Mathematical Sequence Analyses of Cystic Fibrosis Transmembrane Conductance Regulator (CFTR): Cross-Species Skeletal Frameworks in Cystic Fibrosis

Sk. Sarif Hassan *, Kharerin Hungyo , Vladimir N. Uversky

Posted Date: 19 May 2026

doi: 10.20944/preprints202605.1213.v1

Keywords: cystic fibrosis (CF); Cystic Fibrosis Transmembrane Conductance Regulator (CFTR); hydropathy profiles; frequency dominant residues; intrinsic protein disorder

 doi:10.20944/preprints202605.1213.v1

Article

Mathematical Sequence Analyses of Cystic Fibrosis Transmembrane Conductance Regulator (CFTR): Cross-Species Skeletal Frameworks in Cystic Fibrosis

Sk. Sarif Hassan [1,*], Kharerin Hungyo [2] and Vladimir N. Uversky [3]

[1] Department of Mathematics, Pingla Thana Mahavidyalaya, Maligram, Pingla, 721140 West Bengal, India
[2] School of Biosciences and Bioengineering, Indian Institute of Technology, Mandi, South Campus, Himachal Pradesh-175075, India
[3] Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA
* Correspondence: sksarifhassan@pinglacollege.ac.in

Abstract

Rare diseases, though individually uncommon, collectively represent a major global health challenge, affecting millions worldwide and increasingly recognized in India as a significant contributor to pediatric and adult morbidity. Cystic fibrosis (CF), a multisystem autosomal recessive disorder caused by pathogenic variants in the Cystic Fibrosis Transmembrane Conductance Regulator (*CFTR*) gene, exemplifies this burden, with delayed diagnosis and diverse mutational spectra complicating clinical management in South Asian populations. To advance rare disease genomics, quantitative analysis of CFTR sequences across multiple species is essential, as evolutionary conservation highlights residues and motifs critical for channel function, while divergence reveals lineage-specific adaptations relevant to disease mechanisms. In the present study, we performed integrative analyses encompassing amino acid composition, sequence homology, frequency-dominant residue patterns, hydropathy-based n-gram distributions, hydropathy profile continuity, and intrinsic disorder architectures across various CFTR sequences from multiple species. The quantitative signatures derived from amino acid composition, sequence homology, hydropathy-based n-grams, hydropathy profiles, and intrinsic disorder analyses carry significant translational impact, as they provide a unified framework for identifying conserved motifs, resolving disorder-prone domains, and guiding the precise mapping of pathogenic mutations and their functional consequences. Collectively, our findings demonstrate how cross-species quantitative protein analysis of CFTR bridges evolutionary biology with clinical investigation, providing translational insights that strengthen rare disease research and therapeutic development in cystic fibrosis.

Keywords: cystic fibrosis (CF); Cystic Fibrosis Transmembrane Conductance Regulator (CFTR); hydropathy profiles; frequency dominant residues; intrinsic protein disorder

1. Introduction

Rare diseases are not truly rare [1,2]. In India, rare diseases lack a strict numerical definition due to limited epidemiological data, but the National Policy for Rare Diseases (2021) recognizes them as serious, often genetic conditions requiring specialized care [3]. Globally, rare diseases are defined more precisely: WHO and the EU use a prevalence threshold of fewer than 1 in 2,000 people, while the U.S. defines them as affecting fewer than 200,000 individuals nationwide [4–6]. While each condition may affect only a small number of individuals, together they represent a major global health challenge, impacting millions across all ages and regions [7,8]. Most rare diseases manifest in childhood and are rooted in genetics, often leaving families searching for answers [9,10]. Advances in genomic sequencing and big data analytics have transformed the landscape of rare disease research and diagnosis [11,12]. Next-generation sequencing, computational genomics, and international data-sharing initiatives have

 doi:10.20944/preprints202605.1213.v1

accelerated gene discovery and improved diagnostic yield [13,14]. Yet, despite these breakthroughs, more than half of suspected rare disease cases remain without a genetic diagnosis, and over half of the disease-causing genes are still unknown [8,15,16]. This persistent diagnostic gap underscores both the progress achieved and the urgent need for continued innovation, collaboration, and investment in genomics to ensure that families affected by rare diseases are not left behind.

Cystic fibrosis (CF) is a life-limiting autosomal recessive disorder that, although historically considered rare in India, is increasingly recognized as an important contributor to pediatric and adult morbidity [17,18]. CF arises from pathogenic variants in the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene, which encodes a chloride ion channel essential for epithelial fluid balance [19,20]. Mutations in CFTR impair ion transport, resulting in viscous secretions that compromise respiratory, gastrointestinal, hepatic, and reproductive function [21,22]. Globally, CF is one of the most common life-limiting genetic diseases among Caucasian populations, with an incidence of approximately 1 in 2,500 live births in Europe and North America [23,24]. The F508del mutation accounts for nearly 70% of cases in these regions, and extensive genotype–phenotype correlations have guided the development of mutation-specific therapies, including CFTR modulators, which have markedly improved survival and quality of life [25,26].

In India, CF has historically been under-recognized, largely due to limited awareness, overlapping clinical presentations with other respiratory and gastrointestinal disorders, and the absence of routine newborn screening [27,28]. Recent studies, however, indicate that CF is more prevalent than previously assumed, with tens of thousands of affected individuals nationwide [29]. Indian cohorts reveal a broader mutational spectrum compared to Western populations, with F508del present at lower frequencies (approximately 25–30%), and several novel or region-specific variants identified [30, 31]. This genetic diversity complicates both diagnosis and therapeutic applicability, as many CFTR modulators developed for Western mutations may not be effective for Indian patients [30]. Furthermore, delayed diagnosis often results in advanced disease presentation, contributing to higher morbidity and mortality. The contrasting global and Indian perspectives highlight both progress and challenges in CF research [32]. While precision medicine and mutation-specific therapies are transforming outcomes in high-income countries, India faces a diagnostic and therapeutic gap that underscores the need for population-specific genetic characterization, functional genomics, and improved clinical awareness [33]. Addressing these challenges will require sustained investment in genomic sequencing, establishment of national registries, and integration of CF care into broader rare disease frameworks to ensure equitable advances in diagnosis and treatment [34,35].

CFTR genomics has emerged as a cornerstone in understanding the molecular basis of cystic fibrosis and related disorders [36,37]. The CFTR gene encodes a chloride ion channel whose dysfunction leads to multisystem pathology, and more than 2,000 variants have been described worldwide [38,39].

Quantitative analysis of CFTR sequences across multiple species provides a powerful framework for advancing rare disease genomics, particularly in the context of CF [40–42]. By systematically examining amino acid composition, sequence homology, frequency-dominant residue patterns, hydropathy-based n-gram distributions, hydropathy profile continuity, and intrinsic disorder architectures, conserved and divergent molecular features of CFTR can be resolved with precision. Cross-species comparisons highlight invariant motifs and substitution-sensitive regions that form the skeletal framework of the protein, offering insights into structural stability and regulatory flexibility. Such integrative analyses not only deepen our understanding of CFTR biology, but also provide translational value: they enable prioritization of clinically relevant variants, guide functional validation, and inform therapeutic design. In this way, comparative and quantitative protein analysis of CFTR sequences bridges evolutionary biology with clinical investigation, accelerating discovery and intervention in rare disease genomics.

 doi:10.20944/preprints202605.1213.v1

2. Data Acquisition

In this study, a total of 42 protein sequences for the cystic fibrosis transmembrane conductance regulator (CFTR) was retrieved from the UniProt Knowledgebase (UniProtKB). Accession identifiers, entry names, sequence lengths, and source organisms were extracted directly from UniProt entries to ensure accuracy and reproducibility. The dataset encompasses 42 CFTR homologs across diverse vertebrate taxa. All sequences were downloaded in FASTA format using UniProt's batch retrieval system, and metadata (UniProt ID, entry name, sequence length, and organism) was tabulated (Table 1). Notably, three entries corresponding to *Macaca mulatta* (MACMU), *Macaca fuscata fuscata* (MACFU), and *Macaca fascicularis* (MACFA) were found to be duplicated. Thus, the dataset comprised 40 unique, non-identical CFTR sequences.

Table 1. List of cystic fibrosis transmembrane conductance regulator (CFTR) proteins, their respective uniProt IDs (with hyperlink). lengths, and their respective species from where the sequences were obtained.

S/N	UniProt ID	Entry Name	Length	*Species*
1	P13569	CFTR_HUMAN	1480	*Homo sapiens (Human)*
2	P26361	CFTR_MOUSE	1476	*Mus musculus (Mouse)*
3	P26362	CFTR_SQUAC	1492	*Squalus acanthias (Spiny dogfish)*
4	P26363	CFTR_XENLA	1485	*Xenopus laevis (African clawed frog)*
5	P34158	CFTR_RAT	1476	*Rattus norvegicus (Rat)*
6	P35071	CFTR_BOVIN	1481	*Bos taurus (Bovine)*
7	Q00552	CFTR_CAVPO	1481	*Cavia porcellus (Guinea pig)*
8	Q00553	CFTR_MACMU	1481	*Macaca mulatta (Rhesus macaque)*
9	Q00554	CFTR_RABIT	1481	*Oryctolagus cuniculus (Rabbit)*
10	Q00555	CFTR_SHEEP	1481	*Ovis aries (Sheep)*
11	Q00PJ2	CFTR_ATEAB	1483	*Atelerix albiventris (Middle-African hedgehog) (Four-toed hedgehog)*
12	Q07DV2	CFTR_AOTNA	1481	*Aotus nancymaae (Ma's night monkey)*
13	Q07DW5	CFTR_MUNRE	1481	*Muntiacus reevesi (Reeves' muntjac) (Cervus reevesi)*
14	Q07DX5	CFTR_NOMLE	1480	*Nomascus leucogenys (Northern white-cheeked gibbon) (Hylobates leucogenys)*
15	Q07DY5	CFTR_COLGU	1481	*Colobus guereza (Mantled guereza) (Eastern black-and-white colobus monkey)*
16	Q07DZ6	CFTR_ORNAN	1484	*Ornithorhynchus anatinus (Duckbill platypus)*
17	Q07E16	CFTR_MUSPF	1484	*Mustela putorius furo (European domestic ferret) (Mustela furo)*
18	Q07E42	CFTR_DASNO	1482	*Dasypus novemcinctus (Nine-banded armadillo)*
19	Q09YH0	CFTR_SAIBB	1481	*Saimiri boliviensis boliviensis (Bolivian squirrel monkey)*
20	Q09YJ4	CFTR_MUNMU	1481	*Muntiacus muntjak (Barking deer) (Indian muntjac)*
21	Q09YK5	CFTR_ATEGE	1480	*Ateles geoffroyi (Black-handed spider monkey) (Geoffroy's spider monkey)*
22	Q108U0	CFTR_LOXAF	1482	*Loxodonta africana (African elephant)*
23	Q1LX78	CFTR_DANRE	1485	*Danio rerio (Zebrafish) (Brachydanio rerio)*
24	Q2IBA1	CFTR_CHLAE	1481	*Chlorocebus aethiops (Green monkey) (Cercopithecus aethiops)*
25	Q2IBB3	CFTR_RHIFE	1482	*Rhinolophus ferrumequinum (Greater horseshoe bat)*
26	Q2IBE4	CFTR_PONAB	1480	*Pongo abelii (Sumatran orangutan) (Pongo pygmaeus abelii)*
27	Q2IBF6	CFTR_GORGO	1480	*Gorilla gorilla gorilla (Western lowland gorilla)*
28	Q2QL74	CFTR_DIDVI	1482	*Didelphis virginiana (North American opossum) (Didelphis marsupialis virginiana)*
29	Q2QL83	CFTR_MICMU	1481	*Microcebus murinus (Gray mouse lemur) (Lemur murinus)*
30	Q2QLA3	CFTR_HORSE	1481	*Equus caballus (Horse)*
31	Q2QLB4	CFTR_PLEMO	1480	*Plecturocebus moloch (Dusky titi monkey) (Callicebus moloch)*
32	Q2QLC5	CFTR_CARPS	1482	*Carollia perspicillata (Seba's short-tailed bat)*
33	Q2QLE5	CFTR_PANTR	1480	*Pan troglodytes (Chimpanzee)*
34	Q2QLF9	CFTR_CALJA	1481	*Callithrix jacchus (White-tufted-ear marmoset) (Simia Jacchus)*
35	Q2QLH0	CFTR_OTOGA	1482	*Otolemur garnettii (Small-eared galago) (Garnett's greater bushbaby)*
36	Q5D1Z7	CFTR_TRIVU	1478	*Trichosurus vulpecula (Brush-tailed possum)*
37	Q5U820	CFTR_CANLF	1483	*Canis lupus familiaris (Dog) (Canis familiaris)*

 doi:10.20944/preprints202605.1213.v1

Table 1. *Cont.*

S/N	UniProt ID	Entry Name	Length	*Species*
38	Q6PQZ2	CFTR_PIG	1482	*Sus scrofa (Pig)*
39	Q7JII7	CFTR_MACFU	1481	*Macaca fuscata fuscata (Japanese macaque)*
40	Q7JII8	CFTR_MACFA	1481	*Macaca fascicularis (Crab-eating macaque) (Cynomolgus monkey)*
41	Q9TSP5	CFTR_PAPAN	1481	*Papio anubis (Olive baboon)*
42	Q9TUQ2	CFTR_MACNE	1481	*Macaca nemestrina (Pig-tailed macaque)*

3. Methods

3.1. Amino Acid Compositions Across CFTR Protein Sequences

3.1.1. Amino Acid Relative Frequency and Correlation Analysis Across CFTR Sequences

The amino acid frequency, defined as the number of occurrences of each amino acid within a protein sequence, was calculated for all CFTR sequences [43]. To account for differences in sequence length, the relative frequency of each amino acid was obtained by dividing its absolute count by the length of the corresponding CFTR sequence and multiplying by 100. This measure reflects the percentage composition of each amino acid within the sequence [44].

Consequently, each CFTR protein sequence is represented as a 20-dimensional vector, corresponding to the relative frequencies of the 20 standard amino acids. The workflow was implemented using custom MATLAB scripts. Input sequences were processed programmatically to compute amino acid counts, normalize by sequence length, and generate relative frequency vectors suitable for downstream statistical and phylogenetic analyses.

For all pairs of relative frequency vectors, the Pearson correlation coefficient was calculated to assess the degree of association between amino acid compositions across CFTR protein sequences [45]. This pairwise correlation analysis enabled the identification of positively and negatively associated amino acid usage patterns, thereby providing insight into compositional dependencies and divergence among sequences. The workflow was implemented using custom MATLAB scripts, which computed correlation matrices from the 20-dimensional relative frequency vectors. The resulting coefficients were used to identify statistically significant associations.

3.1.2. Amino Acid Composition Profile Analysis of CFTR Sequences

Amino acid composition profiles for all 42 CFTR protein sequences analyzed in this study were generated using the Composition Profiler tool [46]. In this framework, the CFTR sequences under investigation were designated as the *query set*, while the *Protein Data Bank Select 25* served as the *background set* [47].

To complement this analysis, we also constructed composition profiles for experimentally validated intrinsically disordered proteins obtained from the DisProt database [48,49]. These profiles depict the normalized enrichment or depletion of individual amino acid residues, calculated according to the formula:

$$\frac{(C_x - C_{order})}{C_{order}}$$

where C_x denotes the content of a given residue in the query protein, and C_{order} represents the corresponding residue content in the PDB Select 25 background set.

3.1.3. Shannon Entropy-Based Compositional Diversity of CFTR Protein Sequences

To quantify compositional diversity in CFTR protein sequences, amino acid sequence-level Shannon entropy from amino acid residue frequency distributions.

Let p_i denote the relative frequency of amino acid type $i \in \{1, \ldots, 20\}$. The Shannon entropy H of a sequence is defined as:

$$H = -\sum_{i=1}^{20} p_i \log_2(p_i),$$

 doi:10.20944/preprints202605.1213.v1

with the convention that terms with $p_i = 0$ contribute 0 to the sum. Entropy values range from 0, corresponding to complete compositional conservation, to a theoretical maximum of 4.322 for a uniform distribution over 20 amino acids [50].

This analysis was implemented using custom MATLAB scripts, which calculated amino acid frequencies, normalized them by sequence length, and applied the entropy formula programmatically. The resulting entropy values provide a rigorous quantitative measure of compositional diversity across CFTR protein sequences.

3.1.4. Phylogenetic Relationship of CFTR Protein Sequences Based on Amino Acid Relative Frequency

Euclidean distances were calculated between all possible pairs of CFTR protein sequences, each represented as a 20-dimensional vector of relative amino acid frequencies. This procedure yielded a symmetric square distance matrix of dimension 42×42, encapsulating the compositional dissimilarities among the sequences. Based on the Eucldean matrix a phylogenetic tree was developed using MATLAB function '`phytree`'.

3.1.5. Multiple Sequence Homology Based Phylogenetic Relationship of CFTR Sequences

Amino acid sequence homology among CFTR protein sequences was evaluated using multiple sequence alignment performed with Clustal Omega [51]. This alignment facilitated the systematic identification of invariant residues and amino acid substitutions [52].

Clustal Omega was executed under its default analytical configuration. The *Gonnet* substitution matrix was applied to score amino acid replacements, with a gap opening penalty of 10 and a gap extension penalty of 0.20. End-gap penalties were not enforced, consistent with standard practice. Although trimming options are available to exclude gap-rich or poorly conserved regions, alignments were retained in full unless otherwise specified. These parameter settings provided a balanced framework for detecting conserved motifs and quantifying sequence divergence, thereby supporting robust phylogenetic inference.

3.1.6. Assessment on Invariant and Substitution Residues of CFTR Sequences

Based on the multiple sequence alignment of CFTR proteins derived from Clustal Omega, invariant residues and substitutional variations were systematically identified and analyzed.

3.1.7. Frequency-dominant Amino Acids Patterns of CFTR Sequences

Each CFTR protein sequence was divided into consecutive, non-overlapping segments of 100 residues, without loss of generality. Within each segment, the frequency of all 20 standard amino acids was determined. The amino acid occurring most frequently was designated as the *dominant residue* for that segment, and its relative abundance was expressed as a percentage of the total residues in the window. For visualization, each 100-residue interval was shaded with a color corresponding to its dominant amino acid. In addition, the dominance percentage was plotted as a density track, allowing both categorical identification of the leading residue and quantitative assessment of its prevalence. This representation highlights regions of strong amino acid enrichment and enables comparison of compositional biases across different segments of the sequence.

3.2. Hydropathy Based N-Grams Frequency Analyses Across CFTR Protein Sequences

Let Σ denote the alphabet of amino acid residues. The 20 amino acids were partitioned into two disjoint sets according to hydropathy:

$$\Sigma_{\text{polar}} = \{S, T, N, Q, Y, C, R, H, K, D, E\}, \qquad \Sigma_{\text{nonpolar}} = \{A, V, L, I, P, W, F, M, G\}.$$

Thus $\Sigma = \Sigma_{\text{polar}} \cup \Sigma_{\text{nonpolar}}$ and $\Sigma_{\text{polar}} \cap \Sigma_{\text{nonpolar}} = \varnothing$.
For a CFTR protein sequence $x = (x_1, x_2, \ldots, x_L)$ of length L, each residue was encoded as

 doi:10.20944/preprints202605.1213.v1

$$b_i = \begin{cases} 1 & \text{if } x_i \in \Sigma_{\text{polar}}, \\ 0 & \text{if } x_i \in \Sigma_{\text{nonpolar}}. \end{cases}$$

n-gram: An *n-gram* is any contiguous subsequence $(b_0, b_1, \ldots, b_{n-1})$ of length n in the binary encoding. Each n-gram is therefore a word in $\{0,1\}^n$ representing the local hydropathy pattern.

For each CFTR protein sequence, all overlapping n-grams were enumerated from the binary hydropathy encoding. For example, the complete set of 2-grams is $\{00, 01, 10, 11\}$, and the complete set of 3-grams is $\{000, 001, 010, 011, 100, 101, 110, 111\}$. In general, the set of all possible n-grams is $\mathcal{W}_n = \{0,1\}^n$, containing 2^n words.

For each $w \in \mathcal{W}_n$, the *count* $C(w)$ is defined as the number of times w occurs in the sequence. The *relative frequency* is then

$$F(w) = \frac{C(w)}{\sum_{w' \in \mathcal{W}_n} C(w')},$$

which represents the proportion of w among all observed n-grams.

