Evolutionary and Structural Conservation of Pulmonary Surfactant Proteins: Implications for Lung Resilience

Dhrishit Khandhar[1], **Nirupma Singh**[2]

Received July 30, 2025
Accepted February 28, 2026
Electronic access May 31, 2026

Lung fibrosis is a persistent scarring of the lung epithelium that worsens over time. The latest research underscores surfactant proteins (SPs) as having a dual effect on lung health. The pulmonary-associated SPs maintain alveolar cohesion and regulate immunological responses. The complex between these proteins and the lipids found in the lung helps decrease surface tension and prevents the alveolar system from collapsing. This project aims to investigate the evolutionary conservation, structural variation, and mutation hotspots within pulmonary-associated SPs across multiple vertebrate species using bioinformatics tools. Multiple sequence alignment (MSA) was performed for mRNA and protein sequences using ClustalOmega, and the results were visualised with Jalview to identify conserved and variable regions across the species. Conservation was further analysed using the construction of phenetic clustering trees to trace evolutionary divergence. Known pathogenic variants (e.g., SFTPA1 p.Arg219Trp, SFTPC p.Ile73Thr) from ClinVar and UniProt were mapped onto the aligned sequences to assess their structural and functional significance. For the protein sequences, structural modelling using PDB representations of SPs and domain annotation in PyMOL was carried out to identify several disease-associated mutations that occur in highly conserved and functionally important domains. The changes in the structures of conserved regions across different species were analysed. The study concludes that disease-linked mutations frequently coincide with evolutionarily conserved regions, supporting their pathogenic role. This evolutionary and structural framework offers insights into how surfactant protein dysfunction contributes to fibrotic lung diseases, paving the way for more targeted genetic diagnostics and therapeutic research.

Introduction

Lung fibrosis is a persistent scarring of the lung epithelium that causes the respiratory structures to become distorted and disordered within the thoracic cavity[1]. Due to the disfiguration and thickening of the lung tissue, the bloodstream is impaired, and the diffusion of carbon dioxide and oxygen exchange is hindered. This leads to intense respiratory failure, causing shortness of breath and limitations in activity[2]. Pulmonary fibrosis may be triggered by previous, unresolved cases of lung inflammation, where disease-causing fibrotic tissue gradually accumulates within the lung over time, weakening the lung's flexibility and strength[3]. However, other forerunners can also trigger fibrosis, including autoimmune diseases and various types of inflammation[4]. According to a recent study conducted between 2009 and 2020, prevalent and existing cases of idiopathic pulmonary fibrosis have increased due to an ageing population[5]. The lungs remain inflamed due to the disregard of disease in several areas of the world, where interstitial disease continues to aggravate[6]. In countries where antifibrotic treatments are inaccessible or unaffordable, pulmonary disease becomes fatal[7]. However, it remains a relatively rare disease, affected by factors such as age, sex, and other specific conditions[8].

The pulmonary surfactant proteins SP-A, SP-B, SP-C, and SP-D play a vital role in the lung's defence mechanism against severe inflammation[9]. These proteins serve two purposes: they maintain alveolar cohesion and regulate immunological responses. Together, these proteins and surfactant lipids form a complex system that lowers surface tension at the lungs' air-liquid interface, preventing alveolar collapse during respiration[10]. In addition to their biophysical functions, SPs, which bind to pathogens and regulate inflammatory responses, are crucial for innate immunity[11]. Given that these proteins predate the development of vertebrate lungs, their evolutionary history suggests that their primary role may have been in innate protective mechanisms before they were adapted for respiratory functions[12].

Recent studies in the literature highlight that SPs have two sides to their impact on lung health. SP-A inhibits TGF-β1-driven fibrotic pathways, whereas SP-D reduces fibrosis by regulating macrophage activity[13]. On the other hand, SFTPC and SFTPA2 mutations cause proteostasis to be upset, which

[1] *Jamnabai Narsee International School*
[2] *Department of Biotechnology, Faculty of Technology, University of Delhi, Delhi, India*

leads to endoplasmic reticulum stress and alveolar epithelial cell death, two major causes of familial PF[14]. These results are consistent with biomarker studies that demonstrate a correlation between increasing fibrosis and higher pro-SFTPB levels in serum extracellular vesicles[15]. However, there is a significant knowledge gap in the evolutionary context because the majority of current research focuses on human and murine models. There is currently no thorough investigation comparing surfactant protein sequences across vertebrates to trace evolutionary adaptations against fibrosis, even though SP genes like SFTPB appeared in early vertebrates and regulatory changes, rather than additional genes, drove lung evolution[16]. Methodological constraints are the cause of this information gap.

