	30.00	30.72	29.81	29.81	30.19	30.00	31.25	30.56	35.91	38.32	38.60	38.51	38.65	38.16	38.16	46.20	41.78	40.00	42.00	35.71	36.68	35.71	36.92	36.92	43.35	49.37	49.31	89.00	100.00	47.03	48.06	33.43	30.65
31: seq8	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	33.33	39.88	39.88	39.88	39.88	46.15	46.15	41.21	46.97	41.67	44.44	50.00	50.00	50.00	50.00	50.00	47.17	41.67	28.57	48.02	47.03	100.00	53.70	34.09	30.77
32: seq7	30.65	31.67	31.67	31.67	31.67	31.67	34.91	33.33	33.33	37.59	38.80	38.80	38.59	39.79	43.28	42.54	47.15	45.45	42.74	43.55	43.75	43.75	43.75	43.75	43.75	49.01	45.39	42.00	48.90	48.06	53.70	100.00	71.13	63.16
33: seq10	28.95	32.31	32.52	32.86	32.86	32.56	31.47	32.17	32.17	35.84	40.17	41.07	40.00	40.72	52.00	52.00	32.99	31.18	38.30	38.30	37.67	37.67	37.67	38.50	38.50	36.19	44.12	38.53	34.82	33.43	34.09	71.13	100.00	96.30
34: seq21	16.67	50.00	50.00	50.00	50.00	50.00	16.67	33.33	33.33	34.35	35.29	35.29	35.29	35.29	64.29	64.29	32.45	22.22	41.67	41.67	31.82	31.82	31.82	31.82	31.82	39.13	44.83	41.88	31.16	30.65	30.77	63.16	96.30	100.00

Fig. 3 Percentage Identity Matrix of the sequences.

1) Construction of Markov Chain

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 for each nucleotide.

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

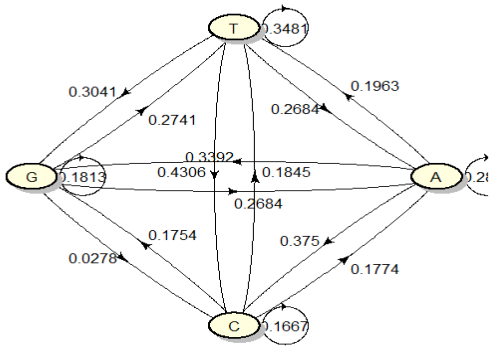

Fig. 4 Transition digraph of Markov Chain.

The bar diagram below illustrates the compositional bias of each nucleotide and dinucleotide for the Markov Chain.