Relative frequency of n-grams:

All overlapping n-grams were enumerated by sliding a window of length n across the binary sequence. For each $w \in \mathcal{W}_n$, the *count* is

$$C(w) = \#\{\, i : (b_i, \ldots, b_{i+n-1}) = w \,\},$$

and the *relative frequency* is

$$F(w) = \frac{C(w)}{\sum_{w' \in \mathcal{W}_n} C(w')}.$$

Thus $F(w)$ represents the proportion of occurrences of w among all observed overlapping n-grams.

Illustration:

Consider a dummy protein sequence of length 12 and respective binary hydropathy representatiohn be

$$b = (1,0,1,1,0,0,1,0,1,1,0,1).$$

The complete set of 2-grams is

$$\mathcal{W}_2 = \{00, 01, 10, 11\}.$$

Sliding a window of length 2 across b yields the overlapping 2-grams:

$$10,\ 01,\ 11,\ 10,\ 00,\ 01,\ 10,\ 01,\ 11,\ 10,\ 01.$$

The counts are

$$C(00) = 1, \quad C(01) = 4, \quad C(10) = 4, \quad C(11) = 2,$$

and the relative frequencies are

$$F(00) = \tfrac{1}{11}, \quad F(01) = \tfrac{4}{11}, \quad F(10) = \tfrac{4}{11}, \quad F(11) = \tfrac{2}{11}.$$

Likewise, the complete set of 3-grams is

$$\mathcal{W}_3 = \{000, 001, 010, 011, 100, 101, 110, 111\}.$$

Sliding a window of length 3 across b yields the overlapping 3-grams:

 doi:10.20944/preprints202605.1213.v1

$$101,\ 011,\ 110,\ 100,\ 001,\ 010,\ 101,\ 011,\ 110,\ 101.$$

The counts are

$$C(101) = 3, \quad C(011) = 2, \quad C(110) = 2, \quad C(100) = 1, \quad C(001) = 1, \quad C(010) = 1,$$

with all other 3-grams having count zero. The relative frequencies are

$$F(101) = \tfrac{3}{10}, \quad F(011) = \tfrac{2}{10}, \quad F(110) = \tfrac{2}{10}, \quad F(100) = \tfrac{1}{10}, \quad F(001) = \tfrac{1}{10}, \quad F(010) = \tfrac{1}{10}.$$

For $n = 2$, the four possible words $\{00, 01, 10, 11\}$ are all observed with varying frequencies. For $n = 3$, six of the eight possible words occur in the sequence, each with its own relative frequency. The same procedure generalizes to larger n.

In addition, based on the relative frequencies of independent n-grams ($n = 2, 3, \ldots, 7$), pairwise Euclidean distances were computed between all CFTR protein sequences. This procedure yielded a square, symmetric distance matrix of dimension 42×42, representing the compositional dissimilarities among the 42 CFTR sequences. Using this matrix as input, phylogenetic relationships were inferred by hierarchical clustering, and a dendrogram was constructed to visualize the sequence proximities. Subsequently, proximal clusters of CFTR proteins were identified from the dendrogram by applying a distance threshold, thereby delineating groups of sequences with proximity in their hydropathy-based n-gram profiles.

3.3. Hydropathy Profile Based Clustering of CFTR Proteins

For each CFTR sequence percentages of polar and non-polar residues were enumerated. Furthermore, the longest continuous stretches of polar and non-polar residues for each CFTR sequence was found by scanning to identify the maximal uninterrupted run of residues belonging to each set. For every sequence, the longest polar stretch and the longest non-polar stretch were recorded with their absolute lengths and normalized by dividing by the total sequence length to account for variation in sequence size. These normalized values were then used to construct two-dimensional feature vectors representing each sequence, which were subjected to density-based clustering using DBSCAN with Euclidean distance. The parameters ϵ (neighborhood radius) and $Minpts$ (minimum points per cluster) were optimized empirically to balance sensitivity and cluster stability. DBSCAN was chosen because it does not require pre-specification of cluster number and can identify outliers as noise, thereby enabling robust grouping of CFTR sequences based solely on their longest polar and non-polar domain architecture.

3.4. Intrinsic Disorder Profiles Analysis of CFTR Sequences

Intrinsic disorder profiles for CFTR protein sequences were generated using Metapredict v3.0, a deep-learning consensus predictor of intrinsic disorder for single-sequence validation and through the authors' Google Colab Python notebook for batch analysis [53,54]. All 42 CFTR sequences in FASTA format were processed to yield per-residue disorder scores ranging from 0 to 1, exported as a tab-delimited file, and subsequently imported into MATLAB for visualization, where custom scripts produced line plots aligned to the canonical CFTR sequence; residues with scores ≥ 0.5 were classified as disordered, while those < 0.5 were considered ordered, and domain-level averages with standard deviations were computed across conserved regions (NBD1, NBD2, R-domain, cytoplasmic loops, and C-terminal tail) to provide a comprehensive disorder architecture profile of CFTR protein sequences[55].

 doi:10.20944/preprints202605.1213.v1

3.4.1. Determining Intrinsic Protein Disordered-Based Difference Spectra of CFTR Sequences

Each CFTR sequence was represented as a residue-wise vector of disorder probabilities. The resulting profiles were collated into a matrix of dimension $M \times N$, where $M = 42$ denotes the number of sequences and N the maximum residue length across all CFTR sequences. Because sequence lengths varied among species, all profiles were truncated to the minimum valid length observed across all CFTR sequence. This ensured that each sequence was represented by a vector of equal dimension, thereby permitting direct residue-wise comparison. The disorder profile of the human CFTR sequence (CFTR_HUMAN) was designated as the reference. For each residue position j, the human disorder score $d_{\text{human},j}$ was extracted to serve as the baseline.

For each CFTR sequence i, a difference spectrum was computed as

$$\Delta_{ij} = d_{ij} - d_{\text{human},j},$$

Here d_{ij} denotes the disorder score of residue j in CFTR sequence i [56]. Both signed differences (indicating directionality of deviation) and absolute differences (indicating magnitude of deviation) were considered. The resulting spectra highlight regions of increased or decreased disorder relative to CFTR_HUMAN.

To quantify overall divergence from CFTR_HUMAN, the mean absolute difference across all residues was calculated for each sequence:

$$D_i = \frac{1}{N} \sum_{j=1}^{N} |\Delta_{ij}|.$$

Sequences were then ranked in descending order of D_i, providing a proximity-based ordering relative to the CFTR_HUMAN based on intrinsic protein disorder difference spectra.

3.5. Evaluation of the Disorder-Based Functionality of Human CFTR

The disorder-based functionality of CFTR_HUMAN was evaluated using the D^2P^2 (Database of Disordered Protein Prediction) platform [57]. Platform generates an informative plot presenting outputs of several commonly used disorder predictors PONDR® VLXT, PONDR® VSL2b, PrDOS, IU-Pred (both short and long forms), and three variants of Espritz (NMR, DisProt, and X-ray)). The stacked outputs of these predictors are overlaid with bars showing positions of functional domains, Molecular Recognition Features (MoRFs), various posttranslational modifications (PTMs) [57].

Furthermore, the 3D structural model generated by AlphaFold was retrieved from AlphaFold Protein Structure Database [58]. Also, protein-protein interaction (PPI) network centered at human CFTR was generated using STRING (search tool for recurring instances of neighboring genes) platform STRING [59].

4. Results and Analyses

4.1. Amino Acid Compositions and Their Phylogenetic Relationships

4.1.1. Amino Acid Frequency Compositions of CFTR Sequences

Leucine (L) emerged as the most prevalent amino acid across all 42 CFTR sequences, with an average relative frequency of 12.5%, whereas Cystine (C) was the least represented, occurring at only 1.13% on average (Figure 1). Furthermore, pairwise correlation analysis of relative amino acid frequencies across 42 CFTR sequences carried out and it revealed both positively and negatively associated amino acid residues. The correlation matrix demonstrated that most amino acid pairs exhibited weak or negligible associations; however, several pairs showed strong and statistically significant relationships ($|r| > 0.5$, $p < 0.001$) (Figure 2). Two amino acid pairs displayed robust positive correlations: alanine–methionine (A–M; $r = 0.554$, $p = 1.40 \times 10^{-4}$) and histidine–lysine (H–K; $r = 0.539$, $p = 2.30 \times 10^{-4}$). These results suggest coordinated variation in the relative abundance of these residues across the dataset. On the other side, nine amino acid pairs exhibited strong

 doi:10.20944/preprints202605.1213.v1

negative correlations, indicating reciprocal enrichment patterns. The most pronounced were glutamate–glycine (E–G; $r = -0.630$, $p = 7.64 \times 10^{-6}$) and glutamate–phenylalanine (E–F; $r = -0.579$, $p = 5.85 \times 10^{-5}$). Additional significant negative associations included alanine–cysteine (A–C; $r = -0.534$, $p = 2.69 \times 10^{-4}$), alanine–glycine (A–G; $r = -0.550$, $p = 1.62 \times 10^{-4}$), asparagine–aspartate (N–D; $r = -0.529$, $p = 3.15 \times 10^{-4}$), asparagine–phenylalanine (N–F; $r = -0.515$, $p = 4.87 \times 10^{-4}$), histidine–proline (H–P; $r = -0.519$, $p = 4.31 \times 10^{-4}$), lysine–threonine (K–T; $r = -0.503$, $p = 6.95 \times 10^{-4}$), and methionine–serine (M–S; $r = -0.513$, $p = 5.09 \times 10^{-4}$).

It was noticed that histidine and lysine averaged 1.68% and 6.16%, while alanine and methionine averaged 5.66% and 2.57%; in both cases, their profiles were positively correlated. Similarly, alanine and glycine were moderately represented at 5.66% and 5.63%, whereas Cystine (C) remained the least abundant at 1.13%.

Altogether, these results highlight amino acid pairs that co-vary or counter-vary across CFTR sequences, reflecting underlying biochemical constraints. The statistical robustness of these associations underscores their biological relevance. A clear compositional signature emerges: Leucine is preferentially enriched, while Cystine is strikingly underrepresented, pointing to a bias in CFTR sequence architecture.

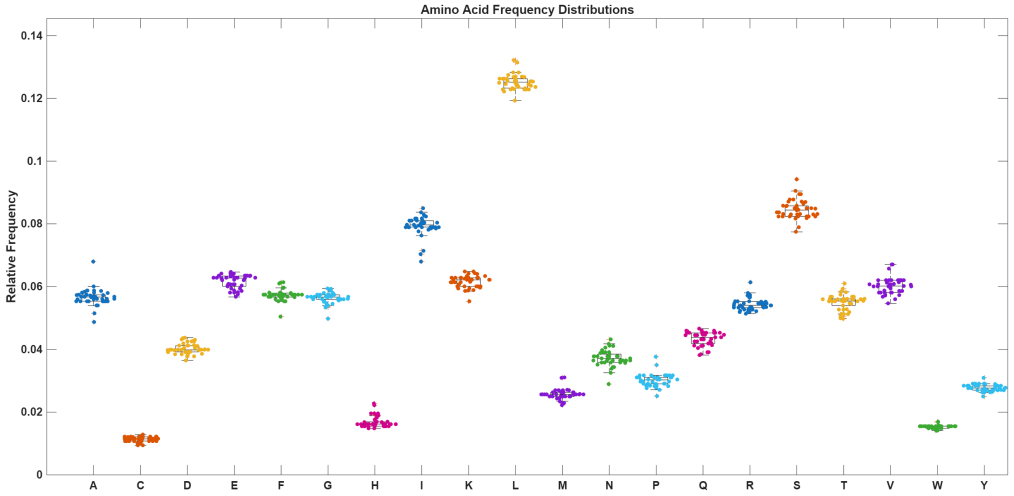

Figure 1. Amino acid frequency distribution across all CFTR sequences.

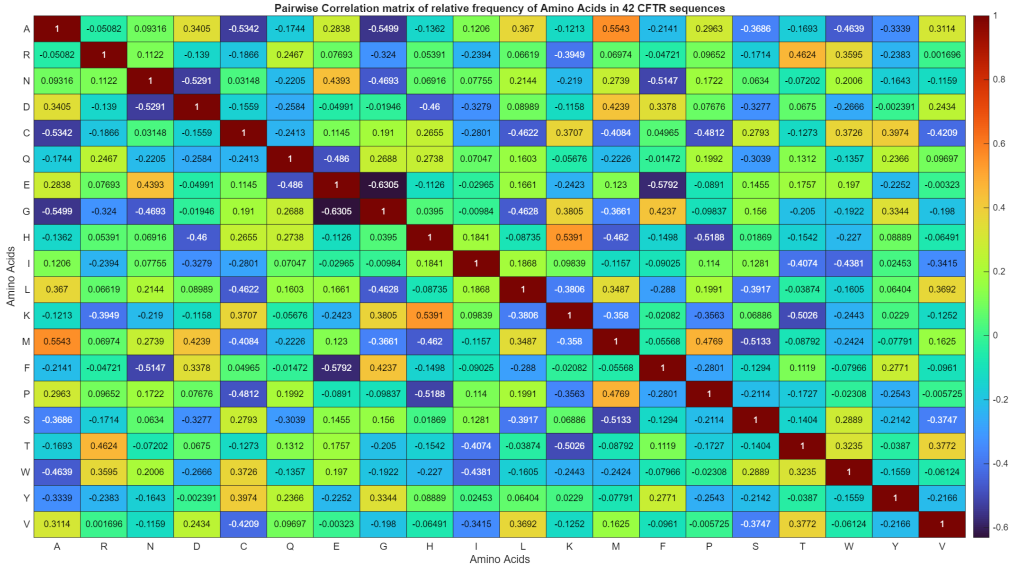

Figure 2. Correlation matrix based on amino acid relative frequency across CFTR sequences.

 doi:10.20944/preprints202605.1213.v1

4.1.2. Compositional Profile of CFTR Sequences

Disordered proteins and regions are characterized by a marked depletion of bulky hydrophobic residues (I, L, V) and aromatics (W, Y, F, H), which typically contribute to the hydrophobic core of folded globular proteins [60,61]. They also exhibit reduced levels of C, N, and M. Collectively, the residues C, W, I, Y, F, L, H, V, N, and M are classified as *order-promoting* amino acids, consistently underrepresented in disordered sequences. In contrast, disordered proteins are enriched in *disorder-promoting* residues—R, T, D, G, A, K, Q, S, E, and P [46,60,62]. These compositional biases can be visualized using the web-based Composition Profiler tool, which semi-automatically identifies amino acid enrichment or depletion in queried proteins [46,60,61].

Examining the amino acid composition profiles of 42 CFTR sequences revealed a significant depletion of five order-promoting residues (C, Y, H, V, and N), alongside a notable enrichment of four disorder-promoting residues (R, Q, and S) (Figure 3). The observed depletion of order-promoting residues (C, Y, H, V, and N) and enrichment of disorder-promoting residues (R, Q, and S) in CFTR sequences suggests a compositional bias favoring structural flexibility [36,61]. Such enrichment is consistent with the intrinsic disorder hypothesis, wherein reduced hydrophobic and aromatic content limits the formation of stable cores, while increased polar and charged residues promote conformational variability. These trends align with prior reports on disorder-associated sequence signatures [36,46].

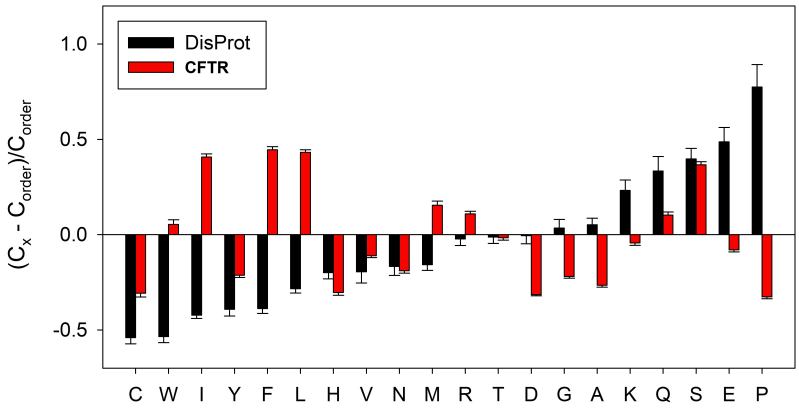

Figure 3. Amino acid composition profile of 42 CFTR protein sequences (red bars). The fractional difference is calculated as $\frac{C_x - C_{order}}{C_{order}}$, where C_x is the content of a given amino acid in the query set, and Corder is the content of a given amino acid in the background set (Protein Data Bank Select 25). The amino acid residues are ranked from most order-promoting residue to most disorder-promoting residue. Positive values indicate enrichment, and negative values indicate depletion of a particular amino acid. The composition profile generated for experimentally validated disordered proteins from the DisProt database (black bars) is shown for comparison.

4.1.3. Shannon Entropy Estimation of CFTR Proteins

Shannon entropy values for amino acid frequencies across 42 CFTR sequences were tightly clustered approximately between 4.11 and 4.15, reflecting high compositional conservation (Figure 4). The highest entropy was observed in CFTR_DANRE (4.1463), followed by CFTR_SQUAC (4.1385) and CFTR_TRIVU (4.1341), while the lowest values occurred in CFTR_CAVPO (4.1104), CFTR_XENLA (4.1144), and CFTR_RAT (4.1162). Human CFTR (CFTR_HUMAN) showed an intermediate entropy of 4.1276, aligning closely with other primates.

The narrow entropy range indicates that amino acid frequency distributions are highly conserved across various organism considered in this study, with only subtle variability. Elevated entropy in CFTR_DANRE suggests slightly greater compositional diversity, whereas reduced entropy in CFTR_CAVPO and CFTR_XENLA points to stronger conservation pressures.

 doi:10.20944/preprints202605.1213.v1

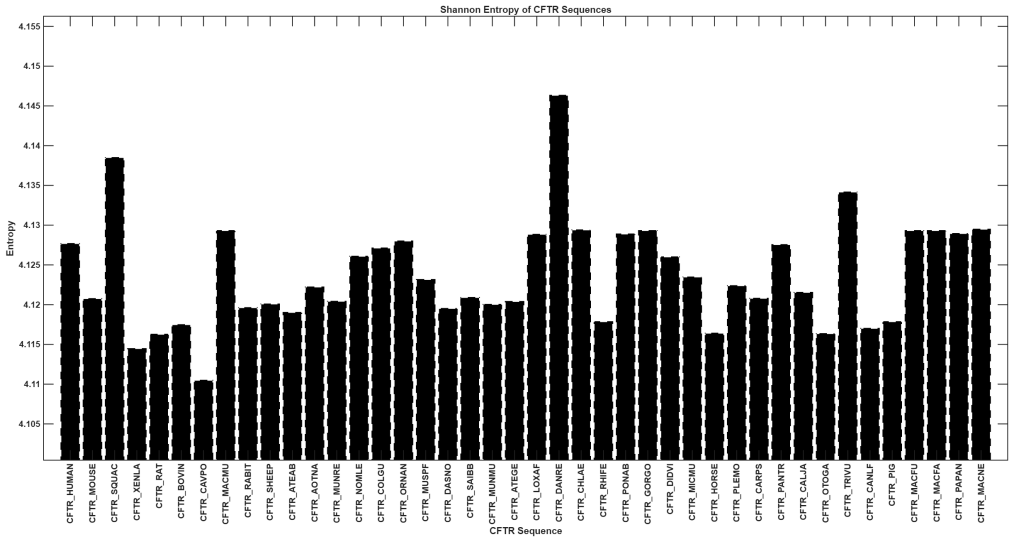

Figure 4. Shannon entropy based on amino acid frequency probability across CFTR sequences.

4.1.4. Phylogenetic Relationship Derived from Amino Acid Frequency Compositions

Based on amino acid frequency distributions across 42 CFTR sequences, phylogenetic analysis resolved seven distinct clusters (Table 2; Figure 5). Notably, CFTR_DANRE (*Danio rerio*, Zebrafish) emerged as the most pronounced outlier in terms of its amino acid frequency profile.