The majority of phylogenetic studies only examine single taxa rather than combining genomic, proteomic, and functional data, and traditional multiple sequence alignment techniques have trouble performing deep evolutionary comparisons of hydrophobic proteins like SP-B/SP-C[17]. Many studies focus solely on either evolutionary phylogeny or clinical genetics. A unified approach that integrates these disciplines is required to understand how evolutionary constraints shape disease susceptibility in human surfactant proteins[18]. While previous studies have examined the evolutionary relationships of surfactant proteins, there is a lack of integration between evolutionary conservation and the growing catalog of human disease-associated variants[19]. A combined evolutionary clinical framework can reveal whether pathogenic mutations disproportionately affect evolutionarily conserved residues, thereby directly linking molecular evolution to clinical outcomes[20]. Therefore, it is still unknown if the antifibrotic qualities of SPs developed when lungs adapted to terrestrial settings or if contemporary vulnerability is a result of lost ancestral defences[21]. To answer these problems, new methods that combine ancestral protein reconstruction, cross-species sequence alignment, and functional tests are needed. This multifaceted approach is currently not present in the literature. Thus, in this study, we aim to investigate the evolutionary conservation, structural variation, and mutation hotspots within pulmonary-associated SPs across multiple vertebrate species using bioinformatics tools.

Methodology

Data collection

The nucleotide and amino acid sequences for surfactant proteins A, B, C, and D were collected for five representative vertebrate species: the wild boar (*Sus scrofa*), the house mouse (Mus musculus), the brown rat (*Rattus norvegicus*), cattle (*Bos taurus*) and humans (*Homo sapiens*). Species were selected to represent three major mammalian clades: Primates (*Homo sapiens*), Rodents (*Mus musculus, Rattus norvegicus*), and Artiodactyla (*Bos taurus, Sus scrofa*). This selection allows comparison across distinct mammalian lineages while maintaining the availability of well-annotated genomic and proteomic data. The sequences were retrieved from the NCBI GenBank database by searching for each protein's gene name in combination with the respective species[22]. Only complete coding sequences with verified annotations were included to ensure data quality and comparability. Only sequences annotated as 'complete CDS' in the GenBank record were included in this study. For each candidate sequence, the FEATURES section of the GenBank file was manually inspected to confirm the presence of a full-length coding region without truncations or internal stop codons. Verified coding sequences corresponding to the genes SFTPA1, SFTPA2, SFTPB, SFTPC, and SFTPD were retained for downstream analysis. The NCBI Protein database was also used to retrieve protein sequences for each surfactant protein in each species. Lastly, the protein structures were retrieved from both the PDB and AlphaFold databases to obtain the most comprehensive set of structures available online.

Multiple sequence alignment

To compare the evolutionary conservation and divergence of SPs among the selected species, multiple sequence alignment was performed using Clustal Omega software with gap opening and extension penalties set automatically by the algorithm, BLOSUM62 substitution matrix for protein alignments, and iterative refinement enabled[23]. Only coding sequences (CDS) annotated as 'complete CDS' in GenBank were included. Non-coding regions and untranslated regions were excluded prior to alignment. Separate alignments were generated for each protein family (SP-A, SP-B, SP-C, and SP-D), allowing for the identification of conserved and variable regions within each group. SFTPA1 and SFTPA2 were analyzed as distinct genes throughout all sequence alignments and evolutionary analyses. This was achieved by compiling each surfactant protein into a single file to prepare it for the Clustal Omega analysis. Then, the 'RNA' option was selected, along with 'ClustalW with character counts' for the output format. The results view link was then pasted into the Jalview software, where the sequence was colour-coded based on the conserved areas[24]. This process was repeated with the protein sequences as well: the 'Protein' option was selected on the Clustal Omega website to receive data, and the results view link was pasted into the Jalview software, where the sequence was colour-coded by the 'By conservation' setting and the 'Clustal' option, using Zappo coloring for protein physicochemical grouping. A conserved position was defined as one exhibiting $\geq 80\%$ amino-acid identity across aligned species. Percent conservation was calculated as the fraction of

conserved residues over the total aligned length.

Phylogenetic tree construction

To facilitate the interpretation of sequence alignments and highlight regions of conservation and variability among the surfactant proteins, the aligned sequences were visualised using Jalview software. Jalview was utilised to generate colour-coded alignment maps, enabling clear identification of the average species distances based on surfactant protein sequences. This was done by selecting all the species on the software, selecting the 'Calculate' option and choosing the 'Calculate Tree, PCA or PaSiMap…' option. To produce phylogenetic trees with evolutionary distances, the 'Average Distance' option was selected under the 'Tree' option. The trees generated using Jalview's Average Distance method represent similarity-based clustering rather than model-based phylogenetic reconstruction. These analyses are therefore interpreted as phenetic relationships rather than true evolutionary histories. Similarity-based trees were generated using average distance clustering in Jalview. Trees were midpoint-rooted, and branch lengths represent relative sequence divergence. Arbitrary distance labels were removed to improve visual clarity.