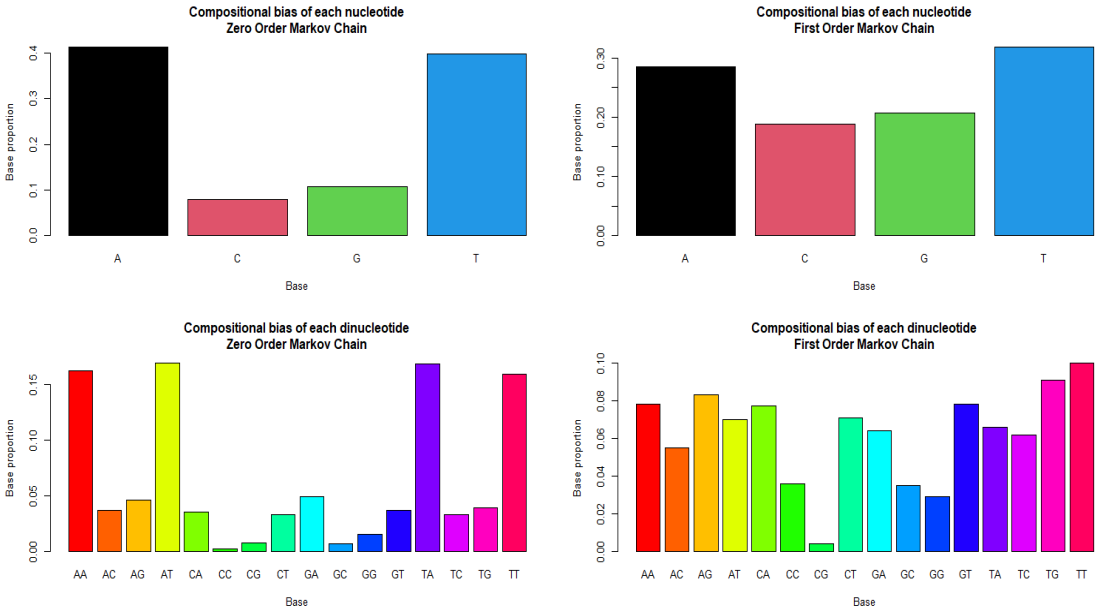

Fig. 5 Compositional bias.

The transition probability of the system switching from one adenine nucleotide to another is represented by P_{11}=0.2857 in the transition probability matrix. The transition probabilities of the consensus sequence transit from the adenine nucleotide to the cytosine, guanine, and thymine nucleotides are P_{12}=0.1774, P_{13}=0.2684, and P_{14}=0.2684 correspondingly. Similar to how successive components of the TP matrix reflect the likelihood of switching from one sequence to another or the same sequence.

The nucleotide is typically chemically altered by methylation whenever the dinucleotide CG occurs in the human genome. CpG dinucleotides are frequently less prevalent in the genome than would be anticipated given the independent probabilities of C and G because there is a reasonably high possibility that this methyl-C will transform into a T.

To determine if the CpG island Markov Chain is present in the sequences acquired, a fragment from a consensus sequence was retrieved, a_{st}^{+} and a_{st}^{-} are computed; where

$$a_{st}^{+} = \frac{c_{st}^{+}}{\sum_{t'} c_{st'}^{+}}, \text{ similarly, } a_{st}^{-} = \frac{c_{st}^{-}}{\sum_{t'} c_{st'}^{-}}.$$

For the fragment route of CGCG, the resultant table of Transition Probabilities is computed. The equation for log-odds ratio may be used to generate the models for discrimination. The CpG island cannot exist in this data because S(x)<0. The matrix P shown in (1) is then the one that should be used for additional computations.

For the investigation of the CpG island occurrence, a fragment of the deduced consensus sequence would be extracted randomly. The positive and negative analogous of transition probabilities such as, $a_{st}^{+} = \frac{c_{st}^{+}}{\sum_{t'} c_{st}^{+}}$, $a_{st}^{-} = \frac{c_{st}^{-}}{\sum_{t'} c_{st}^{-}}$ are calculated. From the analogues, the loglikelihood ratio $\beta_{x_{i-1}x_i}$ is determined. Since the odds ratio, S(x) = $\beta_{x_{i-1}x_i}$ = -2.214, which is lesser than 0, the CpG islands would not occur in the sequences taken for the study.

The Markov Chain is also known to follow the Geometric Distribution (by Feller, standard statement). Let T_i serve as a discrete random variable that expresses the transition of a system to the same nucleotide in order to test this criterion. As a result, T_i = k designates a system deposited during the course of the k subsequent transitions in which the system is still in state i.

The consensus sequence's nucleotides all adhere to the geometric distribution with the following characteristics, it may be inferred.

T_A ~ Geo (0.7327) with mean = 1.3648 and variance = 0.498

T_C ~ Geo (0.6573) with mean = 1.5214 and variance = 0.7932

T_G ~ Geo (0.7365) with mean = 1.3578 and variance = 0.4858

$T_T \sim$ Geo (0.7584) with mean = 1.3186 and variance = 0.42.

The sequence data is multinomially distributed, it might be said. The data are multinomially distributed, and the Markov chain is geometrically distributed. Therefore, the standard Markov Chain has to be changed to the embedded Markov Chain.

