

https://dspace.org/

Unveiling the genetic tapestry Rare disease genomics of spinal muscular atrophy and phenylketonuria proteins

Debaleena Nawn; Sk. Sarif Hassan; Elrashdy M. Redwan; Tanishta Bhattacharya; Pallab Basu ...

ELSEVIER SCIENCE BV
https://hdl.handle.net/10539/40834
WOS

Downloaded from a DSpace repository. https://dspace.org/

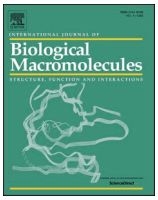
International Journal of Biological Macromolecules 269 (2024) 131960



Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

journal homepage: www.elsevier.com/locate/ijbiomac





Unveiling the genetic tapestry: Rare disease genomics of spinal muscular atrophy and phenylketonuria proteins

Debaleena Nawn [a,1], Sk. Sarif Hassan [b,1], Elrashdy M. Redwan [c,d,e], Tanishta Bhattacharya [f], Pallab Basu [g,h], Kenneth Lundstrom [i], Vladimir N. Uversky [j,*]

[a] Indian Research Institute for Integrated Medicine (IRIIM), Unsani, Howrah 711302, West Bengal, India
[b] Department of Mathematics, Pingla Thana Mahavidyalaya, Maligram, Paschim Medinipur, West Bengal, India
[c] Department of Biological Science, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia
[d] Centre of Excellence in Bionanoscience Research, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[e] Therapeutic and Protective Proteins Laboratory, Protein Research Department, Genetic Engineering and Biotechnology Research Institute, City of Scientific Research and Technological Applications, New Borg EL-Arab 21934, Alexandria, Egypt
[f] Developmental Genetics (Dept III), Max Planck Institute for Heart and Lung Research, Ludwigstrabe 43, 61231, Bad Nauheim, Germany
[g] School of Physics, University of the Witwatersrand, Johannesburg, Braamfontein, 2000, South Africa
[h] Adjunct Faculty, Woxsen School of Sciences, Woxsen University, Hyderabad 500 033, Telangana, India
[i] PanTherapeutics, Rte de Lavaux 49, CH1095 Lutry, Switzerland
[j] Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA

ARTICLE INFO

Keywords:
Phenylalanine 4-monooxygenase (PAH)
Survival motor neuron (SMN1)
Agglomerated proximal sequences
Quantitative genomics
Phylogeny
Rare disease

ABSTRACT

Rare diseases, defined by their low prevalence, present significant challenges, including delayed detection, expensive treatments, and limited research. This study delves into the genetic basis of two noteworthy rare diseases in Saudi Arabia: Phenylketonuria (PKU) and Spinal Muscular Atrophy (SMA). PKU, resulting from mutations in the phenylalanine hydroxylase (PAH) gene, exhibits geographical variability and impacts intellectual abilities. SMA, characterized by motor neuron loss, is linked to mutations in the survival of motor neuron 1 (SMN1) gene. Recognizing the importance of unveiling signature genomics in rare diseases, we conducted a quantitative study on PAH and SMN1 proteins of multiple organisms by employing various quantitative techniques to assess genetic variations. The derived signature-genomics contributes to a deeper understanding of these critical genes, paving the way for enhanced diagnostics for disorders associated with PAH and SMN1.

1. Introduction

The prevalent definitions of rare diseases revolve around the criteria employed by pharmaceutical regulatory bodies to facilitate the development of therapies for rare conditions, commonly referred to as orphan drugs [1,2]. These definitions exhibit regional variations. For instance, in the United States, a rare disease is delineated by the 1983 Orphan Drug Act as a condition impacting fewer than 200,000 individuals. As per the National Institutes of Health (NIH), around 7000 rare diseases impact an estimated 25 to 30 million Americans, translating to one in every ten Americans. The equivalent legislation implemented in the European Union in 2000 designates a disease as rare, when it afflicts <1 in 2000 people [3]. Furthermore, some diseases are 'rare' in some demographics or regions, but not in others [4–7]. Notably, over 80 % of these rare diseases have a discernible genetic basis, and roughly 75 % of these conditions impact children, according to the Saudi Ministry of Health [8]. Recent statistics from 2023 report 7362 cases mentioned in the OMIM and Orphanet databases, highlighting the substantial burden of rare diseases [9]. Saudi Arabia alone grapples with an estimated 6000–8000 distinct rare diseases, a higher prevalence likely stemming from an increased occurrence of pathogenic alleles linked to consanguineous marriages [8]. This elevated number of rare diseases poses significant challenges, including delayed detection and diagnosis, costly treatment options, and a dearth of pertinent research and scientific studies [10].

Recognizing the growing global importance of understanding and

* Corresponding author.
E-mail addresses: debaleena.nawn@gmail.com (D. Nawn), sksarifhassan@pinglacollege.ac.in (Sk.S. Hassan), lradwan@kau.edu.sa (E.M. Redwan), tanishta18@iiserbpr.ac.in (T. Bhattacharya), pallabbasu@gmail.com (P. Basu), lundstromkenneth@gmail.com (K. Lundstrom), vuversky@usf.edu (V.N. Uversky).
[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.ijbiomac.2024.131960
Received 7 January 2024; Received in revised form 30 March 2024; Accepted 27 April 2024
Available online 1 May 2024
0141-8130/© 2024 Elsevier B.V. All rights reserved.

Table 1
List of identical sequences of PAH and SMN1 proteins from various organisms.

Serial No.		List of 100 % identica	l PAH sequences		Remarks
1	tr|A0A0D9S1L9|PAH_CHLSB	tr|A0A2K6BHX1|PAH_MACNE	tr|A0A2K5NEI1|PAH_CERAT	tr|A0A2K5ZL20|PAH_MANLE	100 % Identical
2	tr|A0A2I3S5K6|PAH_PANTR	tr|A0A2R8ZJB8|PAH_PANPA)			100 % Identical
3	tr|A0A2I3SWI0|PAH_PANTR	tr|A0A2R8ZCJ2|PAH_PANPA	tr|G3QF53|PAH_GORGO		100 % Identical
4	tr|A0A2K5HWD5|PAH_COLAP	tr|A0A8C9GZW7|PAH_9PRIM			100 % Identical
5	tr|A0A2K5NDZ8|PAH_CERAT	tr|A0A2K5W2Q3|PAH_MACFA	tr|A0A2K5ZKY3|PAH_MANLE		100 % Identical
6	tr|A0A2R8ZCI5|PAH_PANPA	tr|H2Q6R0|PAH_PANTR			100 % Identical
7	tr|A0A8C4MZ99|PAH_EQUAS	tr|F7BKF9|PAH_HORSE			100 % Identical
	List of 100 % identical SMN1 sequences				
1	tr|A0A2I3MMK8|SMN_PAPAN	tr|A0A2K6CHD7|SMN_MACNE	tr|A0A2K5UX29|SMN_MACFA		100 % Identical
2	tr|A0A2K6LS96|SMN_RHIBE	tr|A0A2K6QRN8|SMN_RHIRO			100 % Identical
3	tr|A0A8I5R6V5|SMN_PAPAN	sp|Q4R4F8|SMN_MACFA	tr|F6V985|SMN_MACMU		100 % Identical

diagnosing rare diseases, various initiatives have been established [11]. For instance, the Orphanet database offers comprehensive information about rare diseases, encompassing their epidemiological distribution, causative factors, and currently available treatment strategies. Additionally, scientific symposiums dedicated to rare diseases have been organized to foster knowledge exchange [12]. Recent years have witnessed pivotal research in smaller cohorts and patient-based studies, shedding light on common mutations underlying various rare diseases, particularly those rooted in genetics [13]. Furthermore, these studies have underscored the role of genetic variations, environmental factors, and geographic disparities in survival in shaping the occurrence and prevalence of rare diseases [14].

Among the rare diseases prevalent in Saudi Arabia, two noteworthy examples are Phenylketonuria (PKU) and Spinal Muscular Atrophies (SMAs). PKU is an autosomal recessive genetic disorder resulting from mutations in the phenylalanine hydroxylase (PAH) gene [15]. The PAH gene plays a vital role in converting Phe to Tyr, and mutations in this gene can disrupt this conversion process, leading to the accumulation of Phe in the bloodstream and the formation of phenyl ketone bodies, typically excreted in urine [16–19].

PKU exhibits geographical variability in its prevalence, affecting approximately 450,000 individuals worldwide [20]. Italy, with a prevalence of 1 in 4000, and Ireland, with a prevalence of 1 in 4545, exhibited higher rates than Iran and Jordan, both at 1 in 5000, as well as Turkey at 1 in 6667. In contrast, Saudi Arabia (1 in 14,245), Iraq (1 in 14,286), and the United Arab Emirates (1 in 14,493) had lower prevalence rates of PKU [20]. Individuals with PKU are characterized by lower intellectual abilities, and during the neonatal stage, it can lead to developmental delays, epilepsy, altered psychological behavior, tremors, eczema, and an unusual odor [21,22]. PKU can be categorized into different stages based on the concentration of Phe in the blood, including mild PKU, moderate PKU, and classic PKU, which represents the severe phenotype [17,23]. Past studies have revealed that the coding sequence of the PAH gene consists of 452 amino acids, spans 1356 base pairs, and has a molecular weight of approximately 52KDa [22]. The PAH gene comprises three distinct domains [24]:

- The Regulatory N-terminal domain (amino acids 1-142)
- The Catalytic central domain (amino acids 143-410)
- The Oligomerization domain, C-terminal, which includes dimerization and tetramerization motifs (amino acids 411-452).

Spinal Muscular Atrophy (SMA) is also an autosomal recessive disorder characterized by the loss of motor neurons in the spinal cord and lower brainstem, leading to muscle weakening and atrophy [25–27]. Recent research has indicated a significant prevalence of SMA in Saudi Arabia, where the numbers hover around 32 in 100,000 births, exceeding global rates [28,29]. SMA, like many rare diseases, has a genetic basis [30]. Mutations in the survival of motor neuron 1 (SMN1) gene, particularly the deletion of exon 7, can trigger the development of this condition [31,32]. SMN1 plays a crucial role in RNA processing and transportation [33]. Depending on the age of onset and the level of motor function that can be achieved, SMA can be categorized into four phenotypes: SMA I, SMA II, SMA III, and SMA IV, with decreasing severity [34]. SMA I accounts for nearly 60 % of all SMA cases and can be diagnosed through genetic testing of SMN1 and copy number testing of SMN2 [35]. The clinical severity of Spinal Muscular Atrophy (SMA) is inversely correlated with the copy number of the SMN2 gene. It ranges from severe weakness and paraplegia in infancy to a milder proximal weakness observed in adulthood [36]. In some instances, specific sequence variations within SMN2, such as the c.859G > C substitution in exon 7, have been found. This substitution creates a new exonic splicing enhancer element, leading to an increase in full-length transcripts and resulting in less severe SMA phenotypes. Therefore, SMN2, while generally known for its dose-dependent influence, can also be subject to sequence variations that modify the disease's severity [37].

Quantitative genomics is pivotal for advancing the diagnosis and understanding of rare diseases, such as PKU and SMA, which have a genetic basis [38,39]. Utilizing advanced technologies like next-generation sequencing, genomic analysis enables the identification of specific genetic variations responsible for these conditions by pinpointing mutations or alterations in genes [40]. This precision in genetic diagnosis not only enhances accuracy but also paves the way for personalized and targeted treatment strategies [41]. Additionally, genomics significantly contributes to early detection and intervention in rare diseases, allowing clinicians to diagnose conditions before symptoms appear [42]. Early diagnosis is crucial for timely interventions, potentially altering disease trajectories and improving outcomes [43]. Furthermore, genomic data from rare disease patients contribute to ongoing research, fostering collaborations and expediting the de- velopment of novel therapies tailored to specific genetic aberrations [44]. In essence, genomics is a cornerstone in the comprehensive understanding, diagnosis, and treatment of rare diseases, offering a route to more precise and personalized medical care [45]. Quantitative genomics of PAH contributes to a more accurate assessment of enzyme activity levels and the severity of PKU [46]. For SMN1, quantitative genomics aids in assessing the copy number variations of the SMN1 gene, crucial for determining the severity of SMA [47,48].

In this present study, we focused on deriving basic signature-genomics of SMN1 and PAH genes collected from various organisms to gain valuable insights into their fundamental characteristics and variations. The derived signature genomics not only contributes to our knowledge of these critical genes, but also sets the stage for future investigations into targeted therapies, diagnostics, and personalized

Table 2
List of 51 PAH and 45 SMN1 proteins from various organisms with their associated Uniprot ID (Hyperlinked with respective Uniprot).

PAH (Phenylalanine 4-monooxygenase)	Renamed as	SMN1 (Survival motor neuron protein)	Renamed as
tr|A0A5F7ZF01|PAH_MACMU	PAH_1	tr|A0A8I5TVG4|SMN_PONAB	SMN1_1
tr|A0A7N9DEA3|PAH_MACFA	PAH_2	tr|A0A0A0MXR7|SMN_PONAB	SMN1_2
tr|A0A8I3X8J4|PAH_CALJA	PAH_3	sp|Q8HYB8|SMN_FELCA	SMN1_3
tr|A0A2K6U7I0|PAH_SAIBB	PAH_4	sp|P97801|SMN_MOUSE	SMN1_4
tr|A0A8C4MZ99|PAH_EQUAS	PAH_5	sp|O35876|SMN_RAT	SMN1_5
tr|A0A8I3RVD0|PAH_CANLF	PAH_6	tr|A0A2I3GD31|SMN_NOMLE	SMN1_6
tr|A0A5F5PJG0|PAH_HORSE	PAH_7	tr|A0A2I3HCL7|SMN_NOMLE	SMN1_7
tr|A0A8I4A1R1|PAH_CALJA	PAH_8	tr|G1QWS5|SMN_NOMLE	SMN1_8
tr|A0A8I3N1J8|PAH_CANLF	PAH_9	tr|A0A2I3G9J7|SMN_NOMLE	SMN1_9
tr|A0A2I3HWG9|PAH_NOMLE	PAH_10	tr|A0A2K6QRM5|SMN_RHIRO	SMN1_10
tr|A0A2I3SWI0|PAH_PANTR	PAH_11	tr|A0A2K6LS96|SMN_RHIBE	SMN1_11
tr|J3KND8|PAH_HUMAN	PAH_12	tr|A0A2K6A6E8|SMN_MANLE	SMN1_12
tr|H2NIF5|PAH_PONAB	PAH_13	tr|A0A2K6A681|SMN_MANLE	SMN1_13
tr|A0A2K6G6V1|PAH_PROCO	PAH_14	tr|A0A2K5MQ61|SMN_CERAT	SMN1_14
tr|A0A2K5CN98|PAH_AOTNA	PAH_15	tr|A0A2K5MQ51|SMN_CERAT	SMN1_15
tr|G3S964|PAH_GORGO	PAH_16	tr|A0A8D2K555|SMN_THEGE	SMN1_16
tr|A0A2R8ZCI5|PAH_PANPA	PAH_17	tr|A0A8D2FSA3|SMN_THEGE	SMN1_17
tr|A0A2I3S5K6|PAH_PANTR	PAH_18	tr|A0A2K5UX34|SMN_MACFA	SMN1_18
sp|P00439|PAH_HUMAN	PAH_19	tr|A0A2K6CHC9|SMN_MACNE	SMN1_19
tr|G1R3M2|PAH_NOMLE	PAH_20	tr|A0A2I3MMK8|SMN_PAPAN	SMN1_20
tr|A0A2I3GWX3|PAH_NOMLE	PAH_21	tr|A0A8I5R6V5|SMN_PAPAN	SMN1_21
tr|A0A8D2ETB6|PAH_THEGE	PAH_22	tr|A0A2K6CHF1|SMN_MACNE	SMN1_22
tr|A0A2K6KXK8|PAH_RHIBE	PAH_23	tr|A0A1D5RJY4|SMN_MACMU	SMN1_23
tr|A0A2K5HWD5|PAH_COLAP	PAH_24	tr|A0A096MVQ6|SMN_PAPAN	SMN1_24
tr|A0A2K6QV60|PAH_RHIRO	PAH_25	tr|G3SHH7|SMN_GORGO	SMN1_25
tr|A0A1D5QG49|PAH_MACMU	PAH_26	tr|A0A2I3T3Q2|SMN_PANTR	SMN1_26
tr|F7HMW9|PAH_MACMU	PAH_27	tr|A0A2I3TMK9|SMN_PANTR	SMN1_27
tr|A0A8I5NBG4|PAH_PAPAN	PAH_28	tr|H2QR14|SMN_PANTR	SMN1_28
tr|A0A0D9S1L9|PAH_CHLSB	PAH_29	tr|A0A2J8L0U3|SMN_PANTR	SMN1_29
tr|A0A8I5R1B8|PAH_PAPAN	PAH_30	sp|Q16637|SMN_HUMAN	SMN1_30
tr|A0A2I3NB59|PAH_PAPAN	PAH_31	tr|E7EQZ4|SMN_HUMAN	SMN1_31
tr|A0A2J8XML5|PAH_PONAB	PAH_32	tr|G3RPQ4|SMN_GORGO	SMN1_32
tr|A0A2K5CNA0|PAH_AOTNA	PAH_33	tr|A0A0D9RT05|SMN_CHLSB	SMN1_33
tr|F7I717|PAH_CALJA	PAH_34	tr|A0A8C9GM82|SMN_9PRIM	SMN1_34
tr|A0A2K6U7J8|PAH_SAIBB	PAH_35	tr|A0A8C9GL66|SMN_9PRIM	SMN1_35
tr|A0A6J3JMG7|PAH_SAPAP	PAH_36	tr|A0A2K5JKT2|SMN_COLAP	SMN1_36
tr|A0A2K6KXJ8|PAH_RHIBE	PAH_37	tr|A0A2K5JKX7|SMN_COLAP	SMN1_37
tr|A0A8C9GYN8|PAH_9PRIM	PAH_38	tr|A0A2K5JKZ1|SMN_COLAP	SMN1_38
tr|A0A2K6QV23|PAH_RHIRO	PAH_39	tr|A0A2J8VTK4|SMN_PONAB	SMN1_39
tr|A0A8D2EUN4|PAH_THEGE	PAH_40	sp|Q5RE18|SMN_PONAB	SMN1_40
tr|A0A096NE64|PAH_PAPAN	PAH_41	tr|A0A8I5YK49|SMN_PONAB	SMN1_41
tr|A0A2K5NDZ8|PAH_CERAT	PAH_42	tr|A0A2J8VTQ8|SMN_PONAB	SMN1_42
tr|A0A2K6BHW6|PAH_MACNE	PAH_43	sp|Q9W6S8|SMN_DANRE	SMN1_43
tr|A0A2K6BHY9|PAH_MACNE	PAH_44	sp|O18870|SMN_BOVIN	SMN1_44
tr|A0A2K5HWG8|PAH_COLAP	PAH_45	sp|O02771|SMN_CANLF	SMN1_45
tr|A0A8I5NX30|PAH_PAPAN	PAH_46		
tr|A0A8C0LJS0|PAH_CANLU	PAH_47		
tr|A0A8C0LPC7|PAH_CANLU	PAH_48		
tr|A0A2K6BHW9|PAH_MACNE	PAH_49		
tr|A0A2K6KXK6|PAH_RHIBE	PAH_50		
tr|A0A8C9KM68|PAH_PANTA	PAH_51		

Table 3
List of features of dimension 1162 calculated by I-features.