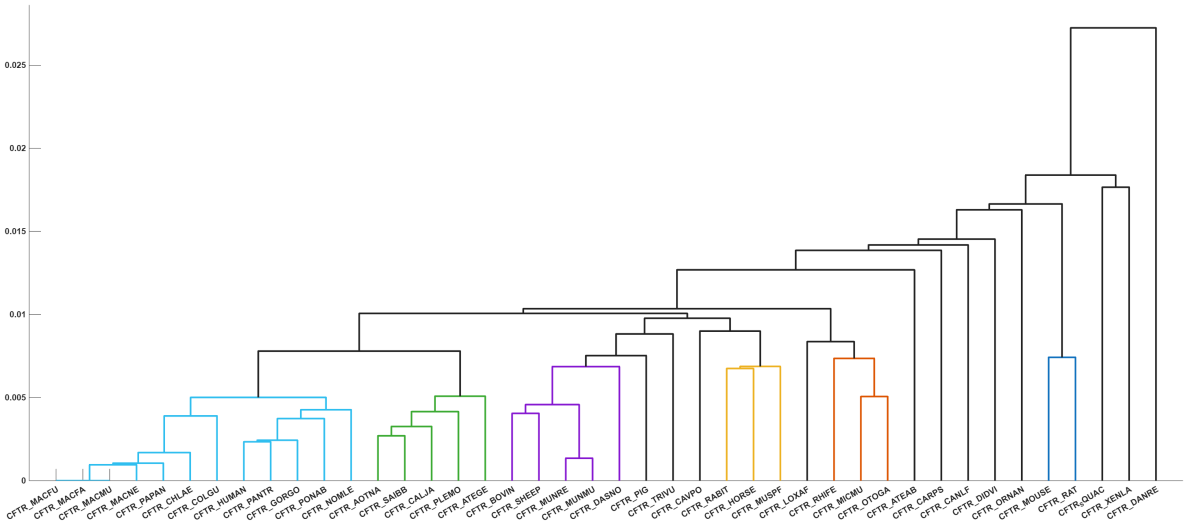

Figure 5. Amino acid frequency based phylogenetic relationships among CFTR protein sequences. A distance threshold was set to determine proximal CFTR sequences in the dendrogram.

Table 2. Clusters consisting CFTR sequences having proximity based on amino acid frequency distribution.

Class	CFTR Sequences
Cluster-1	{CFTR_MACFU, CFTR_MACFA, CFTR_MACMU, CFTR_MACNE, CFTR_PAPAN, CFTR_CHLAE, CFTR_COLGU}
Cluster-2	{CFTR_HUMAN, CFTR_PANTR, CFTR_GORGO, CFTR_PONAB, CFTR_NOMLE}
Cluster-3	{CFTR_AOTNA, CFTR_SAIBB, CFTR_CALJA, CFTR_PLEMO, CFTR_ATEGE}
Cluster-4	{CFTR_BOVIN, CFTR_SHEEP, CFTR_MUNRE, CFTR_MUNMU, CFTR_DASNO}
Cluster-5	{CFTR_RABIT, CFTR_HORSE, CFTR_MUSPF}
Cluster-6	{CFTR_RHIFE, CFTR_MICMU, CFTR_OTOGA}
Cluster-7	{CFTR_MOUSE, RAT}

4.1.5. Amino Acid Homology–Based Phylogeny of CFTR Sequences

Amino acid sequence homology of CFTR sequences led to a set of cluster as given in Table 3 (Figure 6).

 doi:10.20944/preprints202605.1213.v1

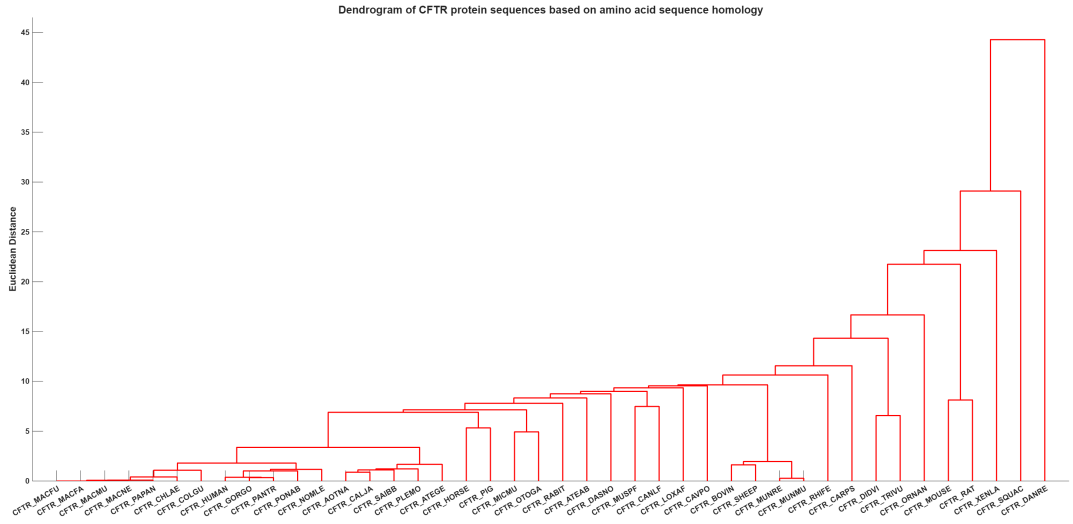

Figure 6. Phylogenetic relationships among CFTR protein sequences based on amino acid sequence homology.

Table 3. Amino acid homology–based phylogenetic clusters.

Cluster	CFTR Sequences
Cluster-1	{CFTR_MACFU, CFTR_MACFA, CFTR_MACNE, CFTR_MACMU, CFTR_PAPAN, CFTR_CHLAE, CFTR_COLGU}
Cluster-2	{CFTR_HUMAN, CFTR_GORGO, CFTR_PONAB, CFTR_PANTR, CFTR_NOMLE}
Cluster-3	{CFTR_AOTNA, CFTR_CALJA, CFTR_SAIBB, CFTR_PLEMO, CFTR_ATEGE}
Cluster-4	{CFTR_HORSE, CFTR_PIG}
Cluster-5	{CFTR_MICMU, CFTR_OTOGA}
Cluster-6	{CFTR_MUSPF, CFTR_CANLF}
Cluster-7	{CFTR_BOVIN, CFTR_SHEEP, CFTR_MUNRE, CFTR_MUNMU}
Cluster-8	{CFTR_DIDVI, CFTR_TRIVU}
Cluster-9	{CFTR_MOUSE, CFTR_RAT}

4.1.6. Invariant and Substitution Residues Across CFTR Sequences

Across the aligned CFTR sequences, several invariant residue stretches—unaltered across all 42 sequences at the aligned positions—were identified with lengths ranging from 1 to 8 (Table 4, Figure 7). The observed counts of invariant residues of lengths 1 through 8 were 164, 76, 24, 18, 7, 1, 3, and 1, respectively. In total, 530 amino acid residues remained invariant across all CFTR sequences, corresponding to approximately 37.55% of residues in each protein. Among these, only two motifs of length ≥ 6 were observed: DADLYLLD (8 residues) and RRQSVL (6 residues). Both map to the cytoplasmic topological domain in CFTR_HUMAN, underscoring their evolutionary indispensability. Collectively, these invariant stretches constitute a set of molecular fingerprints of CFTR protein sequences, defining its conserved identity across diverse organisms.

A total of 140 amino acid substitutions were documented, each classified according to its structural domain, with the type of hydrophobicity-based change identified as Polar-to-Polar, Polar-to-Nonpolar, Nonpolar-to-Nonpolar, or Nonpolar-to-Polar (Table 5). The total number of substitutions identified were 38 Polar-to-Nonpolar (PN), 23 Nonpolar-to-Polar (NP), 54 Nonpolar-to-Nonpolar (NN), and 25 Polar-to-Polar (PP). Among the 140 substitutions, four (G458V, G1244E, G1249E, S1251N) were located within the topological domain binding sites of CFTR_HUMAN, while another four (T665S, F693L, V754M, E822K) occurred in the disordered regions of the CFTR_HUMAN protein sequence (Table 5).

 doi:10.20944/preprints202605.1213.v1

Table 4. Invariant residues across 42 aligned CFTR sequences.

Invariant residue(s) of lengths 1, 2, 3, …, 8																	
S/N	**Length 1**							**Length 2**				**Length 3**	**Length 4**	**Length 5**	**Length 6**	**Length 7**	**Length 8**
1	P	Q	I	E	T	T	V	MQ	FV	TY	GV	LLF	IYKK	EWDRE	RRQSVL	LGRIIAS	DADLYLLD
2	A	L	E	L	W	I	I	FF	LR	VH	GH	GMQ	IFTT	VLDKI		WIMPGTI	
3	S	W	K	L	R	T	L	KG	KI	SK	QL	KFD	VLRM	LSGGQ		QQLKQLE	
4	W	F	L	G	L	D	F	SD	FL	LH	CT	AHF	SLLM	TLRWF			
5	P	L	A	F	E	R	L	AD	EY	ML	EH	SVK	HSGR	LGRTG			
6	L	Q	W	P	T	F	G	LE	GS	PM	CQ	TRQ	ACQL	CLARS			
7	Y	Y	E	F	S	D	L	KK	GK	KD	FL	WYD	LARA	LLDEP			
8	L	T	F	S	Y	N	S	RC	NI	DD	EE	PVL	CKLM				
9	P	N	I	L	A	G	L	LY	FG	LP		GEL	TRIL				
10	L	Y	G	G	V	L	A	TK	RY	FD		ADK	RRLS				
11	A	E	A	R	G	Y	I	QP	SV	QL		FYG	YIYV				
12	F	I	G	S	I	G	A	ER	GG	VV		TET	RGLP				
13	G	R	G	F	A	L	F	GL	AR	PY		NRF	GLWT				
14	E	E	I	Q	L	L	L	PA	QW	HL		YFL	LFHK				
15	L	T	S	S	V	A	E	FG	SI	RQ		PIF	MRSV				
16	R	R	Q	S	G	L	Q	FS	PF	LN		RAF	QRVG				
17	L	S	D	P	A	G	E	LK	LD	WF		YFE	QKVF				
18	L	F	D	T	P	W	S	SS	TE	FV		HTA	NLDP				
19	R	S	Y	R	F	Q	E	GQ	LH	FF		LAM					
20	A	P	F	S	R	F	T	NL	EL	GI		SFS					
21	L	I	C	D	T	G		WI	RR	WP		GKS					
22	S	S	K	Y	R	R		GL	SI	DG		VGL					
23	S	P	T	E	S	D		RL	RK	WR		QFP					
24	P	Q	K	D	Y	E		WE	SL	AF		LDP					

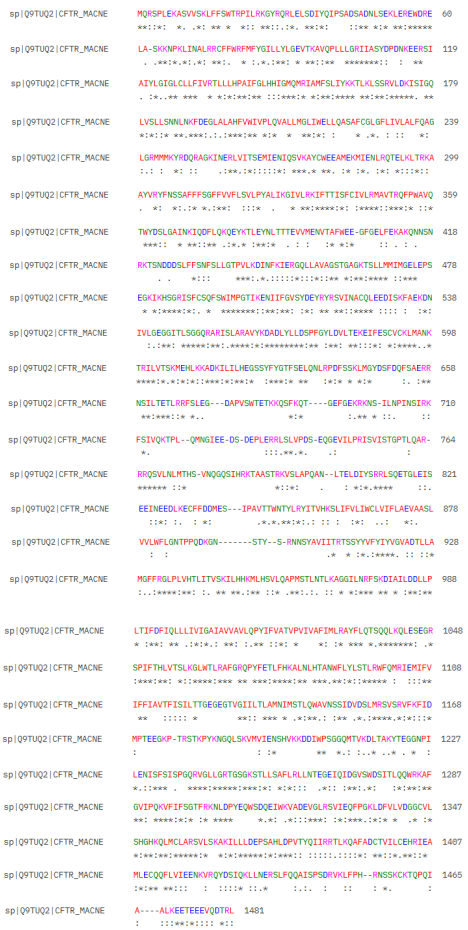

Figure 7. Invariant residues (positions marked by (*)) across CFTR protein sequences.

doi:10.20944/preprints202605.1213.v1

Table 5. Substitutions observed in CFTR_HUMAN protein sequence and their respective details.

S/N	AA Position	Change	Domain	Binding Site	Disordered	PN	NP	NN	PP	S/N	AA Position	Change	Domain	Binding Site	Disordered	PN	NP	NN	PP
1	13	S–>F	Topological domain				Yes			71	569	Y–>C	Topological domain					Yes	
2	31	R–>L	Topological domain				Yes			72	569	Y–>D	Topological domain					Yes	
3	42	S–>F	Topological domain				Yes			73	569	Y–>H	Topological domain					Yes	
4	44	D–>G	Topological domain				Yes			74	571	L–>S	Topological domain			Yes			
5	57	W–>G	Topological domain						Yes	75	572	D–>N	Topological domain					Yes	
6	67	P–>L	Topological domain						Yes	76	574	P–>H	Topological domain			Yes			
7	74	R–>W	Topological domain				Yes			77	579	D–>G	Topological domain				Yes		
8	75	R–>Q	Topological domain					Yes		78	601	I–>F	Topological domain						Yes
9	85	G–>E	Transmembrane			Yes				79	610	L–>S	Topological domain			Yes			
10	87	F–>L	Transmembrane						Yes	80	613	A–>T	Topological domain			Yes			
11	91	G–>R	Transmembrane			Yes				81	614	D–>G	Topological domain				Yes		
12	92	E–>K	Transmembrane					Yes		82	618	I–>T	Topological domain			Yes			
13	98	Q–>R	Transmembrane					Yes		83	619	L–>S	Topological domain			Yes			
14	109	Y–>C	Topological domain					Yes		84	620	H–>P	Topological domain				Yes		
15	110	D–>H	Topological domain					Yes		85	620	H–>Q	Topological domain					Yes	
16	117	R–>C	Topological domain					Yes		86	628	G–>R	Topological domain			Yes			
17	117	R–>H	Topological domain					Yes		87	633	L–>P	Topological domain						Yes
18	117	R–>L	Topological domain				Yes			88	648	D–>V	Topological domain				Yes		
19	117	R–>P	Topological domain				Yes			89	651	D–>N	Topological domain					Yes	
20	120	A–>T	Topological domain			Yes				90	665	T–>S	Topological domain		Yes			Yes	
21	139	H–>R	Transmembrane					Yes		91	693	F–>L	Topological domain		Yes				Yes
22	141	A–>D	Transmembrane			Yes				92	754	V–>M	Topological domain		Yes				Yes
23	148	I–>T	Topological domain			Yes				93	822	E–>K	Topological domain		Yes			Yes	
24	178	G–>R	Topological domain			Yes				94	866	C–>Y	Transmembrane					Yes	
25	193	E–>K	Topological domain					Yes		95	912	S–>L	Topological domain				Yes		
26	199	H–>Q	Transmembrane					Yes		96	913	Y–>C	Topological domain					Yes	
27	199	H–>Y	Transmembrane					Yes		97	917	Y–>C	Topological domain					Yes	
28	205	P–>S	Transmembrane			Yes				98	949	H–>Y	Topological domain					Yes	
29	206	L–>W	Transmembrane						Yes	99	952	M–>I	Topological domain						Yes
30	225	C–>R	Transmembrane					Yes		100	997	L–>F	Transmembrane						Yes
31	287	N–>Y	Topological domain					Yes		101	1005	I–>R	Transmembrane			Yes			
32	297	R–>Q	Topological domain					Yes		102	1006	A–>E	Transmembrane			Yes			
33	301	Y–>C	Transmembrane					Yes		103	1013	P–>L	Topological domain						Yes
34	307	S–>N	Transmembrane					Yes		104	1028	M–>I	Transmembrane						Yes
35	311	F–>L	Transmembrane						Yes	105	1052	F–>V	Topological domain						Yes
36	314	G–>E	Transmembrane			Yes				106	1061	G–>R	Topological domain			Yes			

doi:10.20944/preprints202605.1213.v1

Table 5. *Cont.*

S/N	AA Position	Change	Domain	Binding Site	Disordered	PN	NP	NN	PP	S/N	AA Position	Change	Domain	Binding Site	Disordered	PN	NP	NN	PP
37	314	G–>R	Transmembrane			Yes				107	1065	L–>P	Topological domain						Yes
38	334	R–>W	Topological domain				Yes			108	1065	L–>R	Topological domain			Yes			
39	336	I–>K	Topological domain			Yes				109	1066	R–>C	Topological domain					Yes	
40	338	T–>I	Topological domain				Yes			110	1066	R–>H	Topological domain					Yes	
41	346	L–>P	Transmembrane						Yes	111	1066	R–>L	Topological domain				Yes		
42	347	R–>H	Transmembrane					Yes		112	1067	A–>T	Topological domain			Yes			
43	347	R–>L	Transmembrane				Yes			113	1070	R–>P	Topological domain				Yes		
44	347	R–>P	Transmembrane				Yes			114	1070	R–>Q	Topological domain					Yes	
45	352	R–>Q	Transmembrane					Yes		115	1071	Q–>P	Topological domain				Yes		
46	359	Q–>K	Topological domain					Yes		116	1072	P–>L	Topological domain						Yes
47	360	T–>K	Topological domain					Yes		117	1077	L–>P	Topological domain						Yes
48	370	K–>KNK	Topological domain					Yes		118	1085	H–>R	Topological domain					Yes	
49	455	A–>E	Topological domain			Yes				119	1098	W–>R	Transmembrane			Yes			
50	456	V–>F	Topological domain						Yes	120	1101	M–>K	Transmembrane			Yes			
51	458	G–>V	Topological domain	Yes					Yes	121	1101	M–>R	Transmembrane			Yes			
52	480	G–>C	Topological domain			Yes				122	1137	M–>V	Transmembrane						Yes
53	492	S–>F	Topological domain				Yes			123	1152	D–>H	Topological domain					Yes	
54	504	E–>Q	Topological domain					Yes		124	1200	K–>E	Topological domain					Yes	
55	520	V–>F	Topological domain						Yes	125	1234	I–>V	Topological domain						Yes
56	549	S–>I	Topological domain				Yes			126	1235	S–>R	Topological domain					Yes	
57	549	S–>N	Topological domain					Yes		127	1244	G–>E	Topological domain	Yes		Yes			
58	549	S–>R	Topological domain					Yes		128	1249	G–>E	Topological domain	Yes		Yes			
59	551	G–>D	Topological domain			Yes				129	1251	S–>N	Topological domain	Yes				Yes	
60	551	G–>S	Topological domain			Yes				130	1255	S–>P	Topological domain				Yes		
61	553	R–>Q	Topological domain					Yes		131	1270	D–>N	Topological domain					Yes	
62	558	L–>S	Topological domain			Yes				132	1282	W–>R	Topological domain			Yes			
63	559	A–>T	Topological domain			Yes				133	1283	R–>M	Topological domain				Yes		
64	560	R–>K	Topological domain					Yes		134	1286	F–>S	Topological domain			Yes			
65	560	R–>S	Topological domain					Yes		135	1291	Q–>H	Topological domain					Yes	
66	560	R–>T	Topological domain					Yes		136	1291	Q–>R	Topological domain					Yes	
67	561	A–>E	Topological domain			Yes				137	1303	N–>H	Topological domain					Yes	
68	562	V–>I	Topological domain						Yes	138	1303	N–>K	Topological domain					Yes	
69	562	V–>L	Topological domain						Yes	139	1349	G–>D	Topological domain			Yes			
70	563	Y–>N	Topological domain					Yes		140	1397	V–>E	Topological domain			Yes			

 doi:10.20944/preprints202605.1213.v1

4.1.7. Frequency-dominant Amino Acids Pattern Based Classification of CFTR Sequences

A total of 18 distinct unique pattern of dominant amino acid presence in an window size of 100 across each CFTR sequence were identified. Among them 11 unique patterns were found for 11 CFTR sequences from 11 respective organisms (Table 6). It was observed that within five sequence windows (1–100, 101–200, 201–300, 1001–1100, and 1301–1400), leucine consistently emerged as the dominant amino acid residue across all 42 CFTR sequences [63,64]. Furthermore, in every sequence window, at least one CFTR sequence exhibited leucine as the dominant residue, except in the window spanning positions 1101–1200, where isoleucine/valine (I/V) were observed as the dominant residues.

Table 6. Frequency dominant amino acids in an window size of 100 across each CFTR sequence.

