Investigating the structure and evolution of keratin intermediate filaments

A Thesis

submitted to

Indian Institute of Science Education and Research Pune in partial fulfilment of the requirements for the BS-MS Dual Degree Programme

by

Asita Singh

Indian Institute of Science Education and Research Pune

Dr. Homi Bhabha Road,

Pashan, Pune 411008, INDIA.

Date: 26 March 2025

Under the guidance of

Supervisor: Dr MS Madhusudhan

Affiliation of Supervisor

From May 2024 to March 2025

INDIAN INSTITUTE OF SCIENCE EDUCATION AND RESEARCH PUNE

Certificate

This is to certify that this dissertation 'Investigating the structure and evolution of keratin intermediate filaments' towards the partial fulfilment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Asita Singh at Indian Institute of Science Education and Research under the supervision of Dr MS Madhusudhan, Professor, Department of Biology, during the academic year 2024-2025.

डॉ. एम. एस. मधुसूदन/Dr. M. S. Madhusudhan
आचार्य, जीवविज्ञान विभाग/Professor, Biology Department
भारतीय विज्ञान शिक्षा एवं अनुसंधान संस्थान
Indian Institute of Science Education & Research
पुणे / Pune- 411 008, भारत / India

Dr MS Madhusudhan

Committee:

Dr MS Madhusudhan

Dr Durba Sengupta

This thesis is dedicated to my childhood dog. Rest in peace, Johnny.

Declaration

I hereby declare that the matter embodied in the report entitled “Investigating the structure, sequence and evolution of keratin intermediate filaments” are the results of the work carried out by me at the Department of Biology, Indian Institute of Science Education & Research (IISER) Pune, under the supervision of Dr MS Madhusudhan, and the same has not been submitted elsewhere for any other degree. Wherever others contribute, every effort is made to indicate this clearly, with due reference to the literature and acknowledgement of collaborative research and discussions.

Asita Singh

20201007

Date: 26 March 2025

Table of Contents

List of Tables

List of Figures

Abstract

Intermediate filaments (IFs) are one of the three major types of cytoskeletal filaments present in cells. The monomers of IFs consist of three distinct regions – the head, the coiled-coil domain, and the tail. The heads contain serine and glycine repeats and are essential for the formation of the filament. These monomers form coiled-coil dimers, which then form higher-order structures. As IFs form complex filamentous structures, it is hard to experimentally determine their structure. In a previous work, Dr Neelesh Soni used integrative modelling to model the K5K14 filament, but this did not contain any heads or tails. Here, I have used homology modelling using this K5K14 filament as a template. I built a structure of the K8K18 filament, both with and without heads. I carried out molecular dynamics simulations to analyse the stability and flexibility of the proteins. In order to validate this structure, I analysed a dataset of human intermediate filament variants to study the inter and intra-dimer interactions. I discovered that certain types of mutations are more prevalent at specific heptad positions. Volume and charge are two of the key factors that affect dimer assembly, and changes in these due to mutations will lead to keratinopathies. Using this information, we will be able to predict mutations that disrupt the filament and those that do not, and test our findings experimentally. Additionally, by analysing keratin sequences from different organisms, I found that they can be divided into two distinct types. I also discovered that keratins are well-conserved across species. We can build structures of keratin filaments in different organisms, and identify conserved structural regions, and find interacting proteins as well.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Dr MS Madhusudhan, for his invaluable guidance and support over the past two years. His belief in my abilities has allowed me to become more confident in my abilities. His constant support has helped me with my research, as well as my academic growth.

I am also incredibly grateful to Dr Gayathri Pananghat, who first gave me the opportunity to explore the field of structural biology. She encouraged me to take the leap when I expressed my interested in computational biology, which led me to my current research interests.

I would like to extend my gratitude to all my lab members, both current and past. They have helped me navigate everything – from basic Linux issues to complex research challenges. I would especially like to thank Saksham, who helped me dual boot my laptop; Harshita, for her help with APBS; Atreyi, for her help with my keratin gene regulation work; and last of all, Mukundan, for being a great sounding board and coffee buddy. The past two years would have been infinitely harder without the help of these people as well as other (former) lab members not mentioned here.

I am deeply grateful to my mother, who has supported me in everything I have done for as long as I can remember. I would also like to thank my siblings, Ajita and Manvendra, who are always there for me.

I would like to thank my friends – both present and past – who have made the past five years bearable. A special thanks to my friends Kashika, Gaurav and Akash for supporting me through my master's thesis. Most of all, I would like to thank Sharanya for her unwavering support and for pushing me out of my comfort zone.

The past several years would not have been possible without the support of the wonderful people mentioned here, as well as the many more who are not. To everyone who has helped me on my journey – thank you.

Contributions

Contributor Name	**Contributor Role**
Dr MS Madhusudhan, Asita Singh	Conceptualization Ideas
Dr MS Madhusudhan, Asita Singh	Methodology
Asita Singh	Software
Asita Singh	Validation
Asita Singh	Formal analysis
Asita Singh	Investigation
Dr MS Madhusudhan	Resources
Asita Singh	Data Curation
Asita Singh	Writing - original draft preparation
Dr MS Madhusudhan	Writing - review and editing
Asita Singh	Visualization
Dr MS Madhusudhan	Supervision
Dr MS Madhusudhan	Project administration
Dr MS Madhusudhan	Funding acquisition

Chapter 1: Introduction

The cytoskeleton

Metazoans contain three major types of cytoskeletal filaments that provide mechanical support to the cells – microfilaments (MFs), microtubules (MTs) and intermediate filaments (IFs). Each of these form distinctive networks which interact with each other, and with other elements of the cytoskeleton.

For decades, the cytoskeleton was believed to be present only in eukaryotes. However, it was discovered that bacteria contain actin and tubulin homologs as well (Bork et al., 1992; RayChaudhuri & Park, 1992). Still, most prokaryotes lack IF-like genes, with *Caulobacter crescentus* being an exception. It has an IF-like gene called *CreS* which imparts the bacteria with its characteristic crescent shape (Liu et al., 2024).

Discovery of keratins

Keratins are found in various appendages such as hair, nails, wool, etc., and were thus the first intermediate filament to be extensively studied. In the 1930s, Astbury discovered that unlike crystals of globular proteins, keratins had a surprisingly simple diffraction pattern (Astbury, 1933). He found that the X-ray diffraction pattern in hair changed upon application of mechanical stress. Hair in the 'ground' or non-stressed state was referred to as α-keratin and in mechanically stressed state was called β-keratin. This transformation was referred to as α-to-β transition (Astbury, 1933). Using Pauling's theory that proteins from either beta sheets or alpha helices, Crick proposed that keratins had a 'coiled-coil' structure with two alpha helices wrapped around each other (Crick, 1952; Pauling & Corey, 1953).

Holtzer's group detected the presence of smooth filaments that co-existed with actin and myosin, and were ~10 nm in diameter (Ishikawa et al., 1968). This criterion was used to investigate other tissues for presence of these 10 nm filaments.

Diversity of IFs

Based on sequence homology, intermediate filaments have been grouped into six types as described in **Error! Not a valid bookmark self-reference.** (Conway & Parry, 1988).

Table 1: Types of intermediate filaments present in humans.

Class	Proteins	Count
I	K9 to K20 (epithelial keratins)	28
	K23 to K40 ('hard' keratins)	
II	K1 to K8 (epithelial keratins)	26
	K71 to K86 ('hard' keratins)	
III	Vimentin	5
	GFAP	
	Syncoilin	
	Desmin	
	Peripherin	
IV	NF-L, NF-M, NF-H	7
	α-internexin	
	Nestin	
	Synemin α, Synemin β	
V	Lamins (A, B1, B2, C1, C2)	5
Orphans (VI)	Filensin	2
	Phakinin	

These two classes can be further divided into 'soft' keratins – present in epithelial cells, and 'hard' keratin – present in appendages such as hair and nails. Due to the diversity of IFs, we are limiting our study to these two groups.

In humans, type I genes are present in a cluster on chromosome 17, while type II genes are present on chromosome 12. The only exception is K18, which is present on chromosome 12 with the type II genes (Kalabusheva et al., 2023).

Type I and type II IFs are keratins and have specific interaction pairs that form heterodimers. Some of the known canonical pairs include K5K14, K1K10, K8K18, K3K12, K6K16, K6K17

etc. It is possible for non-canonical pairs to form. For example, in the absence of K14, K5 can pair up with K16 to rescue some skin blistering in mice (Yamada et al., 2002).

Figure 1 shows the localisation of different types of keratins in various different organs and tissues (Ho et al., 2022). It is known that a particular type of keratin appears in more than one type of tissue, and a type of tissue can have more than one type of keratin pair. Also, currently, there is no filament-level data available about the stiffness of specific keratins. Therefore, it is hard to make inferences about the structural properties of different types of keratins.

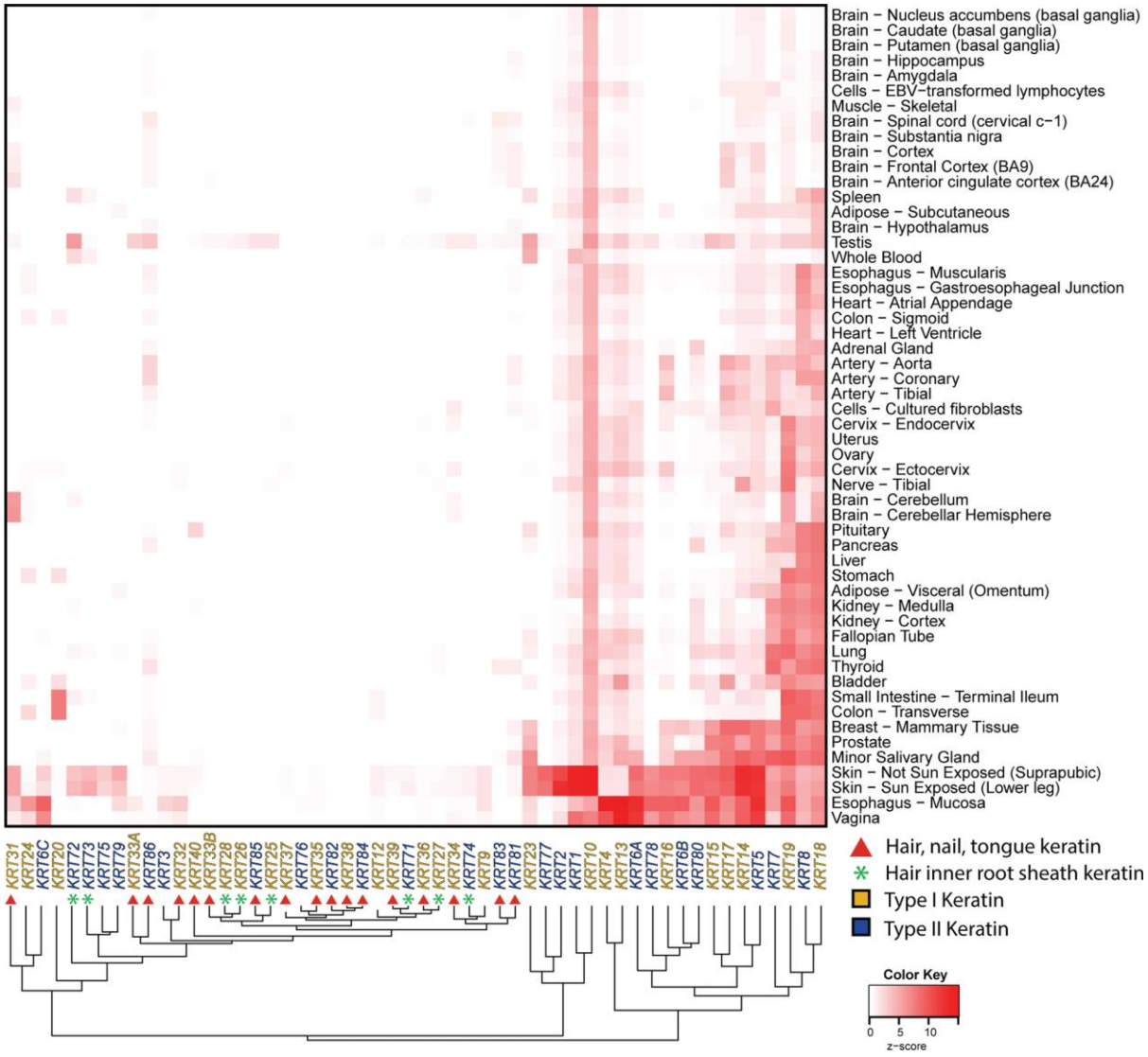


Figure 1: Localisation of human keratins in various tissues and organs. Via Ho et al., 2022.

Mutations in keratin genes can lead to various hereditary keratinopathies such as Epidermolysis Bullosa Simplex (EBS), epidermolytic ichthyosis, etc. Cytokeratin mutations may lead to epithelial to mesenchymal transition and the upregulation of several keratins is linked to multiple types of cancers as well (Francou & Anderson, 2020). In order to understand the

molecular basis of these diseases, we can look at the types of interactions that are disrupted in the variants. To this end, I have analysed the human IF variant dataset obtained from interfil.net (Szeverenyi et al., 2008) to study the filament-disrupting nature of these mutations.

Cytoskeletal filaments: A comparison

Actin filaments are made up of actin monomers which hydrolyse ATP to form microfilaments of diameter ~7 nm. Different organisms have different numbers of actin isoforms, however, they are structurally conserved (Dominguez & Holmes, 2011).