List of various features (1162 dimensional) calculated by I-features		
Descriptor groups	Descriptor	Dimension
Amino acid composition	Dipeptide composition (DPC)	400
Grouped amino acid composition	Grouped Amino Acid Composition (GAAC)	5
	Grouped Dipeptide Composition (GDC)	25
Autocorrelation	Moran Autocorrelation	240
C/T/D	C/T/D Composition (CTDC)	39
Conjoint Triad	Conjoint Triad	343
Quasi-sequence-order	Sequence-Order-Coupling Number (SoC number)	60
Pseudo-amino acid composition	Pseudo-Amino Acid Composition (PAAC)	50

Another 8710 features extracted using ProtrWeb, which were deployed in the present study is summarized in Table 4 [73].

Table 4
List of various features (8710 dimensional) calculated by ProtrWeb

List of various features (8710 dimensional) calculated by ProtrWeb	Dimension
Tripeptide Composition	8000
Normalized Moreau-Broto Autocorrelation	240
Geary Autocorrelation	240
Quasi-Sequence-Order Descriptors	100
Pseudo-Amino Acid Composition	50
Amphiphilic Pseudo-Amino Acid Composition	80

medicine strategies.

2. Data acquisition

For both the proteins Phenylalanine 4-monooxygenase (PAH) and Survival motor neuron (SMN1), BLAST was made using Human PAH (P00439) and Human SMN1 (Q16637), 62 PAH and 50 SMN1 sequences were obtained with at least 95 % similarity with Human PAH (P00439) and Human SMN1 (Q16637), respectively. A list of identical (100 % similarity) sequences is given in Table 1. Among them, 51 PAH and 45 SMN1 unique proteins were obtained from the UniProt database as listed in Table 2.

3. Methods

3.1. Composition profiler of PAH and SMN1 proteins

The Composition Profiler was used to generate an amino acid composition profile of all the PAH and SMN1 proteins analyzed in this study [49]. This set of amino acid sequences was the query set and the 'Protein Data Bank Select 25' was the background set. We also generated a composition profile for experimentally validated disordered proteins from the DisProt [50]. The generated profiles represent plots showing normalized enrichment or depletion of a given residue calculated as $\frac{(Cx-Corder)}{Corder}$, where $\mathscr{C}_x$ is the content of a given residue in its query protein, and $\mathscr{C}_{order}$ is the content of the same residue in the PDB Select 25.

3.2. Determining amino acid frequency composition of PAH and SMN1 proteins

The count of every amino acid in a sequence, termed amino acid frequency, was computed for all PAH and SMN1 proteins [51–53]. Furthermore, the percentage of amino acids in a sequence (obtained from dividing the amino acid frequencies by the length of that sequence and multiplied by 100) is termed the relative frequency of amino acids in that sequence. Relative frequency of 20 amino acids represents a 20-dimensional vector for each protein sequence.

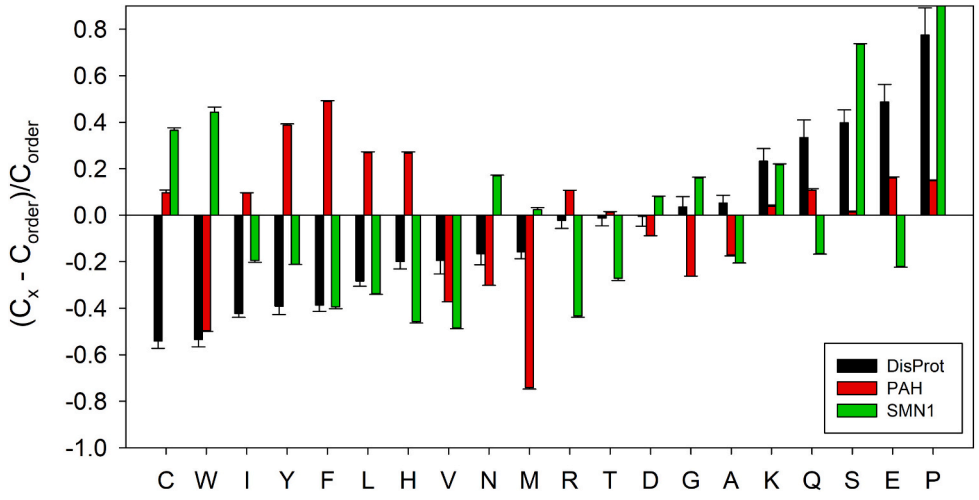


Fig. 1. Amino acid composition profile of 51 PAH (red bars)and 45 SMN1 proteins (green bars). The fractional difference is calculated as $\frac{C_x-C_{order}}{C_{order}}$ where C_x is the content of a given amino acid in the query set (PAH/SMN1 proteins or known intrinsically disordered proteins), order and C_{order} 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. In both cases, error bars correspond to standard deviations over 10,000 bootstrap iterations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

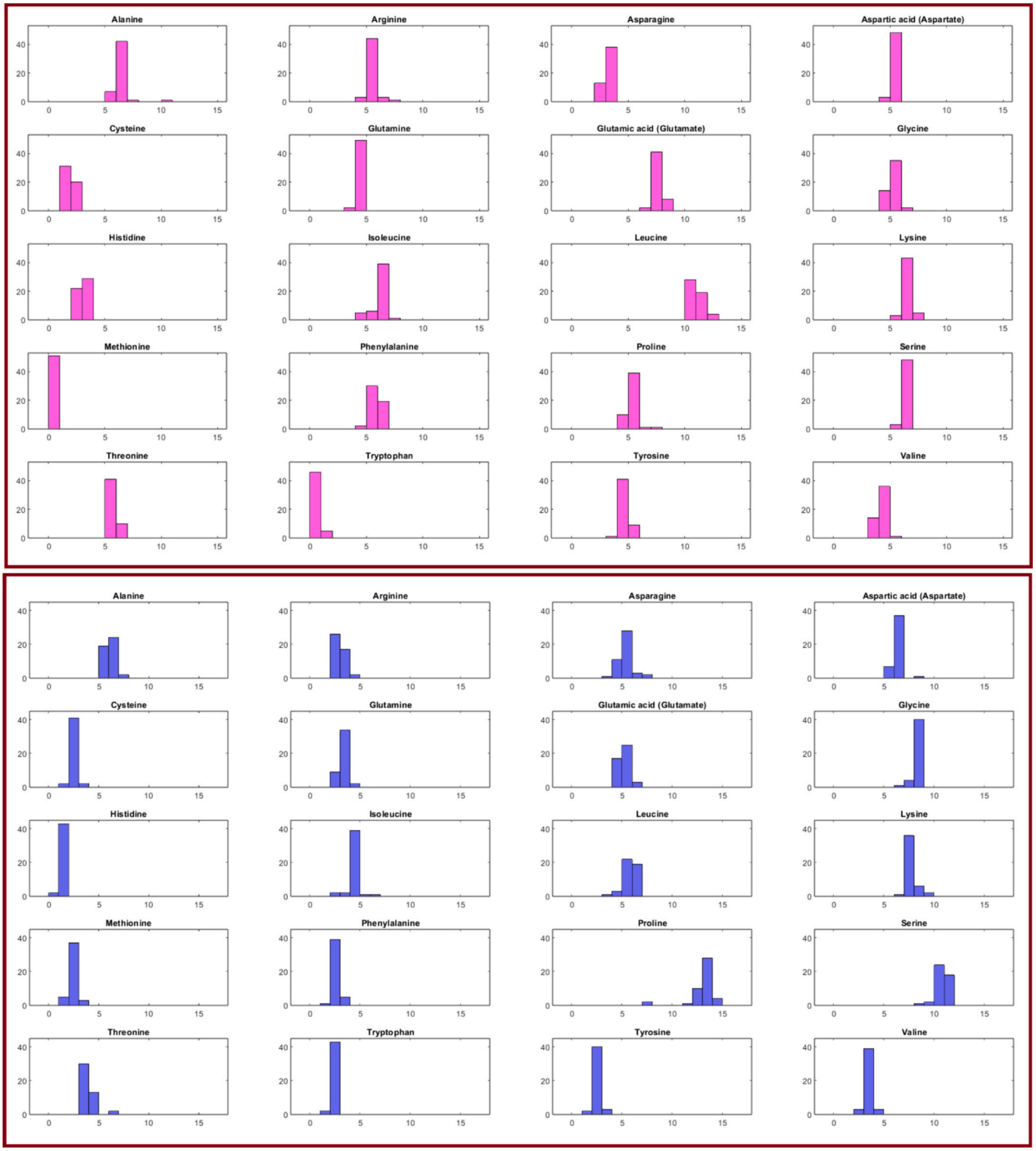


Fig. 2. Histogram of the relative frequency of each amino acid in PAH proteins (Top), and SMN1 proteins (Bottom). X-axis denotes the percentage and Y-axis denotes the number of sequences. In the top histogram, which represents PAH proteins, the X-axis corresponds to the percentage of each amino acid, while the Y-axis denotes the number of sequences. As we move along the X-axis, we observe the distribution of amino acids within the PAH proteins. Peaks in the histogram indicate amino acids that occur frequently, while valleys represent less common ones. By analyzing this histogram, we gain insights into the prevalence and composition of amino acids in PAH proteins. Similarly, in the bottom histogram, which pertains to SMN1 proteins, the X-axis also denotes the percentage of each amino acid, and the Y-axis continues to represent the number of sequences. The histogram for SMN1 proteins provides a distinct profile of amino acid frequencies specific to this gene. Comparing it with the PAH histogram allows us to identify unique patterns or variations between the two protein sets. These two sets of histograms serve as valuable tools for understanding the amino acid composition in PAH and SMN1 proteins, aiding in unraveling their genetic significance and potential implications for rare diseases.

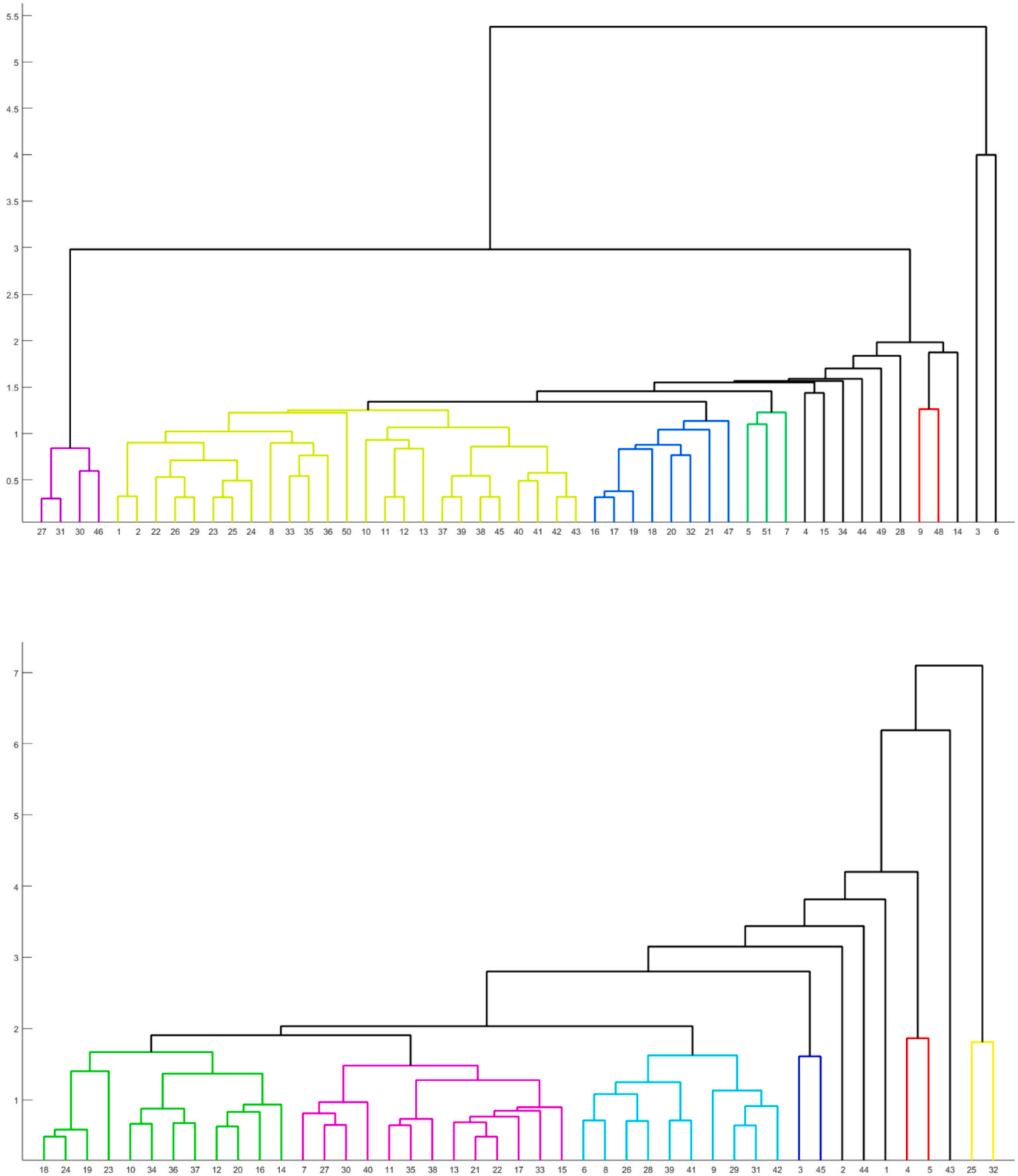


Fig. 3. (Top): Phylogenetic relationship among the PAH proteins based on the relative frequency of amino acids. (Bottom): Phylogenetic relationship among the SMN1 proteins based on relative frequency of amino acids. Note that every color (other than black) in the phylogeny designates a cluster.

Table 5
Clusters derived from the relative frequency of amino acids in PAH and SMN1 proteins.

Clusters	Sequences (PAH)
Cluster-1	{27, 31, 30, 46}
Cluster-2	{1, 2, 22, 26, 29, 23, 25, 24, 8, 33, 35, 36, 50, 10, 11, 12, 13, 37, 39, 38, 45, 40, 41, 42, 43}
Cluster-3	{16, 17, 18, 19, 20, 32, 21, 47}
Cluster-4	{5, 51, 7}
Cluster-5	{9, 48}

Clusters	Sequences (SMN1)
Cluster-1	{18, 24, 19, 23, 10, 34, 36, 37, 12, 20, 16, 14}
Cluster-2	{7, 27, 30, 40, 11, 35, 38, 13, 21, 22, 17, 33, 15}
Cluster-3	{6, 8, 26, 28, 39, 41, 9, 29, 31, 42}
Cluster-4	{3, 45}
Cluster-5	{4, 5}
Cluster-6	{25, 32}

3.2.1. Evaluating Shannon entropy of PAH and SMN1 sequences

Shannon entropy (SE) is a measure of the information content in a system [54]. SE of each protein sequence is evaluated by the formula:

$$SE = -\sum_{i=1}^{20} (p_i * log_2 p_i)$$

pi is the count of amino acid *i* in that sequence divided by the length of that sequence [51]. *SE* reflects the degree of randomness in the amino acid count in a given sequence. A higher value of *SE* indicates greater diversity, while a lower value indicates less diversity.

3.3. Amino acid frequency-based Shannon variability of position of PAH and SMN1 proteins

Shannon entropy is deployed to estimate the variability of amino acid residues at each residue position across all aligned PAH and SMN1 sequences, respectively. The Shannon variability ($\mathscr{H}_v$) for every position is defined as follows:

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

Here, $\mathscr{p}_i$ represents the fraction of residues of amino acid type i at a particular position. $\mathscr{H}_v$ ranges from 0 (only one residue is present at that position) to 4.322 (all 20 residues are equally represented in that position) [55]. Typically, positions with $\mathscr{H}_v > 2$ are considered *variable*, whereas those with $\mathscr{H}_v < 2$ are considered *conserved*. *Highly conserved* positions are those with $\mathscr{H}_v < 1$ [55].

By analyzing per-residue variability using Shannon entropy, one can identify functionally important residues within the receptor protein family. In addition, per-residue variability analysis using Shannon entropy can help identify potential drug targets within the receptor protein family. Residues that are highly variable among sequences may be more amenable to small molecule binding, as they may have more flexibility and plasticity in their binding pockets [56].

3.4. Determining the homogeneous poly-string frequency of amino acids in PAH and SMN1 proteins

A homogeneous poly-string of length n is defined as n consecutive occurrence of a particular amino acid [57,58]. For example, $\mathscr{KKKWWKKWW}$ represents one homogeneous poly-string of K (Lys) with length 3, one homogeneous poly-string of K with length 2, and two homogeneous poly-strings of W (Trp) with length 2. Note that while counting the number of a homogeneous poly-string of length n, only the exclusive/exact occurrence of length n would be taken into consideration. The maximum lengths of homogeneous poly-strings considering all amino acids across all sequences were computed and accordingly, counts of homogeneous poly-strings of all possible lengths (starting from 1 to maximum length) for each amino acid present in a given protein sequence were enumerated [57].

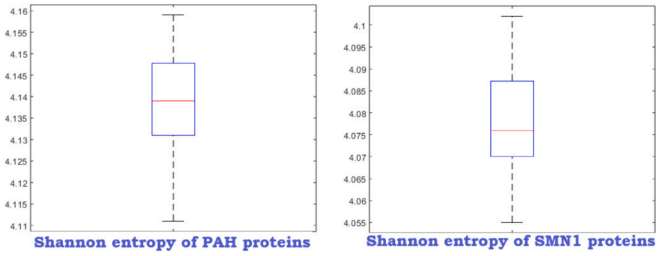


Fig. 4. Box plot depicting the variation of Shannon entropy among the PAH and SMN1 protein sequences.

Table 6
Shannon entropy among the PAH and SMN1 protein sequences.