The repeating identical nucleotides can be swapped out for a single nucleotide symbol to create the Embedded Markov Chain. The Embedded Markov Chain Transition Probability Matrix is,

$$R = \begin{pmatrix} 0.0000 & 0.2726 & 0.3636 & 0.3636 \\ 0.4306 & 0.0000 & 0.0834 & 0.4862 \\ 0.3996 & 0.2358 & 0.000 & 0.3645 \\ 0.3325 & 0.3207 & 0.4103 & 0.0000 \end{pmatrix}$$

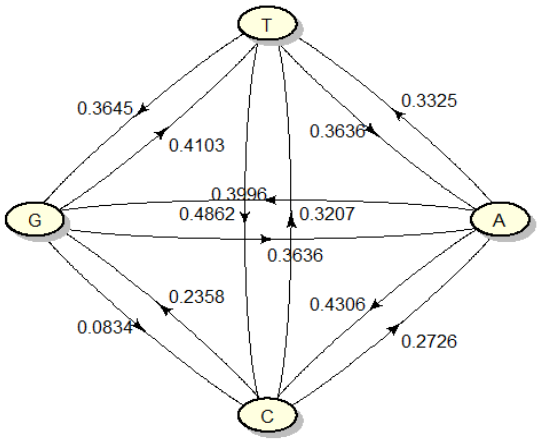

Fig. 6. Transition digraph of the Embedded Markov Chain

V. Conclusion

The homogeneity of 35 mutant CFTR gene sequences was examined. Hierarchical clustering was used to align the uniform gene sequences. The transition frequencies for each nucleotide in the deduced consensus sequence are recorded, and Markov chains and TPM are built. As a transition digraph, the irreducible, ergodic Markov chain is shown.

This study will be beneficial in analysing the genome, identifying, and detecting the gene mutations that cause cystic fibrosis. Early detection of the gene mutation will make it a cost-effective main preventative measure against cancer and medication resistance. While "wholesome genome" selection is a pricy, drawn-out, and complicated issue, the model created by this work employing stochastic approaches will make the teaching of gene sequences quicker, cheaper, and simpler. The time and cost reductions will result in significant savings for the policymaker and society when we take into account the population that needs this testing.

Acknowledgment

The authors would like to express their gratitude to the editor and learned reviewers for their valuable comments and suggestions to improve the earlier version of this manuscript. There is no conflict of interest as declared by the authors.

Funding

This work was financially supported by DST-INSPIRE, Government of India under the grant DST/INSPIRE Fellowship/2019/IF190881.