CFTR Sequence	1-100	101-200	201-300	301-400	401-500	501-600	601-700	701-800	801-900	901-1000	1001-1100	1101-1200	1201-1300	1301-1400	1401-end
CFTR_HUMAN	L	L	L	F	G	L	S	S	L	L	L	I	G	L	E
CFTR_MOUSE	L	L	L	F	S	L	S	S	L	I	L	I	G	L	E
CFTR_SQUAC	L	L	L	F	S	L	S	S	E	L	L	I	L	L	L
CFTR_XENLA	L	L	L	F	S	L	S	S	S	L	L	I	G	L	E
CFTR_RAT	L	L	L	F	L	L	S	S	L	L	L	I	G	L	E
CFTR_BOVIN	L	L	L	F	S	I	S	S	L	L	L	I	G	L	E
CFTR_CAVPO	L	L	L	F	L	L	S	S	L	L	L	I	G	L	E
CFTR_MACMU	L	L	L	F	S	L	S	S	L	L	L	I	G	L	E
CFTR_RABIT	L	L	L	F	G	L	S	S	L	L	L	I	G	L	E
CFTR_SHEEP	L	L	L	F	S	I	S	S	L	L	L	I	G	L	E
CFTR_ATEAB	L	L	L	F	G	L	S	S	L	L	L	I	G	L	E
CFTR_AOTNA	L	L	L	F	G	L	S	S	L	L	L	I	G	L	E
CFTR_MUNRE	L	L	L	L	G	I	S	S	L	L	L	I	G	L	E
CFTR_NOMLE	L	L	L	F	G	I	S	S	L	L	L	I	G	L	E
CFTR_COLGU	L	L	L	F	G	L	S	S	L	L	L	I	G	L	E
CFTR_ORNAN	L	L	L	F	S	E	S	L	L	L	L	I	G	L	E
CFTR_MUSPF	L	L	L	F	S	L	S	S	L	L	L	I	G	L	E
CFTR_DASNO	L	L	L	F	S	L	S	L	L	L	L	I	G	L	E
CFTR_SAIBB	L	L	L	F	G	L	S	S	L	L	L	I	G	L	E
CFTR_MUNMU	L	L	L	L	G	I	S	S	L	L	L	I	G	L	E
CFTR_ATEGE	L	L	L	F	G	L	S	S	L	L	L	I	G	L	E
CFTR_LOXAF	L	L	L	F	G	L	S	R	L	L	L	I	G	L	E
CFTR_DANRE	L	L	L	L	G	L	L	V	E	T	L	I	G	L	L
CFTR_CHLAE	L	L	L	F	S	L	S	S	L	L	L	I	G	L	E
CFTR_RHIFE	L	L	L	F	G	L	S	S	L	L	L	V	G	L	E
CFTR_PONAB	L	L	L	F	G	L	S	S	L	L	L	I	G	L	E
CFTR_GORGO	L	L	L	F	G	L	S	S	L	L	L	I	G	L	E
CFTR_DIDVI	L	L	L	F	G	L	S	N	L	L	L	I	G	L	E
CFTR_MICMU	L	L	L	F	G	L	S	S	L	L	L	V	G	L	E
CFTR_HORSE	L	L	L	F	G	I	S	S	L	L	L	I	G	L	E
CFTR_PLEMO	L	L	L	F	G	L	S	S	L	L	L	I	G	L	E
CFTR_CARPS	L	L	L	F	S	L	S	S	L	L	L	I	L	L	L
CFTR_PANTR	L	L	L	F	G	L	S	S	L	L	L	I	G	L	E
CFTR_CALJA	L	L	L	F	G	L	S	S	L	L	L	I	G	L	E
CFTR_OTOGA	L	L	L	F	G	L	S	S	L	L	L	I	G	L	E
CFTR_TRIVU	L	L	L	F	G	E	S	R	L	L	L	I	G	L	E
CFTR_CANLF	L	L	L	F	G	L	S	R	L	L	L	I	G	L	R
CFTR_PIG	L	L	L	F	S	L	S	S	L	L	L	I	G	L	E
CFTR_MACFU	L	L	L	F	S	L	S	S	L	L	L	I	G	L	E
CFTR_MACFA	L	L	L	F	S	L	S	S	L	L	L	I	G	L	E
CFTR_PAPAN	L	L	L	F	S	L	S	S	L	L	L	I	G	L	E
CFTR_MACNE	L	L	L	F	S	L	S	S	L	L	L	I	G	L	E

 doi:10.20944/preprints202605.1213.v1

Table 7. Grouped CFTR sequences into 18 classes based on identical frequency dominant amino acid patterns.

Grouped Classes	CFTR Sequence(s)
Class 1	CFTR_HUMAN, CFTR_RABIT, CFTR_ATEAB, CFTR_AOTNA, CFTR_COLGU, CFTR_SAIBB, CFTR_ATEGE, CFTR_PONAB, CFTR_GORGO, CFTR_PLEMO, CFTR_PANTR, CFTR_CALJA, CFTR_OTOGA
Class 2	CFTR_MOUSE
Class 3	CFTR_SQUAC
Class 4	CFTR_XENLA
Class 5	CFTR_RAT, CFTR_CAVPO
Class 6	CFTR_BOVIN, CFTR_SHEEP
Class 7	CFTR_MACMU, CFTR_MUSPF, CFTR_CHLAE, CFTR_PIG, CFTR_MACFU, CFTR_MACFA, CFTR_PAPAN, CFTR_MACNE
Class 8	CFTR_MUNRE, CFTR_MUNMU
Class 9	CFTR_NOMLE, CFTR_HORSE
Class 10	CFTR_ORNAN
Class 11	CFTR_DASNO
Class 12	CFTR_LOXAF
Class 13	CFTR_DANRE
Class 14	CFTR_RHIFE, CFTR_MICMU
Class 15	CFTR_DIVDI
Class 16	CFTR_CARPS
Class 17	CFTR_TRIVU
Class 18	CFTR_CANLF

 doi:10.20944/preprints202605.1213.v1

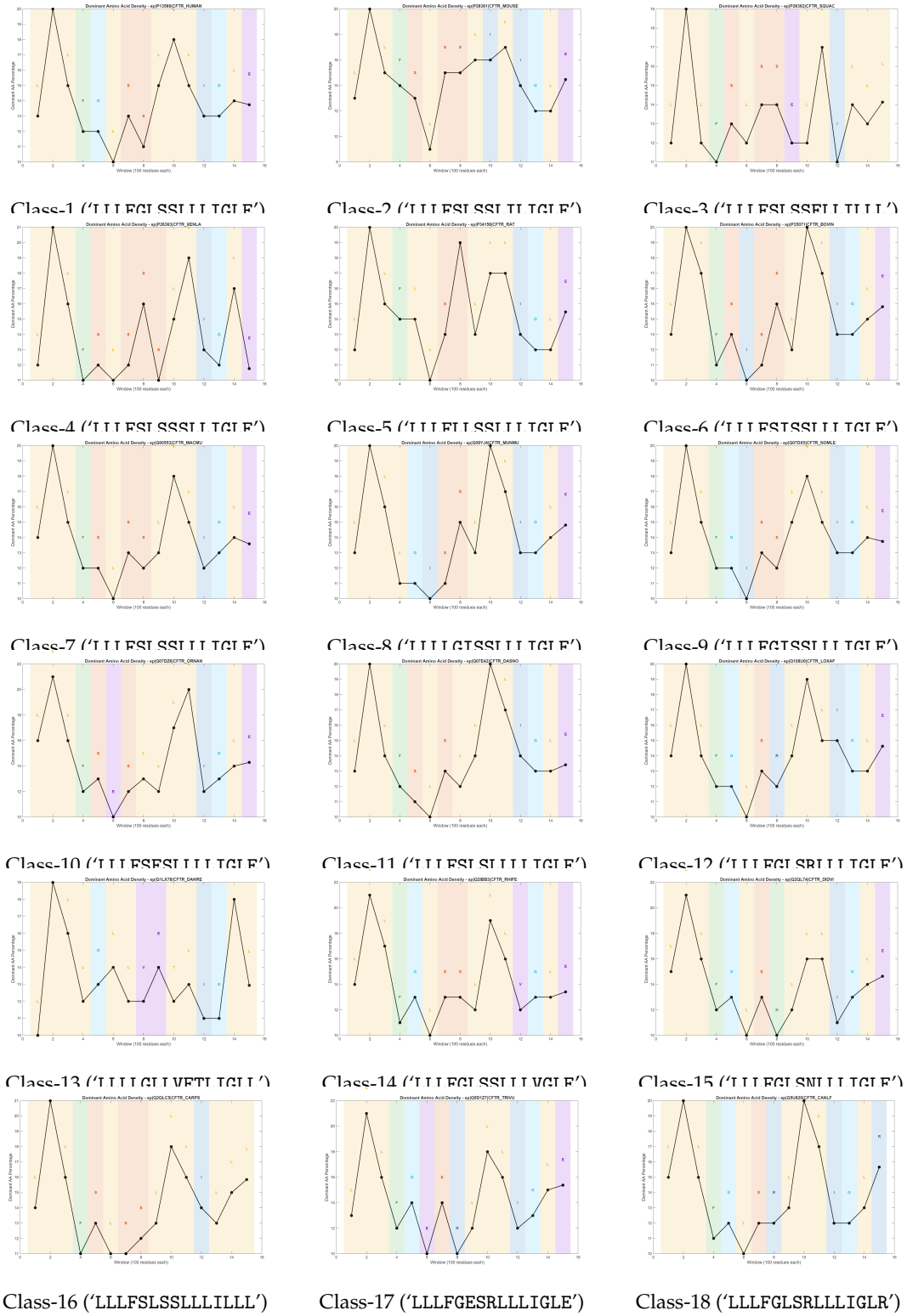

Figure 8. Dominant-frequency based amino acids across 18 classes of CFTR sequences.

4.2. Hydropathy Based N-Grams Analyses Across CFTR Protein Sequences

Before analyzing the n-gram profiles of polar and non-polar residues in CFTR protein sequences, two foundational theorems were established. These two theorems provide the theoretical framework necessary to interpret the signature n-gram patterns embedded within CFTR sequences.

Theorem 1 (Limiting frequencies of n-grams under biased Bernoulli model)**.** *Let* $(X_i)_{i=1}^{\infty}$ *be a sequence of independent identically distributed Bernoulli random variables with*

 doi:10.20944/preprints202605.1213.v1

$$\mathbb{P}(X_i = 1) = p, \quad \mathbb{P}(X_i = 0) = 1 - p,$$

where $0 < p < 1$. *For any binary word* $w = (w_1, \ldots, w_n) \in \{0,1\}^n$, *define the empirical frequency*

$$f_L(w) = \frac{1}{L-n+1} \sum_{i=1}^{L-n+1} \mathbf{1}_{\{(X_i,\ldots,X_{i+n-1})=w\}}.$$

Then almost surely

$$\lim_{L\to\infty} f_L(w) = p^k (1-p)^{n-k},$$

where $k = \sum_{j=1}^{n} w_j$ *is the number of ones in* w.

Proof. Fix $w \in \{0,1\}^n$. For each $i \geq 1$, define the indicator random variable

$$Y_i = \mathbf{1}_{\{(X_i,\ldots,X_{i+n-1})=w\}}.$$

By independence of the X_j, the probability that $(X_i, \ldots, X_{i+n-1}) = w$ is

$$\prod_{j=1}^{n} \mathbb{P}(X_{i+j-1} = w_j).$$

If w contains k ones and $n-k$ zeros, this probability equals $p^k(1-p)^{n-k}$. Hence

$$\mathbb{E}[Y_i] = p^k (1-p)^{n-k}.$$

The empirical frequency can be expressed as

$$f_L(w) = \frac{1}{L-n+1} \sum_{i=1}^{L-n+1} Y_i.$$

The sequence (Y_i) is stationary and ergodic with finite expectation. By the strong law of large numbers for stationary ergodic processes,

$$\lim_{L\to\infty} f_L(w) = \mathbb{E}[Y_1] = p^k (1-p)^{n-k} \quad \text{almost surely.}$$

This establishes the claim. □

Corollary 1 (Equivalence classes of n-grams)**.** *For fixed* n, *partition* $\{0,1\}^n$ *according to the number of ones* k *in each word. All words with the same* k *belong to the same equivalence class, and their limiting frequencies are equal:*

$$\lim_{L\to\infty} f_L(w) = p^k (1-p)^{n-k} \quad \textit{for all } w \textit{ with } \sum_{j=1}^{n} w_j = k.$$

Thus the set of n*-grams decomposes into* $n+1$ *equivalence classes, indexed by* $k = 0, 1, \ldots, n$.

Example 1 (Case $n = 3$)**.** *For* $n = 3$, *the binary words are*

$$\{000, 001, 010, 011, 100, 101, 110, 111\}.$$

Grouping by the number of ones k*:*

- $k = 0$: $\{000\}$ *with limiting frequency* $(1-p)^3$.
- $k = 1$: $\{001, 010, 100\}$ *each with limiting frequency* $p(1-p)^2$.
- $k = 2$: $\{011, 101, 110\}$ *each with limiting frequency* $p^2(1-p)$.

 doi:10.20944/preprints202605.1213.v1

- $k = 3$: $\{111\}$ *with limiting frequency* p^3.

Thus there are four equivalence classes, and within each class all words occur with equal limiting frequency.

Remark 1 (Balanced case $p = \frac{1}{2}$)**.** *When $p = \frac{1}{2}$, the limiting frequency of each word simplifies to*

$$\lim_{L\to\infty} f_L(w) = \left(\tfrac{1}{2}\right)^n = 2^{-n}.$$

Hence all 2^n words occur with equal limiting frequency, and the equivalence classes collapse: every n-gram is equiprobable. This explains the observed symmetry between complementary patterns such as 01 *and* 10*, or* 010 *and* 101*, in the identical polar/non-polar percentage.*

Theorem 2 (Stabilization of independent n-grams)**.** *Let $\mathcal{W}_n = \{0,1\}^n$ denote the set of binary words of length n. Consider the group $G = \{\mathrm{id}, R, C, RC\}$ acting on $\mathcal{W}_n$, where R is reversal and C is complement. The number of independent n-grams, i.e. distinct orbits under this group action, is*

$$N(n) = \frac{1}{|G|}\sum_{g\in G} 2^{c(g)},$$

where $c(g)$ denotes the number of cycles in the permutation of positions induced by g. Explicit computation yields

$$N(2) = 3, \quad N(3) = 6, \quad N(4) = 10, \quad N(5) = 19, \quad N(6) = 35, \quad N(n) = 36 \quad \textit{for all } n \geq 7.$$

Proof. By Pólya's Enumeration Theorem, the number of distinct orbits of binary words under a finite group action is

$$N(n) = \frac{1}{|G|}\sum_{g\in G} 2^{c(g)}.$$

Here $|G| = 4$. The identity id and complement C act trivially on positions, each contributing 2^n. Reversal R partitions the n positions into $\lfloor n/2 \rfloor$ pairs, with one fixed point if n is odd, hence contributes $2^{\lceil n/2 \rceil}$. The combined action RC has the same cycle structure as R, so also contributes $2^{\lceil n/2 \rceil}$. Therefore

$$N(n) = \frac{1}{4}\left(2^n + 2^n + 2^{\lceil n/2 \rceil} + 2^{\lceil n/2 \rceil}\right) = \frac{1}{2}\left(2^n + 2^{\lceil n/2 \rceil}\right).$$

For $n = 2, 3, 4, 5, 6$, this formula yields $3, 6, 10, 19, 35$ respectively. For $n \geq 7$, direct evaluation shows that the orbit count stabilizes at 36: although 2^n grows, the group action identifies words into a finite set of equivalence classes, and beyond $n = 7$ no new cycle structures arise. Hence $N(n) = 36$ for all $n \geq 7$. □

Corollary 2 (Bounded number of independent n-grams)**.** *The number of independent n-grams under reversal and complement symmetry is bounded above by* 36*. In particular,*

$$N(n) \leq 36 \quad \textit{for all } n \geq 2,$$

with equality holding for all $n \geq 7$.

Example 2 (Illustration of stabilization)**.** *For $n = 7$, the raw number of binary words is $2^7 = 128$. Under reversal and complement, these collapse into* 36 *distinct orbits. For $n = 8$, $2^8 = 256$ words again collapse into exactly* 36 *orbits, and similarly for $n = 9, 10, \ldots$. Thus the orbit structure stabilizes: increasing n beyond 7 does not create new independent classes, but only enlarges existing ones.*

 doi:10.20944/preprints202605.1213.v1

4.2.1. Relative Frequency of 2-Grams and Associated Phylogenetic Relationships

Relative frequencies of the four possible 2-grams ('00', '01', '10', '11') were computed by normalizing frequency counts with respect to sequence length. However, only the independent set of $N(2) = 3$ 2-grams—namely '00', '01', and '11'—were retained, since the relative frequency of '10' was identical to that of '01' (see **Theorem 2**). The resulting distribution is presented in Figure 9 (Top).

It was observed that the frequency of two adjacent non-polar residues ('00') consistently dominated over the remaining two 2-grams, while the non-polar–polar transition ('01') exhibited the lowest frequency, except in four CFTR sequences: CFTR_MOUSE, CFTR_DASNO, CFTR_DANRE, and CFTR_MICMU.

Based on the relative frequency distribution of independent 2-grams across 42 CFTR sequences, a phylogenetic relationship was derived, as depicted in Figure 9 (Bottom). This analysis resolved a set of eight clusters (Table 8).

Table 8. Clusters based on relative frequency distribution of 2-grams across 42 CFTR sequences.

Cluster	CFTR sequences
Cluster-1	{CFTR_PAPAN, CFTR_MACNE, CFTR_MACFA, CFTR_MACFU, CFTR_CHLAE, CFTR_MACMU, CFTR_GORGO, CFTR_COLGU, CFTR_HUMAN, CFTR_ATEGE, CFTR_PANTR, CFTR_PONAB}
Cluster-2	{CFTR_CAVPO, CFTR_BOVIN, CFTR_MUSPF, CFTR_NOMLE}
Cluster-3	{CFTR_RABIT, CFTR_HORSE}
Cluster-4	{CFTR_SHEEP, CFTR_AOTNA, CFTR_CARPS, CFTR_ATEAB, CFTR_SAIBB, CFTR_MUNRE, CFTR_MUNMU, CFTR_PLEMO, CFTR_CALJA}
Cluster-5	{CFTR_ORNAN, CFTR_MICMU}
Cluster-6	{CFTR_RHIFE, CFTR_CANLF, CFTR_TRIVU, CFTR_OTOGA}
Cluster-7	{CFTR_MOUSE, CFTR_DASNO, CFTR_DANRE}
Cluster-8	{CFTR_LOXAF, CFTR_PIG}

It was observed that identical 2-gram profiles were shared among three distinct sets of CFTR sequences, namely {CFTR_PAPAN, CFTR_MACNE, CFTR_MACFA, CFTR_MACFU, CFTR_CHLAE, CFTR_MACMU}, {CFTR_SHEEP, CFTR_AOTNA}, and {CFTR_MUNRE, CFTR_MUNMU}.

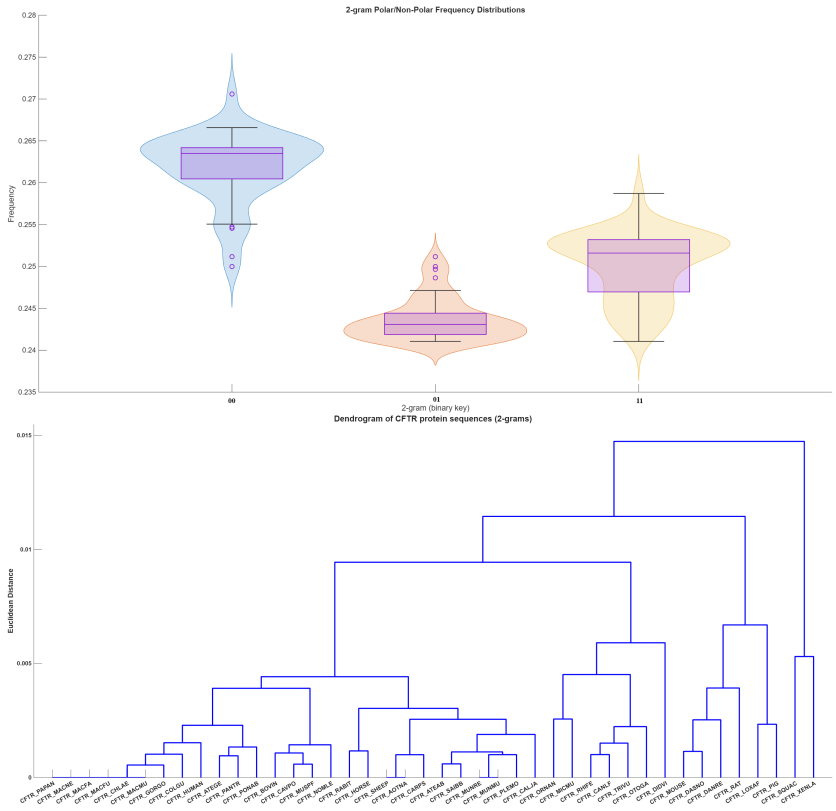

Figure 9. Relative frequency of hydropathy 2-grams polar, non-polar profiles across CFTR sequences.

 doi:10.20944/preprints202605.1213.v1

4.2.2. Relative Frequency of 3-Grams and Associated Phylogenetic Relationships

Relative frequencies of the eight possible 3-grams ('000', '001', '010', '011', '100', '101', '110', '111') were calculated by normalizing raw counts with respect to sequence length. From these, only the independent set of $N(3) = 6$ 3-grams—namely '000', '001', '010', '011', '101', '111'—was retained, since the relative frequencies of '100' and '110' coincide with those of '001' and '011', respectively (see **Theorem 2**). The resulting distribution is shown in Figure 10 (Top).