PAH						SMN1					
PAH Sequences	*SE*	*PAH Sequences*	*SE*	*PAH Sequences*	*SE*	*SMN1 Sequences*	*SE*	*SMN1 Sequences*	*SE*	*SMN1 Sequences*	*SE*
PAH_12	4.16	PAH_32	4.14	PAH_46	4.13	SMN1_1	4.10	SMN1_11	4.08	SMN1_16	4.06
PAH_11	4.16	PAH_17	4.14	PAH_29	4.13	SMN1_4	4.10	SMN1_21	4.08	SMN1_8	4.06
PAH_47	4.16	PAH_16	4.14	PAH_33	4.13	SMN1_43	4.10	SMN1_27	4.08	SMN1_26	4.06
PAH_10	4.16	PAH_20	4.14	PAH_36	4.13	SMN1_3	4.10	SMN1_33	4.08	SMN1_10	4.06
PAH_18	4.15	PAH_9	4.14	PAH_1	4.13	SMN1_5	4.09	SMN1_24	4.08	SMN1_20	4.06
PAH_44	4.15	PAH_14	4.14	PAH_50	4.13	SMN1_44	4.09	SMN1_32	4.08	SMN1_12	4.06
PAH_7	4.15	PAH_8	4.14	PAH_27	4.13	SMN1_25	4.09	SMN1_13	4.07	SMN1_28	4.06
PAH_40	4.15	PAH_35	4.14	PAH_2	4.13	SMN1_38	4.09	SMN1_22	4.07		
PAH_45	4.15	PAH_42	4.14	PAH_31	4.13	SMN1_40	4.09	SMN1_39	4.07		
PAH_13	4.15	PAH_4	4.14	PAH_30	4.12	SMN1_15	4.09	SMN1_18	4.07		
PAH_21	4.15	PAH_51	4.14	PAH_28	4.12	SMN1_23	4.09	SMN1_37	4.07		
PAH_6	4.15	PAH_22	4.14	PAH_49	4.12	SMN1_30	4.09	SMN1_29	4.07		
PAH_38	4.15	PAH_34	4.14	PAH_3	4.11	SMN1_7	4.09	SMN1_2	4.07		
PAH_39	4.15	PAH_43	4.14			SMN1_17	4.08	SMN1_6	4.07		
PAH_41	4.15	PAH_24	4.13			SMN1_35	4.08	SMN1_14	4.07		
PAH_5	4.15	PAH_48	4.13			SMN1_42	4.08	SMN1_19	4.07		
PAH_15	4.15	PAH_25	4.13			SMN1_31	4.08	SMN1_36	4.07		
PAH_19	4.15	PAH_26	4.13			SMN1_45	4.08	SMN1_41	4.07		
PAH_37	4.15	PAH_23	4.13			SMN1_9	4.08	SMN1_34	4.07		

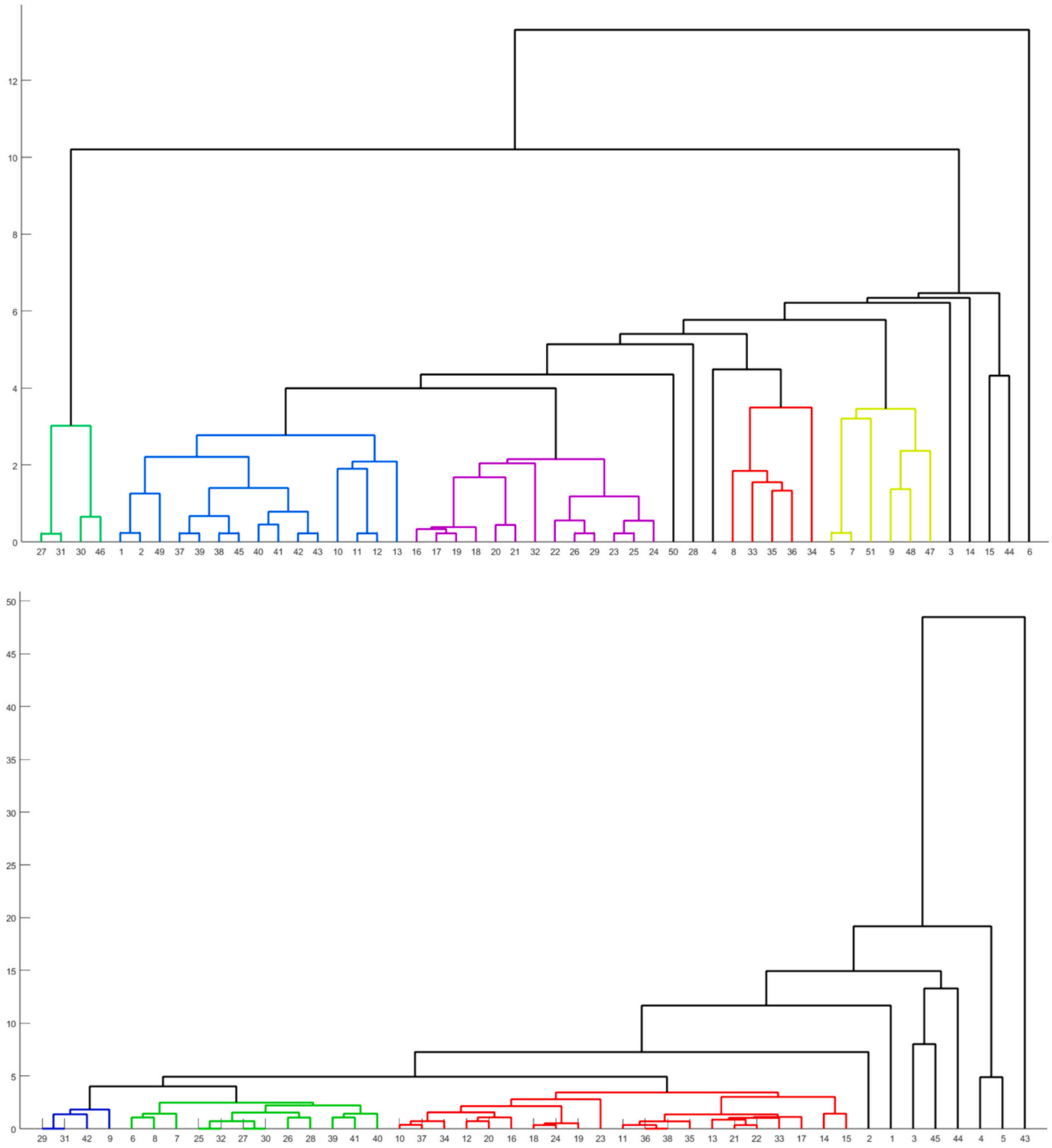


Fig. 5. (Top): Phylogenetic relationship among the PAH proteins based on sequence homology. (Bottom): Phylogenetic relationship among the SMN1 proteins based on sequence homology. Note that every color (other than black) in the phylogeny designates a cluster.

Table 7
Clustering of PAH and SMN1 proteins based on sequence homology.

Clusters	Sequences (PAH)
Cluster-1	{27, 31, 30, 46}
Cluster-2	{1, 2, 49, 37, 39, 38, 45, 40, 41, 42, 43, 10, 11, 12, 13}
Cluster-3	{16, 17, 18, 19, 20, 21, 32, 22, 26, 29, 23, 25, 24}
Cluster-4	{8, 33, 35, 36, 34}
Cluster-5	{5, 7, 51, 9, 48, 47}

Clusters	Sequences (SMN1)
Cluster-1	{29, 31, 42, 9}
Cluster-2	{6, 8, 7, 25, 32, 27, 30, 26, 28, 39, 41, 40}
Cluster-3	{10, 37, 34, 12, 20, 16, 18, 24, 19, 23, 11, 36, 38, 35, 13, 21, 22, 33, 17, 14, 15}

3.5. Evaluating polar, non-polar residue profiles of PAH and SMN1 proteins

Every amino acid in a given protein sequence was identified as polar (P) or non-polar(Q). Thus, every protein sequence became a binary sequence with two symbols: P and Q. Through this binary P-Q profile, a spatial assembly of polar and non-polar residues over a protein sequence was depicted [57,59,60,61].

3.5.1. Change response sequences based on polar, non-polar residue profiles

There are four possible changes between two consecutive residues of a polar and non-polar profile, namely Polar to Polar (PP), Polar to Non-polar (PN), Non-polar to Non-Polar (NN), and Non-polar to Polar (NP) [57]. Such changes were accounted in the form of a sequence according to the spatial sequential arrangement of polar-non-polar residues in a given P-Q binary profile. We call this sequence "P-Q Change Response Sequence ($\mathscr{CRS}_{\mathscr{PQ}}$)". The frequency of each of the four changes was enumerated from the binary polar-non-polar profile corresponding to each PAH/SMN1 protein [57].

3.6. Evaluating acidic, basic, neutral residue profiles of PAH and SMN1 proteins

Every amino acid in a given PAH/SMN1 protein sequence was identified as acidic (A), basic (B), and neutral (N). Thus, every protein sequence became a ternary valued sequence (A-B-N profiles) with three symbols: A, B, and N [57].

3.6.1. Change response sequences based on acidic-basic-neutral residue profiles

There are nine possible changes between two consecutive residues of an A-B-N profile, namely Acidic to Acidic (AA), Acidic to Basic (AB), Acidic to Neutral (AN), Basic to Acidic (BA), Basic to Basic (BB), Basic to Neutral (BN), Neutral to Acidic (NA), Neutral to Basic (NB), Neutral to Neutral (NN). Such changes were accounted in the form of a sequence according to the spatial sequential arrangement of acidic, basic, and neutral residues in a given A-B-N ternary profile. We designate this sequence "A-B-N Change Response Sequence ($\mathscr{CRS}_{\mathscr{ABN}}$)". The frequency of each of the nine changes was counted for a given ternary A-B-N profile corresponding to each PAH/SMN1 protein [64].

3.7. Evaluating intrinsic protein disorder of PAH and SMN1 proteins

Predisposition for intrinsic disorder of all the PAH and SMN1 proteins analyzed in this study was determined using a set of commonly used per-residue disorder predictors, such as PONDR® VLS2, PONDR® VL3, PONDR® VLXT, PONDR® FIT, IUPred-Long, IUPred-Short [62–67]. A web platform called Rapid Intrinsic Disorder Analysis Online (RIDAO) was used to gather results from each predictor in bulk [68]. The percent of predicted intrinsically disorder residues (PPIDR) for each protein was used to classify each protein based on its level of disorder. A residue was considered to be disordered if it had a value of 0.5 or higher. Generally, a PPIDR value of <10 % is taken to correspond to a highly ordered protein, PPIDR between 10 % and 30 % is ascribed to a moderately disordered protein, and PPIDR >30 % corresponds to a highly disordered protein [69,70]. In addition to PPIDR, the mean disorder score (MDS) was calculated for each query protein as a protein length-normalized sum of all the per-residue disorder scores. The per-residue disorder score ranges from 0 to 1, where a score of 0 indicates fully ordered residues and a score of 1 indicates fully disordered residues. Residues with scores above the threshold of 0.5 were considered *disordered residues*. Residues with disorder scores between 0.25 and 0.5 were categorized as *highly flexible*, while those with scores between 0.1 and 0.25 were classified as *moderately flexible* [67]. We also utilized two binary predictors of disorder, the charge-hydropathy (CH) plot and the cumulative distribution function (CDF) (both accessible at Predictor of Natural Disordered Regions), to assess intrinsic disorder at the whole protein level [71,72].

3.7.1. Change response sequences based on intrinsic protein disorder residues

There are sixteen possible changes between two consecutive residues where residues were denoted as disordered (D), highly flexible (HF), moderately flexible (MF), and other (O) in PAH and SMN1 protein sequences namely disordered to disordered (D_D), disordered to highly flexible (D_HF), disordered to moderately flexible (D_MF), disordered to

Table 8
Invariant residues across PAH (SMN1) variants with reference to tr|A0A8C9KM68|PAH_PANTA (sp|O02771|SMN_CANLF).

PAH						SMN1			
AA Position(s)	*AA residues(s)*	*AA Position(s)*	*AA residues(s)*	*AA Position(s)*	*AA residues(s)*	*AA Position(s)*	*AA residues(s)*	*AA Position(s)*	*AA residues(s)*
20	G	148–150	FAD	327–342	I-Q	67–68	RK	134–137	S-L
30	E	152–153	AY	344–346	CLS	72–74	KNK	157	E
33	G	155	Y	349	P	87	W	161–162	ST
39	L	157–162	H-P	353–354	PL	90–91	GD	165	S
44–46	END	164–166	VEY	356–360	L-A	93	C	169	S
48–68	N-T	168–178	*E*-F	362	Q	97–101	W-G	186	W
70–71	LD	180–192	T-E	364	Y	104	Y	192–193	PP
73–74	RS	194–206	N-F	367–371	T-P	106–108	ATI	195	P
76	P	208	E	373–379	Y-F	110	S		
78	L	210–215	N-E	381–387	D-R	112	D		
80–85	N-L	217–222	V-Q	396–397	PF	117–120	T-V		
88–136	D-F	224–260	C-S	404	Y	122–123	YT		
138–142	D-R	262–290	P-S	406	Q	125–127	YGN		
144–146	RRK	311–325	I-D	420	L	129–132	E-N		

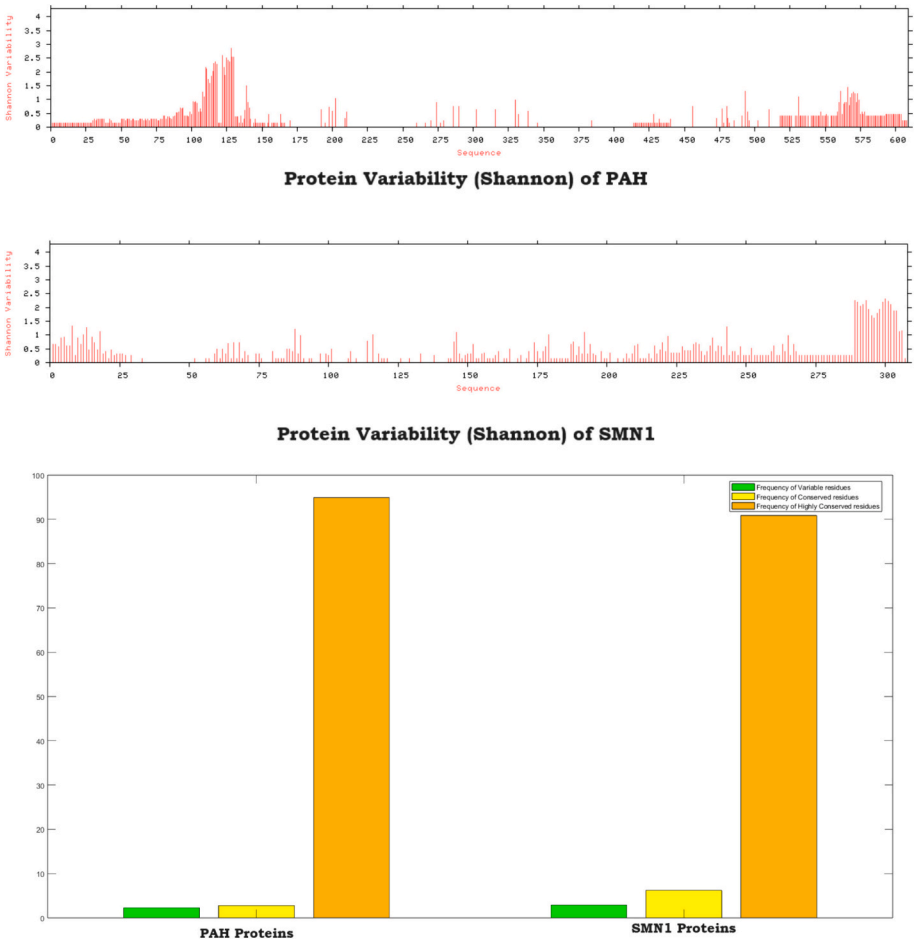


Fig. 6. Shannon variability of amino acid residues in PAH and SMN1 proteins.

Table 9
Frequency of homogeneous poly-string of length 1, 2, … 10 for each protein sequence.

PAH	Homogenous polystring of length l					SMN1	Homogenous polystring of length l								
	l = 1	*l = 2*	*l = 3*	*l = 4*	*l = 8*		*l = 1*	*l = 2*	*l = 3*	*l = 4*	*l = 5*	*l = 6*	*l = 8*	*l = 9*	*l = 10*
PAH_1	395	18	2	0	0	SMN1_1	230	18	1	0	2	0	0	0	1
PAH_2	395	18	2	0	0	SMN1_2	221	16	1	0	2	0	0	0	1
PAH_3	453	25	5	1	1	SMN1_3	233	16	2	0	2	0	0	1	0
PAH_4	408	19	2	0	0	SMN1_4	227	19	1	1	2	1	0	0	0
PAH_5	389	19	2	0	0	SMN1_5	225	19	2	0	2	0	0	0	1
PAH_6	485	27	2	0	0	SMN1_6	218	19	2	0	2	0	0	0	1
PAH_7	419	21	2	0	0	SMN1_7	230	19	2	0	2	0	0	0	1
PAH_8	433	17	2	0	0	SMN1_8	218	19	2	0	2	0	0	0	1
PAH_9	401	21	2	0	0	SMN1_9	228	20	2	0	2	0	0	0	1
PAH_10	405	18	2	0	0	SMN1_10	222	17	2	0	2	0	0	0	1
PAH_11	405	18	2	0	0	SMN1_11	234	17	2	0	2	0	0	0	1
PAH_12	405	18	2	0	0	SMN1_12	219	18	2	0	2	0	0	0	1
PAH_13	403	19	2	0	0	SMN1_13	231	18	2	0	2	0	0	0	1
PAH_14	408	19	2	0	0	SMN1_14	220	18	2	0	2	0	0	0	1
PAH_15	408	20	2	0	0	SMN1_15	232	18	2	0	2	0	0	0	1
PAH_16	410	18	2	0	0	SMN1_16	218	19	2	0	2	0	0	0	1
PAH_17	410	18	2	0	0	SMN1_17	230	19	2	0	2	0	0	0	1
PAH_18	391	18	2	0	0	SMN1_18	228	20	2	0	2	0	0	0	1
PAH_19	410	18	2	0	0	SMN1_19	230	19	2	0	2	0	0	0	1
PAH_20	410	18	2	0	0	SMN1_20	220	18	2	0	2	0	0	0	1
PAH_21	415	19	2	0	0	SMN1_21	232	18	2	0	2	0	0	0	1
PAH_22	409	18	2	0	0	SMN1_22	232	18	2	0	2	0	0	0	1
PAH_23	410	18	2	0	0	SMN1_23	233	19	2	0	2	0	0	0	1
PAH_24	410	18	2	0	0	SMN1_24	228	20	2	0	2	0	0	0	1
PAH_25	410	18	2	0	0	SMN1_25	188	13	2	0	1	0	0	0	0
PAH_26	410	18	2	0	0	SMN1_26	219	19	2	1	2	0	1	0	0
PAH_27	429	19	2	0	0	SMN1_27	231	19	2	1	2	0	1	0	0
PAH_28	416	18	2	0	0	SMN1_28	219	19	2	1	2	0	1	0	0
PAH_29	410	18	2	0	0	SMN1_29	229	20	2	1	2	0	1	0	0
PAH_30	433	19	2	0	0	SMN1_30	230	19	2	0	2	0	0	0	1
PAH_31	429	19	2	0	0	SMN1_31	228	20	2	0	2	0	0	0	1
PAH_32	408	19	2	0	0	SMN1_32	174	14	2	0	1	0	0	0	0
PAH_33	408	19	2	0	0	SMN1_33	234	17	2	0	2	0	0	0	1
PAH_34	413	17	2	0	0	SMN1_34	224	16	2	0	2	0	1	0	0
PAH_35	410	18	2	0	0	SMN1_35	236	16	2	0	2	0	1	0	0
PAH_36	408	19	2	0	0	SMN1_36	218	17	2	0	2	0	1	0	0
PAH_37	405	18	2	0	0	SMN1_37	222	17	2	0	2	0	1	0	0
PAH_38	405	18	2	0	0	SMN1_38	234	17	2	0	2	0	1	0	0
PAH_39	405	18	2	0	0	SMN1_39	221	19	1	0	2	0	0	0	1
PAH_40	404	18	2	0	0	SMN1_40	233	19	1	0	2	0	0	0	1
PAH_41	405	18	2	0	0	SMN1_41	221	19	1	0	2	0	0	0	1
PAH_42	405	18	2	0	0	SMN1_42	229	21	1	0	2	0	0	0	1
PAH_43	405	18	2	0	0	SMN1_43	222	21	2	0	1	1	0	0	0
PAH_44	403	19	2	0	0	SMN1_44	232	14	1	1	1	1	0	1	0
PAH_45	405	18	2	0	0	SMN1_45	227	17	1	0	3	0	1	0	0
PAH_46	417	19	2	0	0										
PAH_47	400	20	2	0	0										
PAH_48	395	19	2	0	0										
PAH_49	399	20	2	0	0										
PAH_50	387	20	2	0	0										
PAH_51	395	19	2	0	0										