Conflict of Interest

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Bell SC, Mall MA, Gutierrez H, Macek M, Madge S, Davies JC, et al. The future of cystic fibrosis care: a global perspective. *The Lancet Respiratory Medicine*. 2020; 8(1): 65-124.
[2] Winikates K. Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Gene. *Embryo Project Encyclopedia*. 2012.
[3] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*. 1989; 77(2): 257-86.
[4] Eddy SR. Multiple alignment using hidden Markov models. *InIsmb.* 1995; 3: 114-120.
[5] Hughey R, Krogh A. Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Bioinformatics*. 1996; 12(2): 95-107.
[6] Schuster-Böckler B, Bateman A. An introduction to hidden Markov models. *Current protocols in Bioinformatics.* 2007; 18(1): A-3A.
[7] Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press; 1998.
[8] Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*. 2006; 7(1): 1.
[9] Simossis V, Kleinjung J, Heringa J. An overview of multiple sequence alignment. *Current protocols in bioinformatics*. 2003; 3(1): 3-7.
[10] Böer J. Multiple alignment using hidden Markov models. *Proteins.* 2016; 4: 14.
[11] Yoon BJ. Hidden Markov models and their applications in biological sequence analysis. *Current Genomics.* 2009; 10(6): 402-15.
[12] 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; 5(3): e01299.
[13] Karuppusamy T. Biological Gene Sequence Stucture Analysis Using Hidden Markov Model. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*. 2021; 12(4): 1652-66.
[14] Meng Y, Fei J. Hidden service publishing flow homology comparison using profile-hidden markov model. *International Journal of Intelligent Systems.* 2022; 37(2): 1081-112.
[15] Jeniffer SD, Senthamarai Kannan K. Stochastic modelling for identifying malignant diseases. *Advances and Applications in Mathematical Sciences*. 2021; 20(9): 1923-1936.
[16] Procter JB, Carstairs G, Soares B, Mourão K, Ofoegbu TC, Barton D, et al. Alignment of biological sequences with Jalview. *InMultiple Sequence Alignment.* 2021: 203-224.
[17] Sarkar BK. Entropy Based Biological Sequence Study. In Entropy and Exergy in Renewable Energy *IntechOpen*. 2021.
[18] Roth C. *Statistical methods for biological sequence analysis for DNA binding motifs and protein contacts*. Ph.D. Thesis, Georg-August-Universität Göttingen; 2021.
[19] Li J, Lee JY, Liao L. A new algorithm to train hidden Markov models for biological sequences with partial labels. *BMC Bioinformatics.* 2021; 22(1): 1-21.
[20] Sasidharan SK, Thomas C. ProDroid-An Android malware detection framework based on profile hidden Markov model. *Pervasive and Mobile Computing*. 2021; 72: 101336.
[21] Meng Y, Fei J. Hidden service publishing flow homology comparison using profile-hidden markov model. *International Journal of Intelligent Systems.* 2022; 37(2): 1081-112.
[22] Kannan KS, Jeniffer SD. Hidden Markov Modelling for Biological Sequence. *In Proceedings of International Conference on Computational Intelligence: ICCI*. 2022: 383.
[23] Kumar S, Gadagkar SR. Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. *Genetics.* 2001; 158(3): 1321-7.
[24] Gagniuc PA. Markov chains: from theory to implementation and experimentation. John Wiley & Sons; 2017.
[25] Modica G, Poggiolini L. A first course in probability and Markov Chains. John Wiley & Sons; 2012.

S. D. Jeniffer is a Research Scholar in Statistics. Her research area of interest is Stochastic Processes, Bioinformatics and Machine Learning. She graduated from Govindammal Aditanar College, Tiruchendur, India. Her post graduate degree in Statistics from Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli, India. Now, She is conducting research in Department of Statistics, Manonmaniam Sundaranar University.

She is awarded the DST-INSPIRE fellowship from Government of India with the grant number DST/INSPIRE Fellowship/2019/IF190881 for conducting the Research. She published two research articles and several articles are under review in peer-reviewed journals. Currently, she is working with the genomic data to develop Stochastic Models in order to predict chronic diseases.

K. Senthamarai Kannan studied his M.Sc.,(1987), M.Phil.,(1989), and Ph.D.,(1994) degree courses in Statistics at Annamalai University. He joined the Department of Statistics, Manonmaniam Sundaranar University as Lecturer in the year 1991 and was selected to the post of Reader in April, 1998. Subsequently, he was promoted as Professor in April 2006. Now he is the Senior Professor and Head of the Department of Statistics, Manonmaniam Sundaranar University, Tirunelveli. His area of specialization is "Stochastic Processes and their Applications".

Dr. K. Senthamarai Kannan was awarded doctoral degree for his dissertation on 'General Bulk Queues'. The results of his dissertation have applications in many real-life situations, which enable the service providers to reduce the waiting time, queue length and to improve their services optimally. His contributions on this important topic have been referred and cited by many scholars and attracted a number of young researchers to this field.

Dr. K. Senthamarai Kannan has published 211 research papers in various International and National Journals and Proceedings and presented papers 181 in conferences, both at National and International levels. He has authored and edited 10 books.