Unlike actin filaments, microtubules consist of αβ-tubulin heterodimers which utilize GTP to form hollow filaments ~25 nm in diameter. Eukaryotes contain multiple isoforms of α/β tubulins as well, and sequence analysis reveals that they are generally well-conserved with extreme divergence limited to the C-terminal region (McKean et al., 2001)

Both MTs and MFs can rapidly transition between growing and shrinking states – this is known as dynamic instability (Mitchison & Kirschner, 1984). However, IFs do not utilise substrates to from filaments and, consequently, do not display dynamic instability. These filaments self-assemble under physiological conditions (Pollard & Goldman, 2018).

Microtubules and actin filaments also serve as tracks for molecular motors. These motors use these tracks to transport intermediate filaments (Pollard & Goldman, 2018).

All three types of cytoskeletal filaments form distinct networks as they have separate functions and regions of localisation in the cells. Still, there is extensive crosstalk between them both directly and via intermediary proteins. Therefore, to fully understand the mechanical properties of the cell, it is essential to study these networks.

Evolution of keratins

Keratins are found in numerous species across many phyla, from arthropods to chordates. In order to compare these proteins from different organisms, it is important to identify their orthologs, i.e., genes that diverged over time due to speciation. We used the UniRef database (Suzek et al., 2015) to obtain keratin orthologs in several model organisms in order to study the evolution of these genes. However, this set of sequences could be incomplete and there is also the possibility of gene polymorphisms (Kapli et al., 2020).

Architecture of IFs

Coiled-coil proteins

Due to their exposed hydrophobic side chains, alpha helices are not very stable in solution. In order to minimise these interactions, two alpha helices can be wound around each other in a supercoil. This is called a 'coiled-coil' (Branden & Tooze, 2012). Crick demonstrated that in this motif, the residues per turn in each helix are reduced from 3.6 to 3.5, leading to the pattern repeating after every seven residues – a heptad repeat (Crick, 1952, 1953).

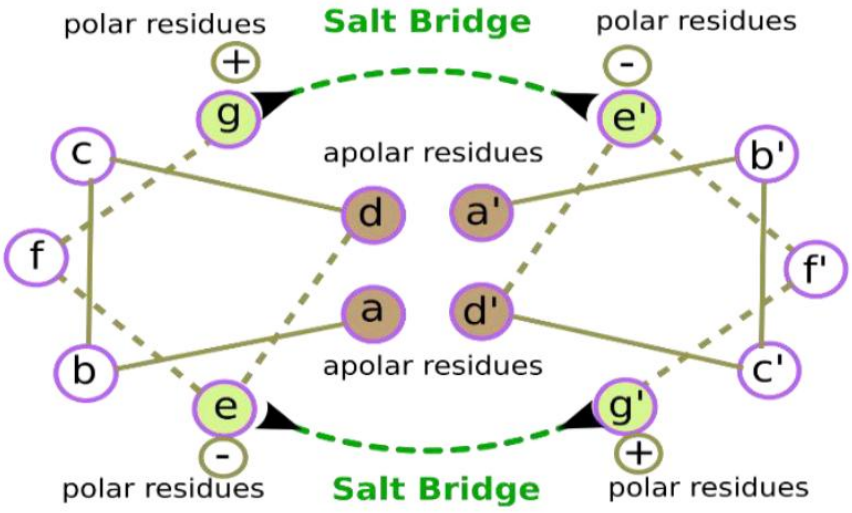


Figure 2: Schematic of residues in a coiled-coil dimer. Residues follow a heptad repeat of $[abcdefg]_n$. Via Soni, 2019.

Figure 2 shows a transection of a coiled-coil dimer with a heptad repeat. The positions are labelled from a to g for one helix and a' to g' for the second helix in the dimer. Residues at positions 'a' and 'd' are hydrophobic and form the core of the coiled-coil. Positions 'e' and 'g' consist of charged residues that form inter-helical salt-bridges, while residues are 'b', 'c' and 'f' positions interact with solvent molecules and are therefore likely to be polar (Ludwiczak et al., 2019).

Structure of the monomer

Intermediate filaments monomers can be divided into distinct regions – an unstructured head at the N-terminal, followed by a coiled-coil region, and an unstructured tail at the C-terminal. The coiled-coil region is not continuous as there are three 'stutters' or linker regions that divide

the coiled-coil region into four parts. These sub-domains are referred to as 1A, 1B, 2A and 2B, and the linkers are labelled as L1, L12 and L2 (Fuchs & Weber, 1994). A schematic of the monomer is shown in Figure 3.

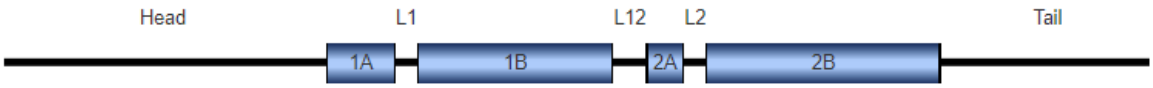


Figure 3: Schematic of an IF monomer. Via interfil.net.

The coiled-coil domain of monomers of type I to IV are ~310 amino acids in length, while lamins are longer at ~350 amino acids. The heads and tails of IFs are highly variable in length and composition, while the coiled-coil heptad region is sequentially conserved (Fuchs & Weber, 1994).

Filament assembly

Depending on the type of IF, the monomers can form either heterodimers or homodimers. Both types form coiled-coil dimers in a parallel manner, where the heads and tails of each monomer point in the same direction. These dimers self-organise to form staggered anti-parallel tetramers, with heads and tails of the two dimers pointing in opposite directions. These tetramers further come together using different types of interactions to form the unit length filament (ULF). One ULF consists of eight tetramers (i.e., it is a 32-mer) and forms the base unit for the formation of these IFs.

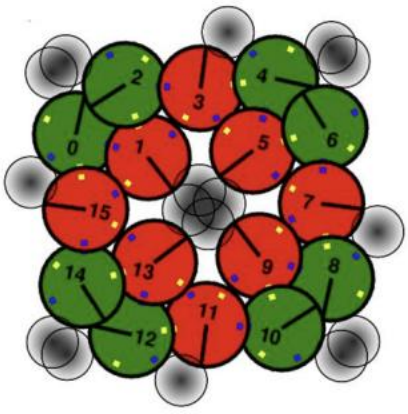


Figure 4: Intersection of an IF. Red circles represent the inner dimers and green circles represent the outer dimers. Shaded circles represent the heads and tails.

The work reported in this thesis will focus on type I and type II IFs which form keratin heterodimers. A cross-section model of a filament is shown in Figure 4. Based on volume calculations, we hypothesise that the heads and tails of the inner dimers are present in the hollow of the filament, while those of the outer dimers are on the external surface. These heads and tails are low-complexity regions consisting of serine and glycine repeats. These serines are phosphorylation sites, which lead to the disassembly of the filament.

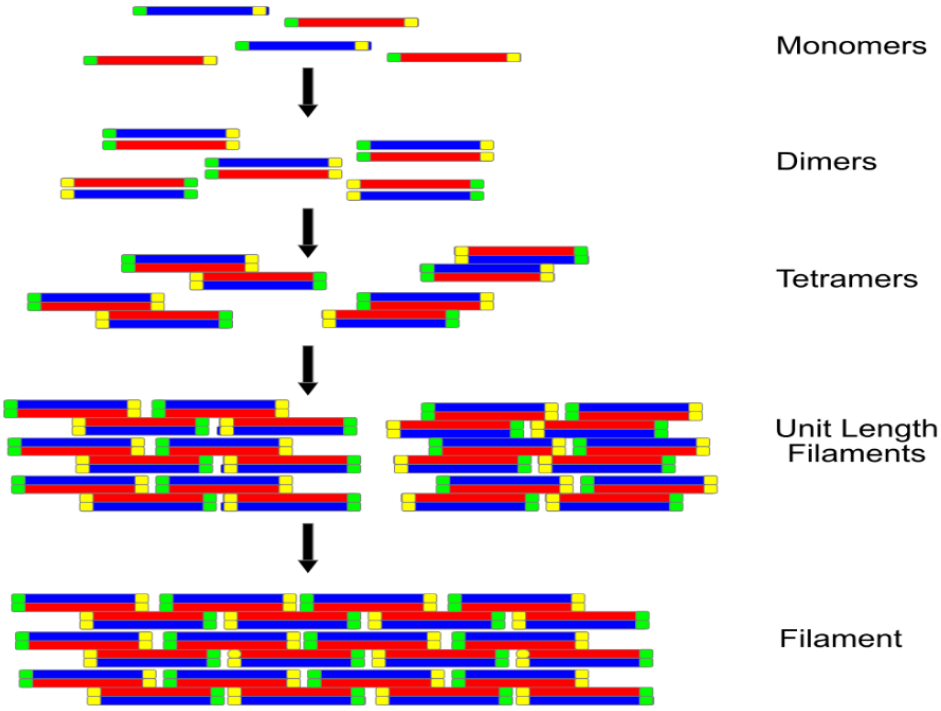


Figure 5: Schematic showing the process of assembly of IFs. Via interfil.net.

Mechanical properties of the filament

IFs are biopolymers and can undergo different types of deformation – bending, stretching, compression and twisting are the main modes. The mode of twisting of IFs has not been studied. Compression has been studied in vimentin filaments, which increase in stiffness upon application of force (Rölleke et al., 2023). This is due to their persistence length, or the bending stiffness, above which the molecule behaves like an elastic rod. It is in the order of 0.5 – 2 μm, depending on the type of IF (Mücke et al., 2004). It is an order of magnitude smaller than the persistence length of actin filaments. AFM and optical tweezer experiments have been used to measure the degree of stretchability of IFs. These experiments suggest that some filaments can

stretch up to 3.6 times their original length before they unravel. It is believed that upon stretching, the alpha helices in the filament undergo transition to form beta sheets (Rölleke et al., 2023). We want to study the mechanism of stretching of IFs.

Experimental Structure Determination

Parts of the coiled-coil dimer in various different IFs have been studied using crystallography in the past decades. However, it has been difficult to study the structure of the higher-order filaments and the interactions driving their assembly are not well-understood. Chemical cross-linking data is available for several IFs. In combination with other experimental techniques, these data can be used to build an integrative model of the filament. This method was used by Neelesh Soni to build a structure of the K5-K14 keratin IF (Soni, 2019). A structure of vimentin, a type III intermediate filament, determined using cryo-EM is available as well (Eibauer et al., 2024) (shown in Figure 6).

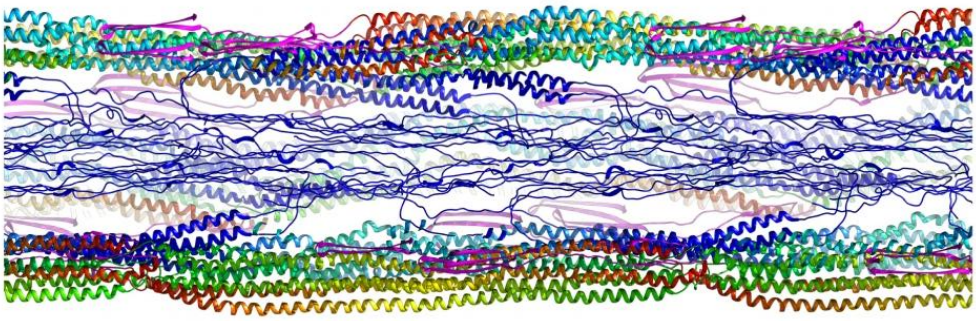


Figure 6: Vertical section of a vimentin filament. Via Eibauer et al.

Objectives

1. To build a structure of a keratin IF with heads and tails folded and study the mechanism of stretching of the filament using steered molecular dynamics.
2. Analyse the human keratin variant dataset and rationalise the mutations present, and use this to validate filament structure.
3. Analyse keratins found in model organisms and compare with human keratin IFs.

Chapter 2: Methodology

Building a structure of a keratin IF

Homology modelling was used to build a structure of the chosen IF (K8K18) with no heads and tails. The 3ULF structure with heads folded is built as per the workflow described below:

1. Using the K5K14 structure as a template, build a 1ULF structure of the chosen keratin IF (with no heads or tails)
2. Add a short stretch of head sequence, apply distance restraints and build homology model
3. Energy minimise the model to reduce clashes
4. Add the next stretch of sequence and use the model build in the previous iteration as the template
5. Repeat until all heads and tails are built
6. Superimpose three 1ULF structures to build a 3ULF filament
7. Re-build structure and energy minimise to reduce clashes

Note that clashes here are defined as a van der Waals radii overlap of >0.6 Å between a pair of atoms, after subtracting 0.4 Å to account for hydrogen bonding.

Homology modelling

MODELLER was used to build the structure of the chosen keratin IF, using the K5K14 structure with no heads and tails as a template (Webb & Sali, 2016). A 1ULF model of the chosen keratin IF with no heads and tails was built first. `salign` was used to align the template and target and `AutoModel` was used to build the structure using homology modelling. `max_molpdf` an internal cutoff function above which model building is aborted. This was increased from the default value of 1e+5 to 1e+8 to allow for model building. `AutoModel.library_schedule` was set to `autosched.slow` for thorough VFTM optimisation and `AutoModel.md_level` was set to `refine.slow` to allow for longer MD optimisation in order to decrease clashes. Optimisation was repeated twice. One model of the target was generated.

This model was used as a template for further structure building by adding a short stretch of the head sequence to the target sequence. Sequence alignment was done manually. `MyModel` class was used to apply distance restraints and select residues only in the head region to refine further. Distance restraints were chosen based on charge similarity of the residues in order to fold in the heads into the hollow of the filament.

`self.residue_range` was used to select the stretch of head added to the model in each iteration. This range was between 8-22 residues in each iteration and 6 iterations were used to build the 3ULF structure with the full heads.