Protein Structure alignment

Following the construction of the phylogenetic trees, the structures of the pulmonary-associated surfactant proteins were obtained from the PDB database[25]. Wild-type structures were identified and downloaded in the Legacy PDB format to ensure that the natural protein in each organism was being studied. These structures were then visualised using the PyMOL application. The solvents were selected and removed to enhance the visualisation. Using the 'Align' function, the SFTPD proteins of each vertebrate in the investigation were compared to that of *Homo sapiens*. The RMSD score was evaluated for the alignment of each surfactant protein type. Structural alignment was performed using the carbohydrate recognition domain (CRD) of SP-D, which is the most functionally conserved domain. Human SP-D was used as a reference structure. Experimentally resolved structures (PDB) were prioritized, and only structures with resolution < 2.5 Å were used. RMSD was calculated over aligned Cα atoms.

Disease and variant mapping

To identify potential mutations and diseases in the surfactant proteins of *Homo sapiens*, the UniProt database was utilised. After each surfactant protein (SFTPA1-D) was found and its UniProt ID was registered, the 'Disease & Variants' section was located, and the variant IDs were noted. The clinical significance of each variant was also analysed using data about the position in the amino acid sequence where the mutation occurred, allowing for the comparison with the results of the previous tests. A general evaluation of the commonly occurring diseases associated with each variant was made.

Results

Multiple Sequence Alignment of RNA Sequences

The multiple sequence alignment of the SFTPA RNA sequences spanned 7,020 nucleotide positions. The alignment of SFTPA RNA sequences for all five species is shown in Figure 1(a). Conserved regions, where all five species aligned, were identified at positions 5,900–6,310 and 4,030–5,710. These conserved segments indicate functional stability across species, suggesting that mutations in these regions could disrupt protein function and contribute to the development of fibrosis. For SFTPB, the MSA covered 14,140 nucleotide positions. The alignment of SFTPB RNA sequences for all five species is shown in Figure 1(b). The longest variable stretches occurred at positions 6,550–14,140 and 1–2,150. Conserved sequences were observed at 2,150–2,390, 2,940–3,020, 3,380–3,510, 4,360–4,540, 4,770–5,060, 5,150–5,260, and 5,800–6,550. The most significant conserved region was 2,220–2,390, representing the longest continuous alignment across all five species. The SFTPC alignment spanned 7,760 positions. The alignment of SFTPC RNA sequences for all five species is shown in Figure 1(c). The longest variable stretch extended from position 1 to 5,520. Conserved sequences were detected at 5,520–5,750, 6,090–6,220, 6,450–6,560, 6,910–7,070, and 7,280–7,500. The most prominent conserved regions were 5,520–5,750 and 6,910–7,070, both of which exhibited the longest continuous alignment. While SFTPC showed lower conservation at the mRNA level due to synonymous nucleotide variation, its protein sequence exhibited high conservation, highlighting translational buffering of functional domains. SFTPD required 1,500 positions for alignment. The alignment of SFTPD RNA sequences for all five species is shown in Figure 1(d). A conserved block spanned positions 70–1,370, while the longest variable stretch was 1,370–1,500. The optimal conserved region was 90–1,140, representing the longest uninterrupted alignment.

Hence, through the analysis of the Jalview visualisations, it can be concluded that SFTPD was the most conserved during the evolution of all five vertebrate species (with 87% of the entire sequence conserved), and was most likely formed via purifying selection. On the other hand, SFTPC had the least conserved length compared to the whole sequence, at around 11%. Therefore, a mutation in SFTPD would most likely lead to the most severe consequences, possibly resulting in idiopathic pulmonary fibrosis.