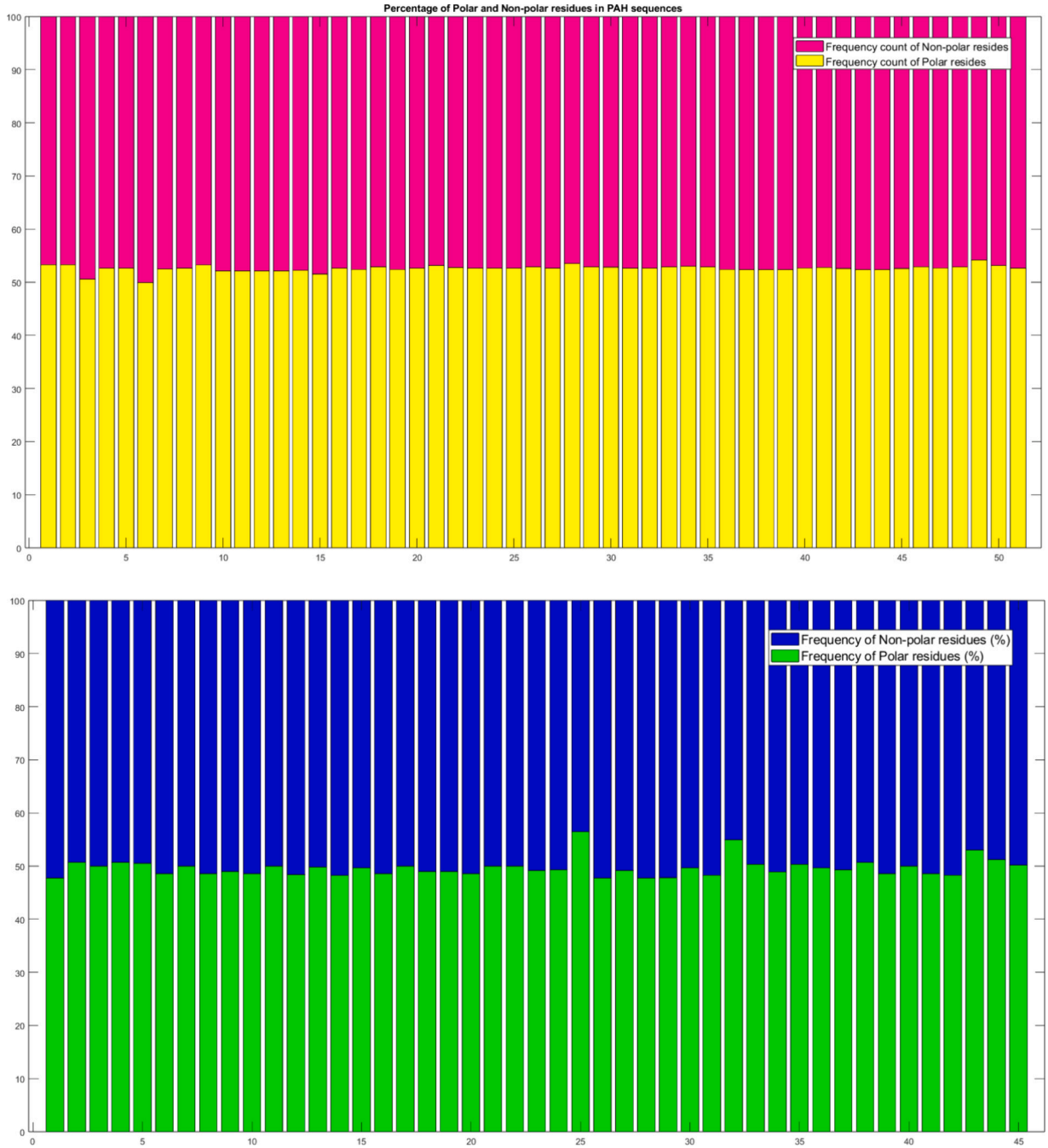


Fig. 7. Percentages of polar, non-polar residues in PAH (Left) and SMN1 (Right) proteins.

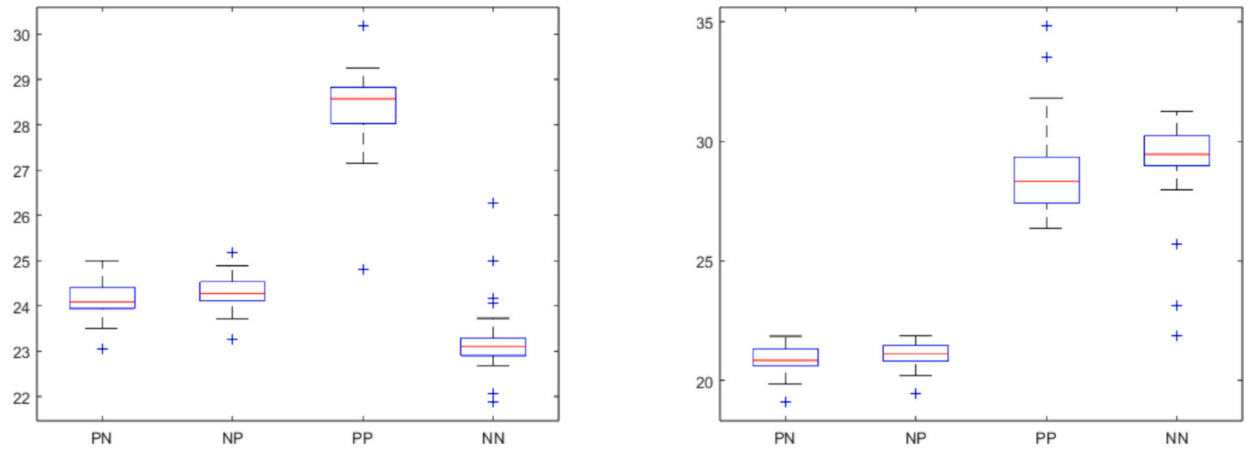


Fig. 8. Box-plot of the relative frequency of PN, NP, PP, and NN changes in PAH (Top) and SMN1 (Bottom) proteins.

other (D_O) and similarly rest twelve HF_D, HF_HF HF_MF, HF_O, MF_D, MF_HF, MF_MF, MF_O, O_D, O_HF, O_MF, and O_O. Such changes were accounted for the form of a sequence according to the spatial sequential arrangement of D, HF, MF, and O residues for each protein sequence [57]. The frequency of each of the sixteen changes was counted from these change response sequences [57].

3.8. Structural and physicochemical features of PAH and SMN1 proteins

Structural and physicochemical descriptors extracted from sequence data have been widely used to characterize sequences and predict structural, functional, expression, and interaction profiles of proteins. iFeature (I-features) and ProtrWeb, two versatile Python-based toolkits were deployed to extract structural and physicochemical properties of the PAH and SMN1 proteins [73].

A list of 1162 features extracted using I-features, which were deployed in the present study was summarized in the following Table 3 [73].

3.9. Formation of distance matrices and dendrograms

Euclidean distance was evaluated between feature vectors of all pairs of PAH/SMN1 protein sequences for each of the following six features: relative frequency of amino acids (dimension 20), the relative frequency of changes obtained from polar-nonpolar profiles (dimension 4), the relative frequency of changes obtained from acidic-basic-neutral profiles (dimension 9), the relative frequency of changes obtained from disordered, highly flexible, moderately flexible, and other residues (dimension 16), structural and physicochemical features (I-feature (dimension 1162) and ProtrWeb (dimension 8710)) [74]. Each of the structural and physicochemical features was normalized in the range of 0 to 100.

Sequence homology-based similarity matrices were obtained from Clustal Omega. Distance matrices were obtained by subtracting each entry of the similarity matrices from 100.

Each feature produces a distance matrix of dimension 51 × 51 (for PAH sequences)/45 × 45 (for SMN1 sequences). Different color thresholds (empirically chosen) were used in different dendrograms. If the color threshold had the value T, then each group of nodes whose linkage was less than T was assigned a unique color in the dendrogram and each color corresponds to a single cluster.

3.10. Agglomerated proximal sets of PAH and SMN1 proteins

A set of sequences 'S' (PAH/SMN1) is called an *'agglomerated proximal set'* if every sequence of the set 'S' belongs to the same cluster for each of the seven dendrograms.

Mathematically, let $\{c_{i,1}, c_{i,2}, c_{i,3}, \ldots c_{i,k}\}$ be a set of k clusters corresponding to an aforementioned feature f_i, where $i = 1,2,\ldots 7$. Note that f_1 stand for the relative frequency of amino acids, f_2 stands for sequence homology, f_3 stands for the relative frequency of changes based on polar, non-polar residues, f_4 stands for the relative frequency of changes based on acidic, basic, and neutral residues, f_5 stands for change response based on intrinsic protein disordered regions, f_6 stands for structural and physicochemical features derived from I-features, and f_7 stands for structural and physico- chemical features derived from ProtrWeb.

A set $\mathcal{S} = \{s_1, s_2, s_3, \ldots, s_{m(m\geq 2)}\}$ is said to be an 'agglomerated proximal set ' if, for some $p \in \{1,2,\ldots k\}$, $\mathcal{S} \subseteq c_{i,p}$ for all $i = 1,2,\ldots 7$. Here, s_j stands for a sequence of PAH/SMN1.

Following this definition, agglomerated proximal sets of PAH and SMN1 protein sequences were derived.

4. Results and analyses

4.1. Similitude and dissimilitude of PAH and SMN1 proteins based on amino acid frequency

4.1.1. Compositional profile of PAH/SMN1 proteins

Characterized by noticeable differences, it has been shown that disordered proteins/regions are significantly depleted in bulky hydrophobic amino acid residues (I, L, and V) and aromatic amino acids (W, Y, F, and H), which are often involved in the formation of the hydrophobic core of a folded globular protein. Disordered proteins/regions also exhibit a low content of C, N, and M residues. These amino acids, namely C, W, I, Y, F, L, H, V, N, and M, which are depleted in disordered proteins and regions, are defined as order-promoting amino acids. On the other hand, disordered proteins and regions are substantially enriched in disorder-promoting amino acids, such as R, T, D, G, A, K, Q, S, E, and P [49,66,75–77]. These biases in the amino acid composition can be visualized using a web-based tool, the Composition Profiler, for the semi-automatic discovery of enrichment or depletion of amino acids in query proteins [49].

Amino acid composition profile analysis of 51 PAH sequences revealed that out of the ten order-promoting residues, four (W, V, N, and M) were significantly depleted in the PAH proteins, whereas four disorder-promoting residues (R, Q, E, and P) were significantly enriched (Fig. 1).

Analysis of the amino acid composition profile for 45 SMN1 sequences unveiled that, among the ten order-promoting residues, six (I, Y, F, L, H, and V) were significantly depleted in SMN1 proteins, while five disorder-promoting residues (D, G, K, S, and P) were significantly enriched (Fig. 1).

Both PAH and SMN1 proteins exhibit a significant depletion of

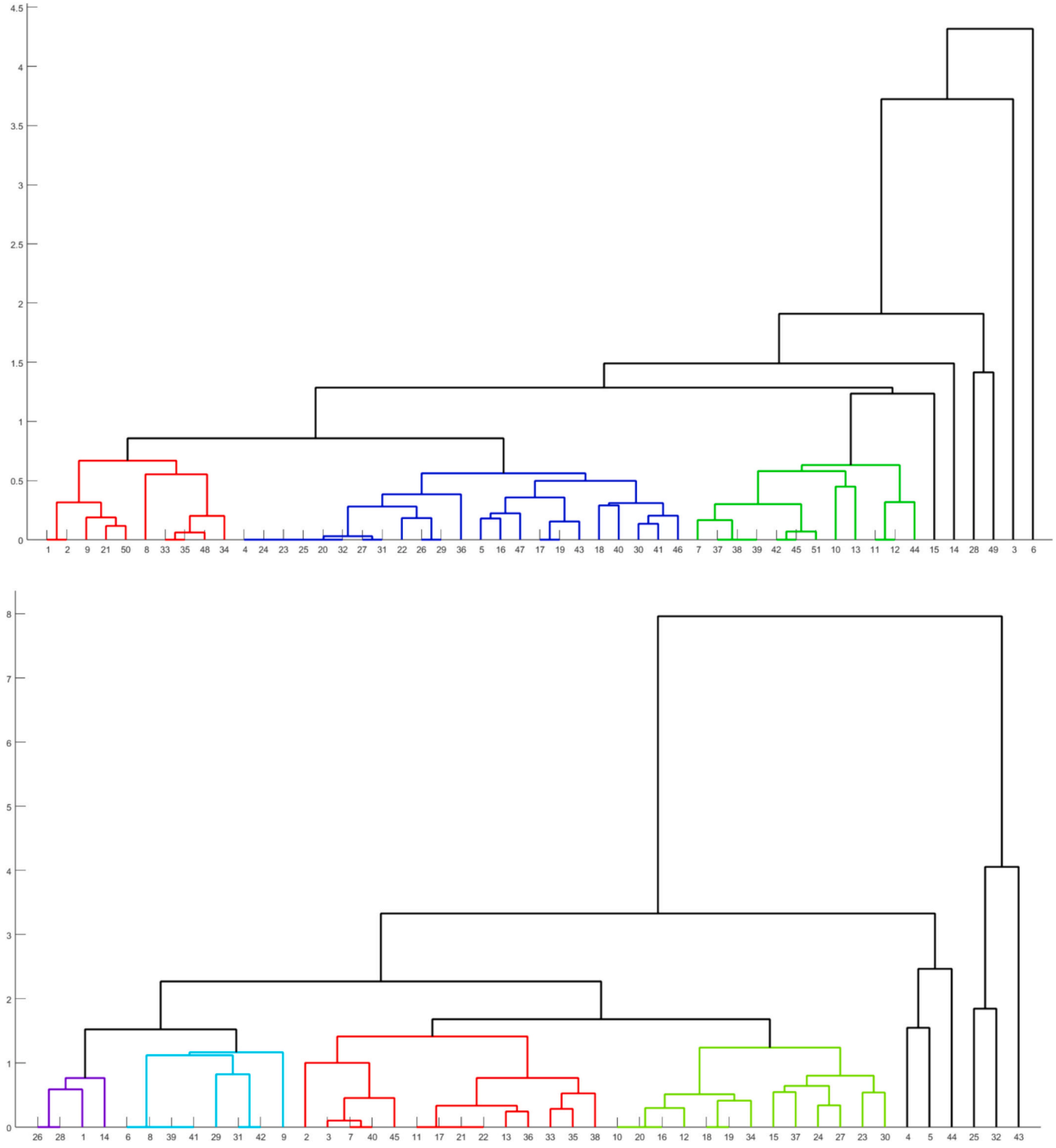


Fig. 9. Phylogenetic relationship among the PAH (Top) and SMN1 proteins (Bottom) based on the relative frequency of PP, NP, PP, and NN changes as obtained from polar, non-polar profiles. Note that every color (other than black) in the phylogeny designates a cluster.

Table 10
Clusters of PAH and SMN1 proteins based on the relative frequency of 'PN', 'NP', 'PP', and 'NN' changes as obtained from Polar, non-polar profiles.

Clusters	Sequences (PAH)
Cluster-1	{1, 2, 9, 21, 50, 8, 33, 35, 48, 34}
Cluster-2	{4, 24, 23, 25, 20, 32, 27, 31, 22, 26, 29, 36, 5, 16, 47, 17, 19, 43, 18, 40, 30, 41, 46}
Cluster-3	{7, 37, 38, 39, 42, 45, 51, 10, 13, 11, 12, 44}

Clusters	Sequences (SMN1)
Cluster-1	{26, 28, 1, 14}
Cluster-2	{6, 8, 39, 41, 29, 31, 42, 9}
Cluster-3	{2, 3, 7, 40, 45, 11, 17, 21, 22, 13, 36, 33, 35, 38}
Cluster-4	{10, 20, 16, 12, 18, 19, 34, 15, 37, 24, 27, 23, 30}

certain order-promoting residues, but the specific amino acids involved differ between the two proteins. Similarly, both proteins show a significant enrichment of disorder-promoting residues, again with differences in the specific amino acids affected.

4.1.2. Relative frequency of amino acids and associated phylogenetic relationship

The analysis of relative amino acid frequencies in PAH and SMN1 sequences, depicted in histograms in Fig. 2, provides valuable insights. Notably, Leu is consistently abundant in all PAH sequences, exceeding a remarkable 10 % frequency. In contrast, Pro stands out as the most frequent amino acid in the majority of SMN1 sequences, except for two instances (SMN_25 and SMN_32, both from the GORGO group) where Ser takes the lead. Within PAH sequences, Met records the lowest frequency, except for four specific sequences (PAH_4, PAH_10, PAH_11, and PAH_12) where Trp assumes the lowest frequency position. Notably, the average frequencies of both Met and Trp in PAH sequences remain under 1 %. In nearly all SMN1 sequences, His consistently displays the lowest frequency, with Met and Cys occupying the lowest frequency positions in SMN_1 and SMN_43, respectively.

These findings unveil distinctive amino acid composition patterns in PAH and SMN1 proteins, reflecting potential structural and functional implications. The prevalence of Leu in PAH sequences and the dominance of Pro in most SMN1 sequences underscore the significance of these amino acids in the respective proteins. Moreover, the variations in amino acid frequencies within specific sequences provide valuable insights into the diversity and unique characteristics of PAH and SMN1 proteins.

Utilizing a distance threshold of 1.3, a total of 51 PAH sequences were grouped into five distinct clusters, as illustrated in the dendrogram (Fig. 3 (Top)). Furthermore, two PAH sequences, namely, PAH_3 and PAH_6 were distant from most the PAH sequences. Similarly, with a distance threshold of 1.87, 45 SMN1 sequences were partitioned into six clusters with SMN_43, SMN_25, and SMN_32 as outliers (Fig. 3 (Bottom)). Notably, the largest clusters, referred to as Cluster-2 in both PAH and SMN1 proteins, consisted of 25 and 13 sequences, respectively, as highlighted in Table 5.

4.1.3. Shannon entropy of receptors

The Shannon entropy (SE) values for PAH and SMN1 sequences were grouped into six and five categories, respectively, based on their identical SEs, as detailed in Table 6. Furthermore, the SE values are notably clustered around 4.14 ± 0.011 for PAH sequences and 4.08 ± 0.011 for SMN1 sequences, as depicted in Fig. 4. These observations indicate that the degree of disorderliness in amino acid frequencies is significantly high for both PAH and SMN1 sequences, as the SE values approach the maximum possible value of 4.322. This high level of disorderliness suggests that the amino acid composition within almost all PAH and SMN1 sequences is characterized by substantial variability and diversity, which may have implications for their structural and functional properties.