`self.restraints` was used to apply upper bound distance restraints with mean of 6 Å and standard deviation of 0.15 Å. So, the upper bound on the distance between the two atoms specified would be 6 Å.

Energy minimisation

Energy minimisation after each iteration was carried out using GROMACS 2023 (Bekker et al., 1993). Steepest descent minimisation scheme was used with a step size of 1 fs for 50,000 steps. Force cutoff was set to <1000 kJ/mol/nm.

Simulation of keratin IF

Molecular dynamics was used to study the stability of the filament. GROMACS 2023 was used with CHARMM 36m force field (Huang et al., 2017). A triclinic box was constructed around the filament with a boundary extending 10 Å around the filament. An explicit SPCE water model was used (Berendsen et al., 1987). The system was neutralised by replacing the solvent molecules with NaCl ions. Particle Mesh Ewald (PME) method was used for estimating long-range electrostatic interactions and LINCS algorithm was used to constrain hydrogen bond lengths. Reference temperature was set to 300 K. Steepest descent algorithm was used to energy minimise the system for 100 ps with a step size of 2 fs and a maximum cutoff of <1000 kJ/mol/nm. NVT equilibration was carried out followed by NPT equilibration for 1 ns each with a step size of 2 fs. Leap-frog algorithm was used for numerical integration at each step. MD production run was simulated for 100 ns and the trajectory was recorded at every 10 ps.

GROMACS as well as MDAnalysis package were used to analyse the trajectory of the filaments (Gowers et al., 2016). The average structure of the protein was generated using the

latter. `gmx trjconv` was used to remove periodic boundary conditions using `-pbc nojump` and fitted the trajectory to the average structure using `-fit rot+trans`. An index file was generated using `-make_ndx` to select only the middle ULF for the analysis. `rms` and `rmsf` commands were used to calculate the RMSD and RMSF respectively.

XMGRACE was utilised for plotting RMSD and RMSF (Turner, 2005). For visualisation and movie-making VMD and PyMol software were used (Humphrey et al., 1996; Schrödinger, LLC, 2015).

Sequence analysis of human keratin variants

Database construction and filtering

Canonical protein sequences of type I and type II IFs as well as the domain information was acquired from interfil.net (Szeverenyi et al., 2008). Sequence variants for the same were also obtained from this database. The data contained information about the cDNA variant and type, protein variant and type, domain, disease, Omim ID, and publication information. Since some mutations are linked to multiple diseases and some mutations were reported from multiple sources, the data was filtered to get only the unique sequence variants.

Sequence alignment

MUSCLE (Edgar, 2004) was used to align all the sequences in fasta format. In order to calculate the pairwise sequence identity and similarity for these proteins, pseqsid package was used to obtain the pairwise identity and similarity for all the sequences. The average of these values was taken as the average identity or similarity. Standard deviation was calculated for this data as well (Edgar, 2004).

Coiled-coil prediction

MARCOIL (Gruber et al., 2006) was used to get the probability of a particular residue being coiled-coil in a sequence along with its predicted heptad index. A cutoff of 50.0 was applied to identify a residue as part of a heptad repeat.

Analysis of model organisms

Database construction and filtering

Sequences of keratins from different organisms were obtained from the UniRef100 database (Suzek et al., 2015). The sequences were filtered by taxonomic ID and keyword ('keratin') and then downloaded in fasta format.

The sequences were filtered based on length and keywords. The minimum length was set to 250 and maximum was set to 750 in order to exclude fragments and non-keratin proteins. Sequences with cartain keywords in the header were remoed. This included 'fragment', 'keratin-associated', 'low quality', 'isoform' and 'splice variant'.

Determining Orthologs

BLAST+ was used to determine the orthologs of human keratins present in our model organisms (Altschul et al., 1990). A database consisting of all human keratin sequence was constructed. All the sequences of the chosen organisms were aligned against this and the sequence with the highest identity was selected as the ortholog to the human keratin IF. Nomenclature of these sequences was standardised and duplicates were removed using percentage identity as a cutoff.

Tree Building

Multiple sequence alignment (MSA) was done using MUSCLE (Edgar, 2004). The fasta output was converted into phylip format using seqmagick package. Maximum likelihood phylogenetic trees were built using IQTree software (Minh et al., 2020). LG amino acid exchange rate matrix was used (Le & Gascuel, 2008). Rate heterogeneity, which accounts for different evolutionary rates across different branches or sites, was specified by the FreeRate model. This relaxes the assumption of Gamma-distributed rates. The number of categories specified was 7.

For smaller trees, bootstrapping was done 100 times and a consensus tree was generated in Newick format. For bigger trees (400+ sequences) ultrafast bootstrapping was used with IQTree. The trees were visualised using Python packages and iTOL webserver (Letunic & Bork, 2024).

Chapter 3: Results

Modelling of K8K18 IF

I chose K8K18 keratin IF for modelling as it has a short head and tail and would therefore be faster to model and simulate. This filament is also well-conserved across organisms and mutations in these keratins are known to cause foetal lethality in mice (Kalabusheva et al., 2023).

Dimers of IFs are staggered. Therefore, we are building a 3ULF model to ensure that the middle ULF is padded on both sides.

K8:

MSIRVTQKSYKVSTSGPRAFSSRSYTSGPGSRISSSSFSRVGSSNFRGGLG
GGYGGASGMGGITAVTVNQSLLSPLVLEVDPNIQAVRTQEKEQIKTLNNKF
ASFIDKVRFLEQQNKMLETKWSLLQQQKTARSNMDNMFESYINNLRRQLET
LGQEKLKLEAELGNMQGLVEDFKNKYEDEINKRTEMENEFVLIKKDVDEAY
MNKVELESRLEGLTDEINFLRQLYEEEIRELQSQISDTSVVLSMDNSRSLD
MDSIIAEVKAQYEDIANRSRAEAESMYQIKYEELQSLAGKHGDDLRRTKTE
ISEMNRNISRLQAEIEGLKGQRASLEAAIADAEQRGELAIKDANAKLSELE
AALQRAKQDMARQLREYQELMNVKLALDIEIATYRKLLEGEESRLESGMQN
MSIHTKTTSGYAGGLSSAYGGLTSPGLSYSLGSSFGSGAGSSSFSRTSSSR
AVVVKKIETRDGKLVSESSDVLPK

K18:

MSFTTRSTFSTNYRSLGSVQAPSYGARPVSSAASVYAGAGGSGSRISVSRS
TSFRGGMGSGGLATGIAGGLAGMGGIQNEKETMQSLNDRLASYLDRVRSLE
TENRRLESKIREHLEKKGPQVRDWSHYFKIIEDLRAQIFANTVDNARIVLQ
IDNARLAADDFRVKYETELAMRQSVENDIHGLRKVIDDTNITRLQLETEIE
ALKEELLFMKKNHEEEVKGLQAQIASSGLTVEVDAPKSQDLAKIMADIRAQ
YDELARKNREELDKYWSQQIEESTTVVTTQSAEVGAAETTLTELRRTVQSL
EIDLDSMRNLKASLENSLREVEARYALQMEQLNGILLHLESELAQTRAEGQ
RQAQEYEALLNIKVKLEAEIATYRRLLEDGEDFNLGDALDSSNSMQTIQKT
TTRRIVDGKVVSETNDTKVLRH

Initially, I tried to build a model of the 1ULF filament all at once. However, this led to too many clashes that were not tolerated by the software. I also tried to build the model by superimposing fully-built dimers to make a filament. However, in addition to clashes, it was hard to put together the dimers due to the high levels of symmetry. After much trial and error, I settled on this workflow where I iteratively built the model by adding short stretch of heads. Table 2 shows the sequence added to the target in each iteration of model building.

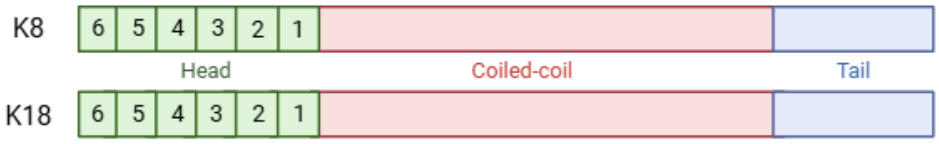


Figure 7: Schematic of K8 and K18 dimer. 1-6 represent the iterations to build the model.

Table 2: Sequence of K8 and K18 added at each iteration of model building.

Iteration	**Dimer Length**	**Sequence added to K8 head**	**Sequence added to K18 head**
1	660	SLLSPLVLEVDPNIQAVRTQEK	GGLATGIAGGLAGMGGIQNEK
2	686	ASGMGGITAVTVNQ	SRSTSFRGGMGS
3	717	VGSSNFRGGLGGGYGG	SVYAGAGGSGSRISV
4	747	TSGPGSRISSSSFSR	VQAPSYGARPVSSAA
5	770	STSGPRAFSSRSY	FSTNYRSLGS
6	790	MSIRVTQKSYKV	MSFTTRST

First, I built a structure of the 1ULF filament with no heads and tails. Then, I built a structure of the filament with the first stretch of the heads, but with no restraints. I used this structure to choose atoms for distance restraints. After trial and error, six atoms were chosen for three distance restraints and the first model was built. The restraints are given in the format `atom:residue_number:chain` in Table 3. These were applied to each dimer. Note that the residue number is different in each iteration due to addition of head sequence.

Table 3: Restraints used to build model of K8K18 IF.

Iteration	Restraints	
	Atom 1	Atom 2
1	OG:1:A	ND2:76:A
	CA:337:B	CA:369:B
	CB:9:A	CG:361:B
2	CA:1:A	CG1:486:B
	CA:347:B	NE2:474:B
3	CA:1:A	OE2:189:A
	CA:363:B	CD:549:B
4	CA:1:A	CE:223:A
	CA:378:B	CB:609:B

From iteration 1 to 4, I chose various distance restraints to make 5-10 alternative models, and each model took ~1 day to generate. I chose the model with the least clashes and carried energy minimisation for this model to further reduce them. When building the initial structure for the 5th iteration, I found that the heads of the inner dimers were buried inside the filament, and those of the outer dimers were close to the filament. Therefore, no restraints were used for the last two rounds. Table 4 shows the number of clashes before and after minimisation, as well as the potential energy of the system after minimisation and the number of steps taken to reach that state.

Table 4: Energy minimisation statistics at each iteration.

Iteration	Clashes (Before)	Clashes (After)	Number of Steps	Potential Energy (in kJ/mol/nm)
1	3464	41	7221	-3.42E+07
2	2649	84	7424	-3.39E+07
3	4224	96	9333	-3.39E+07
4	4367	62	706	-3.17E+07
5	5343	62	9176	-3.36E+07
6	4487	101	1269	-8.61E+08

I used the K5K14 3ULF structure as a template, and three copies of the fully built K8K18 1ULF structure as the target. I superimposed them onto the template to build the full structure and the final model was generated in ~8 days (shown in Figure 9).

The 3ULF model of the K8K18 filament with no heads was built in one step as the clashes were minimal (shown in Figure 8). The model was fully generated in ~3 days.

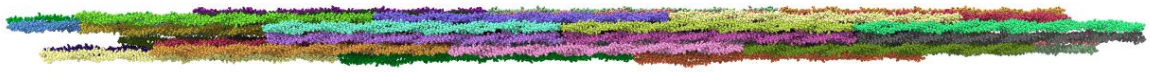


Figure 8:3ULF model of K8K18 filament with no heads or tails.

Figure 9: 3ULF model of K8K18 filament with heads.

Equipotential surface of the dimer

The heads of monomers contain Serines which are phosphorylation sites, and this post-translational modification leads to disassembly of the filament. Figure 11 and Figure 11 show the positions of Serines, and acidic and basic amino acids in K8 and K18 monomer respectively.

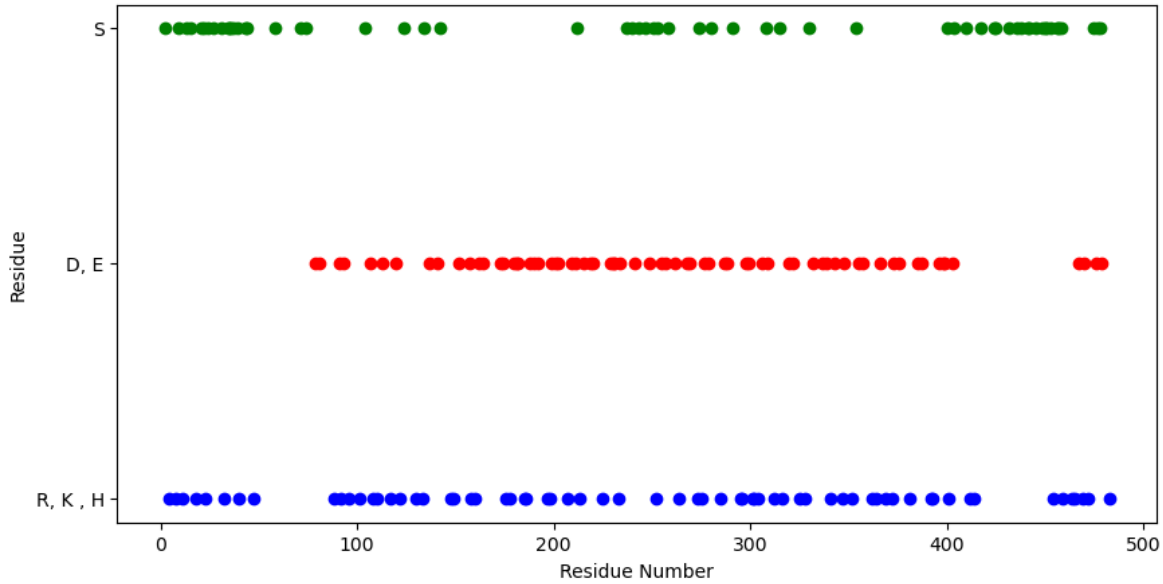


Figure 10: Graph showing the positions of Serines, and acidic and basic amino acids vs the residue position for K8.