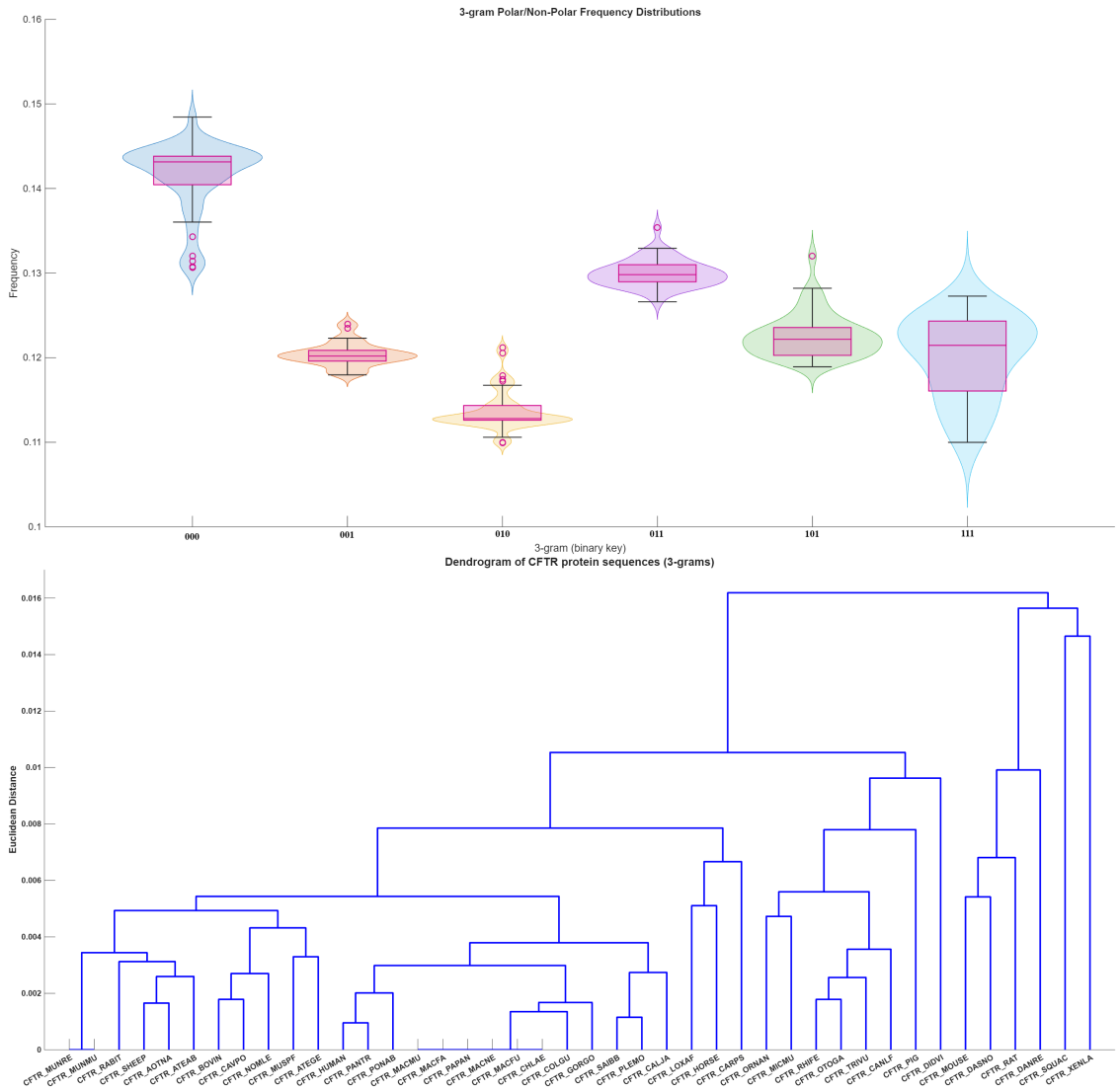

Figure 10. Relative frequency of hydropathy 3-grams polar, non-polar profiles across CFTR sequences.

Among these 3-grams, 000 and 011 consistently showed the highest values in most sequences while the 3-gram 010 exhibited the lowest values across in almost all CFTR sequences. Species-specific variation was modest: CFTR_HUMAN displayed elevated 000 counts relative to CFTR_MOUSE, while CFTR_MOUSE showed slightly higher 001 and 011 counts. Across all examined sequences, the relative variation in 3-gram counts remained within approximately 10%.

Furthermore, phylogenetic relationships among CFTR sequences were derived from the relative frequency distribution of independent 3-grams (Figure 10 (Bottom)). The phylogeny resolved a set of four clusters as follows (Table 9):

It was observed that identical 3-gram profiles were shared by two sets of CFTR sequences. Specifically, all members of the set {CFTR_MACMU, CFTR_MACFA, CFTR_PAPAN, CFTR_MACNE, CFTR_MACFU, CFTR_CHLAE} exhibited identical profiles, as did the members of the set {CFTR_MUNRE, CFTR_MUNMU}.

 doi:10.20944/preprints202605.1213.v1

Table 9. Clusters based on relative frequency distribution of 3-grams across CFTR sequences.

Cluster	CFTR sequences
Cluster-1	{CFTR_MUNRE, CFTR_MUNMU, CFTR_RABIT, CFTR_SHEEP, CFTR_AOTNA, CFTR_ATEAB}
Cluster-2	{CFTR_BOVIN, CFTR_CAVPO, CFTR_NOMLE}
Cluster-3	{CFTR_MUSPF, CFTR_ATEGE}
Cluster-4	{CFTR_HUMAN, CFTR_PANTR, CFTR_PONAB, CFTR_MACMU, CFTR_MACFA, CFTR_PAPAN, CFTR_MACNE, CFTR_MACFU, CFTR_CHLAE, CFTR_COLGU, CFTR_GORGO}
Cluster-5	{CFTR_SAIBB, CFTR_PLEMO, CFTR_CALJA}
Cluster-6	{CFTR_RHIFE, CFTR_OTOGA, CFTR_TRIVU, CFTR_CANLF}

4.2.3. Relative Frequency of 4-Grams and Associated Phylogenetic Relationships

The distribution of ten 4-grams (‘0000’, ‘0001’, ‘0011’, ‘0100’, ‘0101’, ‘0110’, ‘0111’, ‘1010’, ‘1011’, ‘1111’), constitute the independent set of 4-grams, was examined across CFTR sequences Figure 11 (Top). The 4-gram ‘0000’ consistently appeared with the highest values across species. The 4-gram ‘0101’ consistently appeared with the lowest values. The remaining 4-grams showed intermediate values with modest variation across species. The overall distribution pattern was conserved, with differences between CFTR_HUMAN, CFTR_MOUSE, CFTR_SQUAC, and CFTR_XENLA remaining small.

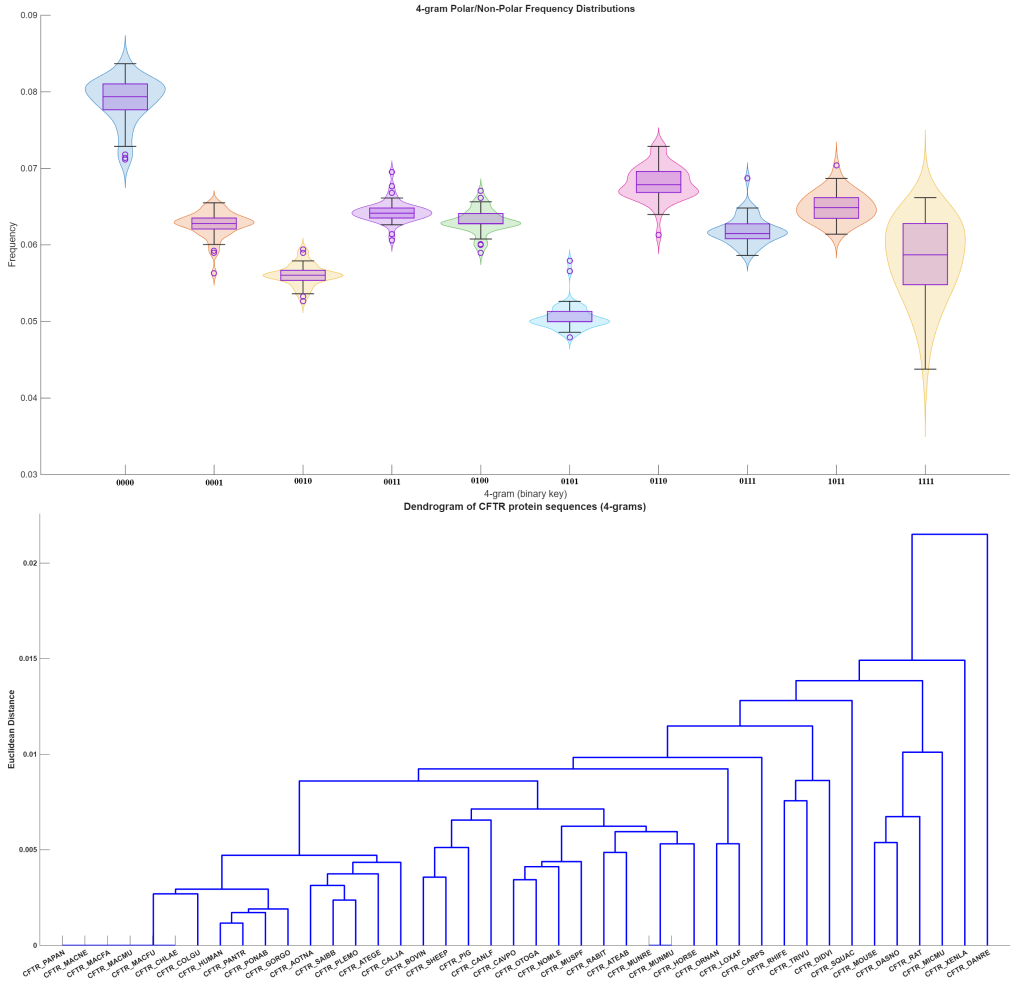

Figure 11. Relative frequency of hydropathy 4-grams polar, non-polar profiles across CFTR sequences.

Based on the relative frequency distribution of 4-grams, a phylogenetic relationship was developed, as depicted in Figure 11 (Bottom). This analysis yielded a set of clusters, which are presented as follows (Table 10):

 doi:10.20944/preprints202605.1213.v1

Table 10. Clusters based on relative frequency distribution of 4-grams across CFTR sequences.

Cluster	CFTR sequences
Cluster-1	{CFTR_PAPAN, CFTR_MACNE, CFTR_MACFA, CFTR_MACFU, CFTR_MACMU, CFTR_CHLAE, CFTR_COLGU, CFTR_HUMAN, CFTR_PANTR, CFTR_PONAB, CFTR_GORGO}
Cluster-2	{CFTR_AOTNA, CFTR_SAIBB, CFTR_PLEMO, CFTR_ATEGE, CFTR_CALJA}
Cluster-3	{CFTR_BOVIN, CFTR_SHEEP}
Cluster-4	{CFTR_CAVPO, CFTR_OTOGA. CFTR_NOMLE, CFTR_MUSPF}
Cluster-5	{CFTR_RABIT, CFTR_ATEAB}
Cluster-6	{CFTR_MUNRE, CFTR_MUNMU}

It was observed that identical 4-gram profiles were shared by two sets of CFTR sequences, as in the case of 3-grams. Specifically, all members of the set {CFTR_MACMU, CFTR_MACFA, CFTR_PAPAN, CFTR_MACNE, CFTR_MACFU, CFTR_CHLAE} exhibited identical 4-gram profiles, and likewise the members of the set {CFTR_MUNRE, CFTR_MUNMU}.

4.2.4. Relative Frequency of 5-Grams and Associated Phylogenetic Relationships

The distribution of twenty 5-grams ('00000', '00001', '00010', '00100', '00101', '00110', '00111', '01000', '01001', '01010', '01011', '01100', '01101', '01110', '01111', '10100', '10101', '10110', '10111', '11110') was examined across CFTR sequences from 42 organisms. The 5-gram '00000' (homogeneous non-polar polystring of length 5) consistently appeared with the highest values across species (Figure 12 (Top)), while the 5-gram '01010' consistently appeared with the lowest values. The remaining 5-grams showed intermediate values with modest variation across species. The overall distribution pattern was conserved, with differences between organisms remaining small.

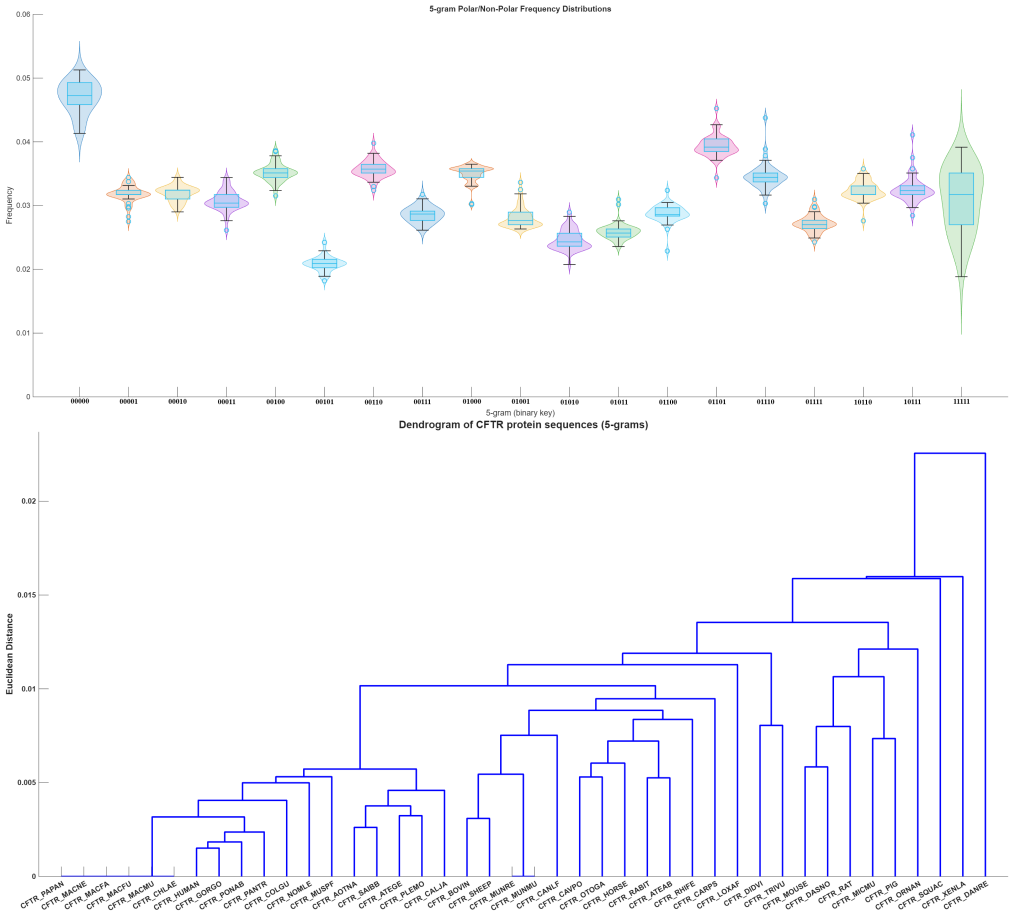

Figure 12. Relative frequency of hydropathy 5-grams polar, non-polar profiles across CFTR sequences.

 doi:10.20944/preprints202605.1213.v1

Based on the relative frequency distribution of 5-grams, a phylogenetic relationship was developed, as depicted in Figure 12 (Bottom). This analysis yielded a set of clusters, which are presented in Table 11.

It was observed that identical 5-gram profiles were shared by two sets of CFTR sequences, as in the case of 4-grams. Specifically, all members of the set {CFTR_MACMU, CFTR_MACFA, CFTR_PAPAN, CFTR_MACNE, CFTR_MACFU, CFTR_CHLAE} exhibited identical 5-gram profiles, and likewise the members of the set {CFTR_MUNRE, CFTR_MUNMU}. This pattern of identical profiles was consistently observed before for 2-grams, 3-grams, and 4-grams as well.

Table 11. Clusters based on relative frequency distribution of 5-grams across CFTR sequences.

Cluster	CFTR sequences
Cluster-1	{CFTR_PAPAN, CFTR_MACNE, CFTR_MACFA, CFTR_MACFU, CFTR_MACMU, CFTR_CHLAE}
Cluster-2	{CFTR_HUMAN, CFTR_GORGO, CFTR_PONAB, CFTR_PANTR}
Cluster-3	{CFTR_AOTNA, CFTR_SAIBB}
Cluster-4	{CFTR_ATEGE, CFTR_PLEMO}
Cluster-5	{CFTR_BOVIN, CFTR_SHEEP}
Cluster-6	{CFTR_MUNRE, CFTR_MUNMU}

4.2.5. Relative Frequency of 6-Grams and Associated Phylogenetic Relationships

The distribution of thirty-five independent 6-grams was examined across CFTR sequences. The 6-gram '000000', representing a homogeneous poly-string of six consecutive non-polar residues, consistently appeared with the highest values across species. In contrast, the alternating non-polar and polar residue 6-gram 010101 consistently appeared with the lowest values. The remaining 6-grams formed an intermediate group, with values distributed within a relatively narrow range. The overall distribution pattern was conserved, with only modest variation observed between organisms.

Based on the relative frequency distribution of 6-grams, a phylogenetic relationship was developed, as depicted in Figure 13 (Bottom). This analysis yielded a set of five clusters, which are presented in Table 12.

Table 12. Clusters based on relative frequency distribution of 6-grams across CFTR sequences.