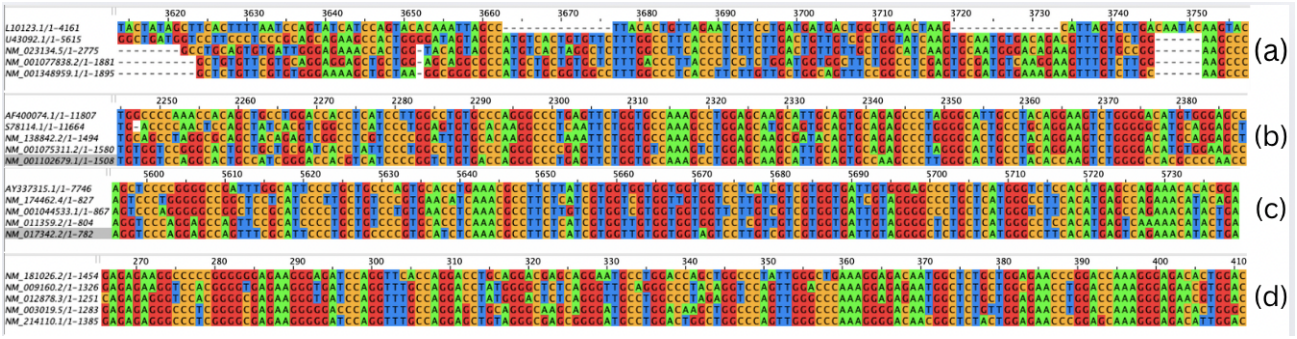

Fig. 1 Multiple sequence alignment of surfactant protein mRNA coding sequences. (a) SFTPA1 alignment showing conserved regions (blue blocks) and variable regions (multicolored) across species: Hs, *Homo sapiens*; Mm, *Mus musculus*; Rn, *Rattus norvegicus*; Bt, *Bos taurus*; Ss, *Sus scrofa*. Conserved regions are predominantly located within the coding region corresponding to functionally constrained domains.

Multiple Sequence Alignment of protein sequences

The multiple sequence alignment of the SFTPA protein sequences spanned 260 amino acid positions. The alignment of SFTPA protein sequences for all five species is shown in Figure 2(a). The longest variable stretch was from the 1st to the 140th position, after which the sequence remained predominantly conserved across all species. Furthermore, the MSA results for the SFTPB amino acid sequences showed that the longest variable stretch was between the 155th and the 170th positions. The alignment of SFTPB protein sequences for all five species is shown in Figure 2(b). The rest of the 380 amino acid-long alignments remained mainly conserved. The conservation of the SFTPC proteins for all species was also high, as there were no long stretches of variable alignment. The alignment of SFTPC protein sequences for all five species is shown in Figure 3(c). The alignment spanned 190 amino acids. The MSA results for the mRNA sequences of SFTPD remain consistent with those of the protein sequences, as SFTPD remains one of the most highly conserved proteins across all species. The alignment of SFTPD protein sequences for all five species is shown in Figure 4(d). Out of the 390 amino acid long alignment, there are no significant patches of variable alignment. Out of each protein, SFTPD had the highest percentage conservation, while SFTPA had the lowest percentage conservation since it had the highest area of variable alignments. SFTPC had the highest area of conserved sequences, however.

Phenetic clustering trees

In the phenetic clustering tree of protein sequences of SP-A, *Bos taurus* [AAI23547.1] and *Sus scrofa* [ACF22969.1] had the most closely related proteins, being the only cluster in the tree. They had a distance of 36.5 from the closest branch point. However, the protein of *Mus musculus* [NP_075623.2] was also found to be related to these two species. *Rattus norvegicus* [EDL90870.1] showed the highest overall divergence in this similarity-based cluster analysis, having a distance of 455.625 from the closest branch point. The protein most closely related to the one in *Homo sapiens* [KAI4076605.1] was *Mus musculus*, with a distance of only 58.5 from each other. In the phenetic clustering tree of protein sequences of SP-B, there was one cluster between *Rattus norvegicus* and *Mus musculus*, and the second cluster was between *Bos taurus* and *Sus scrofa* remains. *Homo sapiens* showed the highest overall divergence in this similarity-based cluster analysis, from the rest of the species, but was the closest to *Bos taurus*, with a distance of 108. In the phenetic clustering tree of protein sequences of SP-C, there were two clusters present: one between *Mus musculus* and *Rattus norvegicus*, and the other between *Bos taurus* and *Sus scrofa*. The protein in *Homo sapiens* was the most diverged from that observed in the other species. The protein in *Sus scrofa* was the most closely related to the one in *Homo sapiens*, with a distance of 24.75. In the phenetic clustering tree of protein sequences of SP-D, *Sus scrofa* and *Bos taurus* had closely related protein sequences, as they shared a cluster. The cluster between *Rattus norvegicus* and *Mus musculus* was also seen in this tree. Hence, *Homo sapiens* showed the highest overall divergence in this similarity-based cluster analysis. The protein in *Rattus norvegicus* was the most closely related to the one in *Homo sapiens*, with a distance of 159.75 in the SP-D phenetic clustering tree.