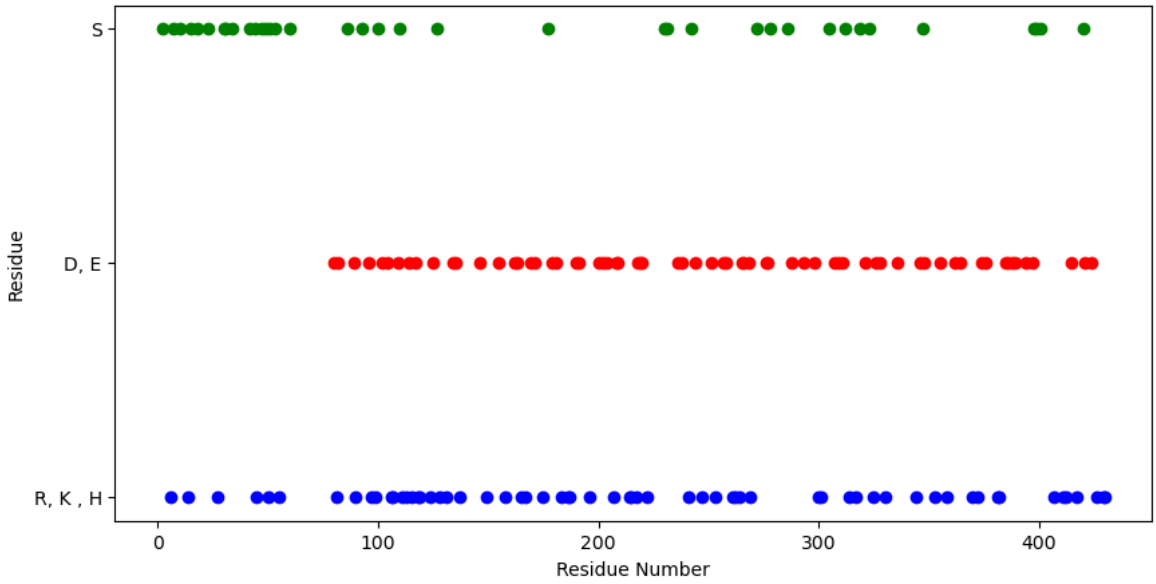


Figure 11: Graph showing the positions of Serines, and acidic and basic amino acids vs the residue position for K18.

I used Applied Boltzmann Poisson Solver (APBS) software (Jurrus et al., 2018) to calculate the equipotential surface of the K5K14 and K8K18 dimers and visualised it via PyMol (Jurrus et al., 2018). Figure 12 shows the equipotential surface of the K8K18 IF dimer with heads and tails unfolded. The unstructured regions are positively charged compared to the coiled-coil region which has high positive charge. This fits in with the theory that the negative-negative charge repulsion due to phosphorylation of serines in the head is responsible for the disassembly of the filament.

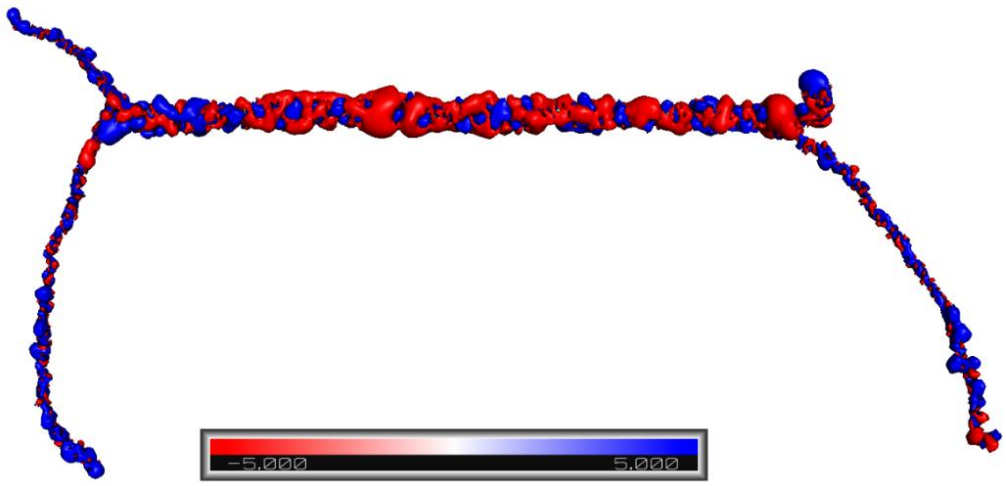


Figure 12: Equipotential surface of a dimer of K8K18. Units are in kB/T.

Molecular Dynamics Simulation of K8K18 filament

Simulation of K8K18 filament with heads

The K8K18 3ULF filament with heads is contained in a triclinic box of volume 45,332.6 nm^3. The system has 4,484,647 atoms, out of which 3,918,390 are solvent atoms. The system has a non-zero total charge of -817.0e, which is neutralised by replacing solvent atoms with Na. Steepest descent energy minimisation converged to $F_{max} < 1000$ kJ/mol/nm in 9578 steps. The potential energy of the system was -7.4840024e+07 kJ/mol.

I computed the root mean square deviation (RMSD) of the protein (Figure 13). The blue curve shows the RMSD over time for the entire protein while the red curve shows the RMSD for only the middle ULF. Additionally, upon calculating the root mean square fluctuation (RMSF) of the C_α atoms in the protein, I found that the middle ULF (highlighted in grey) undergoes less fluctuation over time compared to the flanking ULFs (Figure 14).

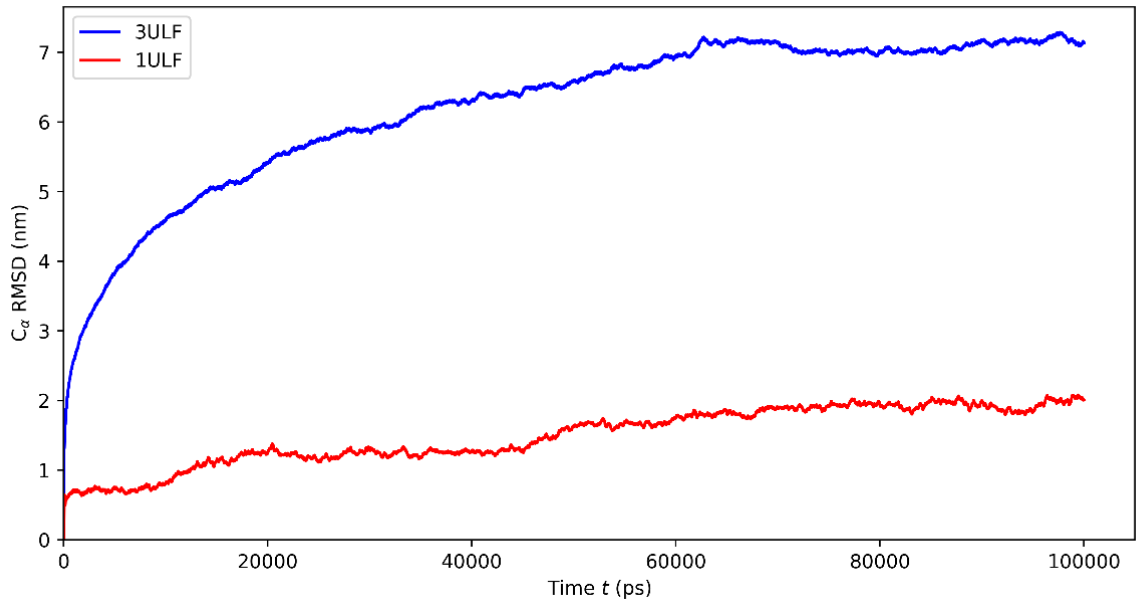


Figure 13: Root mean square deviation (RMSD) of K8K28 filament with heads.

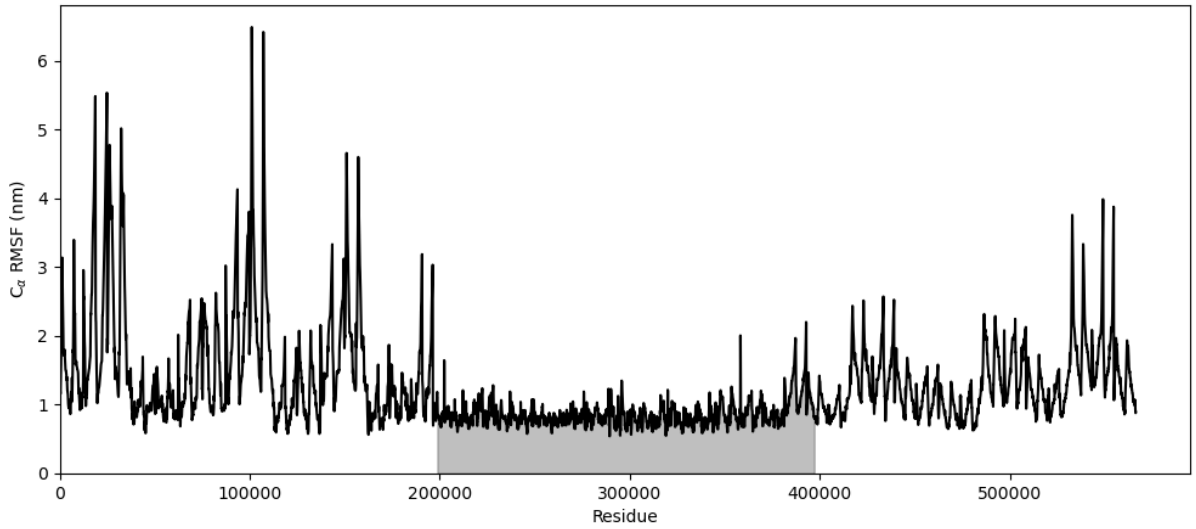


Figure 14: Root mean square fluctuation (RMSF) of K8K18 filament with heads over 100 ns.

Simulation of K8K18 filament with no heads or tails

A triclinic box of volume 25,986.1 nm^3 was constructed around the K8K18 3ULF filament with no heads or tails. The system contained 2,585,500 atoms in total, out of which 2,126,925 were solvent molecules. The system had a non-zero charge of -1423.0e which was neutralised by replacing the solvent molecules with Na ions. Energy minimisation converged in 6297 steps. The potential energy of the system was -4.2949956e+07 kJ/mol.

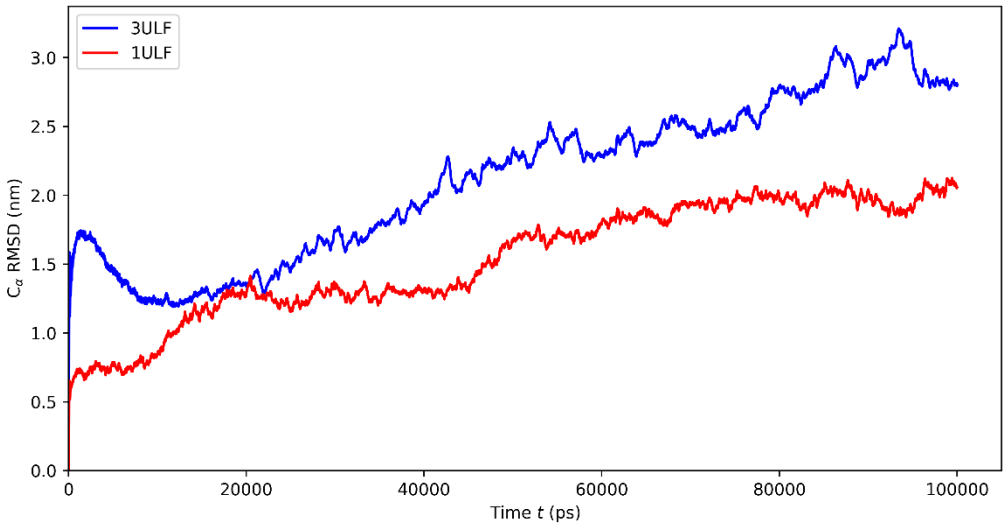


Figure 15: Root mean square deviation (RMSD) of simulation of K8K18 filament with no heads or tails.

Figure 15 shows the RMSD of the entire 3ULF filament in blue and the middle ULF (1ULF) in red. The RMSD of the 3ULF with no heads is lower than that of the 3ULF filament with heads, likely due to the absence of heads. Figure 16 shows the RMSF of the C_α atoms of the filament. Similar to the filament with heads, the middle ULF shows less fluctuation over time compared to the flanking filaments.