4.2. Sequence homology and per-residue Shannon variability

4.2.1. Sequence homology-based phylogenetic relationships and invariant residues

Using a distance threshold of 3.5, PAH sequences were grouped into five distinct clusters, depicted in Fig. 5 (Up). PAH_6 stood out as an outlier. Simultaneously, SMN1 sequences were partitioned into three clusters and SMN_43 became an outlier, as shown in Fig. 5 (Bottom). Notably, the most extensive clusters, identified as Cluster-2 for PAH and Cluster-3 for SMN1, contained 15 and 21 sequences, respectively, as detailed in Table 7.

Invariant residues in PAH(SMN1) sequences were determined through a multiple sequence alignment using Clustal Omega, referencing the tr|A0A8C9KM68|PAH_PANTA (sp|O02771|SMN_CANLF) sequence (see Table 8). We identified 23, 20, and 8 invariant residues of lengths 1, 2, and 3 in the PAH sequences, respectively. Notably, the longest invariant residue, spanning from amino acid position 88 to 136, had a length of 49 within the PAH sequences. Conversely, in the SMN1 sequences (Table 8), the longest invariant residue observed had a length of 5. It was noted that 66.2 %, 63.8 % and 2.4 % residues were invariant in the regulatory N-terminal domain (amino acids 1–142), Catalytic central domain (amino acids 143–410), and oligomerization domain C-terminal amino acids 411–452), respectively.

Invariant residues in PAH (SMN1) sequences across various organisms often correspond to critical functional regions of a protein [78]. These residues are involved in catalytic activity, ligand binding, and so on. As such, they can serve as signatures that define the sequence's function. These invariant residues are conserved through evolution, because any mutation in these positions may disrupt the essential functions of the protein [78]. This conservation over time suggests the importance of these residues and can serve as a signature of evolutionary selection for that particular sequence. Invariant residues can be used as diagnostic markers in disease studies [79].

4.2.2. Shannon variability of amino acid residue positions

The Shannon variability analyses revealed that an impressive 94.9 % of residues in PAH variants and a substantial 90.88 % of residues in SMN1 variants were highly conserved (as depicted in Fig. 6). This high degree of conservation underscores the significance of these conserved residues in the structural and functional aspects of both PAH and SMN1 proteins. Such a level of conservation suggests that these residues likely play critical roles in maintaining the stability and functionality of these proteins, which could have implications for understanding their biological functions and the potential consequences of genetic variations in these conserved regions.

4.3. Homogeneous poly-string frequency of amino acids

The maximum length of a homogeneous poly-string was found to be 8 and 10 considering all amino acids across all PAH and SMN1 sequences, respectively. No poly-string of lengths 5, 6, or 7 was present in any of the 51 PAH sequences and no poly-string of length 7 was present in any of the 45 SMN1 sequences (Table 9). The frequencies of homogeneous poly-strings for each of the twenty amino acids separately are provided in Supplementary file 1.

In the analysis of PAH sequences, a distinct pattern emerges regarding the lengths and amino acid compositions of homogeneous poly-strings. Among all PAH sequences, only PAH_3 had one poly-string of length 4 (of Arg) and another of length 8 (of Ala), whereas the remaining 50 PAH sequences exhibit poly-strings of maximum length 3 underscoring a consistent structural feature. Furthermore, it was noticed that only PAH_3 possessed five poly-strings (of Ala, Pro, Glx, and Lys) of

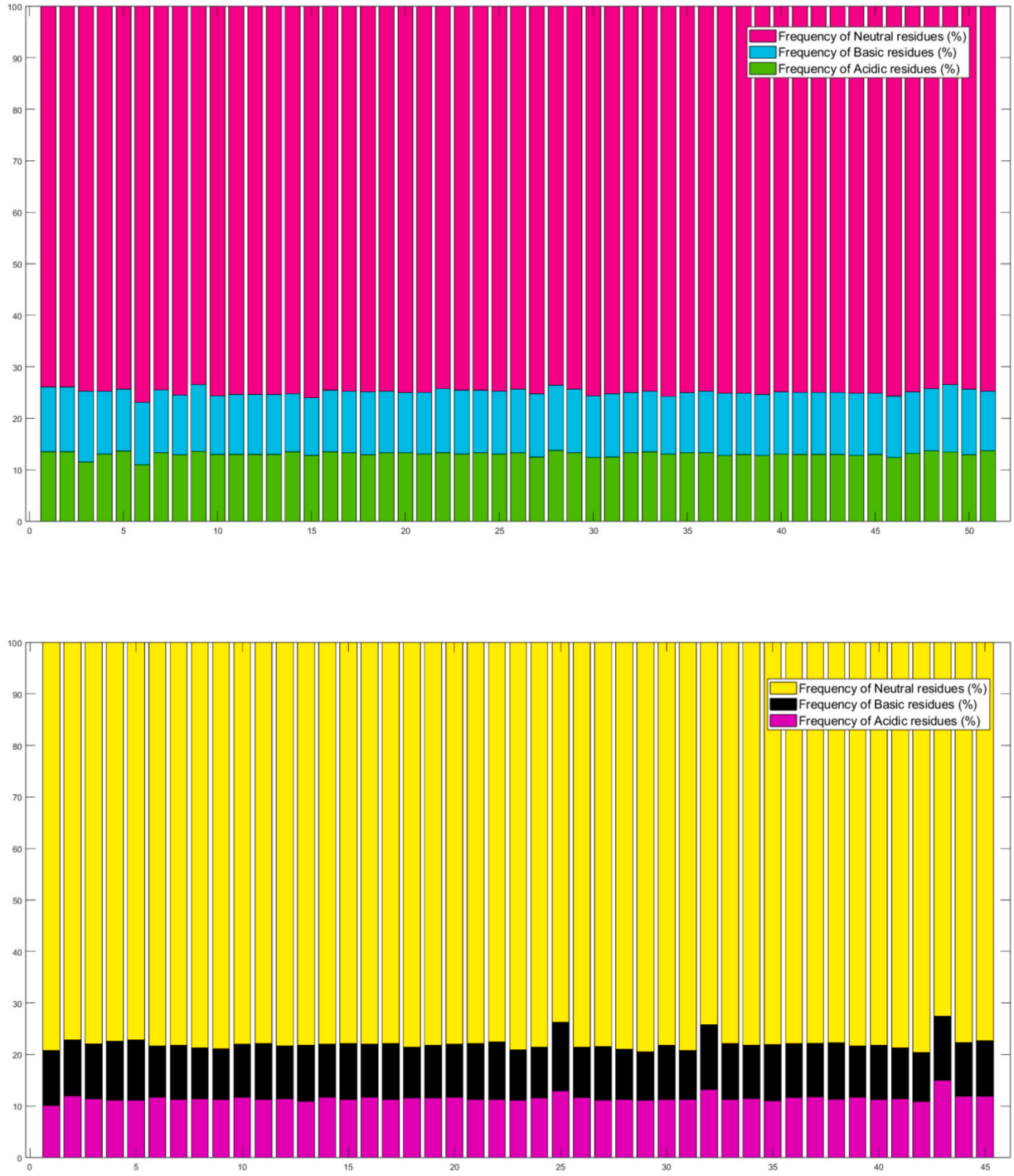


Fig. 10. Percentages of acidic, basic, neutral residues in PAH (Top) and SMN1 (Bottom) proteins.

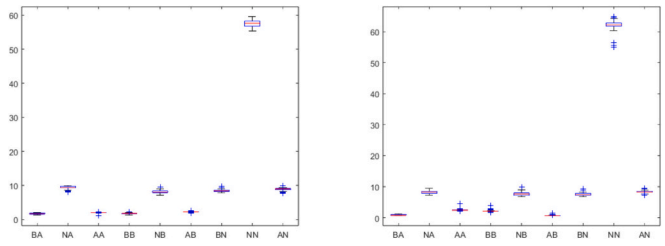


Fig. 11. Box-plot of the relative frequency of all nine changes in PAH (Left) and SMN1 (Right) proteins.

length 3, while each of the remaining 50 PAH sequences contained two poly-strings (one of Glx and another of Lys) of length 3. Unique characteristic of PAH_3 suggests a potential functional significance of this particular sequence. None of the PAH had a poly-string of length 2 consisting of Cis, His or Val except PAH_47, PAH_21, and PAH_14, which had one poly-string of length 2 of Cys, His, and Val, respectively. None of the PAH had poly-string of length 2 containing Asp or Met or Asn or Trp. This may imply a selective constraint or functional constraint in the amino acid composition of these poly-strings. All 51 PAH sequences contained a single occurrence of FF, II, and YY (poly-strings of Phe, Ile, and Tyr with length 2). KK (poly-strings of Lys with length 2) was present with frequency 2 in all PAH, except PAH_49 which had frequency 3

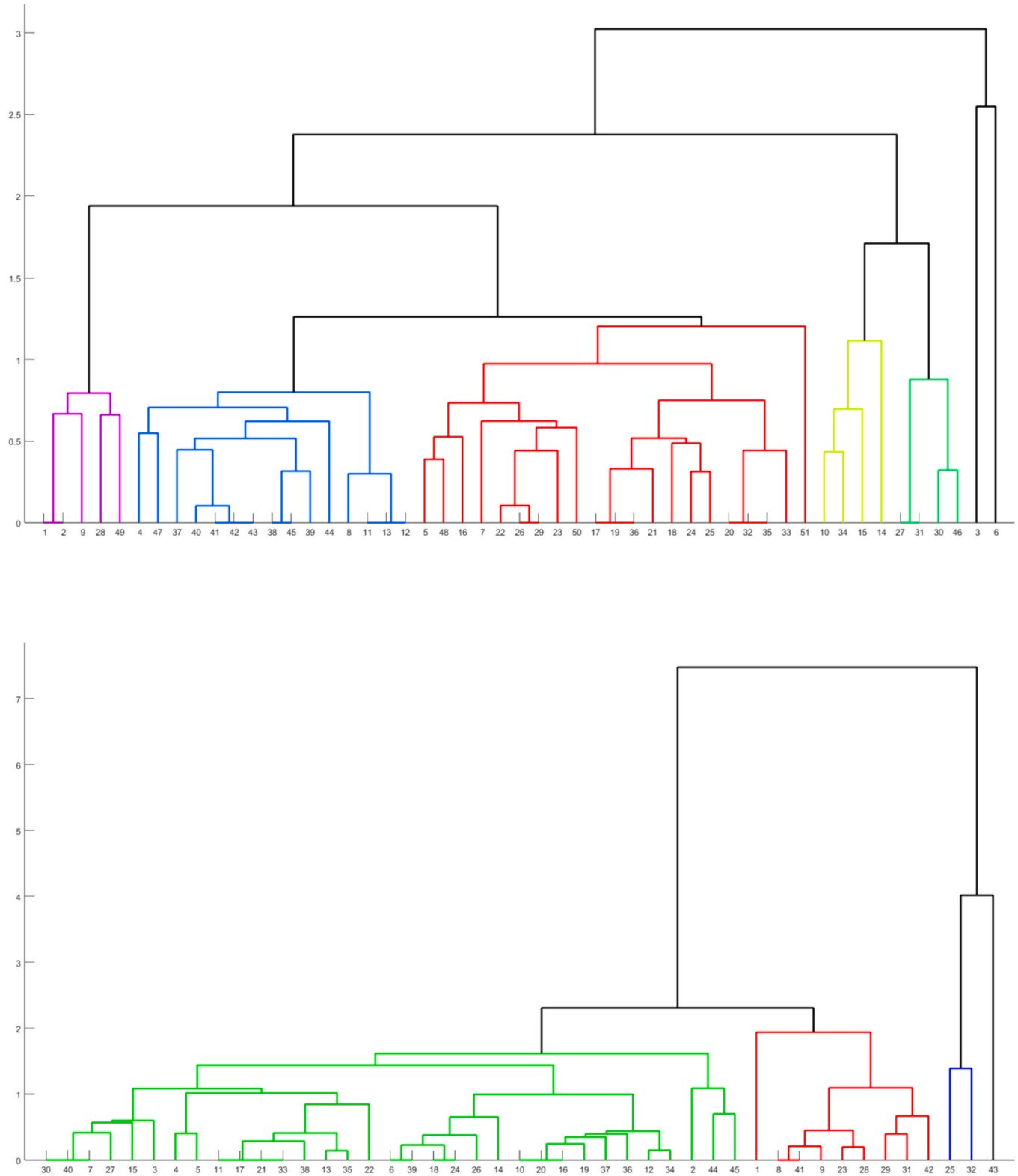


Fig. 12. Phylogenetic relationship among the PAH (Top) and SMN1 (Bottom) proteins based on the relative frequency of BA, NA, AA, BB, NB, AB, BN, NN, and AN changes as obtained from acidic, basic, and neutral profiles. Note that every color (other than black) in the phylogeny designates a cluster.

Table 11
Clusters of PAH and SMN1 proteins based on the relative frequency of BA, NA, AA, BB, NB, AB, BN, NN, and AN changes as obtained from acidic, basic, and neutral profiles.

Clusters	Sequences (PAH)
Cluster-1	{1, 2, 9, 28, 49}
Cluster-2	{4, 47, 37, 40, 42, 43, 38, 45, 39, 44, 8, 11, 13, 12}
Cluster-3	{5, 48, 16, 7, 22, 26, 29, 23, 50, 17, 19, 36, 21, 18, 24, 25, 20, 32, 35, 33, 51}
Cluster-4	{10, 34, 15, 14}
Cluster-5	{27, 31, 30, 46}

Clusters	Sequences (SMN1)
Cluster-1	{30, 40, 7, 27, 15, 3, 4, 5, 11, 17, 21, 33, 38, 13, 35, 22, 6, 39, 18, 24, 26, 14, 10, 20, 16, 19, 37, 36, 12, 34, 2, 44, 45}
Cluster-2	{1, 8, 41, 9, 23, 28, 29, 31, 42}
Cluster-3	{25, 32}

of the same. These findings showcase a consistent pattern across most sequences.

Turning attention to the analysis of SMN1 sequences, a striking dominance of Pro in poly-strings of various lengths is evident. 29 out of 45 SMN1 sequences had single poly-string of length 10 and all of them comprised of proline. Poly-string of length 9 appeared in SMN1_3 and SMN1_44 only (with frequency 1 consisting of proline) and in no other SMN1. Additionally, 10 out of 45 SMN1 sequences had single poly-string of length 8, and all of them were composed of Pro. Sequences that had poly-string of length 10 did not have poly-strings of length 8. SMN1_4, SMN1_43, and SMN1_44 only had poly-string of length 6 (with frequency 1 consisting of Pro, Pro, and Gly, respectively) and no other SMN1 displayed poly-string of length 6. Except for SMN1_45, all SMN1 possessed poly-strings of length 5 with frequency 1 or 2 and made of Pro. Along with Pro, SMN1_45 had another poly-string of length 5 consisting of Gly. Out of 45, only 6 SMN1 sequences (SMN1_4, SMN1_26, SMN1_27, SMN1_28, SMN1_29, and SMN1_44) had poly-string of length 4 (all with frequency 1 and composed of Pro). Single occurrences of VVV (poly-string of Val with length 3) were noticed in all sequences except SMN1_43 in which VVV was absent. SMN1_43 had two poly-strings (one of Ala and the other of Glx) of length 3 and among the rest 44 SMN1 sequences, 35 had a single occurrence of GGG (poly-string of Gly with length 3) along with VVV. WW and VV (poly-string of Trp and Val with length 2) were not identified in any of the 45 SMN1 sequences except SMN1_43 which possessed VV with frequency 2. The unique features in SMN1_43 imply sequence-specific variations that may contribute to distinct structural or functional properties. All SMN1 sequences had a single occurrence of EE and AA (poly-string of Glx and Ala with length 2) underscoring shared structural elements among all SMN1 sequences.

4.4. Polar, non-polar residue profiles of PAH/SMN1 proteins

The percentage distributions of polar and non-polar residues were computed for each PAH Fig. 7(Top) and SMN1 Fig. 7 (Bottom) sequence. The analysis revealed that the average ratio of polar to non-polar residues in PAH sequences was 1.11 ± 0.03 and ranged from 0.99 to 1.18. The average ratio of polar to non-polar residues in SMN1 sequences was 0.99 ± 0.072 and ranged from 0.91 to 1.29. It was noteworthy that the number of polar and non-polar residues was almost identical in each PAH and SMN1 sequence. This uniformity in the frequency of polar and non-polar residues underscores a specific compositional characteristic shared among these sequences.

4.4.1. Change response sequences of polar, and non-polar profiles

Fig. 8 (Left) illustrates the relative frequency distribution of four distinct changes: 'PN', 'NP', 'PP', and 'NN' within the PAH sequences. Notably, the median percentage of 'PP' was the highest among the four categories as PP for all PAH sequences range from 27.15 % to 30.18 %, except PAH_6. It's interesting to observe that among these sequences, PAH_28 exhibited the lowest percentage of 'NN' changes at 21.88 %, while two sequences, PAH_6 and PAH_3, displayed significantly high percentages of 'NN' changes (exceeding 25 %) implying distinct functional or structural features in these sequences. The majority of PAH sequences exhibited a high percentage of polar-to-polar residue changes, indicating a propensity for the conservation of polar interactions within these sequences.

On the other hand, in the case of SMN1 sequences, the relative frequency distribution of the same four changes ('PN', 'NP', 'PP', and 'NN') reveals that the median percentage of 'NN' was the highest indicating a preference for conserving non-polar interactions within this set (see Fig. 8 (Right)). SMN1_25 and SMN1_32 showed the lowest 'NN' changes at 21.88 % and 23.11 %, respectively, possibly reflecting a specific evolutionary adaptation in these sequences. Additionally, percentages of 'PN' and 'NP' changes were comparatively low for two sequences, SMN1_5 and SMN1_4, as indicated by their outlier status in the box plot (see Fig. 8 (Right)).

The analyses of residue changes provide insights into the conservation and variation of specific interactions within the PAH and SMN1 sequences. These observations can be valuable for understanding the functional and structural characteristics of these sequences and may guide further investigations into their biological significance.

Based on a distance threshold of 0.7, 51 PAH sequences formed three clusters (Fig. 9 (Up)), and 45 SMN1 sequences formed four clusters (Fig. 9 (Bottom)) based on the distance threshold of 1.5. The largest clusters (cluster-2 for PAH and cluster-3 for SMN1) had 23 and 14 sequences, respectively, as noted in Table 10. PAH_3 and PAH_6 turned out to be distant from the rest of the PAH sequences while SMN_25, SMN_32, and SMN_43 became outliers among SMN1 sequences.

4.5. Acidic, basic, neutral residue-based phylogenetic relationship

Based on the data presented in Fig. 10, the analysis involved the calculation of the percentages of acidic, basic, and neutral residues for each PAH and SMN1 protein. This analysis encompassed 51 PAH sequences and 45 SMN1 sequences, revealing varying percentages of neutral residues ranging from 73.48 % to 76.88 % for PAH and from 72.59 % to 79.59 % for SMN1. The calculated ratio of acidic to basic residue percentages within the PAH sequences was determined to be 1.08 ± 0.06, while for SMN1 sequences, it was 1.09 ± 0.071. This indicates a relatively balanced distribution between acidic and basic residues in both PAH and SMN1 sequences.