3D Structure Alignment of protein structures

The results of the protein structure alignments were comparable to the multiple sequence alignments of the mRNA and protein sequences. Due to the unavailability of surfactant proteins A, B, and C for all the species, only the alignment of SFTPD was investigated for *Homo sapiens* and *Sus scrofa*. Furthermore, only the A-chain of the SFTPD protein for *Sus scrofa* was found. Regardless, the alignment of these two proteins

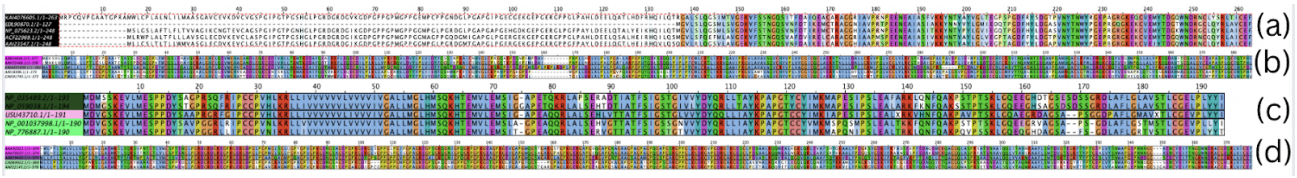

Fig. 2 Jalview protein sequence alignments. This figure displays a collection of multiple sequence alignments for each surfactant protein (SP-A, SP-B, SP-C, and SP-D) across all species (Hs, *Homo sapiens*; Mm, *Mus musculus*; Rn, *Rattus norvegicus*; Bt, *Bos taurus*; Ss, *Sus scrofa*), visualised using Jalview software. They are colour-coded according to the conserved areas. The trend for the mRNA sequence alignments is followed here: SP-A is the least evolutionarily conserved protein, having predominantly variable regions. Consistent with the high sequence conservation, the variant analysis revealed SP-D as one of the most conserved proteins through evolution.

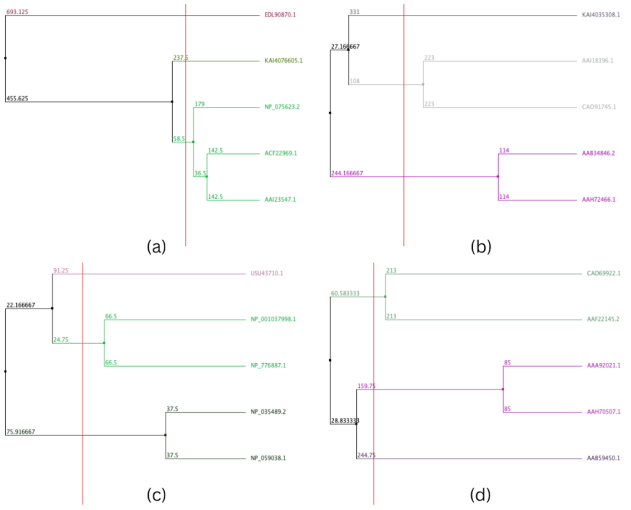

Fig. 3 Phenetic clustering trees for surfactant proteins visualised in Jalview. This figure shows the phenetic clustering trees for each surfactant protein derived from the protein sequence visualisations. (a). This shows the phenetic clustering tree of protein sequences for SP-A. (b) This shows the phenetic clustering tree of protein sequences for SP-B. (c) This shows the phenetic clustering tree of protein sequences for SP-C. (d) This shows the phenetic clustering tree of protein sequences for SP-D. The scale bar indicates relative sequence divergence.

corresponded to the results on the phenetic clustering tree produced on Jalview. The visualisation is seen in Figure 4. The results of the SFTPD proteins in *Homo sapiens* and *Sus scrofa* are shown in Figure 3 (d). It shows the RMSD (root mean square deviation) value of 0.538 Å (ångström). The very low RMSD value of 0.538 Å indicates a high degree of structural conservation between human and Sus scrofa SP-D, despite observable sequence divergence. This supports strong purifying selection acting on the protein's three-dimensional structure. Since an RMSD value from 0-2 Å is considered highly conserved, 2-4 Å is considered moderately similar, and greater than 4 Å means significant structural differences, it was sufficient to conclude that the SFTPD proteins in the selected species were distantly related.