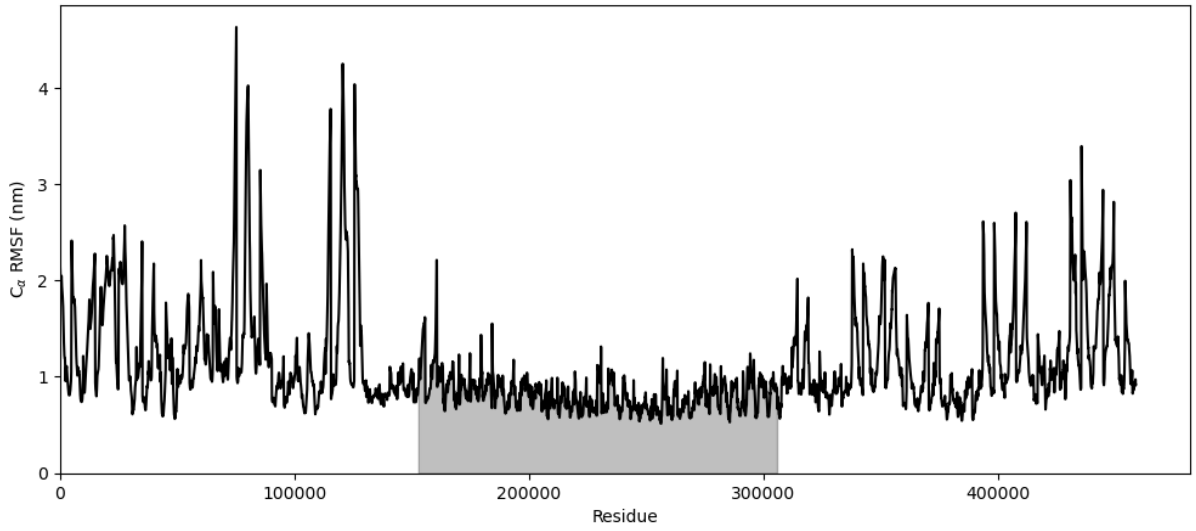


Figure 16: Root mean square fluctuation (RMSF) of K8K18 filament with no head or tail over 100 ns.

Analysis of human keratin variants

Construction of the dataset

Humans have 56 unique keratins, out of which 28 are type I and 26 are type II (Table 5). These categories are further divided into hard and soft keratins. Soft keratins or cytokeratins are found in epithelial cells while hard keratins are found in appendages such as hair, nails, etc. The brackets contain the alternative names for these proteins.

Table 5: List of keratins present in humans.

	Type I Keratins (Acidic)	Type II Keratins (Basic)
Soft Keratins	K9	K1
	K10	K2 (K2e)
	K12	K3
	K13	K4
	K14	K5
	K15	K6
	K16	K7
	K17	K8
	K18	
	K19	
	K20	
Hard Keratins	K23	K71 (K6irs1)
	K24	K72 (K6irs2)
	K25 (K25irs1)	K73 (K6irs3)
	K26 (K25irs2)	K74 (K6irs4)
	K27 (K25irs3)	K75 (K6hf)
	K28 (K25irs4)	K76 (K2p)
	K31 (Ha1)	K77 (K1b)
	K32 (Ha2)	K78 (K5b)
	K33a (Ha3-I)	K79 (K6I)
	K33b (Ha3-II)	K80 (Kb20)
	K34 (Ha4)	K81 (Hb1)
	K35 (Ha5)	K82 (Hb2)
	K36 (Ha6)	K83 (Hb3)
	K37 (Ha7)	K84 (Hb4)
	K38 (Ha8)	K85 (Hb5)
	K39 (Ka35)	K86 (Hb6)
	K40 (Ka36)	

Several factors affect the formation of dimers and filaments of keratins, such as the volume of amino acids at ‘a’ and ‘d’ positions, charge position, hydrophobicity, and more (Soni, 2019).

A total of 567 unique keratin variants were present in the dataset. These are all known disease-causing mutants. The protein variant type distribution of the variants is shown in Table 6.

Table 6: Mutation-type distribution of human keratin variants.

Protein Variant Type	Count
Substitution	467
Frameshift	46
Deletion	30
Unknown	6
Insertion	5
Indel	5
Silent	4
Deletion, Substitution	2
Frameshift, Substitution	1
Frameshift, Deletion	1
Total	567

The majority of the mutations are single mutations of different types. However, there are a few known variants that have two different types of mutations. For my analysis, I will only consider the single-point substitution mutations. Table 7 and Table 8 show the distribution of all the known disease-causing variants for type I and type II keratins respectively.

Table 7: Keratin-wise distribution of type I variants.

Keratin	Count
K9	23
K10	30
K12	19
K13	7
K14	80
K16	19
K17	22
K18	10
K19	1
K25	2
Total	213

Table 8:Keratin-wise distributions of type II variants.

Keratin	Count
K1	38
K2	15
K3	3
K4	2
K5	112
K6a	34
K6b	4
K6c	2
K8	24
K71	1
K74	3
K81	3
K83	1
K85	1
K86	11
Total	254

Domain-wise distribution of keratin variants

The length of the coiled-coil region of keratins, excluding linkers, is 276 aa or ~40 heptad repeats. The length of the heads and tails varies significantly. K3 has the longest head at 198 aa, while K31, K33a, and K33b have the shortest head at only 56 aa. K2 has the longest tail (201 aa) and K19 has the shortest tail (8 aa). On average, the length of the heads is ~116 aa, longer than the average length of the tails at ~71 aa. Out of the three linkers, L12 is longest with a size of 16-17 aa and L2 is the shortest with a size of 8 aa. The size of all the domains and sub-domains of all human keratins can be seen in the Appendix.

Figure 17 shows the domain-wise distribution of mutations. The majority of these mutations are present in the peripheral regions of the proteins (head, 1A, 2B and tail). This could be because mutations in these regions are tolerated better compared to the mutations in the centre of the coiled-coil domain.

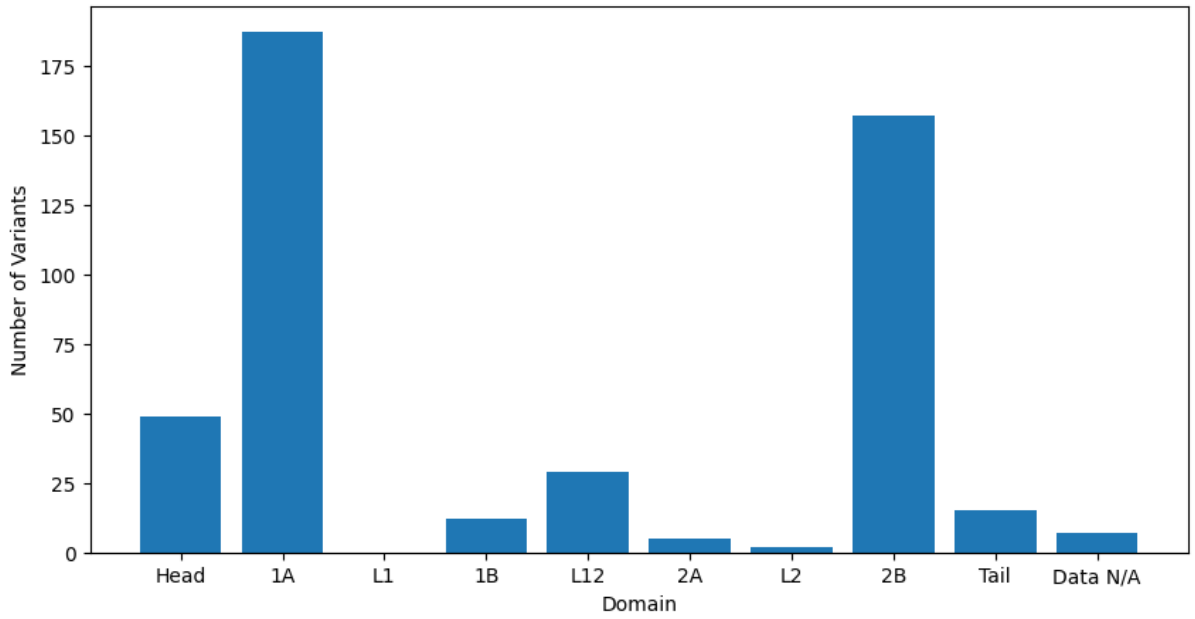


Figure 17: Domain-wise distribution of keratin variants

Figure 18 shows a heatmap of all the different amino acid mutations. Mutations with frequency >20 are all single-point nucleotide mutations. Proline to Leucine mutations are the highest represented in this dataset, followed by Lysine to Glutamic Acid mutations.

Mutations at 'a' and 'd' positions

'a' and 'd' positions contain mostly hydrophobic residues (~71%) since they form the hydrophobic core of the coiled-coil protein. Figure 19 shows the distribution of mutation types, where the majority are hydrophobic to hydrophobic mutations. There are also a significant number of hydrophobic to non-hydrophobic and non-hydrophobic to non-hydrophobic mutations at these positions (Figure 19).

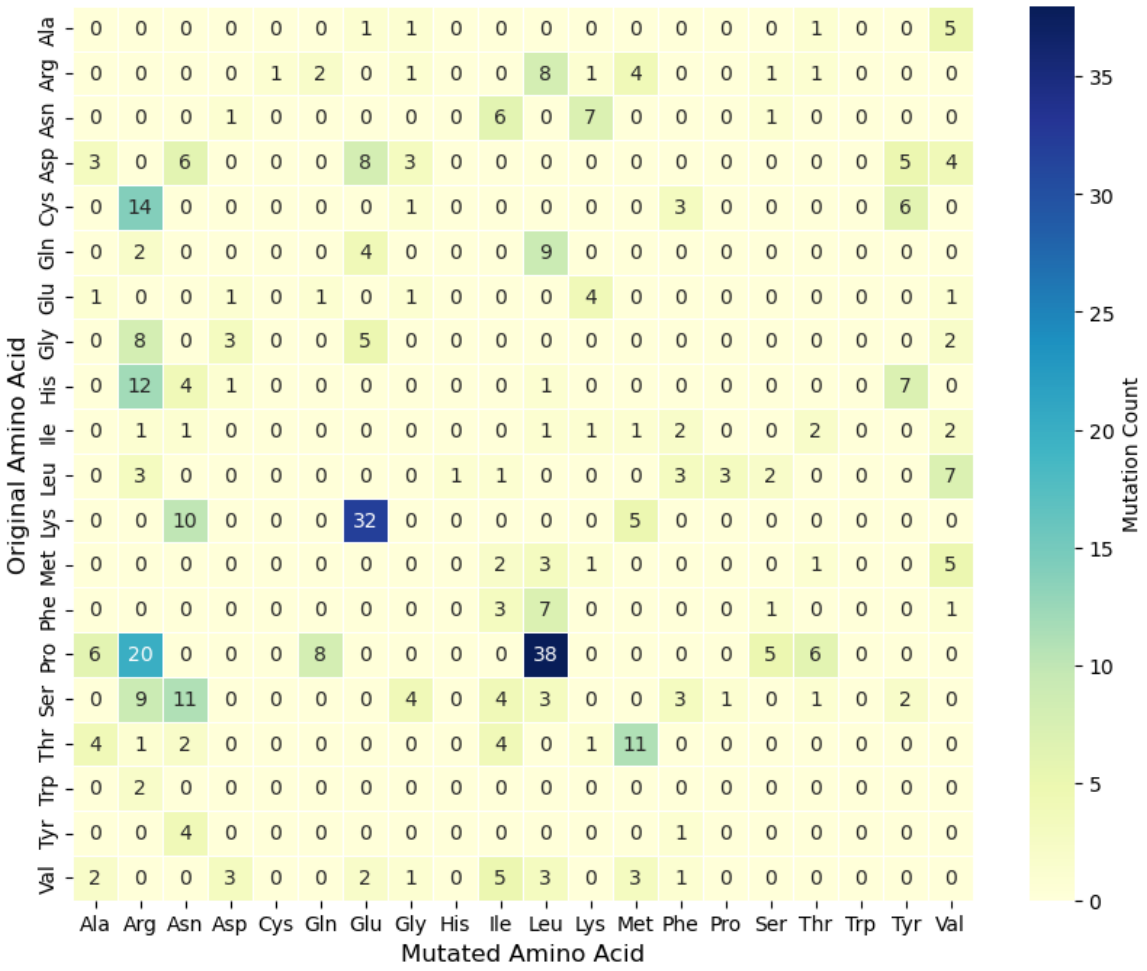


Figure 18: Distribution of all protein variants.

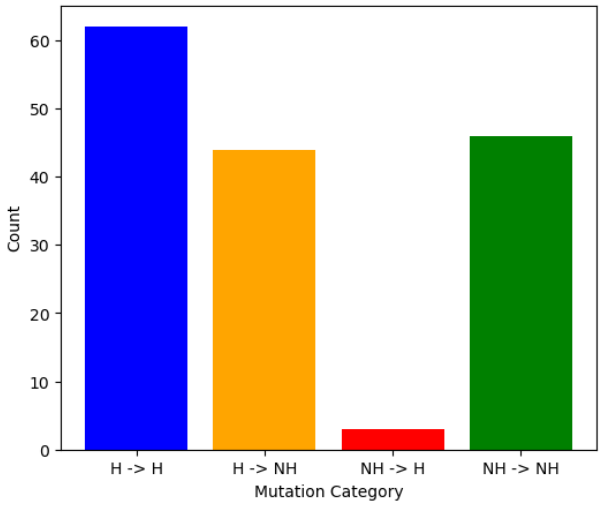


Figure 19: Distribution of variant types at 'a' and 'd' positions. H refers to hydrophobic and NH refers to non-hydrophobic.

Mutations at ‘b’, ‘c’, ‘e’, ‘f’ and ‘g’ positions

Residues at these heptad positions are mostly non-hydrophobic (~60%). ‘e’ and ‘g’ residues form intra-dimer salt bridges, while residues at ‘b’, ‘c’ and ‘f’ positions are responsible for the inter-dimer interactions that hold the filament together. The majority of the mutations at these positions are non-hydrophobic to non-hydrophobic, followed by non-hydrophobic to hydrophobic. Hydrophobic to hydrophobic, and hydrophobic to non-hydrophobic mutations present in the dataset as well, though these are less in number (Figure 20).