4.6. Change response sequences of acidic, basic, and neutral profiles

The relative frequency distribution of nine different residue changes in PAH (SMN1) sequences, denoted as BA, NA, AA, BB, NB, AB, BN, NN, and AN, was visualized in Fig. 11 (Left (Right)). It was found that the percentage of 'NN' change of residues in all PAH (SMN1) sequences was high in comparison to the other eight changes with values ranging from 55.36 % to 59.56 % (from 54.9 % to 64.8 %). Furthermore, it was observed that the percentage of acidic to basic residue changes was lowest in all SMN1 sequences, whereas the percentage of basic to basic (basic to acidic in some sequences) was found to be the minimum among all.

Under a distance threshold of 1.25 (2), the analysis led to the formation of five distinct clusters from PAH and three from SMN1 sequences. This information is visually depicted in the dendrogram (Fig. 12 (Top)) and summarized in Table 11. Notably, the largest cluster for PAH contained 21 sequences, as listed as Cluster-3 in Table 11.

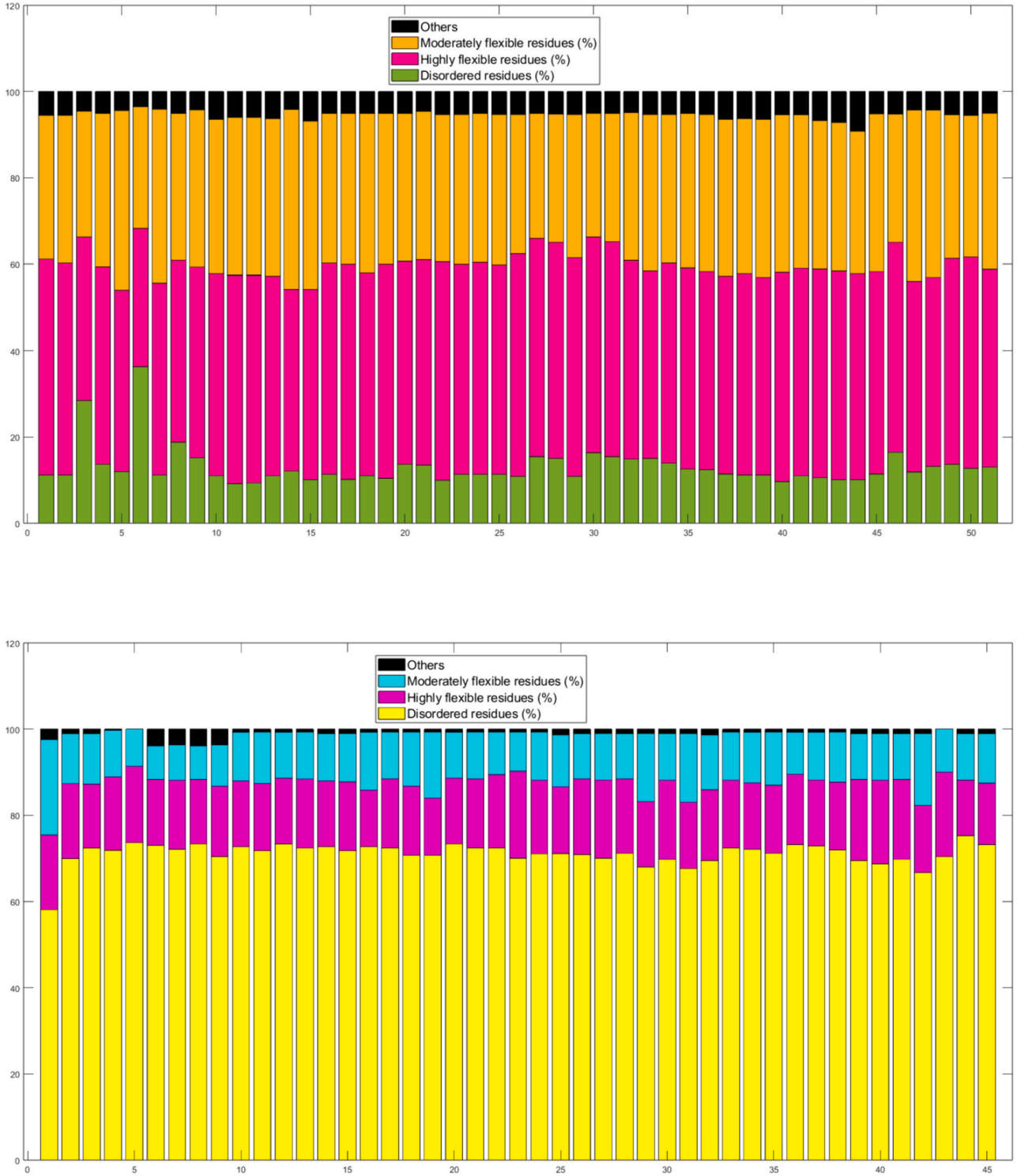


Fig. 13. Percentages of disordered, highly flexible, moderately flexible, and other residues in PAH (Top) and SMN1 (Bottom) proteins.

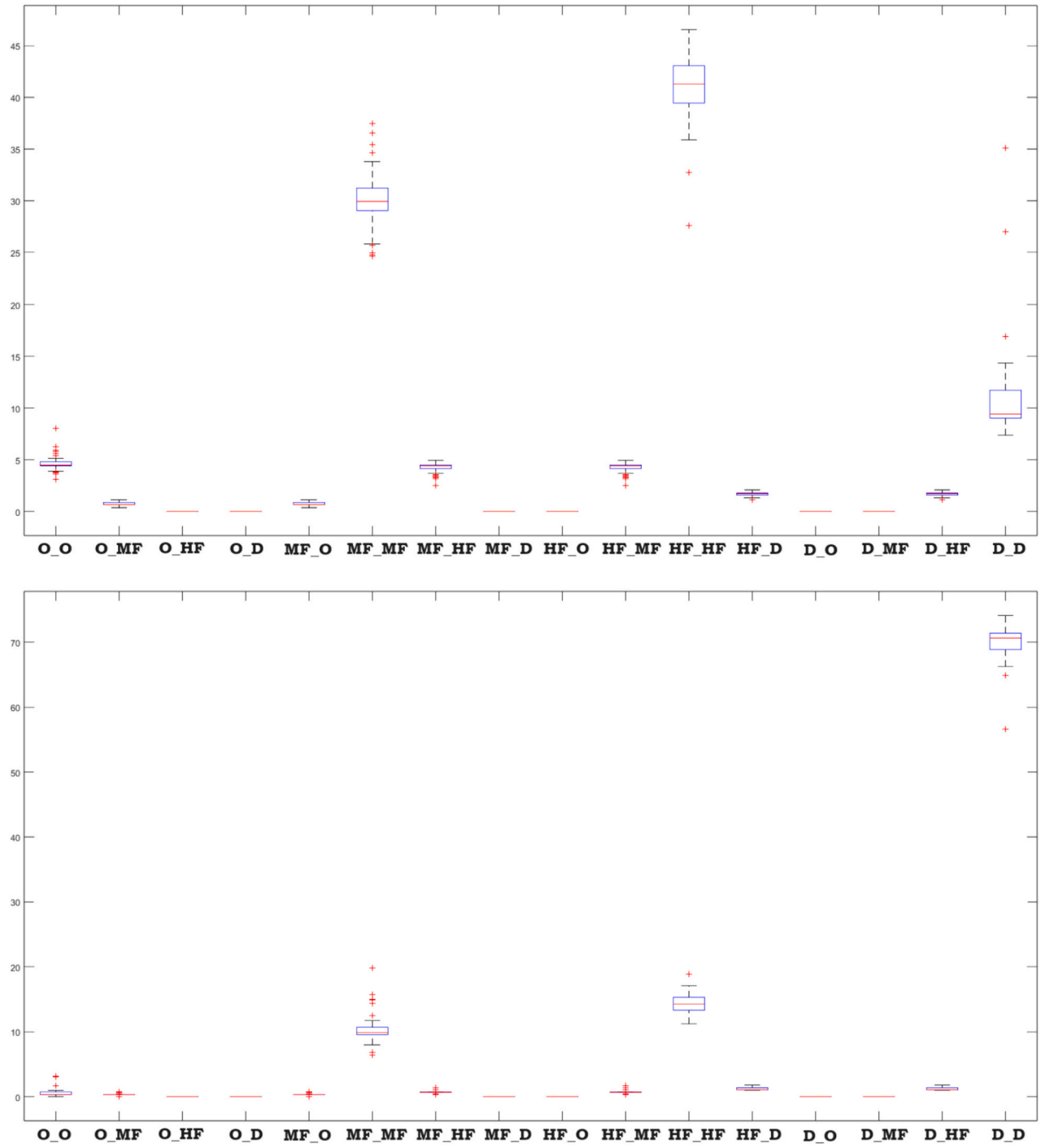


Fig. 14. Box-plot of the percentages of disordered, highly flexible, moderately flexible, and other residues changes in PAH (Top) and SMN1 (Bottom) protein sequences.

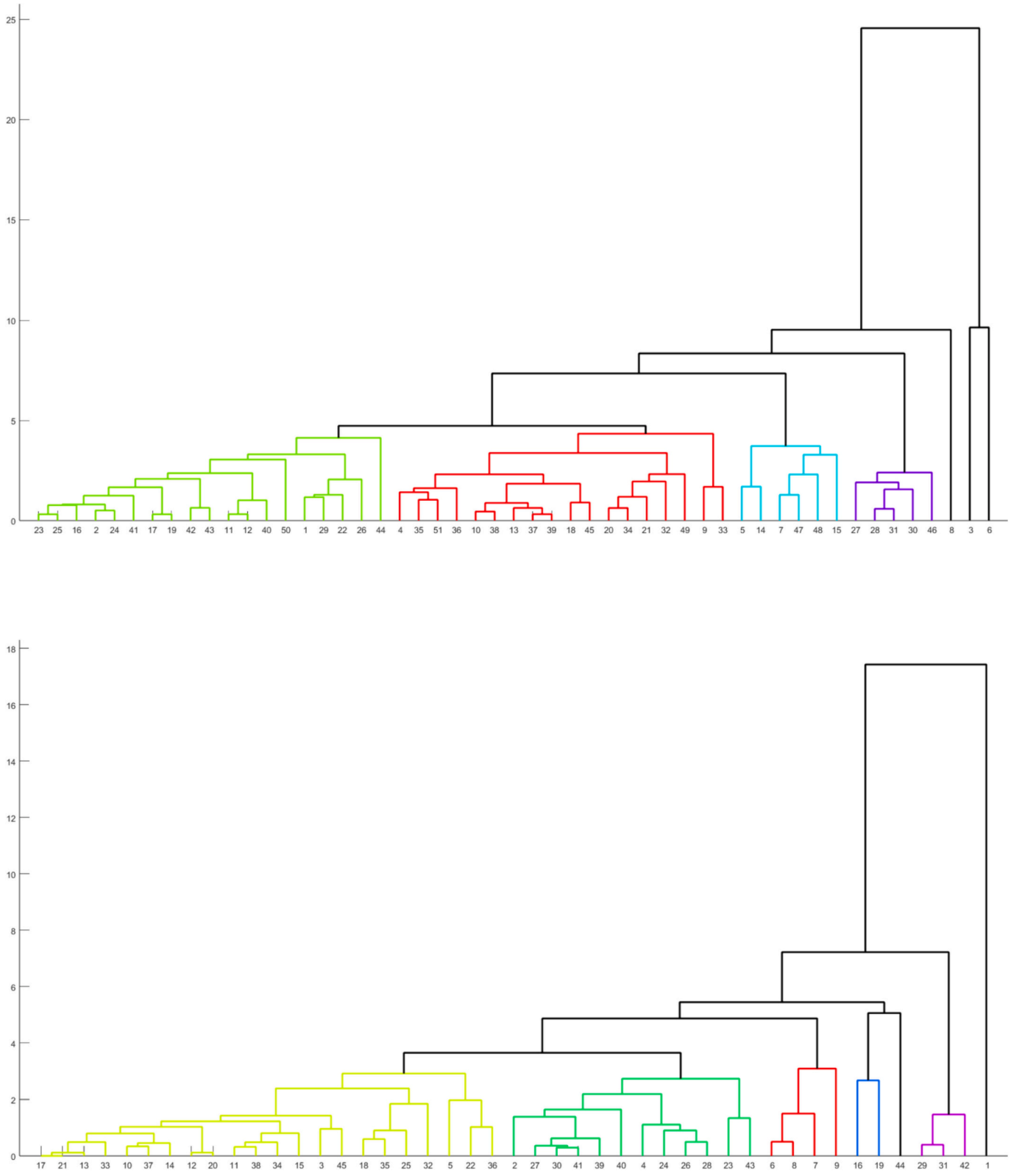


Fig. 15. Phylogenetic relationship among the PAH (Top) and SMN1 (Bottom) proteins based on percentages of disordered, highly flexible, moderately flexible, and other residues. Note that every color (other than black) in the phylogeny designates a cluster.

Table 12
Clusters of PAH and SMN1 proteins based on the change response profiles of disordered, highly flexible, moderately flexible, and other residues in PAH and SMN1 proteins.

Clusters	Sequences (PAH)
Cluster-1	{23, 25, 16, 2, 24, 41, 17, 19, 42, 43, 11, 12, 40, 50, 1, 29, 22, 26, 44}
Cluster-2	{4, 35, 51, 36, 10, 38, 13, 37, 39, 18, 45, 20, 34, 21, 32, 49, 9, 33}
Cluster-3	{5, 14, 7, 47, 48, 15}
Cluster-4	{27, 28, 31, 30, 46}

Clusters	Sequences (SMN1)
Cluster-1	{17, 21, 13, 33, 10, 37, 14, 12, 20, 11, 38, 34, 15, 3, 45, 18, 35, 25, 32, 5, 22, 36}
Cluster-2	{2, 27, 30, 41, 39, 40, 4, 24, 26, 28, 23, 43}
Cluster-3	{6, 8, 7, 9}
Cluster-4	{16, 19}
Cluster-5	{29, 31, 42}

Among the three clusters, cluster-1 and cluster-3 contain the highest (33) and lowest (2) number of SMN1 sequences, respectively. Similar to previous dendrograms, PAH_3 and PAH_6 became outliers among PAH sequences, while SMN_25, SMN_32, and SMN_43 were found to be distant from the rest of the SMN1 sequences.

4.7. Intrinsic protein disorder analysis

In all PAH sequences (Fig. 13 (Top)), a prominent presence of highly flexible residues was observed, with an average percentage of 46.74 ± 3.43. The percentage of disordered residues within each PAH sequence was 13.03, with a standard deviation of 4.54 while the percentage of moderately flexible residues varied between 28.25 and 41.59.

A predominant part of SMN1 sequences (ranging from 58.13 % to 71.14 %) was identified as the disordered region (Fig. 13 (Bottom)). The percentage of highly flexible residues in SMN1 sequences had a mean of 16.24 with a standard deviation of 1.6.

4.7.1. Change response sequences of disordered, highly flexible, moderately flexible, and other residue profiles

In Fig. 14 we present the relative frequency distribution of sixteen changes derived from the change response sequence based on the intrinsic disorder profiles of PAH and SMN1 sequences. Notably, we observed that the following transitions did not occur in any of the PAH or SMN1 sequences: O_HF, O_D, MF_D, HF_O, D_O, and D_MF. This absence of transitions between these specific states is of particular significance.

It is worth mentioning that each PAH protein sequence exhibited the highest proportions of changes from highly flexible to highly flexible residues (HF_HF), with a wide range varying from 27.57 % to 46.56 %. The next highest percentage of changes involved moderately flexible to moderately flexible (self-transitions) residues in PAH sequences, ranging from 24.63 % to 37.47 %. This confirms the prevalent presence of consecutive highly flexible residues as well as moderately flexible residues in PAH protein sequences.

Each SMN1 sequence displayed the highest proportions of changes involving disordered to disordered residues (D_D), with a range spanning from 56.59 % to 74.12 %. Furthermore, it was observed that the percentages of transitions involving moderately flexible (highly flexible) to moderately flexible (highly flexible) residues in SMN1 sequences varied from 6.406 (11.18) to 19.79 (18.919).

Utilizing a distance threshold of 4.5, PAH (SMN1) sequences led to the emergence of four (five) distinct clusters, clearly visible in the dendrogram (Fig. 15), summarized in Table 12. Notably, the two most substantial clusters, denoted Cluster-1 and Cluster-2, encompassed 19 and 18 PAH sequences, respectively (Table 12). The largest cluster (Cluster-1) comprised 22 SMN1 sequences (Fig. 15 (Bottom)). Here PAH_3 and PAH_6 were outliers among PAH sequences, but SMN_1 was found to be distant from the rest of the SMN1 sequences contrary to previous cases.

4.8. Phylogenetic relationship based on structural and physicochemical features

4.8.1. I-features

Through a comprehensive analysis using a distance threshold of 7.5 (6), we unveiled the presence of six (three) distinct clusters among PAH (SMN1) sequences based on I-features. The delineation of these clusters is visually depicted in the dendrogram using distinct colors (Fig. 16 and concisely summarized in Table 13. Remarkably, for PAH sequences, the largest cluster (Cluster 1) encompassed 22 PAH sequences, and for SMN1, the largest cluster (Cluster 2) contained 16 SMN1 sequences as tabulated in 13. Similar to dendrograms based on the relative frequency of amino acid, PQ profile, and ABN profile, PAH_3 and PAH_6 were distant from most of the PAH sequences while SMN_25, SMN_32, and SMN_43 were away from the rest of SMN1 sequences.

4.8.2. ProtrWeb-features

The phylogenetic analysis, incorporating various ProtrWeb features as outlined in Section 3.8, resulted in the iden- tification of four clusters comprising a total of 43 PAH sequences, using a distance threshold of 10 (refer to Fig. 17). Notably, Cluster-1, the largest cluster, includes 28 of these PAH sequences, as detailed in Table 14. This suggests a significant level of similarity among these sequences, possibly indicating a common evolutionary origin or functional similarity. Similarly, 37 SMN1 sequences were grouped into six clusters using a distance threshold of 8, with Cluster-1 also containing the highest number of sequences, 12. This clustering pattern implies shared evolutionary relationships or functional characteristics among the SMN1 sequences within Cluster-1. Furthermore, the identification of PAH_3 and PAH_6 as outliers, along with SMN_43 being distinct from the rest of the SMN1 sequences, highlights the presence of unique sequence variations or functional divergence within these datasets, warranting further investigation.