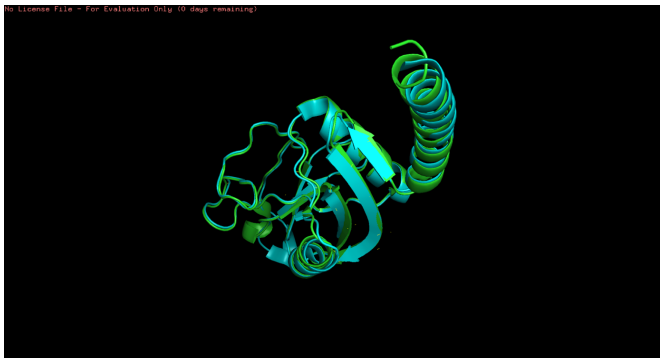

Fig. 4 PyMOL alignment of SFTPD structures. This figure shows the aligned A-chains of the surfactant protein D of *Homo sapiens* and *Sus scrofa.* The RMSD score of 0.538 is seen in the image. The blue A-chain represents the surfactant protein D in wild boars, and the green A-chain represents the surfactant protein D in humans. Both structures are seen to have alpha helices, beta-folded sheets, and triple-helical structures.

Analysis of disease and variant mapping

For the SFTPA1 protein in *Homo sapiens*, the variants as provided by UniProt include VAR_086118, VAR_086119, VAR_086120, and VAR_086121. These occurred in positions 178, 208, 211, and 225 in the amino acid sequences, respectively. For SFTPA1, data on the diseases and their clinical significance were not mentioned on the UniProt website. Hence, conclusions could not be made on the benignity or pathogenicity of the variants. The variants for the SFTPA2 protein were VAR_086122, VAR_086123, VAR_086124, VAR_063519, VAR_063520, VAR_086125, VAR_086126, VAR_086127, VAR_086128, and VAR_086129. The positions of these mutations were 171, 178, 181, 198, 231, 233 [thrice], 238, and 242, respectively. The pathogenic diseases associated with these mutations were interstitial lung disease and pulmonary fibrosis, indicating that SFTPA2 is

more susceptible to lung conditions. Disease-associated variants in SFTPA2 (e.g., positions 231 and 233) were located within highly conserved regions identified by protein MSA, suggesting disruption of evolutionarily constrained domains. In contrast, all variants identified in SFTPD occurred at positions showing greater interspecies variability, consistent with their benign classification. Furthermore, SFTPB had all natural variants, causing most of them to be benign. The variants included VAR_006948, VAR_013099, VAR_006950, VAR_006949, VAR_036856, and VAR_013100. These mutations were present in the 131st, 176th, 228th [twice], 236th, and 272nd positions, respectively. Most of the diseases were linked to some version of surfactant metabolism dysfunctions and alveolar malfunctions, which were predominantly benign but also included likely pathogenic and fully pathogenic variants. The opposing trend was observed in SFTPC, where most mutations were pathogenic to pulmonary factors. However, the types of disease were similar in SFTPB and SFTPC. The latter was associated with pulmonary fibrosis, unlike in SFTPB. The variant IDs were VAR_036855, VAR_026753, VAR_026754, VAR_026755, and VAR_026756. Their positions in the protein's amino acid sequence were 66, 73, 116, 167, and 188, respectively. Lastly, similar to SFTPB, the variants for SFTPD were all naturally occurring ones. They include VAR_020937, VAR_020938, VAR_020939, VAR_020940, and VAR_020941. In the amino acid sequence, they occurred at positions 31, 123, 180, 290, and 309, respectively. While all the disease names were unspecified, the clinical significance of all variants was noted as benign. This data further validates the hypothesis that SFTPD is the most vital pulmonary-associated surfactant protein and has evolved to adapt against mutations.

The evaluation of the disease and variant mapping of the surfactant proteins in Homo sapiens concluded that there was high evolutionary conservation of SFTPD. The lack of pathogenic variants, as seen on UniProt, indicates that the high evolutionary conservation of SFTPD is consistent with strong functional constraint, suggesting that deleterious mutations could have severe consequences. This constraint may explain the absence of reported pathogenic SFTPD variants in living populations, rather than indicating an inherent resistance to mutation. Contrastingly, surfactant proteins A, B, and C allow mutations associated with pulmonary diseases as their sequences were amongst the least conserved across the species chosen for the investigation.

Discussion

This study supports the understanding of evolutionary conservation of surfactant proteins involved in lung functioning across five different species. The study took place at the level of mRNA sequences, amino acid sequences, protein structures, and mutations of surfactant proteins A1, A2, B, C, and D. Multiple sequence alignments and 3D structure alignment was carried out for all the SPs to reveal the evolutionary distance and patterns across the species to highlight the mutation vulnerable regions which could have direct effect on pulmonary SPs. Specialised visualisation software was used for visualisation and phenetic clustering tree construction. Our results indicate that SFTPD is under strong purifying selection, reflecting its non-redundant role in innate immune defense. The absence of pathogenic variants supports this evolutionary constraint. In contrast, the greater variability observed in SFTPA, SFTPB, and SFTPC aligns with their tolerance for disease-associated mutations that manifest clinically rather than lethally.