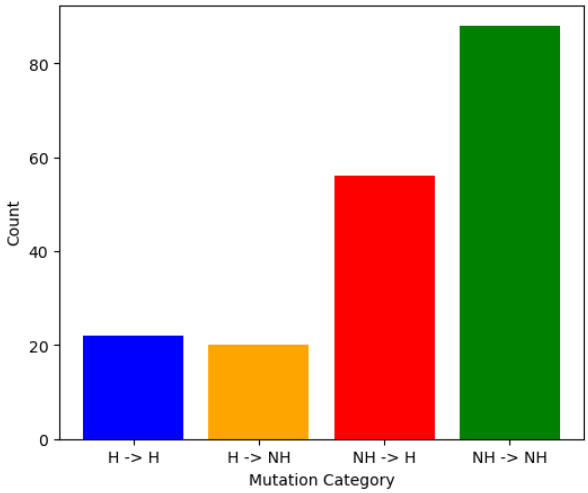


Figure 20: Distribution of variant types at 'b', 'c', ‘e’, ‘f’ and ‘g’ positions. H refers to hydrophobic and NH refers to non-hydrophobic.

Mutations in non-heptad regions

The non-heptad regions, that is, the heads, tails and linkers have all four classes of mutations. The distribution of these mutations in these regions (Figure 21) is similar to the overall distribution of all mutations (Figure 22)

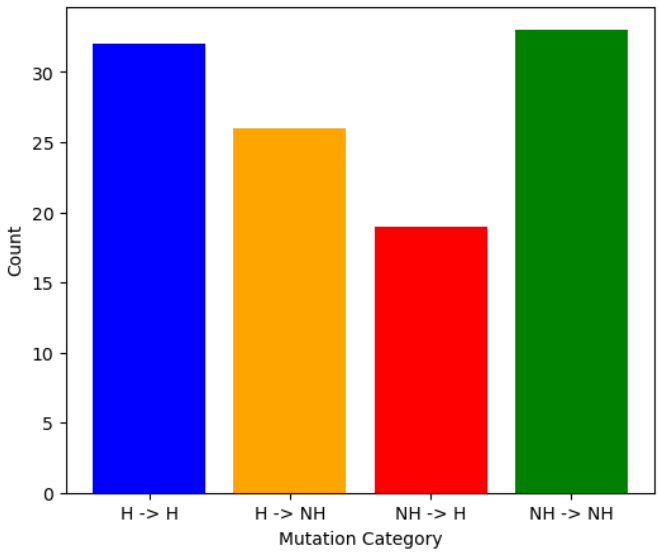


Figure 21: Distribution of mutations in non-heptad regions. H refers to hydrophobic, and NH refers to non-hydrophobic.

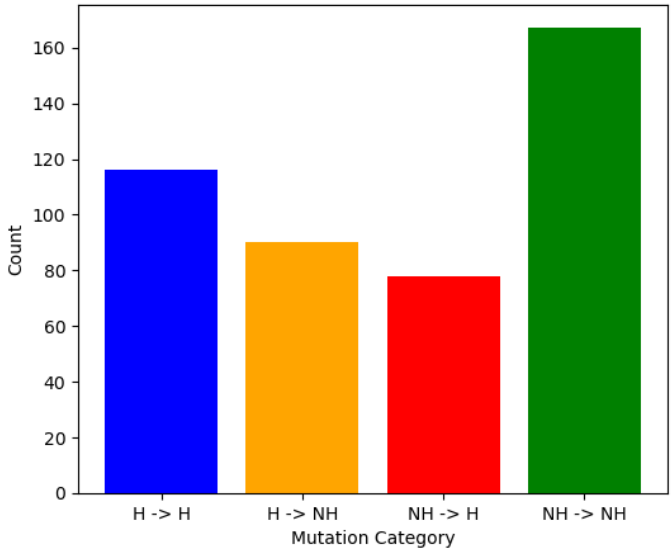


Figure 22: Distribution of all mutations based on change in hydrophobicity. H refers to hydrophobic, and NH refers to non-hydrophobic.

Figure 23 shows the mutations present at non-heptad positions. The highest frequency is seen by Lysine to Glutamic Acid, followed by Proline to Leucine.

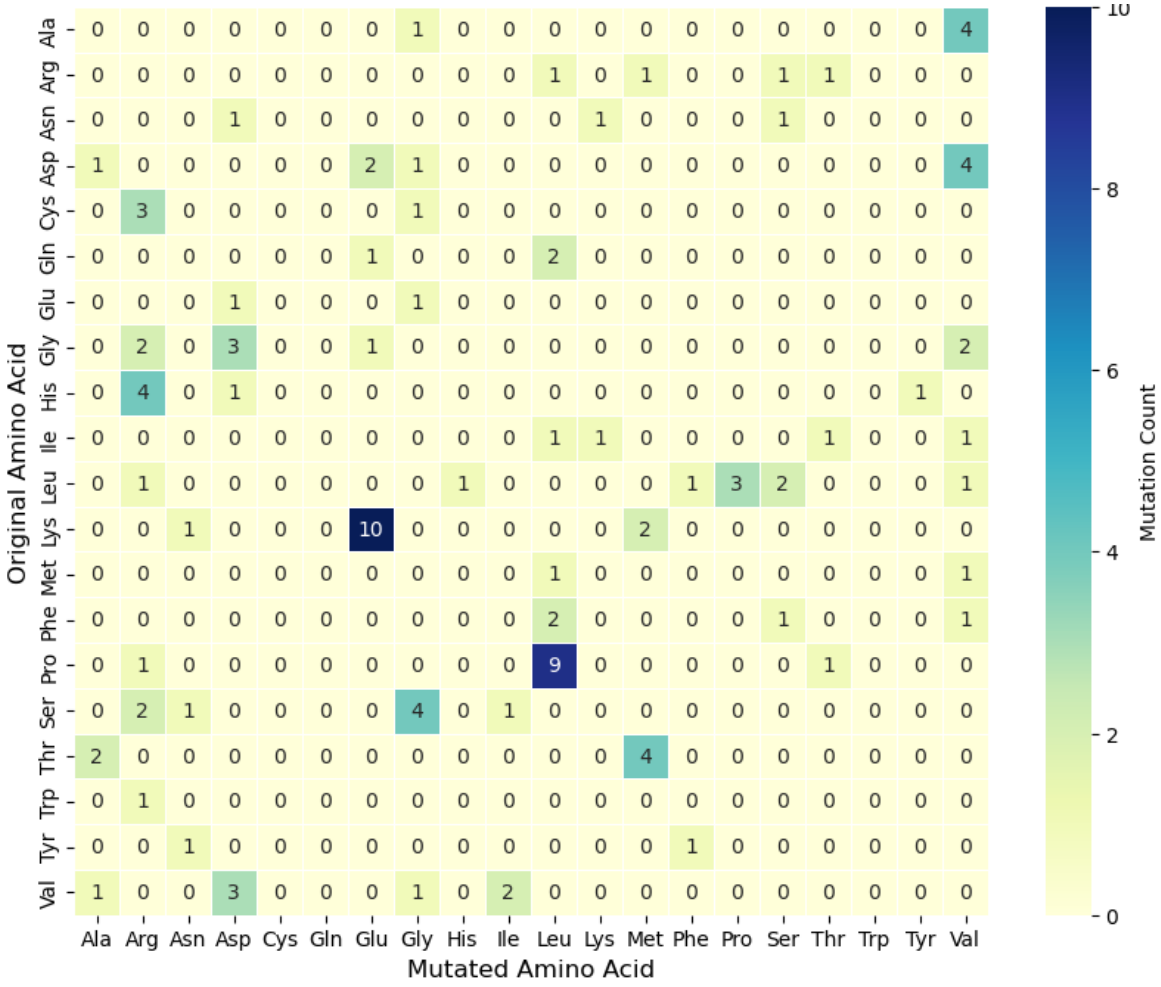
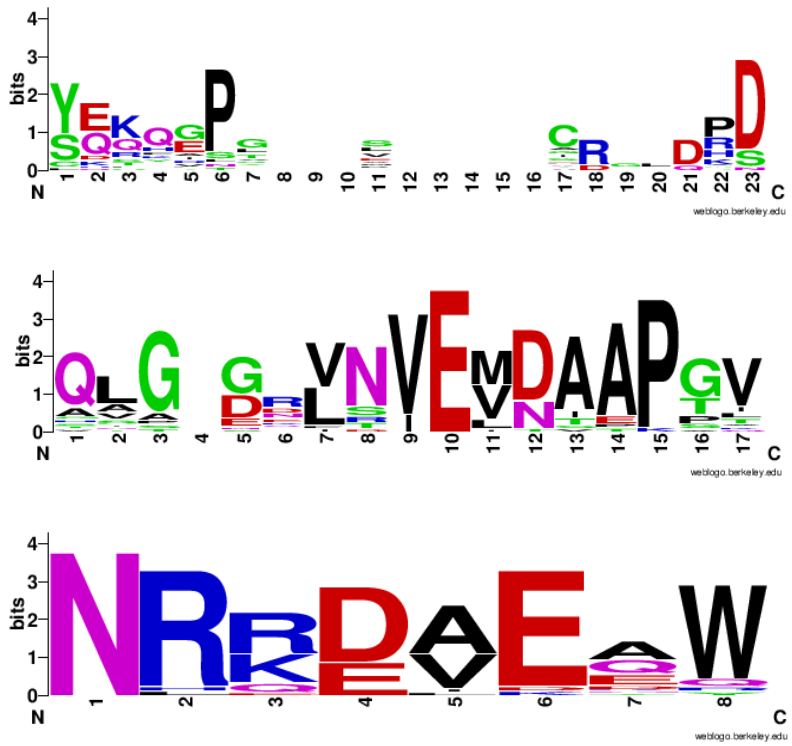


Figure 23: Heatmap of mutations present in non-heptad regions of keratins

Mutations in linkers

As mentioned before, the length of these linkers is similar in different keratins. Multiple sequence alignment of the linkers in human keratins revealed that certain residues are well-preserved in these regions. L1 is least conserved and L2 is most conserved in both type I and type II keratins. However, the functional relevance of these residues is unknown.

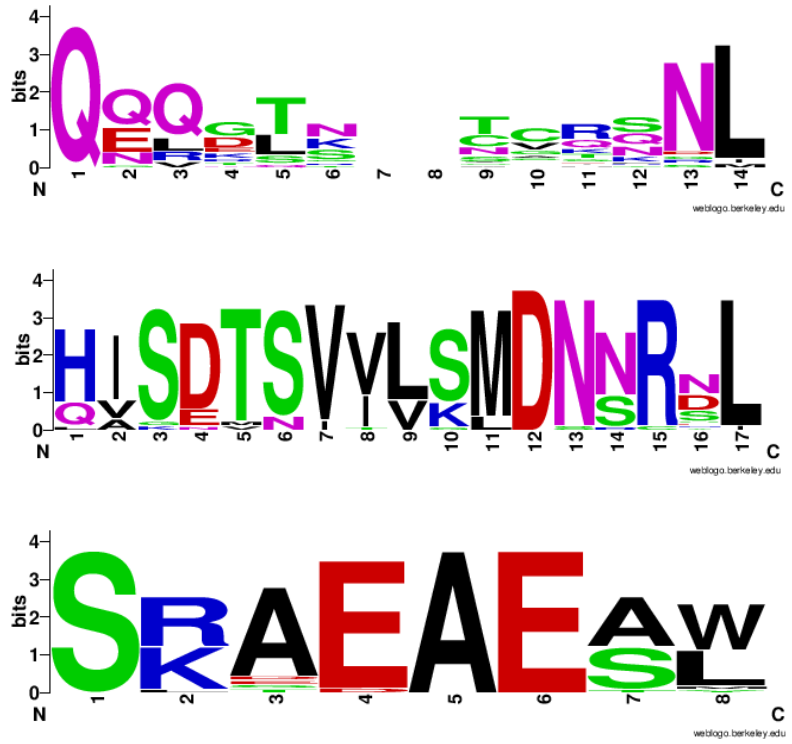


L1

L12

L2

Figure 24: Conservation of linker regions in type I IFs.



L1

L12

L2

Figure 25: Conservation of linkers in type II IFs.

Analysis of keratins in model organisms

Data Curation

The organisms were chosen from different families based on practicality. We wanted to sample organisms that have some known IFs. However, most of these sequences are unreviewed, and domain and localisation data is unavailable.

The following animals were chosen for the analysis:

1. *Balaenoptera musculus* (Blue whale)
2. *Bos taurus* (Buffalo)
3. *Camelus dromedarius* (Camel)
4. *Felis catus* (Cat)
5. *Gallus gallus* (Chicken)
6. *Pan troglodytes* (Chimpanzee)
7. *Canis lupus* (Dog)
8. *Vulpes vulpes* (Fox)
9. *Cavia porcellus* (Guinea Pig)
10. *Equus asinus* (Horse)
11. *Lemur catta* (Lemur)
12. *Macaca mulatta* (Monkey)
13. *Mus musculus* (Mouse)
14. *Pongo abelii* (Orangutan)
15. *Sus scrofa* (Pig)
16. *Columba livia* (Pigeon)
17. *Oryctolagus cuniculus* (Rabbit)
18. *Rattus norvegicus* (Rat)
19. *Tursiops truncatus* (Turtle)
20. *Canis lupus dingo* (Wolf)
21. *Equus quagga* (Zebra)
22. *Danio rerio* (Zebrafish)
23. *Taeniopygia guttata* (Zebrafinch)

After filtering and standardising the nomenclature, I found that different organisms contain different numbers of keratins. Figure 26 shows the keratins present in each organism.