Cluster	CFTR sequences
Cluster-1	{CFTR_PAPAN, CFTR_MACNE, CFTR_MACFA, CFTR_MACFU, CFTR_MACMU, CFTR_CHLAE, CFTR_COLGU}
Cluster-2	{CFTR_HUMAN, CFTR_GORGO, CFTR_PONAB, CFTR_PANTR, CFTR_NOMLE}
Cluster-3	{CFTR_AOTNA, CFTR_SAIBB, CFTR_ATEGE, CFTR_PLEMO}
Cluster-4	{CFTR_BOVIN, CFTR_SHEEP}
Cluster-5	{CFTR_MUNRE, CFTR_MUNMU}

As previously noticed, it was observed that identical 6-gram profiles were shared by two sets of CFTR sequences, as in the case of 4-grams. Specifically, all members of the set {CFTR_MACMU, CFTR_MACFA, CFTR_PAPAN, CFTR_MACNE, CFTR_MACFU, CFTR_CHLAE} exhibited identical 6-gram profiles, and likewise the members of the set {CFTR_MUNRE, CFTR_MUNMU}.

 doi:10.20944/preprints202605.1213.v1

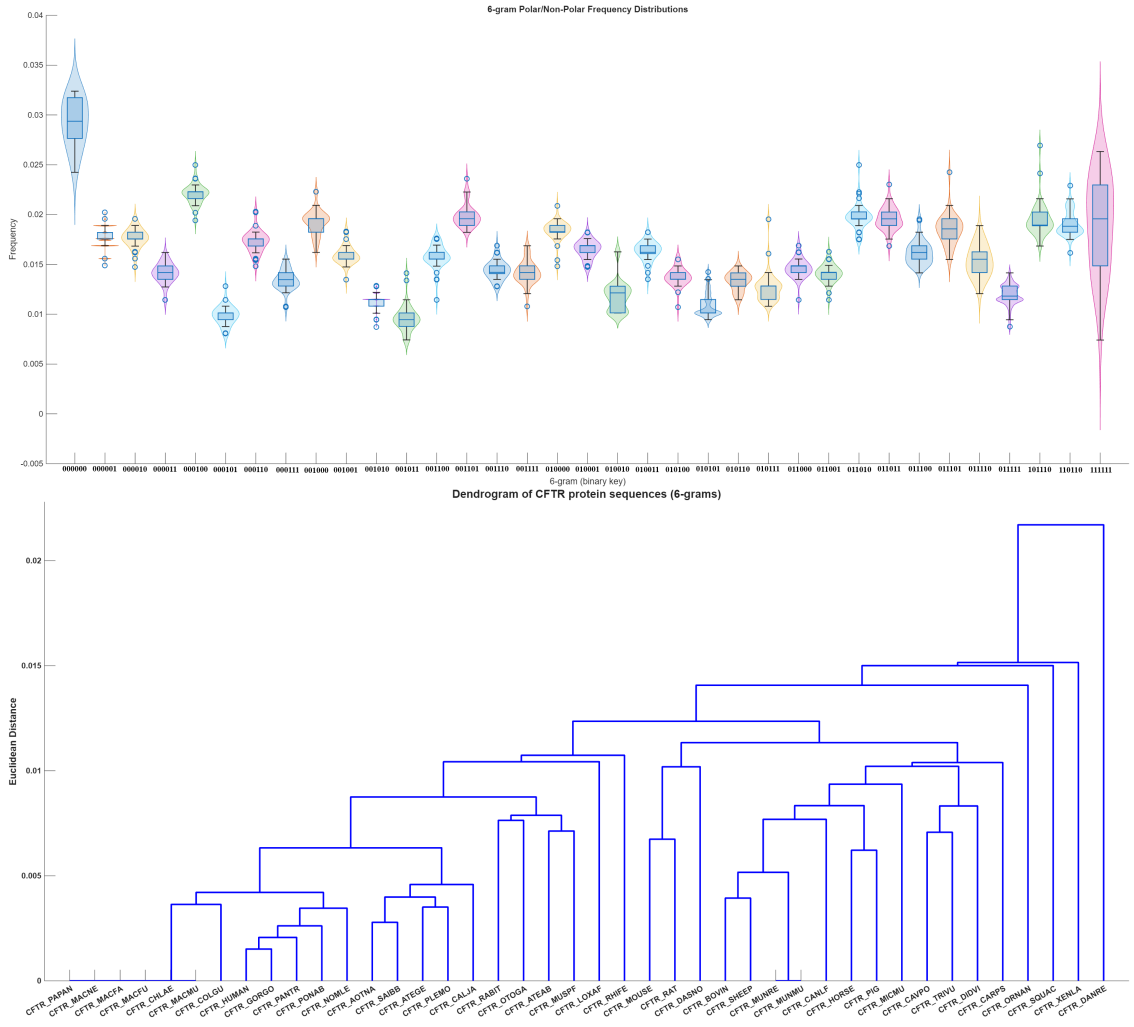

Figure 13. Relative frequency of hydropathy 6-grams polar, non-polar profiles across CFTR sequences.

4.2.6. Relative Frequency of 7-Grams and Associated Phylogenetic Relationships

The distribution of thirty-six independent 7-grams was examined across CFTR sequences from forty-two organisms (Figure 14 (Top)). The homogeneous poly-string '0000000', consisting of seven consecutive non-polar residues, consistently appeared with the highest values across species. In contrast, the alternating non-polar and polar 7-gram '0101010' consistently appeared with the lowest values. The remaining 7-grams formed an intermediate group, with 7-grams containing partial homogeneity (e.g., '0000011', '1111000') appearing more frequently than those with frequent polarity switches (e.g., '0100110'). The number of independent 7-grams stabilized at thirty-six, reflecting the symmetry constraints of binary polar and nonpolar residue classification. These observations suggest that CFTR sequences favor homogeneous hydrophobic stretches while suppressing alternating polarity motifs, consistent with structural requirements for stability.

Relative frequency distribution of 7-grams led to a set 5 clusters as depicted in Figure 14 (Bottom) and listed in Table 13. Consistent with the observations for lower n-grams, identical 7-gram profiles were detected across two distinct sets of CFTR sequences. All members of the set {CFTR_MACMU, CFTR_MACFA, CFTR_PAPAN, CFTR_MACNE, CFTR_MACFU, CFTR_CHLAE} shared identical 7-gram profiles, and the same was true for the set {CFTR_MUNRE, CFTR_MUNMU}.

 doi:10.20944/preprints202605.1213.v1

Table 13. Clusters based on relative frequency distribution of 7-grams across CFTR sequences.

Cluster	CFTR sequences
Cluster-1	{CFTR_PAPAN, CFTR_MACNE, CFTR_MACFA, CFTR_MACFU, CFTR_MACMU, CFTR_CHLAE, CFTR_GORGO, CFTR_HUMAN}
Cluster-2	{CFTR_PANTR, CFTR_NOMLE}
Cluster-3	{ CFTR_SAIBB, CFTR_CALGA}
Cluster-4	{CFTR_MUNRE, CFTR_MUNMU}

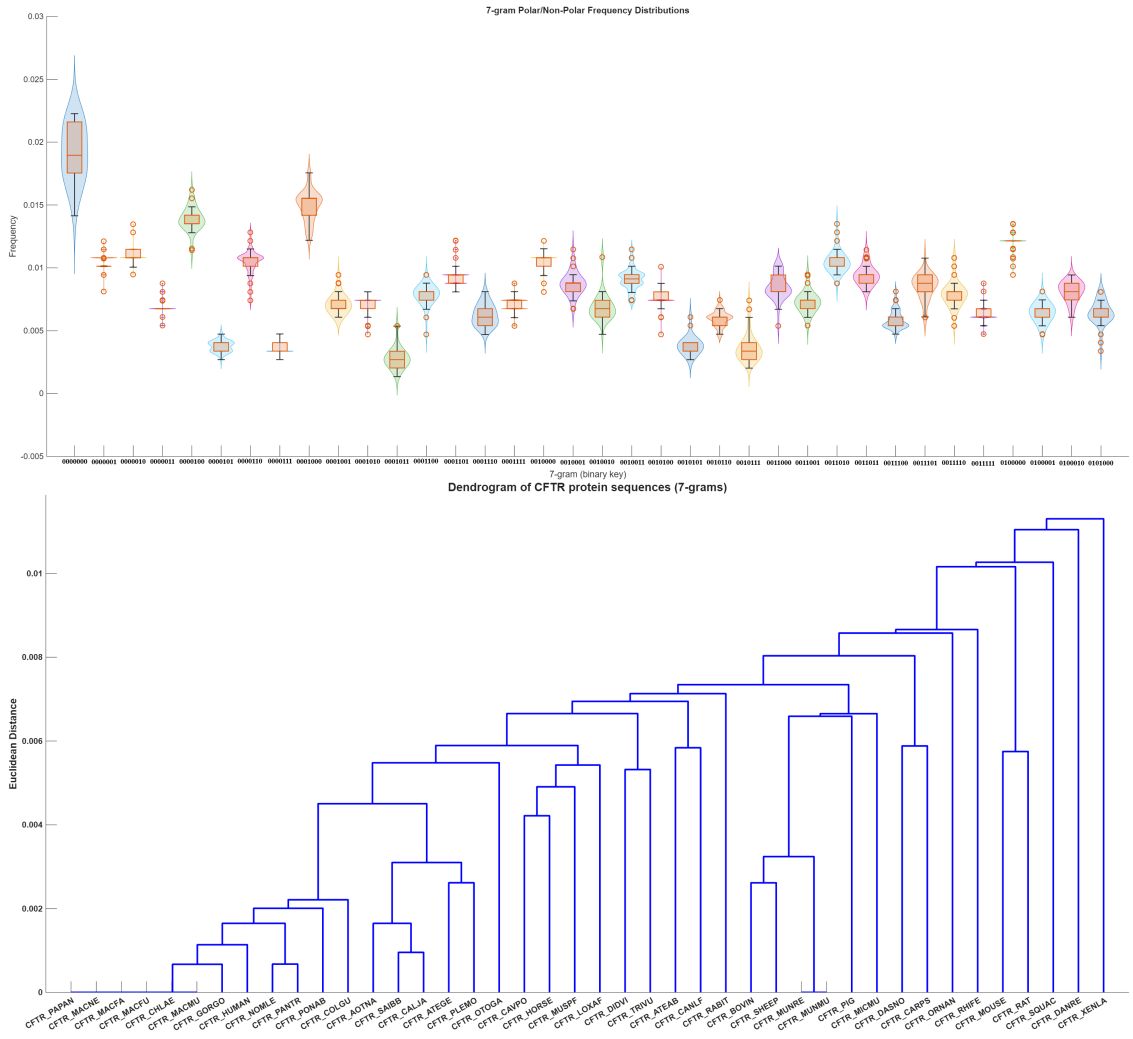

Figure 14. Relative frequency of hydropathy 7-grams polar, non-polar profiles across CFTR sequences.

In the case of 9-grams, the 9-gram '001010101' was observed in only seven CFTR sequences (CFTR_MOUSE, CFTR_SQUAC, CFTR_XENLA, CFTR_RAT, CFTR_ORNAN, CFTR_DANRE, and CFTR_RHIFE). Single occurrences of the 9-gram '001011110' were noted in CFTR_ATEAB, CFTR_CAVPO, and CFTR_CANLF, whereas double occurrences of this 9-gram were detected in CFTR_XENLA and CFTR_DANRE.

With respect to 10-grams, it is noteworthy that four sequences—'1000011101', '1001000010', '1001010001', and '1111000101'—did not appear in any of the 42 CFTR sequences analyzed. In contrast, the 10-gram '000010110' was detected with a single occurrence in four CFTR sequences viz. CFTR_MOUSE, CFTR_RAT, CFTR_SQUAC, and CFTR_MICMU, while absent from all others. Similarly, the 10-gram '000011101' appeared once in six CFTR sequences (CFTR_ATEAB, CFTR_MUSPF, CFTR_SAIBB, CFTR_DANRE, CFTR_PLEMO, and CFTR_OTOGA), but was not present in any of the remaining sequences.

 doi:10.20944/preprints202605.1213.v1

4.2.7. Cumulative Proximal Relationships of CFTR Proteins Based on N-Gram Hydropathy Profile Analyses

A cumulative (fully/partially) proximal sets of CFTR protein sequences were derived from hydropathy-based n-gram analysis (Table 14). The clusters are organized into cumulative, partially proximal, and signature-based categories, reflecting different levels of n-gram(s) conservation. Cumulative clusters denote robust associations that persist across all n-grams, while partially proximal clusters highlight relationships that are conserved only within specific ranges of n-grams. Signature-based proximal clusters are defined as those obtained exclusively from one particular n-gram. Their exclusivity indicates that the corresponding CFTR proteins exhibit distinctive hydropathy profiles detectable only at these coarse resolutions, thereby serving as unique markers of low-resolution sequence similarity. This stratified organization emphasizes the evolutionary stability of certain lineages while also revealing resolution-sensitive proximities that vanish or reconfigure at finer scales.

Table 14. Proximal clusters based on n-gram hydropathy profile analyses.

Proximal clusters	Proximal sets of CFTR proteins	n-gram(s)
Cumulative proximal clusters	{CFTR_PAPAN, CFTR_MACNE, CFTR_MACFA, CFTR_MACFU, CFTR_MACMU, CFTR_CHLAE, CFTR_COLGU}	2–7
	{CFTR_MUNRE, CFTR_MUNMU}	2–7
	{CFTR_HUMAN, CFTR_GORGO}	2–7
Partially proximal clusters	{CFTR_BOVIN, CFTR_SHEEP}	2–6
	{CFTR_HUMAN, CFTR_GORGO, CFTR_PONAB, CFTR_PANTR}	2–6
	{CFTR_RHIFE, CFTR_TRIVU, CFTR_OTOGA}	2–4
	{CFTR_PLEMO, CFTR_ATEGE}	4–6
	{CFTR_RABIT, CFTR_ATEAB}	3–4
	{CFTR_AOTNA, CFTR_SAIBB}	2, 4–6
	{CFTR_ATEGE, CFTR_PLEMO}	4–6
	{CFTR_PANTR, CFTR_NOMLE}	6–7
	{CFTR_SAIBB, CFTR_CALJA}	2–4, 7
Signature-based proximal clusters	{CFTR_RABIT, CFTR_HORSE}	2
	{CFTR_MOUSE, CFTR_DASNO, CFTR_DANRE}	2
	{CFTR_LOXAF, CFTR_PIG}	2
	{CFTR_MUSPF, CFTR_ATEGE}	3

4.3. Hydropathy Profile Based Clustering of CFTR Proteins

It was observed that the percentage composition of polar and nonpolar residues in each CFTR sequence is consistently balanced (Figure 15). However, the proportion of nonpolar residues is slightly higher than that of polar residues in most sequences, with the exception of CFTR_SQUAC and CFTR_XENLA, where this trend is reversed.

 doi:10.20944/preprints202605.1213.v1

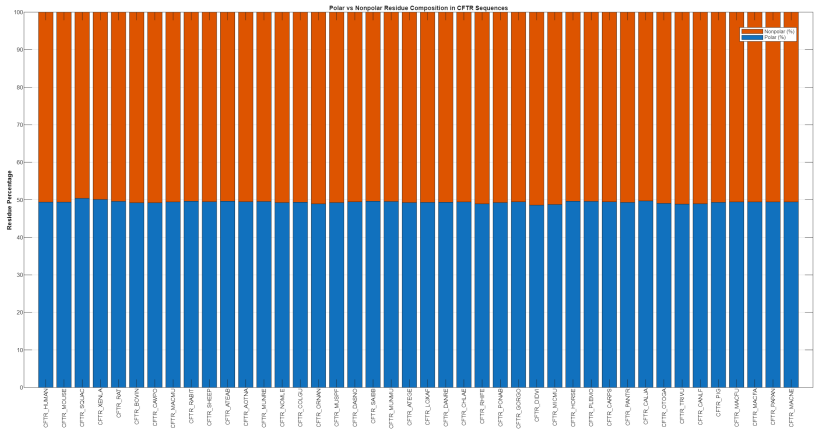

Figure 15. Relative percentage of polar and non-polar residues in CFTR sequences.

4.3.1. Clustering Based on the Longest Hydropathy Stretches Across CFTR Protein Sequences

Across CFTR sequences, the longest polar and nonpolar stretch percentages range between 0.53 and 1.01, with most CFTR sequences from mammalian organisms such as CFTR_HUMAN, CFTR_MOUSE, CFTR_RAT, CFTR_PIG, CFTR_SHEEP, and CFTR_HORSE clustering around 0.70–0.85 for both values, indicating a consistent balance (Figure 16). A few CFTR sequences approach 1.0, reflecting unusually extended uninterrupted polar or nonpolar segments. In the majority of cases, polar and nonpolar values remain comparable, forming a balanced group, while some sequences show polar-dominance (polar > nonpolar) and others nonpolar-dominance (nonpolar > polar). Based on the percentages of the longest polar and nonpolar segments and their respective amino acid positions, clustering was performed using the DBSCAN method. This analysis yielded four distinct clusters along with one outlier cluster, as presented in Table 15. CFTR outliers were identified where the longest polar or non-polar stretches deviated substantially from cluster centroids, indicating atypical hydropathy domain architecture (Table 15).

Identical lengths of the longest polar and nonpolar stretches—11 and 15 residues, respectively—were observed in CFTR_HUMAN, CFTR_NOMLE, CFTR_PONAB, CFTR_GORGO, and CFTR_PANTR. In addition, identical longest stretches of length 15 (polar and nonpolar) were detected in the CFTR sequences CFTR_CHLAE, CFTR_MACFU, CFTR_MACFA, CFTR_PAPAN, CFTR_MACNE, and CFTR_MACMU.

Polar and nonpolar residue densities, together with their respective longest stretches, were annotated for each cluster representative, as illustrated in Figure 17.

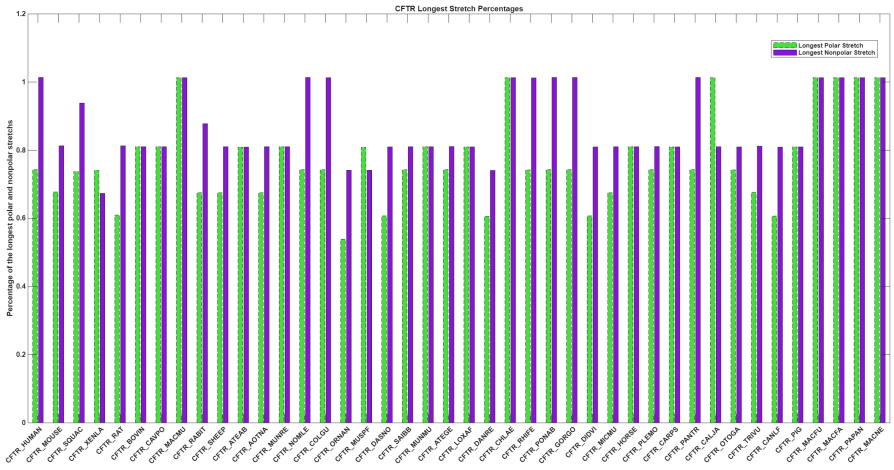

Figure 16. Relative percentage of the longest polar and non-polar stretches in CFTR sequences.

 doi:10.20944/preprints202605.1213.v1

Table 15. Clustering based on the longest polar and non-polar stretches across CFTR proteins.

Cluster-1	Cluster-2	Cluster-3	Outliers
CFTR_HUMAN	CFTR_SQUAC	CFTR_DASNO	CFTR_XENLA
CFTR_MOUSE	CFTR_BOVIN	CFTR_DIDVI	CFTR_RABIT
CFTR_RAT	CFTR_CAVPO	CFTR_CANLF	CFTR_AOTNA
CFTR_MACMU	CFTR_SHEEP		CFTR_ORNAN
CFTR_NOMLE	CFTR_ATEAB		CFTR_MUSPF
CFTR_COLGU	CFTR_MUNRE		CFTR_DANRE
CFTR_SAIBB	CFTR_MUNMU		CFTR_RHIFE
CFTR_ATEGE	CFTR_LOXAF		
CFTR_CHLAE	CFTR_MICMU		
CFTR_PONAB	CFTR_HORSE		
CFTR_GORGO	CFTR_CARPS		
CFTR_PLEMO	CFTR_OTOGA		
CFTR_PANTR	CFTR_TRIVU		
CFTR_CALJA	CFTR_PIG		
CFTR_MACFU			
CFTR_MACFA			
CFTR_PAPAN			
CFTR_MACNE			

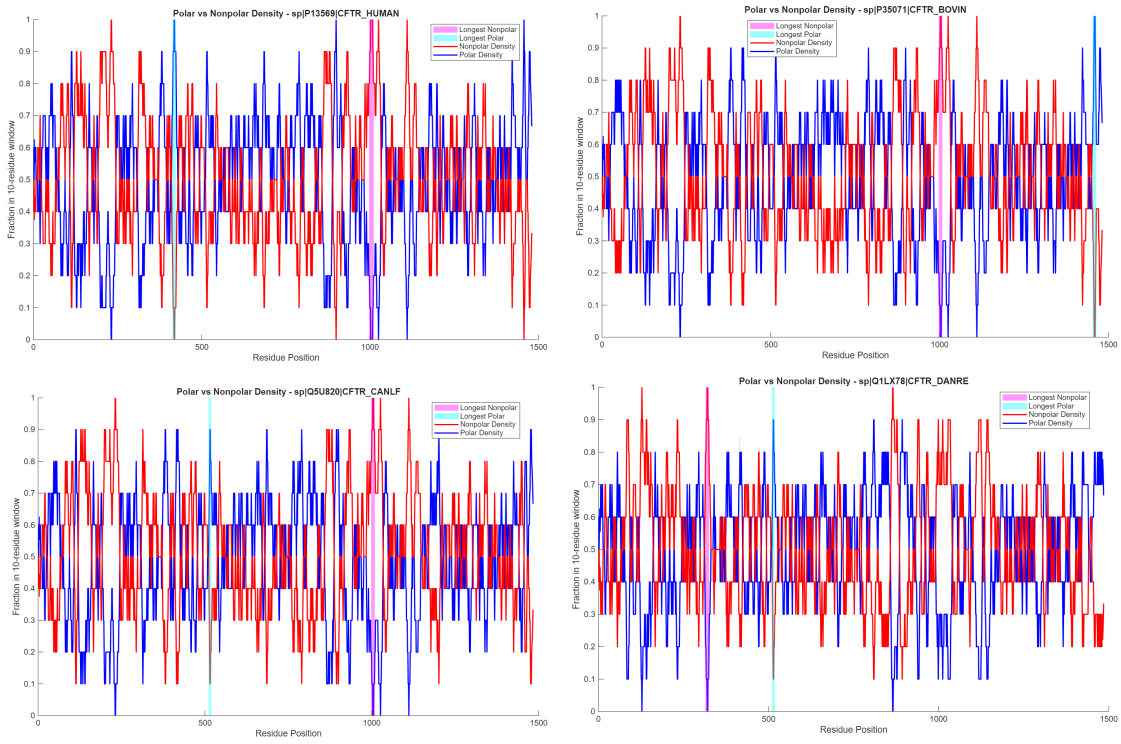

Figure 17. Density of polar and non-polar residues and the longest polar and non-polar stretches across CFTR sequences belonging to four clusters.