In certain previous studies, the function of surfactant protein A (SP-A) in lung fibrosis has been examined by a research group where sample collection, protein quantification, gene expression profiling, and functional assays in mouse models were all part of their approach to evaluate the course of fibrosis[26]. Although it focuses on in vivo experimental models rather than comparative genomics or phylogenetics, this method reflects the integration of sequence analysis and functional validation found in the current effort[15]. In a model of lung damage brought on by bleomycin, Guo et al. investigated the protective properties of surfactant protein D (SP-D). Animal care, histopathological examination, protein measurement, and evaluation of inflammatory indicators were all part of their process. Although they concentrated on fibrosis and inflammation in a disease model, the current study's methodology is similar in that it uses protein quantification and sequence analysis[27]. A recent study examined the phylogeny and comparative genomics of surfactant proteins in vertebrates. They used phylogenetic tree construction, multiple sequence alignment, evolutionary analysis, and sequence retrieval from databases as part of their methodology. They only looked at evolutionary links and sequence conservation, as opposed to the current approach, which combines disease mapping and variant analysis[28]. Thus, overall, this study covers the aspects of evolutionary conservation across different species as well as disease variant mapping for pulmonary surfactant proteins.

However, this study acknowledges some limitations, such as the lack of complete and wild-type surfactant protein sequences in public databases, which limited the choice of vertebrate species. The evaluation's evolutionary depth and breadth may have also been constrained by the fact that not all proteins were included for every species. Furthermore, data gaps in protein structures also hindered extensive analysis. The structural comparison and alignment of certain surfactant proteins in the chosen species were limited by the absence of experimentally characterised, naturally occurring protein structures, which may have decreased the precision of structure-function

interpretations. Lastly, there were some limitations in the disease and variant mapping process as well. The literature and databases that were accessible for use in the variant mapping and clinical significance evaluations did not include all known or novel harmful variants. The future scope of this study involves further investigation of variant-associated genes for the identification of their therapeutic potential in lung-associated diseases. Future studies incorporating reptiles and birds, along with AlphaFold-derived structures, could extend this framework to broader vertebrate evolution. Overall, this study aims to determine how the evolutionary trajectory of surfactant proteins has influenced their protective roles in lung health and whether it has contributed to reducing vulnerability to lung fibrosis, a disorder characterised by excessive scarring and loss of pulmonary function.

References

1 M. Wilson and T. Wynn, Pulmonary fibrosis: pathogenesis, etiology and regulation. Mucosal Immunology, 2, 103–121 (2009). https://doi.org/10.1038/mi.2008.85.

2 T. Maher, Global incidence and prevalence of idiopathic pulmonary fibrosis. Respiratory Research, 22, 197 (2021). https://doi.org/10.1186/s12931-021-01791-z.

3 I. Savin, M. Zenkova and A. Sen'kova, Pulmonary Fibrosis as a Result of Acute Lung Inflammation: Molecular Mechanisms, Relevant In Vivo Models, Prognostic and Therapeutic Approaches. International Journal of Molecular Sciences, 23 (2022). https://doi.org/10.3390/ijms232314959.

4 D. Klay, T. Hoffman, A. Harmsze, J. Grutters and C. Moorsel, Systematic review of drug effects in humans and models with surfactant-processing disease. European Respiratory Review, 27 (2018). https://doi.org/10.1183/16000617.0135-2017.

5 N. Golchin, Incidence and prevalence of idiopathic pulmonary fibrosis: a systematic literature review and meta-analysis. BMC Pulmonary Medicine, 25, 378 (2025). https://doi.org/10.1186/s12890-025-03836-1.

6 M. Althobiani, Interstitial lung disease: a review of classification, etiology, epidemiology, clinical diagnosis, pharmacological and non-pharmacological treatment. Frontiers in Medicine, 11, 1296890 (2024). https://doi.org/10.3389/fmed.2024.1296890.

7 S. Sasikumar and S. Patidar, Progressive fibrotic interstitial lung diseases in India: national challenges and implications for global health policies. Health Research Policy and Systems, 24, 8 (2025). https://doi.org/10.1186/s12961-025-01425-6.

8 B. Guo, Evolutionary genetics of pulmonary anatomical adaptations in deep-diving cetaceans. BMC Genomics, 25, 339 (2024). https://doi.org/10.1186/s12864-024-10263-9.

9 S. Han and R. Mallampalli, The Role of Surfactant in Lung Disease and Host Defense against Pulmonary Infections. Annals of the American Thoracic Society, 12, 765–774 (2015). https://doi.org/10.1513/AnnalsATS.201411-507FR.