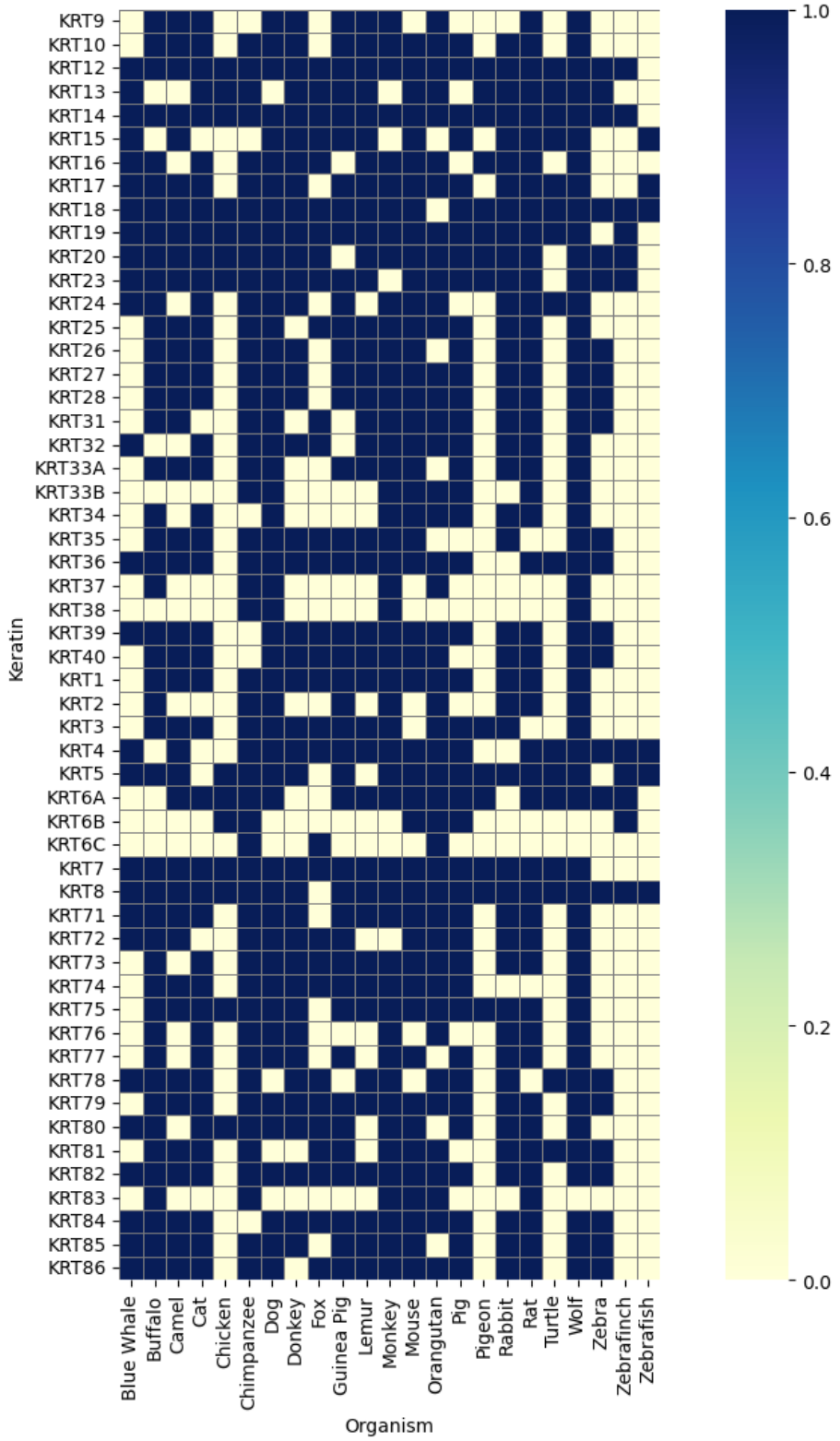


Figure 26: Heatmap showing the keratins present in specific organisms. Blue indicates that the keratin is present and yellow indicates that it is not.

Organism-specific analysis

For each organism, I separated the type I and type II keratins, and then did an MSA. I used this to calculate the mean identity and similarity as well as the standard deviation for these groups of proteins.

Table 9: Similarity and identity for type I IFs in different organisms.

Organism	Number of Sequences	Identity		Similarity	
		Average	Standard Deviation	Average	Standard Deviation
Human	28	46.89	11.62	55.5	10.16
Blue Whale	14	44.79	10.21	53.33	9.54
Buffalo	24	45.36	10.96	53.82	9.75
Camel	20	44.93	10.8	53.59	9.56
Cat	24	47.86	11.95	56.21	10.51
Chicken	8	51.53	12.49	59.66	10.91
Chimpanzee	24	47.9	11.61	56.47	10.15
Dog	28	44.72	15.14	52.88	14.19
Fox	17	48.74	12.04	57.43	10.64
Guinea Pig	20	45.68	11.13	54.1	10.04
Horse	22	43.65	12.93	51.95	12.41
Lemur	23	47.41	10.93	56.23	9.78
Monkey	25	47.29	12.45	55.75	10.85
Mouse	28	49.24	12.54	57.69	11.03
Orangutan	22	46.99	10.73	55.57	9.52
pigeon	21	47.73	12.38	56.12	10.55
Pigeon	8	45.33	7.82	54.47	7.76
Rabbit	23	46.01	11.7	54.31	10.6
Rat	30	46.98	12.56	55.49	10.87
Turtle	9	51.9	9.04	60.38	7.88
Wolf	30	45.9	14.62	54.22	13.76
Zebra	14	44.82	10.14	53.81	8.81
Zebrafinch	6	45.41	9.57	54.55	9.53
Zebrafish	3	53.22	10.35	63.31	9.4

Table 10: Identity and similarity of type II IFs in different organisms.

Organism	Number of Sequences	Identity		Similarity	
		Average	Standard Deviation	Average	Standard Deviation
Human	26	54.17	10.82	62.36	9.22
Blue Whale	12	45.47	9.88	54.19	9.57
Buffalo	23	51.32	10.11	59.97	8.34
Camel	20	52.41	8.21	61.16	6.83
Cat	22	51.76	10.15	59.95	8.83
Chicken	11	60.13	16.94	67.24	14.36
Chimpanzee	27	52.57	11.39	60.61	9.78
Dog	23	52.63	10.46	60.61	9.02
Fox	16	51.25	9.73	59.72	8.5
Guinea Pig	23	50.47	9.99	59.27	8.65
Horse	20	49.83	10.24	57.99	9.29
Lemur	18	53.05	8.89	61.73	7.53
Monkey	25	53.2	10.66	61.63	9.11
Mouse	25	52.58	11.51	61.39	9.93
Orangutan	26	54.93	10.7	63	9.09
pigeon	23	51.98	9.75	60.67	8.42
Pigeon	6	54.15	14.2	62.11	11.43
Rabbit	22	48.9	11.52	57.12	10.68
Rat	28	51.13	12.43	59.72	10.39
Turtle	7	52.68	7.66	61.19	6.95
Wolf	25	51.36	10.52	59.52	8.93
Zebra	11	53.61	10.19	62.35	8.5
Zebrafinch	5	57.82	14.33	64.81	12.49
Zebrafish	3	63.11	12.72	72.97	9.16

When building phylogenetic trees for all the above organisms, I found that all the keratins could be clustered into type I and type II. Figure 27, Figure 28 and Figure 29 show the phylogenetic trees for blue whale, orangutan and pigeon respectively. All type I's are clustered in green and type II's are clustered in orange.

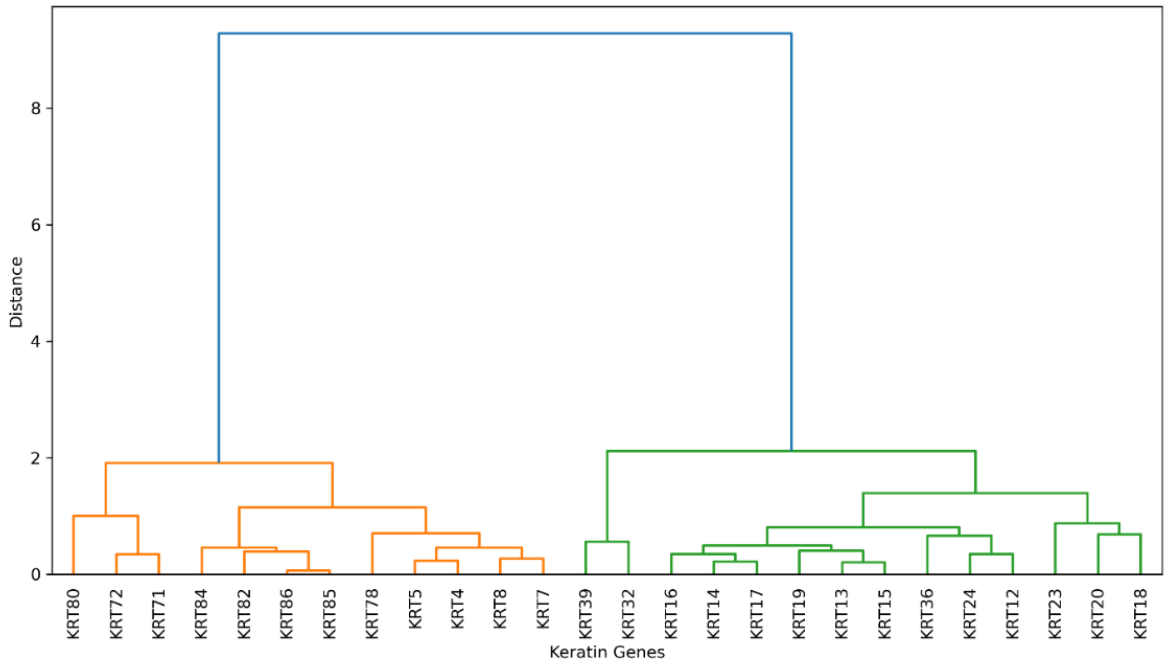


Figure 27: Phylogenetic tree of keratins in blue whale.

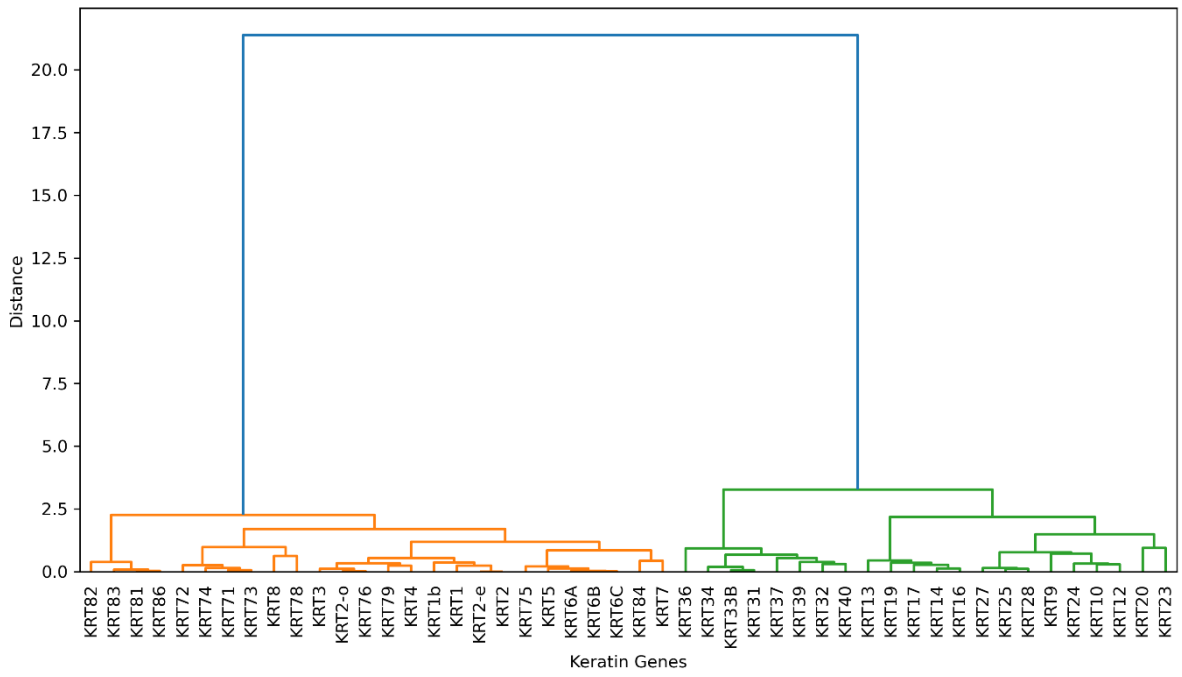


Figure 28: Phylogenetic tree for keratins in orangutan.