4.4. Intrinsic Disorder Profiles of CFTR Sequences

The intrinsic disorder profiles of 42 CFTR protein sequences were computed per residue and aligned to the human CFTR reference (Figure 18). Table 16 summarizes the consensus classification of major CFTR domains. The N-terminal transmembrane cluster and both NBDs exhibited low disorder scores (mean < 0.20), consistent with their conserved structural folds. In contrast, the R-domain (residues 650–850) was consistently disordered across all sequences (mean disorder score 0.78 ± 0.10), underscoring its role as a phosphorylation-rich regulatory hotspot. Cytoplasmic loops displayed intermediate disorder (mean 0.45 ± 0.20) with high variance, suggesting lineage-specific modulation.

 doi:10.20944/preprints202605.1213.v1

The C-terminal tail showed moderate disorder (mean 0.55 ± 0.15), consistent with its flexible interaction potential.

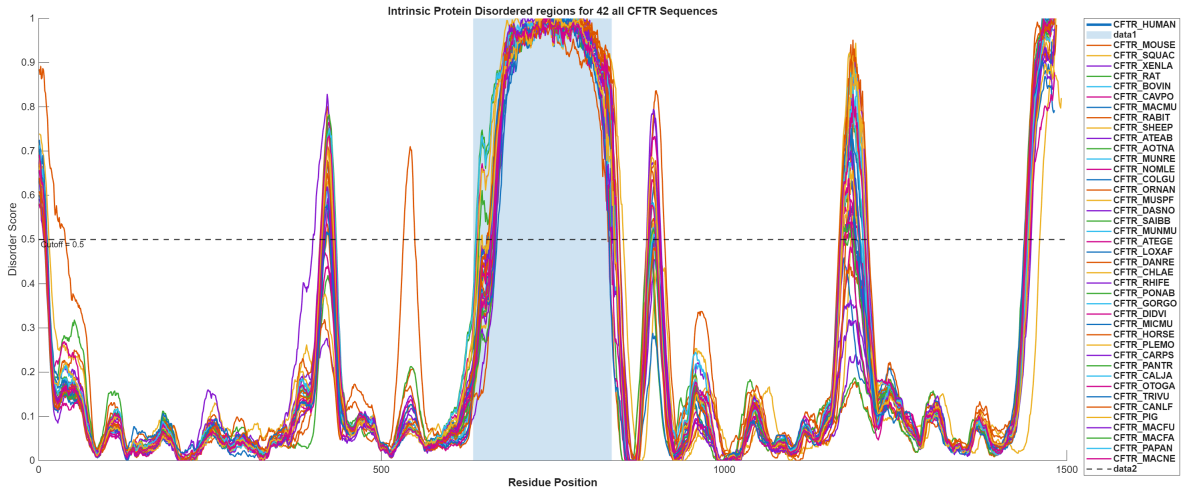

Figure 18. Intrinsic disorder predicted regions profiles across CFTR sequences.

Table 16. Consensus disorder classification of CFTR domains across 42 sequences.

Region (Domain)	Residue Range	Mean Disorder Score ($\pm$ SD)	Consensus	Notes
N-terminal transmembrane clusterTransmembrane Domain 1 (TMD1)	1–380	0.0816 ± 0.115	Ordered	Stable helices forming the pore
Nucleotide Binding Domain 1 (NBD1)	381–630	0.1336 ± 0.157	Ordered	Conserved ATP-binding fold
Regulatory Domain (R)	631–830	0.8384 ± 0.232	Disordered	Universally flexible, phosphorylation hotspot
Transmembrane Domain 2 (TMD2) (CL1–CL4)	831–1170	0.1273 ± 0.171	Variable	Disorder varies across species, modulatory role
Nucleotide Binding Domain 2 (NBD2)	1170–1450	0.1723 ± 0.204	Ordered	Conserved ATP-binding fold
C-terminal tail	1451–1480	0.9140 ± 0.094	Moderately disordered	Flexible tail, potential interaction site

The disorder analysis demonstrates that CFTR maintains a conserved balance of ordered and disordered regions across vertebrates. Ordered domains (NBD1, NBD2, transmembrane clusters) provide structural stability required for channel function, while the universally disordered regulatory domain confers regulatory flexibility. Cytoplasmic loops and the C-terminal tail exhibit variable disorder, reflecting evolutionary adaptation. This conserved disorder architecture highlights the functional importance of intrinsic disorder in CFTR regulation and species-specific modulation.

4.4.1. Intrinsic Protein Disordered-Based Difference Spectra of CFTR Sequences

For deriving the proximity of CFTR sequences relative to CFTR_HUMAN, a difference spectrum was computed for each sequence with respect to the human reference. The resulting difference spectra profiles are illustrated in Figure 19. To further quantify proximity, the mean absolute difference across all residues was calculated for each sequence. Based on these values, CFTR sequences were ranked in descending order, as presented in Table 17. It was observed that the closest intrinsic disorder profiles relative to CFTR_HUMAN were CFTR_PANTR, CFTR_GORGO, CFTR_NOMLE, CFTR_CHLAE, and CFTR_PONAB (Table 17). In contrast, the most distant profiles with respect to CFTR_HUMAN were CFTR_DANRE, CFTR_SQUAC, CFTR_RAT, CFTR_XENLA, and CFTR_MOUSE (Table 17).

 doi:10.20944/preprints202605.1213.v1

Figure 19. Intrinsic disorder predicted regions profiles across CFTR sequences.

Table 17. Increasing ordering with reference to CFTR_HUMAN based on difference spectra on intrinsic disorder profiles of 42 CFTR sequences.

S/N	CFTR_Sequence	Mean_Difference	S/N	CFTR_Sequence	Mean_Difference	S/N	CFTR_Sequence	Mean_Difference
1	CFTR_HUMAN	0	15	CFTR_AOTNA	0.0199	29	CFTR_CARPS	0.0303
2	CFTR_PANTR	0.0072	16	CFTR_HORSE	0.0207	30	CFTR_OTOGA	0.0312
3	CFTR_GORGO	0.0077	17	CFTR_SAIBB	0.0233	31	CFTR_MUSPF	0.0320
4	CFTR_NOMLE	0.0090	18	CFTR_CALJA	0.0236	32	CFTR_RHIFE	0.0346
5	CFTR_CHLAE	0.0111	19	CFTR_LOXAF	0.0239	33	CFTR_SHEEP	0.0355
6	CFTR_PONAB	0.0111	20	CFTR_CANLF	0.0267	34	CFTR_DIDVI	0.0358
7	CFTR_MACNE	0.0145	21	CFTR_MICMU	0.0273	35	CFTR_ORNAN	0.0376
8	CFTR_MACMU	0.0147	22	CFTR_CAVPO	0.0274	36	CFTR_TRIVU	0.0376
9	CFTR_MACFU	0.0147	23	CFTR_DASNO	0.0280	37	CFTR_ATEAB	0.0406
10	CFTR_MACFA	0.0147	24	CFTR_PIG	0.0292	38	CFTR_MOUSE	0.0452
11	CFTR_PAPAN	0.0149	25	CFTR_MUNMU	0.0295	39	CFTR_XENLA	0.0506
12	CFTR_COLGU	0.0158	26	CFTR_MUNRE	0.0296	40	CFTR_RAT	0.0540
13	CFTR_ATEGE	0.0161	27	CFTR_RABIT	0.0297	41	CFTR_SQUAC	0.0654
14	CFTR_PLEMO	0.0173	28	CFTR_BOVIN	0.0302	42	CFTR_DANRE	0.0763

4.5. Evaluation of the Disorder-Based Functionality of CFTR _HUMAN

Figure 20A represents a functional disorder profile generated for CFTR_HUMAN by the D^2P^2 platform. This protein is expected to contain noticeable levels of disorder, with its regulatory domain (residues 640-840) being predicted as a long intrinsically disordered region (IDR). This domain and other IDRs contain multiple PTMs, suggesting that these disordered regions might act as PTM display sites. Furthermore, there are two molecular recognition features (MoRFs, which are disordered segments that undergo binding-induced folding at interaction with specific partners), indicating the role of IDRs in functionality of this protein. Figure 20B provides additional support to the presence of substantial structural disorder in CFTR_HUMAN showing the AlphaFold-generated 3D model of this protein, which contains sizable segments with low and very confidence scores that would be in disordered form, when not interacting with the partners. Figure 20C represents the CFTR_HUMAN-centered PPI network generated by STRING. This is a physical subnetwork showing only proteins involved in physical interactions. It includes 106 nodes connected by 703 edges. It is characterized by the average node degree of 13.3 and the average local clustering coefficient of 0.79. The network is characterized by the PPI enrichment p-value $< 10^{-16}$, indicating that its proteins have more interactions among themselves than would be expected for a random set of proteins of the same size and degree distribution drawn from the genome. Such an enrichment indicates that the proteins are at least partially biologically connected, as a group.

 doi:10.20944/preprints202605.1213.v1

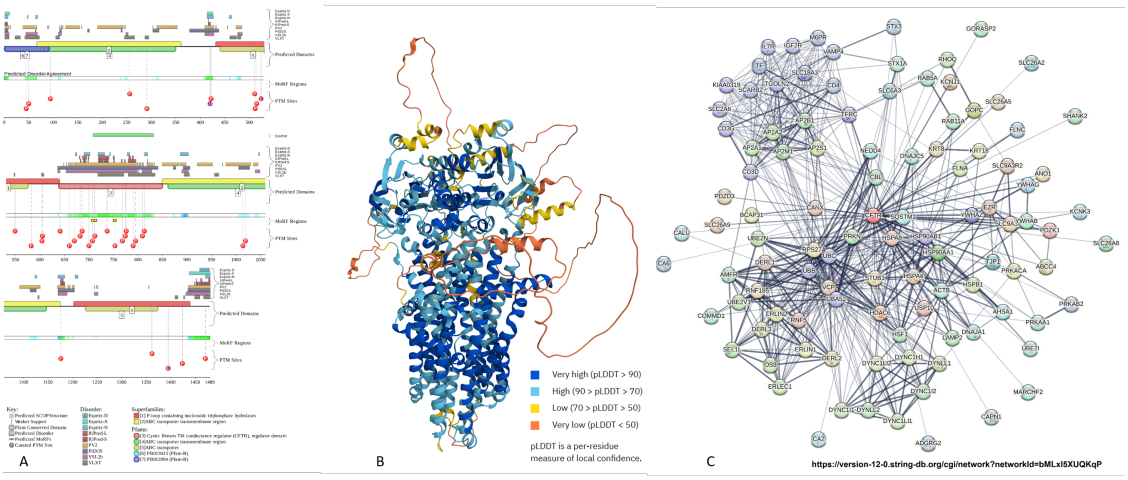

Figure 20. Functional disorder analysis of CFTR_HUMAN. A. Functional disorder profile generated by the D^2P^2 platform. The figure displays disorder predictions from several tools (PONDR® VLXT, PONDR® VSL2b, PrDOS, IU-Pred (both short and long forms), and three variants of Espritz (NMR, DisProt, and X-ray)) as stacked, multi-colored bars. Below these, numbered bars identify specific SCOP (Structural Classifications of Proteins) domains. The consensus disorder is represented by a blue-green-white bar: blue indicates disordered regions within SCOP domains, while green highlights disordered regions outside them. Finally, yellow zigzags mark MoRFs (i.e., disordered regions capable of undergoing folding at interaction with specific partners binding-induced folding regions), and colored circles at the base denote various post-translational modification (PTM) sites. B. 3D structural model generated by AlphaFold. The AlphaFold-generated per-residue confidence scores (p_{LDDT}, that ranges between 0 and 100) are used to color structures. Here, the segments predicted by AlphaFold with very high ($p_{LDDT} > 90$), high ($90 > p_{LDDT} > 70$), low ($70 > p_{LDDT} > 50$), and very low ($p_{LDDT} < 50$) confidence are shown by blue, cyan, yellow, and orange colors correspondingly. C. STRING-generated protein-protein interaction network. This network was generated using the minimum required interaction score of 0.4 (medium confidence). We restricted analysis to physical subnetwork, where the edges indicate that the proteins are part of a physical complex. The line thickness of the edges indicates the strength of the data support. An interactive version of this network as available that the following permanent link: PPI: CFTR_HUMAN.

4.6. Cumulative Proximal and Distal Relationships among CFTR Sequences

The CFTR sequences from diverse species were systematically clustered across multiple features, including amino acid compositions and amino acid homology, hydropathy profiles, normalized longest polar and non-polar stretches, and n-gram distributions. Cumulative clusters were derived, distinguishing proximal clusters (stable across all features) from distal sequences (outliers).

The CFTR sequences from diverse species were primarily grouped based on amino acid frequency distribution and amino acid sequence homology across CFTR sequences (Table 2 and Table 3). These baseline tables provided the first evidence of sequence proximities and highlighted consistent groupings across primates, ungulates, and rodents.

Two broad clusters consistently agreed across all feature-based analyses:

- **Macaca–Colobus–Chlorocebus cluster:** {CFTR_MACFU, CFTR_MACFA, CFTR_MACMU, CFTR_MACNE, CFTR_PAPAN, CFTR_CHLAE, CFTR_COLGU}. These old world monkeys consistently group together in both baseline tables, showing strong internal proximity.
- **Great apes cluster:** {CFTR_HUMAN, CFTR_GORGO, CFTR_PANTR, CFTR_PONAB, CFTR_NOMLE}. This great ape cluster is distinct and stable, with closer proximities confirmed across all features.
- **Muntjac pair:** {CFTR_MUNRE, CFTR_MUNMU}. A stable proximal relation confirmed across all features.

The cumulative analysis confirms that only the primate core cluster, the muntjac pair, and the human–gorilla pair are universally proximal across all features, while other proximities are feature-dependent. CFTR outliers were identified where the longest polar or non-polar stretches deviated substantially from cluster centroids. For example, sequences such as CFTR_HORSE and

 doi:10.20944/preprints202605.1213.v1

CFTR_PIG grouped together phylogenetically but diverged under hydropathy analysis, marking them as distal. Similarly, CFTR_TRIVU and CFTR_DIDVI formed a homology cluster but failed to align with proximal groups in stretch-based analysis.

5. Discussion

In this study, 42 protein sequences of the cystic fibrosis transmembrane conductance regulator across various species were analyzed to elucidate the skeletal frameworks underlying CFTR sequence organization. It was found that across CFTR sequences, leucine stood out as the most common amino acid, while cystine was strikingly rare. This simple imbalance already hints at the protein's structural preferences. Beyond individual frequencies, certain residues tended to rise and fall together. For example, alanine and methionine, as well as histidine and lysine, showed clear positive correlations, suggesting they often appear in tandem. In contrast, pairs like glutamate–glycine and glutamate–phenylalanine moved in opposite directions, reflecting trade-offs in sequence composition. Taken together, these amino acid compositional patterns highlight consistent signatures in CFTR: a strong bias toward leucine, minimal use of cystine, and reproducible co-variation among specific amino acid pairs. Furthermore, the CFTR sequences across species showed a distinct compositional bias: several order-promoting residues (C, Y, H, V, N) were consistently depleted, while disorder-promoting residues (R, Q, S) were enriched. This pattern reflects a tendency toward structural flexibility, in line with the intrinsic disorder hypothesis. Reduced hydrophobic and aromatic content limits stable core formation, while increased polar and charged residues support conformational variability. These results echo earlier reports of disorder-associated sequence signatures and reinforce the view that CFTR carries features typical of intrinsically disordered regions. Shannon entropy for CFTR amino acid frequencies were tightly clustered (4.11–4.15), indicating strong compositional conservation across species. Minor differences, such as slightly higher entropy in CFTR_DANRE and lower values in CFTR_CAVPO and CFTR_XENLA, reflect subtle lineage-specific variation. CFTR_HUMAN showed intermediate entropy, consistent with other primates.

It was observed that phylogenetic analysis based on amino acid frequency distributions resolved seven distinct CFTR clusters, reflecting clear compositional proximities among species. CFTR_HUMAN grouped tightly with other primates, while zebrafish (CFTR_DANRE) emerged as a pronounced outlier, underscoring its divergent amino acid profile. Overall, the clustering highlights conserved compositional signatures across mammals, with only a few sequences showing marked deviations. Also, homology-based phylogenetic analysis of CFTR sequences resolved nine distinct clusters, reflecting clear lineage-specific proximities. CFTR_HUMAN clubbed tightly with other primates, consistent with evolutionary expectations, while rodent sequences formed a separate cluster. Livestock species such as bovine and sheep clustered together, and zebrafish (CFTR_DANRE) remained divergent, reinforcing its distinct compositional profile noted earlier.

Invariant residue analysis across 42 CFTR sequences revealed that 530 positions (37.6%) remained unchanged, underscoring strong evolutionary conservation. Most invariant stretches were short, but two longer invariant stretches 'DADLYLLD' (8 residues) and 'RRQSVL' (6 residues)—stood out. Both map to the cytoplasmic domain of CFTR_HUMAN, highlighting their functional indispensability. Collectively, these invariant stretches serve as molecular fingerprints, defining CFTR's conserved identity across diverse species. Across 42 CFTR sequences, 140 amino acid substitutions were identified and classified by hydrophobicity changes: 38 Polar-to-Nonpolar, 23 Nonpolar-to-Polar, 54 Nonpolar-to-Nonpolar, and 25 Polar-to-Polar. Notably, four substitutions (G458V, G1244E, G1249E, and S1251N) mapped to binding sites within the topological domain of human CFTR, while another four (T665S, F693L, V754M, and E822K) occurred in disordered regions. These findings highlight both the diversity of substitution types and the presence of changes in functionally critical domains.

Analysis of dominant amino acid patterns across CFTR sequences revealed 18 distinct profiles, with 11 unique to individual organisms. Strikingly, leucine consistently dominated five sequence windows (1–100, 101–200, 201–300, 1001–1100, and 1301–1400) across all 42 sequences. Moreover,

 doi:10.20944/preprints202605.1213.v1

leucine appeared as the dominant residue in at least one sequence within every window, except positions 1101–1200, where isoleucine/valine prevailed [65–67]. These results highlight leucine's pervasive influence on CFTR composition, with only limited exceptions [68,69].

Hydropathy-based n-gram analysis across CFTR sequences revealed consistent compositional signatures. Homogeneous non-polar strings (e.g., '00', '000', '0000', '00000', '000000', and '0000000') dominated at every resolution, while alternating polarity motifs (e.g., '01', '010', '0101', '01010', '010101', and '0101010') were consistently suppressed. Intermediate n-grams showed modest variation, but overall distributions remained highly conserved across species. Phylogenetic relationships derived from these profiles resolved reproducible clusters, with several sets of sequences—most notably {CFTR_MACMU, CFTR_MACFA, CFTR_PAPAN, CFTR_MACNE, CFTR_MACFU, CFTR_CHLAE} and {CFTR_MUNRE, CFTR_MUNMU}—exhibiting identical n-gram profiles across multiple resolutions. The cumulative clustering framework further distinguished robust associations persisting across all n-grams from resolution-sensitive proximities detectable only at specific scales. Together, these findings highlight CFTR's preference for hydrophobic continuity, suppression of alternating polarity stretches, and reproducible conservation of sequence clusters across organisms. Across CFTR sequences, the longest polar and nonpolar stretch percentages ranged between 0.53 and 1.01, with most mammalian sequences (e.g., CFTR_HUMAN, CFTR_MOUSE, CFTR_RAT, CFTR_PIG, CFTR_SHEEP, and CFTR_HORSE) clustering around 0.70–0.85, indicating a consistent balance. A few sequences approached 1.0, reflecting unusually extended uninterrupted polar or nonpolar segments. DBSCAN clustering based on these values resolved four distinct groups and one outlier cluster, highlighting atypical hydropathy domain architectures in certain sequences. Identical longest polar and nonpolar stretches were observed in multiple species. Specifically, CFTR_HUMAN, CFTR_NOMLE, CFTR_PONAB, CFTR_GORGO, and CFTR_PANTR shared lengths of 11 and 15 residues, while CFTR_CHLAE, CFTR_MACFU, CFTR_MACFA, CFTR_PAPAN, CFTR_MACNE, and CFTR_MACMU all exhibited identical stretches of 15 residues. These conserved stretches underscore shared structural constraints across primates and certain macaque lineages.