10 P. Nkadi, T. Merritt and D.-A. Pillers, An overview of pulmonary surfactant in the neonate: Genetics, metabolism, and the role of surfactant in health and disease. Molecular Genetics and Metabolism, 97, 95–101 (2009). https://doi.org/10.1016/j.ymgme.2009.01.015.

11 A. Nayak, E. Dodagatta-Marri, A. Tsolaki and U. Kishore, An Insight into the Diverse Roles of Surfactant Proteins, SP-A and SP-D in Innate and Adaptive Immunity. Frontiers in Immunology, 3, 131 (2012). https://doi.org/10.3389/fimmu.2012.00131.

12 I. Hernández-Hernández, Endogenous LXR signaling controls pulmonary surfactant homeostasis and prevents lung inflammation. Cellular and Molecular Life Sciences, 81, 287 (2024). https://doi.org/10.1007/s00018-024-05310-3.

13 M. Cedzyński and A. Świerzko, The Role of Pulmonary Collectins, Surfactant Protein A (SP-A) and Surfactant Protein D (SP-D) in Cancer. Cancers, 16 (2024). https://doi.org/10.3390/cancers16183116.

14 A. Abdel Megeid, Correlating SFTPC gene variants to interstitial lung disease in Egyptian children. Journal of Genetic Engineering and Biotechnology, 20, 117 (2022). https://doi.org/10.1186/s43141-022-00399-0.

15 K. Kim, D. Shin, G. Lee and H. Bae, Loss of SP-A in the Lung Exacerbates Pulmonary Fibrosis. International Journal of Molecular Sciences, 23 (2022). https://doi.org/10.3390/ijms23105292.

16 Y. Aono, Surfactant protein-D regulates effector cell function and fibrotic lung remodeling in response to bleomycin injury. American Journal of Respiratory and Critical Care Medicine, 185, 525–536 (2012). https://doi.org/10.1164/rccm.201103-0561OC.

17 F. Wu, L. Mueller, D. Crouzillat, V. Pétiard and S. Tanksley, Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics, 174, 1407–1420 (2006). https://doi.org/10.1534/genetics.106.062455.

18 P. Gluckman, F. Low, T. Buklijas, M. Hanson and A. Beedle, How evolutionary principles improve the understanding of human health and disease. Evolutionary Applications, 4, 249–263 (2011). https://doi.org/10.1111/j.1752-4571.2010.00164.x.

19 L. Nogee, Genetic causes of surfactant protein abnormalities. Current Opinion in Pediatrics, 31, 330–339 (2019). https://doi.org/10.1097/mop.0000000000000751.

20 J. Dudley, Human genomic disease variants: a neutral evolutionary explanation. Genome Research, 22, 1383–1394 (2012). https://doi.org/10.1101/gr.133702.111.

21 A. Pastva, J. Wright and K. Williams, Immunomodulatory roles of surfactant proteins A and D: implications in lung disease. Proceedings of the American Thoracic Society, 4, 252–257 (2007). https://doi.org/10.1513/pats.200701-018AW.

22 D. Benson, GenBank. Nucleic Acids Research, 41, 36–42 (2013). https://doi.org/10.1093/nar/gks1195.

23 F. Sievers and D. Higgins, Clustal Omega, accurate alignment of very large numbers of sequences. Methods in Molecular Biology, 1079, 105–116 (2014). https://doi.org/10.1007/978-1-62703-646-7_6.

24 J. Procter, Alignment of Biological Sequences with Jalview. Methods in Molecular Biology, 2231, 203–224 (2021). https://doi.org/10.1007/978-1-0716-1036-7_13.

25 H. Berman, The Protein Data Bank. Nucleic Acids Research, 28, 235–242 (2000). https://doi.org/10.1093/nar/28.1.235.

26 C. Scheffzük, D. Biedziak, N. Gisch, T. Goldmann and C. Stamme, Surfactant protein A modulates neuroinflammation in adult mice upon pulmonary infection. Brain Research, 1840, 149108 (2024). https://doi.org/10.1016/j.brainres.2024.149108.

27 M. Murata, Surfactant protein D is a useful biomarker for monitoring acute lung injury in rats. Experimental Lung Research, 42, 314–321 (2016). https://doi.org/10.1080/01902148.2016.1215570.

28 R. Holmes, Comparative and Evolutionary Studies of Vertebrate Arylsulfatase B, Arylsulfatase I and Arylsulfatase J Genes and Proteins: Evidence for an ARSB-like Sub-family. Journal of Proteomics & Bioinformatics, 9 (2016). https://doi.org/10.4172/jpb.1000418.