In summary, structural and physicochemical analyses provide valuable insights into the structural relationships and disparities among the examined PAH and SMN1 sequences. Clustering analyses reveal evolutionary relationships and potential functional similarities, while the identification of structural features informs their roles in biochemical path- ways. Variations in physicochemical properties can signify differences in stability, solubility, and interaction capabilities, potentially linking these to unique functions or disease mechanisms in conditions like PKU and SMA. Additionally, these analyses can identify potential

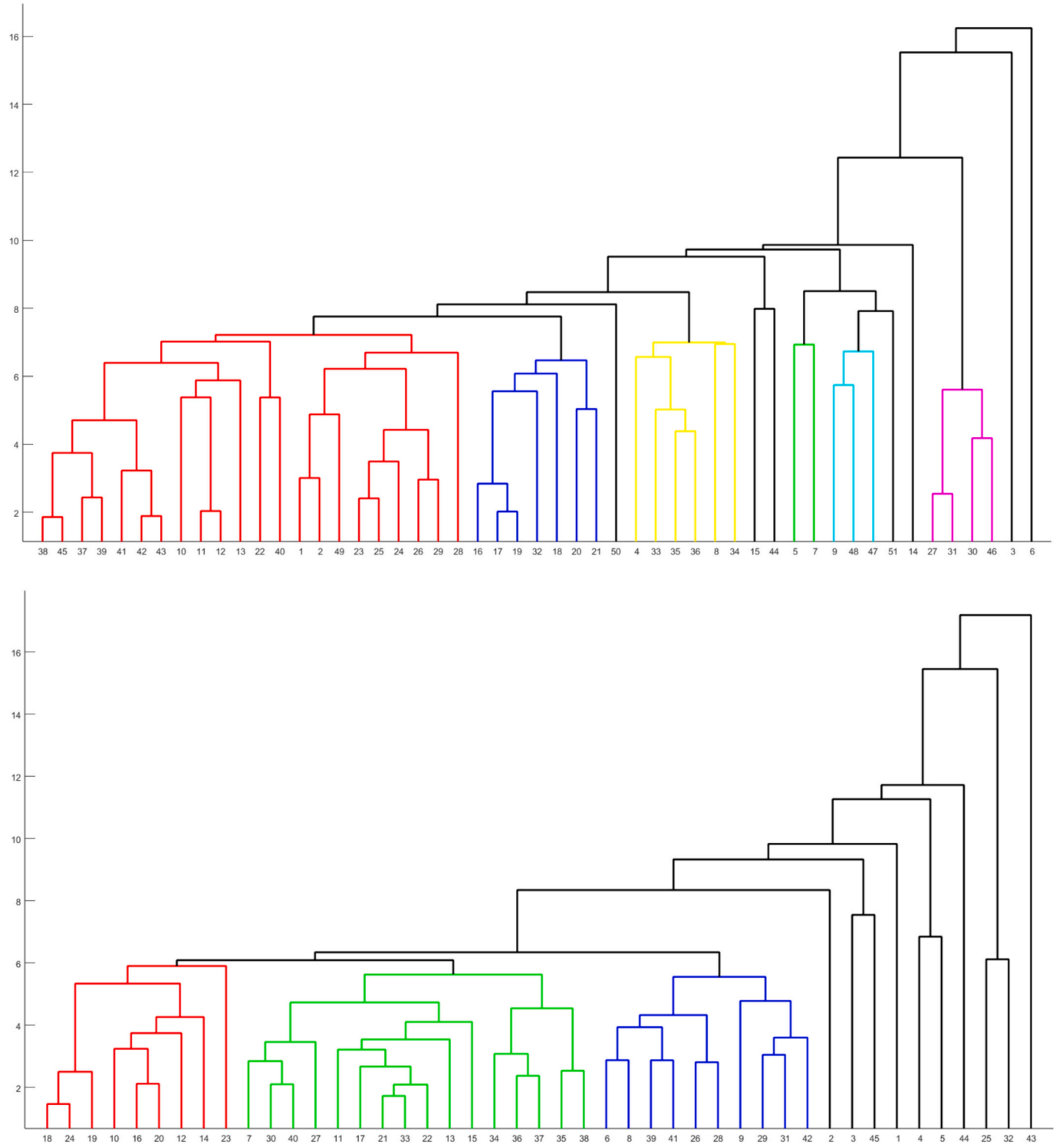


Fig. 16. Phylogenetic relationship among the PAH (Top) and SMN1 (Bottom) proteins based on structural and physicochemical features. Note that every color (other than black) in the phylogeny designates a cluster.

Table 13
Clusters of PAH and SMN1 proteins based on structural and physicochemical features.

Clusters	Sequences (PAH)
Cluster-1	{38, 45, 37, 39, 41, 42, 43, 10, 11, 12, 13, 22, 40, 1, 2, 49, 23, 25, 24, 26, 29, 28}
Cluster-2	{16, 17, 19, 32, 18, 20, 21}
Cluster-3	{4, 33, 35, 36, 8, 34}
Cluster-4	{5. 7}
Cluster-5	{48, 47}
Cluster-6	{27, 31, 30, 46}

Clusters	Sequences (SMN1)
Cluster-1	{18, 24, 19, 10, 16, 20, 12, 14, 23}
Cluster-2	{7, 30, 40, 27, 11, 17, 21, 33, 22, 13, 15, 34, 36, 37, 35, 38}
Cluster-3	{6, 8, 39, 41, 26, 28, 9, 29, 31, 42}

drug targets, guiding drug development efforts.

4.9. Agglomerated proximal sets of PAH and SMN1 proteins

Proximal sets of PAH and SMN1 proteins were derived as discussed in Section 3.10 (Table 15). PAH (SMN1) proteins belonging to the agglomerated proximal sets share a high amount of similarities as extracted from the quantitative features adumbrated in this study.

5. Discussion and conclusions

Rare diseases like PKU and SMA affect millions of people around the world, effective and affordable treatment strate- gies are yet to be developed. In the present study, we focus on understanding the correlation between different amino acid changes and the different PAH and SMN1 variants. For both PAH and SMN1, there was a decrease of some order- promoting amino acids and an increase in specific disorder-promoting residues. This correlates with previous studies, in which the SMN1 protein is predicted to have long disordered regions, which are close to the binding sites of SMN1 [80]. Previous studies on PAH have shown that there are also several regions in the PAH sequence, where the disordered regions can also affect the regulatory domain and the protein activity [81]. In our current study, we also found that among all the PAH sequences, the maximum homogeneous polystrings length is 8 and for SMN1 sequences it is 10. Previous studies have shown that the amino acid repeats are crucial for protein function and play an important role in protein-protein interaction [82]. Homogeneous polystring length also provides useful insight into understanding the protein evolution and functional characteristics [83]. Sequences with long polystrings have a faster evolution [83]. Pro-rich polystrings in SMN1, especially those of lengths 8, 9, and 10, characterize the majority of SMN1 sequences, suggesting a potential role of Pro in the structural stability or functional diversity of SMN1 proteins. Additionally, for both PAH and SMN1 the Shannon entropy was very high, suggesting higher degrees of randomness in terms of amino acid frequency, which means that there are more possibilities for the protein conformation [84]. It is important to note that even though Shannon entropy provides a lot of information about the diversity of both SMN1 and PAH, it is also limited in many ways as it does not take into account the environmental and other external interactions [85].

The current study also highlights that both PAH and SMN1 have similar frequencies of polar and non-polar residues, which can be correlated to the importance of stabilizing reaction especially between the polar and non-polar residues [86]. We also investigated the frequency distribution of the PP, NP, PN, NN changes and it was interesting to note that the PAH sequences have a higher percentage of PP change and for SMN1 the highest was the NN residue change. The distribution of the polar and nonpolar residues can affect the protein aggregation and also plays an important role in functional characterization [60,87]. Changes in this distribution might affect the protein solubility and function and thereby can lead to disease development.

Both acidic and basic residues play an important role in the functional characterization of the sequence such as DNA interaction, activator, and coactivator interaction, etc. [88]. In the current study, we found that there is a balanced distribution between acidic and basic residues. Such distribution studies are very important as conserved residues are under higher evolutionary pressure and any changes in the acidic and basic amino acid distribution can cause sequence changes and lead to disease phenotypes. Such changes in distribution can also affect the interaction of the SMN1 and PAH with other activators and DNA sequences, thereby leading to misregulation in the pathway. Our study also found that the percentage of NN change of residues was the highest, which can be attributed to the presence of a higher number of neutral residues in both SMN1 and PAH. The phylogenetic relationship derived from spatial distribution of acidic, neutral and basic residues lead to the formation of clusters for PAH and SMN1 sequences which can be correlated with similarity between the different sequences depending on the arrangement of acidic, neutral or basic residues in sequences.

Previous studies have shown that PAH has several flexible residues, is highly flexible in a solution state, and can undergo conformational changes [89,90]. In this study, we identified an increased presence of highly flexible residues among all the PAH sequences, which also points towards the importance of these flexible residues in conformational changes which can also be mobilized near the active site residues [91]. For SMN1 sequences, there is an increased presence of disordered residues and previous studies have also shown that SMN1 is a highly disordered protein [92]. Identifying such residues becomes very important as changes in the intrinsic protein disorder can cause changes in the conformation, thereby affecting protein activity [93]. This becomes especially important for identifying the disease-causing changes not just at the sequence level but also in spatial conformation. This investigation also highlights the importance of the absence of transition between the different states in SMN1 and PAH sequence for the transitions O_HF, O_D, MF_D, HF_O, D_O, and D_MF. It can be interesting to look further into why such transitions are absent and its potential implications on the protein function and conformation. Studying these signature-genomics can also help us in understanding the interaction of key residues both at sequence level and also spatially.

Furthermore, nine agglomerated proximal sets were identified and cumulatively they include 29 out of 51 PAH se- quences, while five agglomerated proximal sets were observed among 18 out of 45 SMN1 sequences (Table 15). Sequences within these proximal sets, whether PAH or SMN1, exhibit a significant level of similarity in terms of the various quan- titative features outlined in this study. On the other hand, PAH_3 and PAH_6 were outliers in all dendrograms except dendrograms based on sequence homology where PAH_6 was the only outlier. Among

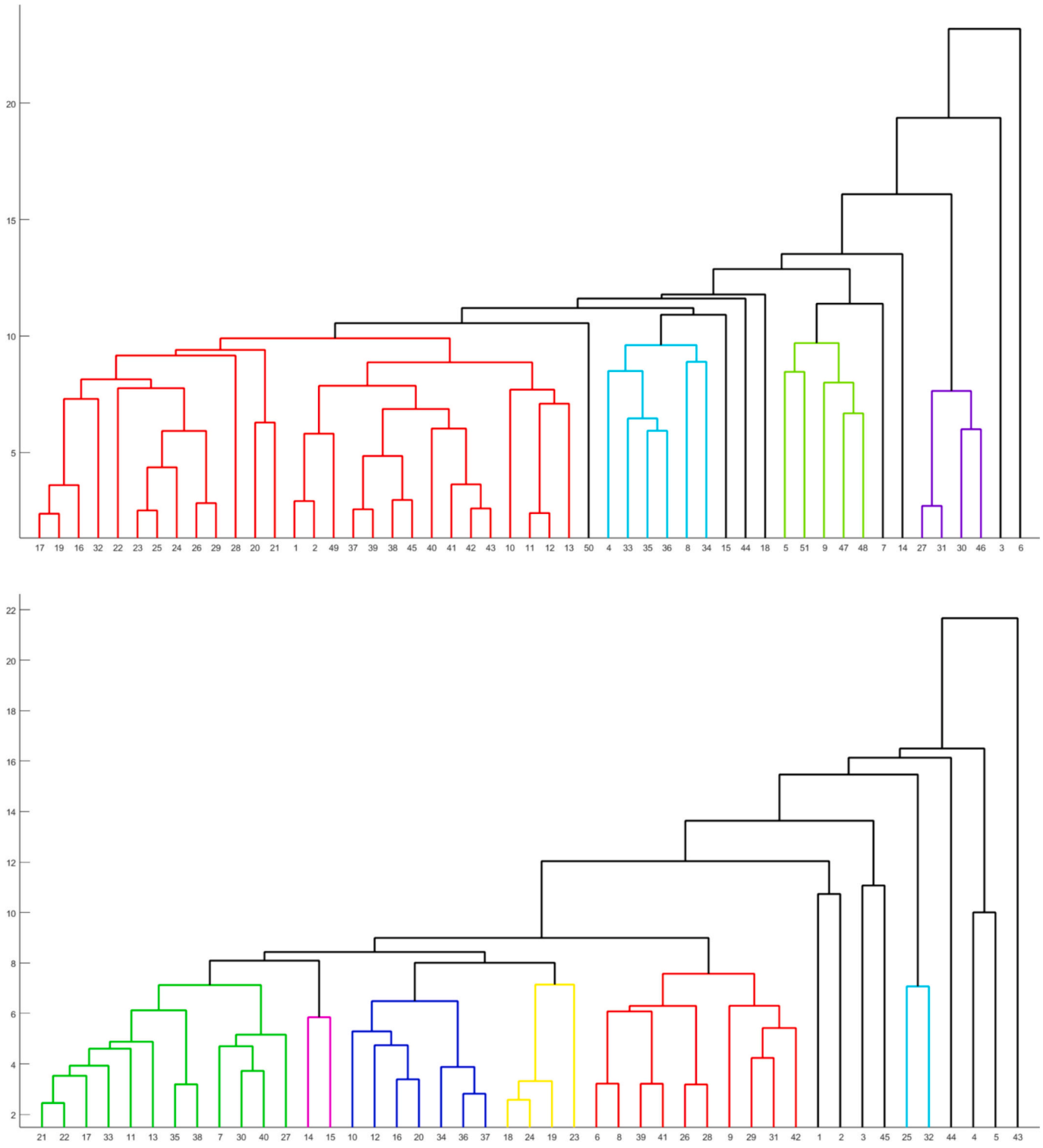


Fig. 17. Phylogenetic relationship among the PAH (Top) and SMN1 (Bottom) proteins based on structural and physicochemical features (ProtrWeb). Note that every color (other than black) in the phylogeny designates a cluster.

Table 14
Clusters of PAH and SMN1 proteins based on structural and physicochemical features (ProtrWeb).

Clusters	Sequences (PAH)
Cluster-1	{17, 19, 16, 32, 22, 23, 24, 25, 26, 29, 28, 20, 21, 1, 2, 49, 37, 39, 38, 45, 40, 41, 42, 43, 10, 11, 12, 13}
Cluster-2	{4, 33, 35, 36, 8, 34}
Cluster-3	{5, 51, 9, 47, 48}
Cluster-4	{27, 31, 30, 46}

Clusters	Sequences (SMN1)
Cluster-1	{21, 22, 17, 33, 11, 13, 35, 38, 7, 30, 40, 27}
Cluster-2	{14, 15}
Cluster-3	{10, 12, 16, 20, 34, 36, 37}
Cluster-4	{18, 24, 19, 23}
Cluster-5	{6, 8, 39, 41, 26, 28, 9, 29, 31, 42}
Cluster-6	{25, 32}

Table 15
List of proximal sets of PAH and SMN1 proteins.

Serial No.	Proximal Sets (PAH)	Serial No.	Proximal Sets (PAH)	Serial No.	Proximal Sets (SMN1)
1	{1,2}	7	{20,32}	1	{11,13,17,21,22,33,35,38}
2	{40,43}	8	{33,35}	2	{27,30}
3	{11,12,42}	9	{27,30,31,46}	3	{10,12,20}
4	{22,23,24,25,26,29}			4	{34,37}
5	{13,37,38,39,45}			5	{29,31,42}
6	{16,17,19}				

Illustration: Consider $\mathscr{S} = \{40, 43\}$ of PAH proteins. The objective is to show that $\mathscr{S}$ is an agglomerated proximal set. Note that, $\mathscr{S} \subseteq c_{1,2}$ ($i = 1, j = 2$) (Table 5), $\mathscr{S} \subseteq c_{2,2}$ ($i = 2, j = 2$) (Table 7), $\mathscr{S} \subseteq c_{3,2}$ ($i = 3, j = 2$) (Table 10), $\mathscr{S} \subseteq c_{4,2}$ ($i = 4, j = 2$) (Table 11), $\mathscr{S} \subseteq c_{5,1}$ ($i = 5, j = 1$) (Table 12), $\mathscr{S} \subseteq c_{6,1}$ ($i = 6, j = 1$) (Table 13), and $\mathscr{S} \subseteq c_{7,1}$ ($i = 7, j = 1$) (Table 14). Hence $\mathscr{S} = \{40, 43\}$ is an agglomerated proximal set and consequently, PAH_40 and PAH_43 were derived to be proximal.

SMN1 sequences SMN_43, SMN_25, and SMN_32 were distant from other SMN1 sequences in four out of seven dendrograms while SMN_43 alone was the outlier in two dendrograms as described in the results section.

The study's focus on understanding the correlation between amino acid changes and variants of PAH and SMN1 proteins provides valuable insights into these specific proteins, yet its limited scope and generalization could be considered as notable limitations. While identifying patterns in amino acid changes and protein properties, the study lacks direct experimental evidence of how these changes affect protein function and does not consider the biological context in which these proteins function. Additionally, the sample size and representativeness of the sequences analyzed are limited, potentially impacting the generalizability of the findings. Acknowledging these limitations strengthens the scientific discourse and guides future investigations on elucidating the genetic landscapes of SMN1 and PAH, laying the foundation for further research aimed at comprehending, diagnosing, and potentially treating disorders associated with these proteins. However, future studies should aim to address these limitations to ensure a more comprehensive understanding of the molecular mechanisms underlying rare diseases and to translate these findings into clinical applications effectively.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijbiomac.2024.131960.

CRediT authorship contribution statement

Debaleena Nawn: Writing – original draft, Validation, Investigation, Formal analysis, Data curation, Conceptualization. **Sk. Sarif Hassan:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Investigation, Formal analysis, Data curation, Conceptualization. **Elrashdy M. Redwan:** Writing – original draft, Validation, Supervision, Investigation, Conceptualization, Data curation, Formal analysis. **Tanishta Bhattacharya:** Writing – original draft, Validation, Investigation, Formal analysis, Data curation. **Pallab Basu:** Writing – original draft, Validation, Investigation, Formal analysis, Data curation. **Kenneth Lundstrom:** Data curation, Formal analysis, Investigation, Validation, Writing – original draft. **Vladimir N. Uversky:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Investigation, Data curation, Formal analysis.

Declaration of competing interest

The authors declare no conflict of interest.

Data availability

Data will be made available on request.