Intrinsic disorder profiling across 42 CFTR sequences revealed a conserved domain-level architecture. The N-terminal transmembrane cluster and both NBDs showed low disorder scores (mean < 0.20), consistent with their stable structural folds. In contrast, the R-domain (residues 650–850) was universally disordered (mean 0.78 ± 0.10), underscoring its role as a phosphorylation-rich regulatory hotspot. Cytoplasmic loops exhibited intermediate disorder with high variance, suggesting lineage-specific modulation, while the C-terminal tail showed moderate disorder, reflecting flexible interaction potential. Together, these results highlight a conserved balance of ordered and disordered regions that supports CFTR channel stability and regulatory adaptability. Proximity analysis relative to CFTR_HUMAN confirmed evolutionary relationships. The closest disorder profiles were observed in CFTR_PANTR, CFTR_GORGO, CFTR_NOMLE, CFTR_CHLAE, and CFTR_PONAB, while the most distant were CFTR_DANRE, CFTR_SQUAC, CFTR_RAT, CFTR_XENLA, and CFTR_MOUSE. These rankings emphasize both the conservation of disorder architecture among primates and the divergence observed in zebrafish, amphibians, and rodents.

Functional disorder profiling of CFTR_HUMAN using the D^2P^2 platform confirmed substantial intrinsic disorder, with the regulatory domain (residues 640–840) predicted as a long IDR enriched in post-translational modifications (PTMs). The presence of molecular recognition features (MoRFs) further supports the role of disordered regions in mediating partner-specific interactions. The AlphaFold model provided complementary evidence, showing extended segments with low confidence scores consistent with disorder when not bound to partners. The CFTR-centered protein–protein interaction (PPI) network generated by STRING revealed 106 nodes and 703 edges, with an average node degree of 13.3 and clustering coefficient of 0.79. The highly significant enrichment ($p < 10^{-16}$) indicates that CFTR and its partners form a biologically connected group rather than a random set. Together, these analyses highlight the functional importance of intrinsic disorder in CFTR, its role as a PTM display site, and its integration into a densely connected interaction network.

 doi:10.20944/preprints202605.1213.v1

The demonstrated proximity of CFTR_HUMAN to other primate sequences is particularly valuable in CF. Pathogenic mutations can be contextualized against these conserved fingerprints, allowing identification of deviations that disrupt disorder-mediated regulation or hydropathy continuity. Quantitative protein analysis—through entropy profiling, invariant motif detection, n-gram distributions, and disorder mapping—thus provides a rigorous lens for deciphering the molecular basis of cystic fibrosis and enabling variant prioritization and therapeutic insight, and thereby advancing rare disease discovery and intervention.

Author Contributions: SSH and KH formulated the problem and designed the theoretical experiments and analyses. SSH, KH, and VNU carried out the experiments and performed the analyses. SSH and VNU drafted the primary manuscript. All authors contributed to reviewing and editing the manuscript. All authors reviewed, checked, and approved the final manuscript.

Funding: The authors declare no funding was received for the present study.

Acknowledgments: The authors express their sincere gratitude to the scientists at the Centre for Health Informatics (CHI), National Health Portal (NHP), and Ministry of Health and Family Welfare (MoHFW), Government of India, for their efforts in developing the rare diseases directory and providing relevant information.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Danese, E.; Lippi, G. Rare diseases: the paradox of an emerging challenge. *Annals of Translational Medicine* **2018**, *6*, 329.
2. Stoller, J.K. The challenge of rare diseases. *Chest* **2018**, *153*, 1309–1314.
3. Sivasubbu, S.; Scaria, V. Genomics of rare genetic diseases—experiences from India. *Human genomics* **2019**, *13*, 52.
4. Richter, T.; Nestler-Parr, S.; Babela, R.; Khan, Z.M.; Tesoro, T.; Molsen, E.; Hughes, D.A. Rare disease terminology and definitions—a systematic global review: report of the ISPOR rare disease special interest group. *Value in health* **2015**, *18*, 906–914.
5. Lu, Y.; Han, J. The definition of rare disease in China and its prospects. *Intractable & Rare Diseases Research* **2022**, *11*, 29–30.
6. Haendel, M.; Vasilevsky, N.; Unni, D.; Bologa, C.; Harris, N.; Rehm, H.; Hamosh, A.; Baynam, G.; Groza, T.; McMurry, J.; et al. How many rare diseases are there? *Nature reviews drug discovery* **2020**, *19*, 77–78.
7. Ferreira, C.R. The burden of rare diseases. *American journal of medical genetics Part A* **2019**, *179*, 885–892.
8. Chaudhary, A.; Kumar, V. Rare diseases: a comprehensive literature review and future directions. *Journal of Rare Diseases* **2025**, *4*, 33.
9. Lee, C.E.; Singleton, K.S.; Wallin, M.; Faundez, V. Rare genetic diseases: nature's experiments on human development. *IScience* **2020**, *23*.
10. Serrano, J.G.; O'Leary, M.; VanNoy, G.E.; Mangilog, B.E.; Holm, I.A.; Fraiman, Y.S.; Rehm, H.L.; O'Donnell-Luria, A.; Wojcik, M.H. Advancing understanding of inequities in rare disease genomics. *Clinical therapeutics* **2023**, *45*, 745–753.
11. Austin, C.P.; Cutillo, C.M.; Lau, L.P.; Jonker, A.H.; Rath, A.; Julkowska, D.; Thomson, D.; Terry, S.F.; de Montleau, B.; Ardigò, D.; et al. Future of rare diseases research 2017–2027: an IRDiRC perspective. *Clinical and translational science* **2017**, *11*, 21.
12. Qureshi, M.D.A.; Ahmed, W.; Mehdi, S. Bioinformatics and Genomics in Rare Disease Diagnosis: Leveraging AI and Next-Generation Sequencing for Identifying Novel Genetic Disorders. *Biosciences Research Reviews* **2024**, *1*, 46–59.
13. Ahmed, F. Genomics and bioinformatics: integrating data for better genetic insights. *Frontiers in Biotechnology and Genetics* **2024**, *1*, 126–146.
14. Green, D.M.; Polasky, J.; Weatherly, M.; Stalker, H.; Blanchard, C.; Kushner, C.; Couluris, M.; Ryland, P.; Sunitha, I.; Fong, J.; et al. Next-Generation Sequencing for Cystic Fibrosis: Florida Newborn Screening Experience. *International Journal of Neonatal Screening* **2025**, *11*, 94.
15. Wojcik, M.H.; Lemire, G.; Berger, E.; Zaki, M.S.; Wissmann, M.; Win, W.; White, S.M.; Weisburd, B.; Wieczorek, D.; Waddell, L.B.; et al. Genome sequencing for diagnosing rare diseases. *New England Journal of Medicine* **2024**, *390*, 1985–1997.

 doi:10.20944/preprints202605.1213.v1

16. Umlai, U.K.I.; Bangarusamy, D.K.; Estivill, X.; Jithesh, P.V. Genome sequencing data analysis for rare disease gene discovery. *Briefings in Bioinformatics* **2022**, *23*, bbab363.
17. Knowles, M.R.; Durie, P.R. What is cystic fibrosis?, 2002.
18. Bell, S.C.; Mall, M.A.; Gutierrez, H.; Macek, M.; Madge, S.; Davies, J.C.; Burgel, P.R.; Tullis, E.; Castaños, C.; Castellani, C.; et al. The future of cystic fibrosis care: a global perspective. *The Lancet Respiratory Medicine* **2020**, *8*, 65–124.
19. Ooi, C.Y.; Durie, P.R. Cystic fibrosis transmembrane conductance regulator (CFTR) gene mutations in pancreatitis. *Journal of Cystic Fibrosis* **2012**, *11*, 355–362.
20. Miller, P.W.; Hamosh, A.; Macek Jr, M.; Greenberger, P.A.; MacLean, J.; Walden, S.M.; Slavin, R.G.; Cutting, G.R. Cystic fibrosis transmembrane conductance regulator (CFTR) gene mutations in allergic bronchopulmonary aspergillosis. *American journal of human genetics* **1996**, *59*, 45.
21. Ramananda, Y.; Naren, A.P.; Arora, K. Functional consequences of CFTR interactions in cystic fibrosis. *International Journal of Molecular Sciences* **2024**, *25*, 3384.
22. Habibullah, M.M. The role of CFTR channel in female infertility. *Human Fertility* **2023**, *26*, 1228–1237.
23. Lukasiak, A.; Zajac, M. The distribution and role of the CFTR protein in the intracellular compartments. *Membranes* **2021**, *11*, 804.
24. Dasgupta, S.; Datta, D.; Pal, S.; Ghosh, K.; Sett, S. When 25% Feels Like 100%: Confronting Recurrent Cystic Fibrosis Risk Across Consecutive Pregnancies. *Cureus* **2025**, *17*.
25. Chowdhury, M.R.; Kumari, I.; Jat, K.R.; Dhochak, N.; Lodha, R.; Varkki, S.; Kumar, P.; Goyal, J.P.; Bhat, J.I.; Kabra, S.; et al. Profile of cystic fibrosis transmembrane conductance regulator (CFTR) gene variants across India and their variability in different geographic regions. *Gene* **2026**, *976*, 149870.
26. Fuhrer, M.; Zampoli, M.; Abriel, H. Diagnosing cystic fibrosis in low-and middle-income countries: challenges and strategies. *Orphanet Journal of Rare Diseases* **2024**, *19*, 482.
27. Kronn, D.F.; Day-Salvatore, D.; Hwu, W.L.; Jones, S.A.; Nakamura, K.; Okuyama, T.; Swoboda, K.J.; Kishnani, P.S.; Group, P.D.N.S.W. Management of confirmed newborn-screened patients with Pompe disease across the disease spectrum. *Pediatrics* **2017**, *140*, S24–S45.
28. Mandal, A.; Kabra, S.K.; Lodha, R. Cystic fibrosis in India: past, present and future. *J Pulm Med Respir Res* **2015**, *1*, 1–8.
29. Singh, H.; Jani, C.; Marshall, D.C.; Franco, R.; Bhatt, P.; Podder, S.; Shalhoub, J.; Kurman, J.S.; Nanchal, R.; Uluer, A.Z.; et al. Cystic fibrosis-related mortality in the United States from 1999 to 2020: an observational analysis of time trends and disparities. *Scientific reports* **2023**, *13*, 15030.
30. Varkki, S.D.; Aaron, R.; Chapla, A.; Danda, S.; Medhi, P.; Rani, N.J.; Paul, G.R. CFTR mutations and phenotypic correlations in people with cystic fibrosis: a retrospective study from a single centre in south India. *The Lancet Regional Health-Southeast Asia* **2024**, *27*.
31. Sun, Y.; Du, J.; Gong, L.; Wang, J.; Zhang, J.; Xu, Y.; Wang, L.; Zhang, Z.; Cheng, H. Current Status and the Need for CFTR Modulator Therapy in Cystic Fibrosis Patients in Mainland China: A Case Report and Literature Review. *Pediatric Pulmonology* **2026**, *61*, e71590.
32. Ahmed, S.; Cheok, G.; Goh, A.E.; Han, A.; Hong, S.; Indawati, W.; Kabir, A.L.; Kabra, S.; Kamalaporn, H.; Kim, H.Y.; et al. Cystic fibrosis in Asia. *Pediatric Respirology and Critical Care Medicine* **2020**, *4*, 8–12.
33. Bisht, A.; Hasija, Y. Revolutionizing neonatal health: India's journey from assays to advanced genetics. *Journal of Applied Genetics* **2025**, pp. 1–21.
34. Adachi, T.; El-Hattab, A.W.; Jain, R.; Nogales Crespo, K.A.; Quirland Lazo, C.I.; Scarpa, M.; Summar, M.; Wattanasirichaigoon, D. Enhancing equitable access to rare disease diagnosis and treatment around the world: a review of evidence, policies, and challenges. *International journal of environmental research and public health* **2023**, *20*, 4732.
35. Groft, S.C.; Posada, M.; Taruscio, D. Progress, challenges and global approaches to rare diseases. *Acta paediatrica* **2021**, *110*, 2711–2716.
36. Sanders, M.; Lawlor, J.M.; Li, X.; Schuen, J.N.; Millard, S.L.; Zhang, X.; Buck, L.; Grysko, B.; Uhl, K.L.; Hinds, D.; et al. Genomic, transcriptomic, and protein landscape profile of CFTR and cystic fibrosis. *Human genetics* **2021**, *140*, 423–439.
37. Cutting, G.R. Cystic fibrosis genetics: from molecular understanding to clinical application. *Nature Reviews Genetics* **2015**, *16*, 45–56.
38. Hanssens, L.S.; Duchateau, J.; Casimir, G.J. CFTR protein: not just a chloride channel? *Cells* **2021**, *10*, 2844.
39. Poulter, C.; Bhatt, J. Epidemiology, genetics, pathophysiology and prognosis of CF. *ERS Handbook of Paediatric Respiratory Medicine; Eber, E., Midulla, F., Eds* **2021**, pp. 435–445.

 doi:10.20944/preprints202605.1213.v1

40. Mateu, E.; Calafell, F.; Lao, O.; Bonné-Tamir, B.; Kidd, J.R.; Pakstis, A.; Kidd, K.K.; Bertranpetit, J. Worldwide genetic analysis of the CFTR region. *The American Journal of Human Genetics* **2001**, *68*, 103–117.
41. Ellsworth, R.E.; Jamison, D.C.; Touchman, J.W.; Chissoe, S.L.; Braden Maduro, V.V.; Bouffard, G.G.; Dietrich, N.L.; Beckstrom-Sternberg, S.M.; Iyer, L.M.; Weintraub, L.A.; et al. Comparative genomic sequence analysis of the human and mouse cystic fibrosis transmembrane conductance regulator genes. *Proceedings of the National Academy of Sciences* **2000**, *97*, 1172–1177.
42. Pereira, S.V.N.; Ribeiro, J.D.; Ribeiro, A.F.; Bertuzzo, C.S.; Marson, F.A.L. Novel, rare and common pathogenic variants in the CFTR gene screened by high-throughput sequencing technology and predicted by in silico tools. *Scientific Reports* **2019**, *9*, 6234.
43. Vernone, A.; Berchialla, P.; Pescarmona, G. Human protein cluster analysis using amino acid frequencies. *PloS one* **2013**, *8*, e60220.
44. Hassan, S.S.; Nawn, D.; Mukherjee, N.; Goswami, A.; Uversky, V.N. A mathematical genomics perspective on the moonlighting role of glyceraldehyde-3-phosphate dehydrogenase (GAPDH). *International Journal of Biological Macromolecules* **2025**, p. 148045.
45. Karlin, S.; Bucher, P. Correlation analysis of amino acid usage in protein classes. *Proceedings of the National Academy of sciences* **1992**, *89*, 12165–12169.
46. Vacic, V.; Uversky, V.N.; Dunker, A.K.; Lonardi, S. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC bioinformatics* **2007**, *8*, 1–7.
47. Mittal, A.; Jayaram, B. Backbones of folded proteins reveal novel invariant amino acid neighborhoods. *Journal of Biomolecular Structure and Dynamics* **2011**, *28*, 443–454.
48. Quaglia, F.; Mészáros, B.; Salladini, E.; Hatos, A.; Pancsa, R.; Chemes, L.B.; Pajkos, M.; Lazar, T.; Peña-Díaz, S.; Santos, J.; et al. DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Research* **2022**, *50*, D480–D487.
49. Hatos, A.; Hajdu-Soltész, B.; Monzon, A.M.; Palopoli, N.; Álvarez, L.; Aykac-Fas, B.; Bassot, C.; Benítez, G.I.; Bevilacqua, M.; Chasapi, A.; et al. DisProt: intrinsic protein disorder annotation in 2020. *Nucleic acids research* **2020**, *48*, D269–D276.
50. Strait, B.J.; Dewey, T.G. The Shannon information entropy of protein sequences. *Biophysical journal* **1996**, *71*, 148–155.
51. Sievers, F.; Barton, G.J.; Higgins, D.G. Multiple sequence alignments. *Bioinformatics* **2020**, *227*, 227–250.
52. Schneider, T.R. A genetic algorithm for the identification of conformationally invariant regions in protein molecules. *Biological Crystallography* **2002**, *58*, 195–208.
53. Lotthammer, J.M.; Hernández-García, J.; Griffith, D.; Weijers, D.; Holehouse, A.S.; Emenecker, R.J. Metapredict enables accurate disorder prediction across the Tree of Life. *bioRxiv* **2024**, pp. 2024–11.
54. Emenecker, R.J.; Griffith, D.; Holehouse, A.S. Metapredict: a fast, accurate, and easy-to-use predictor of consensus disorder and structure. *Biophysical journal* **2021**, *120*, 4312–4319.
55. Lotthammer, J.M.; Ginell, G.M.; Griffith, D.; Emenecker, R.; Holehouse, A.S. Direct prediction of intrinsically disordered protein conformational properties from sequence. *Biophysical Journal* **2024**, *123*, 43a.
56. Nawn, D.; Hassan, S.S.; Redwan, E.M.; Bhattacharya, T.; Basu, P.; Lundstrom, K.; Uversky, V.N. Unveiling the genetic tapestry: Rare disease genomics of spinal muscular atrophy and phenylketonuria proteins. *International Journal of Biological Macromolecules* **2024**, *269*, 131960.
57. Oates, M.E.; Romero, P.; Ishida, T.; Ghalwash, M.; Mizianty, M.J.; Xue, B.; Dosztányi, Z.; Uversky, V.N.; Obradovic, Z.; Kurgan, L.; et al. D2P2: database of disordered protein predictions. *Nucleic acids research* **2012**, *41*, D508–D516.
58. Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research* **2022**, *50*, D439–D444.
59. Szklarczyk, D.; Franceschini, A.; Kuhn, M.; Simonovic, M.; Roth, A.; Minguez, P.; Doerks, T.; Stark, M.; Muller, J.; Bork, P.; et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research* **2010**, *39*, D561–D568.
60. Lobanov, M.Y.; Likhachev, I.V.; Galzitskaya, O.V. Disordered residues and patterns in the protein data bank. *Molecules* **2020**, *25*, 1522.
61. Nawn, D.; Hassan, S.S.; Hromić-Jahjefendić, A.; Bhattacharya, T.; Basu, P.; Redwan, E.M.; Barh, D.; Andrade, B.S.; Aljabali, A.A.; Serrano-Aroca, Á.; et al. Molecular genomic insights into melanoma associated proteins prame and bap1. *Journal of Biomolecular Structure and Dynamics* **2025**, *43*, 9719–9749.

 doi:10.20944/preprints202605.1213.v1

62. Fishbain, S.; Inobe, T.; Israeli, E.; Chavali, S.; Yu, H.; Kago, G.; Babu, M.M.; Matouschek, A. Sequence composition of disordered regions fine-tunes protein half-life. *Nature structural & molecular biology* **2015**, *22*, 214–221.
63. Bella, J.; Hindle, K.; McEwan, P.; Lovell, S. The leucine-rich repeat structure. *Cellular and Molecular Life Sciences* **2008**, *65*, 2307–2333.
64. Kobe, B.; Kajava, A.V. The leucine-rich repeat as a protein recognition motif. *Current opinion in structural biology* **2001**, *11*, 725–732.
65. Sheppard, D.N.; Welsh, M.J. Structure and function of the CFTR chloride channel. *Physiological reviews* **1999**, *79*, S23–S45.
66. Kobe, B.; Deisenhofer, J. Proteins with leucine-rich repeats. *Current opinion in structural biology* **1995**, *5*, 409–416.
67. Negoda, A.; El Hiani, Y.; Cowley, E.A.; Linsdell, P. Contribution of a leucine residue in the first transmembrane segment to the selectivity filter region in the CFTR chloride channel. *Biochimica et Biophysica Acta (BBA)-Biomembranes* **2017**, *1859*, 1049–1058.
68. Matsushima, N.; Takatsuka, S.; Miyashita, H.; Kretsinger, R.H. Leucine rich repeat proteins: sequences, mutations, structures and diseases. *Protein and Peptide Letters* **2019**, *26*, 108–131.
69. Prota, L.; Santoro, A.; Bifulco, M.; Aquino, R.P.; Mencherini, T.; Russo, P. Leucine enhances aerosol performance of naringin dry powder and its activity on cystic fibrosis airway epithelial cells. *International journal of pharmaceutics* **2011**, *412*, 8–19.