References

[1] M.J. Field, T.F. Boat, et al., Profile of Rare Diseases, Rare Diseases and Orphan Products: Accelerating Research and Development, 2010.
[2] A. Schieppati, J.-I. Henter, E. Daina, A. Aperia, Why rare diseases are an important medical and social issue, Lancet 371 (9629) (2008) 2039–2041.
[3] T. Richter, S. Nestler-Parr, R. Babela, Z.M. Khan, T. Tesoro, E. Molsen, D. A. Hughes, Rare disease terminology and definitions—a systematic global review: report of the ispor rare disease special interest group, Value Health 18 (6) (2015) 906–914.
[4] C.P. Austin, C.M. Cutillo, L.P. Lau, A.H. Jonker, A. Rath, D. Julkowska, D. Thomson, S.F. Terry, B. de Montleau, D. Ardigò, et al., Future of rare diseases research 2017–2027: an irdirc perspective, Clin. Transl. Sci. 11 (1) (2018) 21.
[5] A.A. Shafie, A. Supian, M.A. Ahmad Hassali, L.-H. Ngu, M.-K. Thong, H. Ayob, N. Chaiyakunapruk, Rare disease in Malaysia: challenges and solutions, PLoS One 15 (4) (2020) e0230850.
[6] M. Simone Baldovino, M. Domenica Taruscio, M. Dario Roccatello, Rare diseases in europe: from a wide to a local perspective, Isr. Med. Assoc. J. 18 (2016) 359–363.
[7] M.P. de la Paz, A. Villaverde-Hueso, V. Alonso, S. János, Ó. Zurriaga, M. Pollán, I. Abaitua-Borda, Rare diseases epidemiology research, in: Rare Diseases Epidemiology, Springer, 2010, pp. 17–39.
[8] M.F. Alsaqa'aby, N. Ibrahim, An overview about rare diseases in Saudi Arabia and reimbursement of orphan drugs, Glob. J. Med. Therap. 1 (2) (2019) 8–13.
[9] C.R. Ferreira, The burden of rare diseases, Am. J. Med. Genet. A 179 (6) (2019) 885–892.
[10] E.F. Augustine, H.R. Adams, J.W. Mink, Clinical trials in rare disease: challenges and opportunities, J. Child Neurol. 28 (9) (2013) 1142–1150.
[11] K.M. Boycott, A. Rath, J.X. Chong, T. Hartley, F.S. Alkuraya, G. Baynam, A. J. Brookes, M. Brudno, A. Carracedo, J.T. den Dunnen, et al., International

cooperation to enable the diagnosis of all rare genetic diseases, Am. J. Hum. Genet. 100 (5) (2017) 695–705.
[12] S. Dharssi, D. Wong-Rieger, M. Harold, S. Terry, Review of 11 national policies for rare diseases in the context of key patient needs, Orphanet J. Rare Dis. 12 (1) (2017) 1–13.
[13] A. Doyle, M.P. McGarry, N.A. Lee, J.J. Lee, The construction of transgenic and gene knockout/knockin mouse models of human disease, Transgenic Res. 21 (2012) 327–349.
[14] K. Olden, S.L. White, Health-related disparities: influence of environmental factors, Med. Clin. 89 (4) (2005) 721–738.
[15] A. El-Metwally, L. Yousef Al-Ahaidib, A. Ayman Sunqurah, K. Al-Surimi, M. Househ, A. Alshehri, O.B. Da'ar, H. Abdul Razzak, A.N. AlOdaib, The prevalence of phenylketonuria in Arab countries, Turkey, and Iran: a systematic review, Biomed. Res. Int. 2018 (2018).
[16] C.S. Brown, U. Lichter-Konecki, Phenylketonuria (pku): a problem solved?, in: Molecular Genetics and Metabolism Reports 6, 2016, pp. 8–12.
[17] F.J. van Spronsen, N. Blau, C. Harding, A. Burlina, N. Longo, A.M. Bosch, Phenylketonuria, Nat. Rev. Dis. Primers. 7 (1) (2021) 36.
[18] J. Zschocke, Phenylketonuria mutations in europe, Hum. Mutat. 21 (4) (2003) 345–356.
[19] N. Blau, Genetics of phenylketonuria: then and now, Hum. Mutat. 37 (6) (2016) 508–515.
[20] A. Hillert, Y. Anikster, A. Belanger-Quintana, A. Burlina, B.K. Burton, C. Carducci, A.E. Chiesa, J. Christodoulou, M. Đorđević, L.R. Desviat, et al., The genetic landscape and epidemiology of phenylketonuria, Am. J. Hum. Genet. 107 (2) (2020) 234–250.
[21] K. Ashe, W. Kelso, S. Farrand, J. Panetta, T. Fazio, G. De Jong, M. Walterfang, Psychiatric and cognitive aspects of phenylketonuria: the limitations of diet and promise of new treatments, Front. Psychol. 10 (2019) 561.
[22] N.A. Elhawary, I.A. AlJahdali, I.S. Abumansour, E.N. Elhawary, N. Gaboon, M. Dandini, A. Madkhali, W. Alo-saimi, A. Alzahrani, F. Aljohani, et al., Genetic etiology and clinical challenges of phenylketonuria, Hum. Genom. 16 (1) (2022) 1–17.
[23] U. Lichter-Konecki, J. Vockley, Phenylketonuria: current treatments and future developments, Drugs 79 (5) (2019) 495–500.
[24] C.R. Scriver, The pah gene, phenylketonuria, and a paradigm shift, Hum. Mutat. 28 (9) (2007) 831–845.
[25] L.A. Nash, J.K. Burns, J. Warman Chardon, R. Kothary, R.J. Parks, Spinal muscular atrophy: more than a disease of motor neurons? Curr. Mol. Med. 16 (9) (2016) 779–792.
[26] M. Shababi, C.L. Lorson, S.S. Rudnik-Schöneborn, Spinal muscular atrophy: a motor neuron disorder or a multi- organ disease? J. Anat. 224 (1) (2014) 15–28.
[27] B.T. Darras, J.A. Markowitz, U.R. Monani, C. Darryl, Spinal muscular atrophies, in: Neuromuscular Disorders of Infancy, Childhood, and Adolescence, 2015, pp. 117–145.
[28] M. Al Jumah, R. Majumdar, Z. Rehana, S. Al Rajeh, W. Eyaid, A pilot study of spinal muscular atrophy carrier screening in Saudi Arabia, J. Pediatr. Neurol. 5 (03) (2007) 221–224.
[29] Y. AlRuthia, G.S. Almuaythir, H.H. Alrasheed, W.R. Alsharif, M.-H. Temsah, F. Alsohime, I. Sales, M. Alwhaibi, F.A. Bashiri, Proxy-reported quality of life and access to nusinersen among patients with spinal muscular atrophy in Saudi Arabia, in: Patient Prefer. Adherence, 2021, pp. 729–739.
[30] S. Ogino, R.B. Wilson, Genetic testing and risk assessment for spinal muscular atrophy (sma), Hum. Genet. 111 (2002) 477–500.
[31] M.C. Keinath, D.E. Prior, T.W. Prior, Spinal muscular atrophy: mutations, testing, and clinical relevance, Appl. Clin. Genet. (2021) 11–25.
[32] B. Wirth, An update of the mutation spectrum of the survival motor neuron gene (smn1) in autosomal recessive spinal muscular atrophy (sma), Hum. Mutat. 15 (3) (2000) 228–237.
[33] C. Fallini, G.J. Bassell, W. Rossoll, Spinal muscular atrophy: the role of smn in axonal mrna regulation, Brain Res. 1462 (2012) 81–92.
[34] A. D'Amico, E. Mercuri, F.D. Tiziano, E. Bertini, Spinal muscular atrophy, Orphanet J. Rare Dis. 6 (1) (2011) 1–10.
[35] M. Calucho, S. Bernal, L. Alías, F. March, A. Venceslá, F.J. Rodríguez-Álvarez, E. Aller, R.M. Fernández, S. Borrego, J.M. Millán, et al., Correlation between sma type and smn2 copy number revisited: an analysis of 625 unrelated spanish patients and a compilation of 2834 reported cases, Neuromuscul. Disord. 28 (3) (2018) 208–215.
[36] S.J. Kolb, J.T. Kissel, Spinal muscular atrophy, Neurol. Clin. 33 (4) (2015) 831–846.
[37] T.W. Prior, A.R. Krainer, Y. Hua, K.J. Swoboda, P.C. Snyder, S.J. Bridgeman, A. H. Burghes, J.T. Kissel, A positive modifier of spinal muscular atrophy in the smn2 gene, Am. J. Hum. Genet. 85 (3) (2009) 408–413.
[38] C. Gonzaga-Jauregui, J.R. Lupski, Genomics of Rare Diseases: Understanding Disease Genetics Using Genomic Approaches, Academic Press, 2021.
[39] C.F. Wright, D.R. FitzPatrick, H.V. Firth, Paediatric genomics: diagnosing rare disease in children, Nat. Rev. Genet. 19 (5) (2018) 253–268.
[40] J. Bras, R. Guerreiro, J. Hardy, Use of next-generation sequencing and other whole-genome strategies to dissect neurological disease, Nat. Rev. Neurosci. 13 (7) (2012) 453–464.
[41] A.A. Seyhan, C. Carini, Are innovation and new technologies in precision medicine paving a new era in patients centric care? J. Transl. Med. 17 (2019) 1–28.
[42] A.L. Wise, T.A. Manolio, G.A. Mensah, J.F. Peterson, D.M. Roden, C. Tamburro, M. S. Williams, E.D. Green, Genomic medicine for undiagnosed diseases, Lancet 394 (10197) (2019) 533–540.
[43] H.H. Caicedo, D.A. Hashimoto, J.C. Caicedo, A. Pentland, G.P. Pisano, Overcoming barriers to early disease intervention, Nat. Biotechnol. 38 (6) (2020) 669–673.
[44] N.K. Litterman, M. Rhee, D.C. Swinney, S. Ekins, Collaboration for rare disease drug discovery research, in: F1000Research 3, 2014.
[45] S.H. Katsanis, N. Katsanis, Molecular genetic testing and the future of clinical genomics, Nat. Rev. Genet. 14 (6) (2013) 415–426.
[46] J.J. Mitchell, Y.J. Trakadis, C.R. Scriver, Phenylalanine hydroxylase deficiency, Genet. Med. 13 (8) (2011) 607–617.
[47] X. Chen, A. Sanchis-Juan, C.E. French, A.J. Connell, I. Delon, Z. Kingsbury, A. Chawla, A.L. Halpern, R.J. Taft, N. BioResource, et al., Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data, Genet. Med. 22 (5) (2020) 945–953.
[48] M.E. Butchbach, Copy number variations in the survival motor neuron genes: implications for spinal muscular atrophy and other neurodegenerative diseases, Front. Mol. Biosci. 3 (2016) 7.
[49] D.A. Keith, V. Vladimir, L. Stefano, et al., Composition profiler: a tool for discovery and visualization of amino acid composition differences, BMC Bioinformatics 8 (2007) 1–7.
[50] F. Quaglia, B. Mészáros, E. Salladini, A. Hatos, R. Pancsa, L.B. Chemes, M. Pajkos, T. Lazar, S. Peña-Díaz, J. Santos, et al., Disprot in 2022: improved quality and accessibility of protein intrinsic disorder annotation, Nucleic Acids Res. 50 (D1) (2022) D480–D487.
[51] S.S. Hassan, M. Sil, S. Chakraborty, A. Goswami, P. Basu, D. Nawn, V.N. Uversky, Possible functional proximity of various organisms based on the bioinformatics analysis of their taste receptors, Int. J. Biol. Macromol. 222 (2022) 2105–2121.
[52] C. Nacar, The frequencies of amino acids in secondary structural elements of globular proteins, Clin. Exp. Health Sci. 13 (2) (2023) 84–89.
[53] S. Hormoz, Amino acid composition of proteins reduces deleterious impact of mutations, Sci. Rep. 3 (1) (2013) 2919.
[54] B.J. Strait, T.G. Dewey, The shannon information entropy of protein sequences, Biophys. J. 71 (1) (1996) 148–155.
[55] M. Garcia-Boronat, C.M. Diez-Rivero, E.L. Reinherz, P.A. Reche, Pvs: a web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery, Nucleic Acids Res. 36 (suppl_2) (2008) W35–W41.
[56] D. González-Ruiz, H. Gohlke, Targeting protein-protein interactions with small molecules: challenges and perspectives for omputational binding epitope detection and ligand finding, Curr. Med. Chem. 13 (22) (2006) 2607–2625.
[57] D. Nawn, S.S. Hassan, M. Sil, A. Ghosh, A. Goswami, P. Basu, G.W. Dayhoff, K. Lundstrom, V.N. Uversky, The distal-proximal relationships among the human moonlighting proteins: Evolutionary hotspots and darwinian checkpoints, Inter. J. Biol. Macromol. 259 (2023) 128998.
[58] S. Shen, B. Kai, J. Ruan, J.T. Huzil, E. Carpenter, J.A. Tuszynski, Probabilistic analysis of the frequencies of amino acid pairs within characterized protein sequences, Phys. A Stat. Mech. Appl. 370 (2) (2006) 651–662.
[59] M.Y. Karpeisky, V. Ilyin, Analysis of non-polar regions in proteins, J. Mol. Biol. 224 (3) (1992) 629–638.
[60] H. Xiong, B.L. Buckwalter, H.-M. Shieh, M.H. Hecht, Periodicity of polar and nonpolar amino acids is the major determinant of secondary structure in self-assembling oligomeric peptides, Proc. Natl. Acad. Sci. 92 (14) (1995) 6349–6353.
[61] S. Kamtekar, J.M. Schiffer, H. Xiong, J.M. Babik, M.H. Hecht, Protein design by binary patterning of polar and nonpolar amino acids, Science 262 (5140) (1993) 1680–1685.
[62] B. Mészáros, G. Erdős, Z. Dosztányi, Iupred2a: context-dependent prediction of protein disorder as a function of redox state and protein binding, Nucleic Acids Res. 46 (W1) (2018) W329–W337.
[63] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, A.K. Dunker, Exploiting heterogeneous sequence properties im- proves prediction of protein disorder, Proteins Struct. Funct. Bioinform. 61 (S7) (2005) 176–182.
[64] K. Peng, P. Radivojac, S. Vucetic, A.K. Dunker, Z. Obradovic, Length-dependent prediction of protein intrinsic disorder, BMC Bioinform. 7 (2006) 1–17.
[65] K. Peng, S. Vucetic, P. Radivojac, C.J. Brown, A.K. Dunker, Z. Obradovic, Optimizing long intrinsic disorder predictors with protein evolutionary information, J. Bioinforma. Comput. Biol. 3 (01) (2005) 35–60.
[66] P. Romero, Z. Obradovic, X. Li, E.C. Garner, C.J. Brown, A.K. Dunker, Sequence complexity of disordered protein, Proteins Struct. Funct. Bioinform. 42 (1) (2001) 38–48.
[67] B. Xue, R.L. Dunbrack, R.W. Williams, A.K. Dunker, V.N. Uversky, Pondr-fit: a meta-predictor of intrinsically disordered amino acids, Biochim. Biophys. Acta (BBA)-Prot. Proteom. 1804 (4) (2010) 996–1010.
[68] G.W. Dayhoff, V.N. Uversky, Rapid prediction and analysis of protein intrinsic disorder, Protein Sci. 31 (12) (2022) e4496.
[69] K. Rajagopalan, S.M. Mooney, N. Parekh, R.H. Getzenberg, P. Kulkarni, A majority of the cancer/testis antigens are intrinsically disordered proteins, J. Cell. Biochem. 112 (11) (2011) 3256–3267.
[70] V.N. Uversky, Analyzing Idps in Interactomes, Intrinsically Disordered Proteins: Methods and Protocols, 2020, pp. 895–945.
[71] V.N. Uversky, J.R. Gillespie, A.L. Fink, Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins Struct. Funct. Bioinform. 41 (3) (2000) 415–427.
[72] C.J. Oldfield, Y. Cheng, M.S. Cortese, C.J. Brown, V.N. Uversky, A.K. Dunker, Comparing and combining predictors of mostly disordered proteins, Biochemistry 44 (6) (2005) 1989–2000.
[73] Z. Chen, P. Zhao, F. Li, A. Leier, T.T. Marquez-Lago, Y. Wang, G.I. Webb, A.I. Smith, R.J. Daly, K.-C. Chou, et al., Ifeature: a python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (14) (2018) 2499–2502.

[74] M. Forina, C. Armanino, V. Raggio, Clustering with dendrograms on interpretation variables, Anal. Chim. Acta 454 (1) (2002) 13–19.
[75] A.K. Dunker, J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, C. J. Oldfield, A.M. Campen, C.M. Ratliff, K.W. Hipps, et al., Intrinsically disordered protein, J. Mol. Graph. Model. 19 (1) (2001) 26–59.
[76] R. Williams, Z. Obradovic, V. Mathura, W. Braun, E. Garner, J. Young, S. Takayama, C.J. Brown, A.K. Dunker, The protein non-folding problem: amino acid determinants of intrinsic order and disorder, in: Biocomputing 2001, World Scientific, 2000, pp. 89–100.
[77] P. Radivojac, L.M. Iakoucheva, C.J. Oldfield, Z. Obradovic, V.N. Uversky, A. K. Dunker, Intrinsic disorder and functional proteomics, Biophys. J. 92 (5) (2007) 1439–1456.
[78] M. Machovič, Š. Janeček, The invariant residues in the α-amylase family: just the catalytic triad, Biologia 58 (6) (2003) 1127–1132.
[79] G.R. Thomas, J.R. Forbes, E.A. Roberts, J.M. Walshe, D.W. Cox, The Wilson disease gene: spectrum of mutations and their consequences, Nat. Genet. 9 (2) (1995) 210–217.
[80] V.N. Uversky, Targeting intrinsically disordered proteins in neurodegenerative and protein dysfunction diseases: another illustration of the d2 concept, Expert Rev. Proteom. 7 (4) (2010) 543–564.
[81] E.C. Arturo, K. Gupta, M.R. Hansen, E. Borne, E.K. Jaffe, Biophysical characterization of full-length human pheny- lalanine hydroxylase provides a deeper understanding of its quaternary structure equilibrium, J. Biol. Chem. 294 (26) (2019) 10131–10145.
[82] H. Luo, H. Nijveen, Understanding and identifying amino acid repeats, Brief. Bioinform. 15 (4) (2014) 582–591.
[83] M. Alba, P. Tompa, R. Veitia, Amino acid repeats and the structure and evolution of proteins, gene and Protein, Evolution 3 (2007) 119–130.
[84] C.E. Shannon, A mathematical theory of communication, ACM SIGMOBILE Mobile Comput. Commun. Rev. 5 (1) (2001) 3–55.
[85] T.N. Roach, Use and abuse of entropy in biology: a case for caliber, Entropy 22 (12) (2020) 1335.
[86] C.D. Andrew, S. Penel, G.R. Jones, A.J. Doig, Stabilizing nonpolar/polar side-chain interactions in the α-helix, Proteins Struct. Funct. Bioinform. 45 (4) (2001) 449–455.
[87] E. Monsellier, M. Ramazzotti, P.P. De Laureto, G.-G. Tartaglia, N. Taddei, A. Fontana, M. Vendruscolo, F. Chiti, The distribution of residues in a polypeptide sequence is a determinant of aggregation optimized by evolution, Biophys. J. 93 (12) (2007) 4382–4391.
[88] H. Ge, R.G. Roeder, Purification, cloning, and characterization of a human coactivator, pc4, that mediates Tran- scriptional activation of class ii genes, Cell 78 (3) (1994) 513–523.
[89] H. Frauenfelder, S.G. Sligar, P.G. Wolynes, The energy landscapes and motions of proteins, Science 254 (5038) (1991) 1598–1603.
[90] J.E. Fuchs, D. Fuchs, K.R. Liedl, Dynamic regulation of phenylalanine hydroxylase, Pteridines 25 (2) (2014) 33–39.
[91] J. Li, L.J. Dangott, P.F. Fitzpatrick, Regulation of phenylalanine hydroxylase: conformational changes upon phenylalanine binding detected by hydrogen/ deuterium exchange and mass spectrometry, Biochemistry 49 (15) (2010) 3327–3335.
[92] C.L. Lorson, H. Rindt, M. Shababi, Spinal muscular atrophy: mechanisms and therapeutic strategies, Hum. Mol. Genet. 19 (R1) (2010) R111–R118.
[93] J. Habchi, P. Tompa, S. Longhi, V.N. Uversky, Introducing protein intrinsic disorder, Chem. Rev. 114 (13) (2014) 6561–6588.