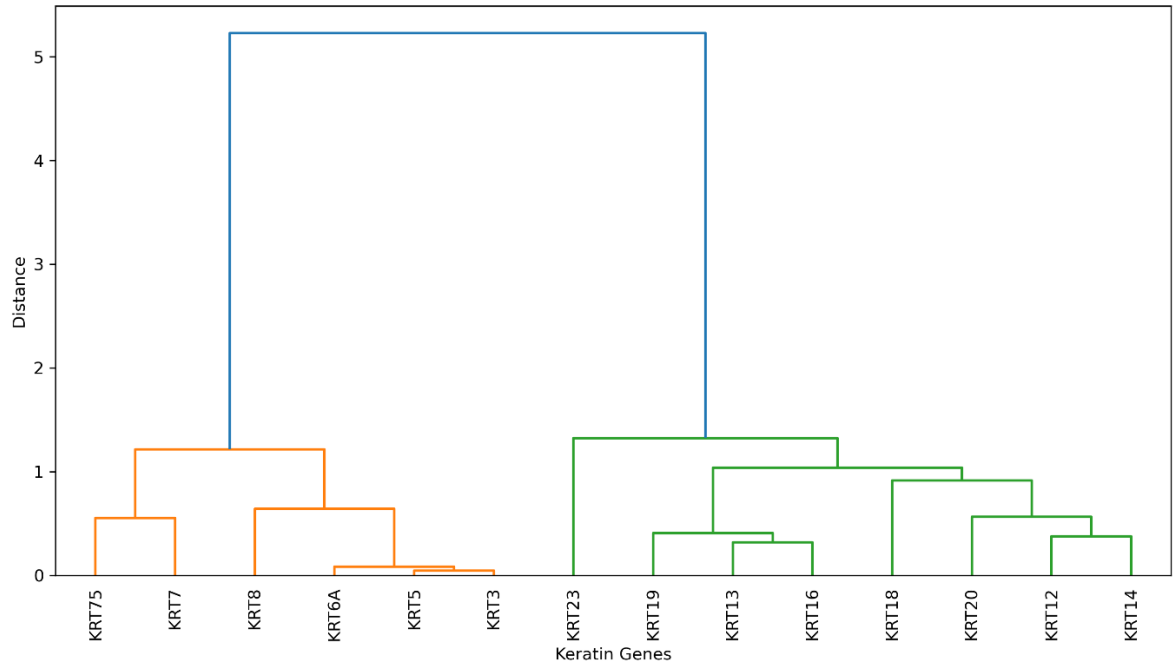


Figure 29: Phylogenetic tree of keratins in pigeon.

Keratin-ortholog analysis

Upon finding orthologs, I did MSA for a particular keratin for all the organisms. To make sure I had a sizable dataset, I chose genes which are represented in at least 15 organisms (excluding humans). There were a total of 28 such keratins. The average percentage identity and similarity, and the standard deviation for the same is shown in

Table 11. KRT85 was found to be the most conserved, with an identity of 89%. KRT20 was the least conserved with an identity of 69%.

Additionally, I decided to look at the identity of different domains of keratins in different organisms. Since domain information is not available for the sequences, I used human keratins as a reference. I aligned the sequences and used the human domain boundaries to infer the head, coiled-coil and tail regions of my sequences. The average identity was calculated for each cluster of sequences. I found that while the average identity was lower for the head region compared to the coiled-coil region, the identity was still higher than 60% for all the chosen keratins. So, we may say that the head region is conserved in keratins across different organisms.

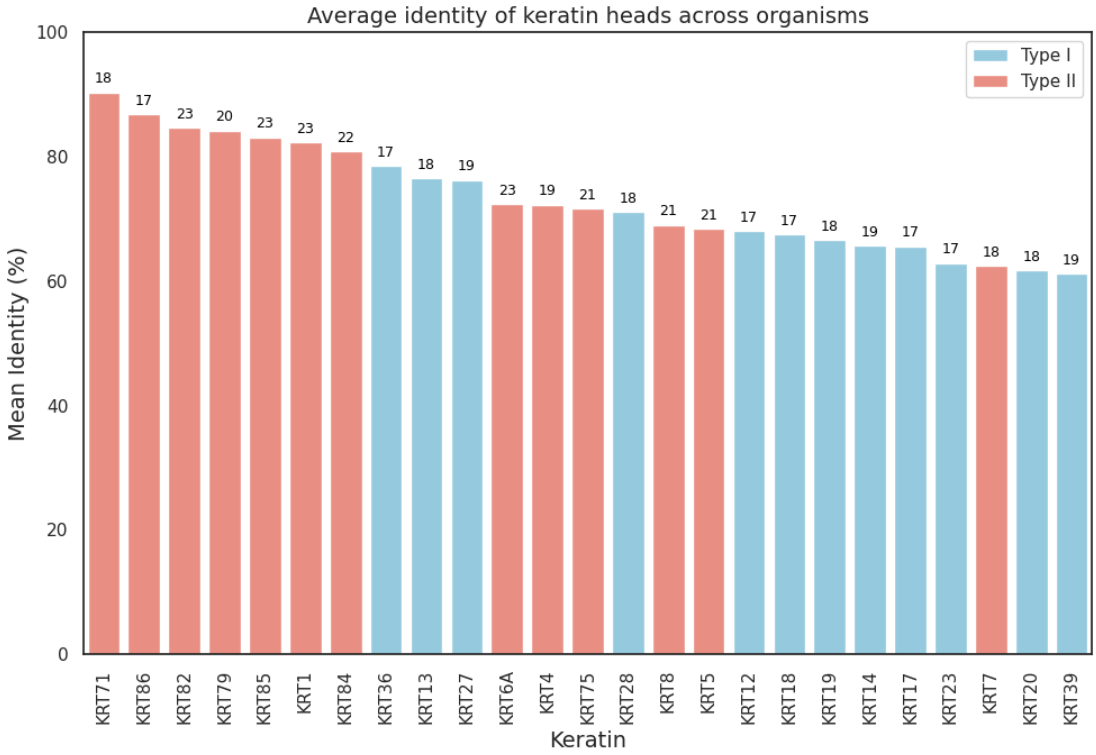


Figure 30: Average identity of head region of keratins present in different organisms. The count at the top of the bar is the number of organisms that the keratin is found in.

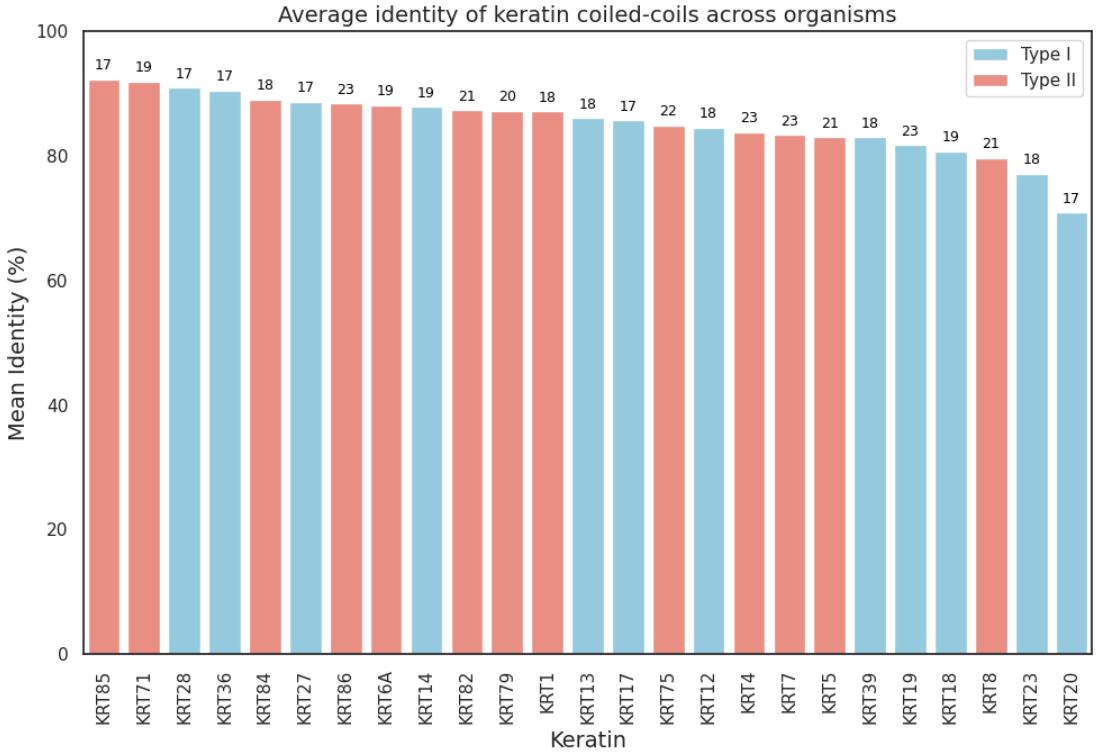


Figure 31: Average identity of coiled-coil region of keratins present in different organisms. The count at the top of the bar is the number of organisms that the keratin is found in.

On building a phylogenetic tree consisting of all of these sequences, we found that keratins from different organisms cluster together due to their high similarity.

Table 11: Similarity and identity of keratins across organisms.

Keratin	Number of Sequences	Identity		Similarity	
		Average	Standard Deviation	Average	Standard Deviation
KRT1	16	81.238	7.614	85.229	6.386
KRT10	16	85.148	5.766	87.052	5.287
KRT12	22	79.131	8.947	82.783	7.462
KRT13	17	80.743	11.000	85.073	9.098
KRT14	22	85.284	11.311	89.518	8.330
KRT16	16	78.486	14.588	82.464	12.632
KRT17	19	85.192	11.521	88.357	9.298
KRT18	22	76.845	14.569	83.110	11.166
KRT19	21	82.684	6.815	86.657	5.626
KRT2	15	81.189	6.048	85.432	4.954
KRT20	20	69.045	11.095	75.392	9.240
KRT23	20	72.521	14.763	78.450	11.421
KRT27	16	89.912	5.480	92.565	4.229
KRT28	16	86.610	4.830	89.676	3.915
KRT36	18	86.792	4.786	89.542	3.882
KRT4	18	77.045	12.285	81.365	9.943
KRT5	19	79.469	14.319	83.950	11.899
KRT6A	18	81.241	11.158	85.809	8.988
KRT7	20	76.962	12.687	81.509	10.998
KRT71	16	89.299	9.098	91.525	8.335
KRT75	17	78.421	10.930	83.147	9.253
KRT78	16	79.253	5.697	83.161	4.614
KRT79	17	85.170	4.834	88.339	3.879
KRT8	22	80.635	10.677	85.177	8.285
KRT82	18	85.416	5.142	88.345	4.173
KRT84	18	82.806	5.392	86.018	4.394
KRT85	16	89.721	7.069	91.665	6.929
KRT86	17	85.935	6.742	89.192	5.622

Chapter 4: Discussion

Keratins in humans: A structural perspective

Actin and tubulin-based cytoskeletons have been preserved in all eukaryotes; however, IFs are limited to metazoans. IF genes emerged ~500 million years ago and diversified rapidly (Ho et al., 2022). Today, they are found in a variety of organisms – sponges, arthropods, amphibians, mammals and more.

Understanding the structure, and mechanism of assembly and disassembly of IFs is key to uncovering why different cells have distinct mechanical properties. Modelling such a large and long coiled-coil structure using structure prediction methods like AlphaFold is challenging. Therefore, we have to rely on other approaches; integrative modelling was used in this case. This method was used to build a model of the K5K14 keratin IF, which forms the basis for my work.

Heads in IFs are crucial for the formation of the filament; if they are cut off, the filament does not assemble. Tail deletion studies of IFs have shown that tails are not necessary for assembly, though they may affect the structure and assembly of the filament (Omary et al., 2006). Therefore, I aimed to build a model of IFs with these regions to understand their mechanism of action.

I built a model of the K8K18 IF due to its short head and tail regions. Using homology modelling, I generated a structure of this protein both with and without its heads. Molecular dynamics simulations of the filaments demonstrated that the middle ULF showed less structural dynamics compared to the flanking regions. The RMSD of the model with the heads is higher than that without the heads, likely due to the unstructured head regions. This could also be due to the staggered nature of the dimers, leaving the “ends” exposed. This may be why the flanking regions in the model without heads also have higher RMSF than the middle ULF.

Several previous studies have shown that the low-complexity head regions form cross-beta interactions that hold the filament together (Eibauer et al., 2024; Omary et al., 2006; Rölleke et al., 2023). By analysing the hydrogen bonds in the MD simulations, we may be able to study such interactions.

Additionally, to compare the two models and see the mechanism of disassembly of the filament, I aim to carry out steered molecular dynamics simulations and apply various different external forces to see how it unravels. The current hypothesis in the field for the mechanism of stretching of the filament states that the alpha helices unravel and form beta sheets upon application of a force (Rölleke et al., 2023). Upon removal of the force, the protein may re-fold, given that the maximum force was below a certain threshold (Sapra & Medalia, 2021). I aim to study the helical properties of the dimers in the filament using MDs to test this hypothesis. Simulations have been carried out to test this, however, these were done at the dimer level and not the filament level (Rölleke et al., 2023).

Building a model of the K8K18 filament has helped us understand how to build models for several different IFs. This can be used to compare their structures and mechanisms of assembly. Using these models, we may be able to identify how other cytoskeletal proteins, such as microtubules, plakins, etc., interact with the IF network. This is key to understanding the mechanical properties of cells.

Human keratin variants: Structure to sequence

To verify the model, we can employ two approaches; we can explain existing data with this model, or make predictions that can be tested experimentally. We have first used the former method and tried to explain existing substitution mutations leading to some disease phenotypes. This dataset contains details regarding variants of multiple different human keratins.

Coiled-coil proteins follow a “knobs-into-holes” packing, where a residue from one helix (knob) packs into a cavity formed by the other monomer (hole). These motifs are highly stable and follow regular patterns. Soni determined several features that affect the stability of coiled-coil dimers of IFs.

The presence of a non-hydrophobic amino acid at ‘a’ and ‘d’ positions will disrupt the van der Waals interactions in the hydrophobic core and lead to instability. In addition to this, the volume of the hydrophobic amino acids at these positions is key for proper packing of the dimer. Change in volume due to hydrophobic to hydrophobic mutations may also lead to keratinopathies (Soni, 2019). Consequently, we see a large number of hydrophobic to

hydrophobic and hydrophobic to non-hydrophobic disease-causing mutations in the hydrophobic core.

~70% of all residues at these positions are hydrophobic. Non-hydrophobic residues can be tolerated via various mechanisms. The presence of oppositely charged residues can form a hydrogen bond; mutation at these positions to similarly charged amino acids may lead to repulsion that destabilises the dimer. Long-chain charged residues may be tolerated due to the hydrophobic interactions of the R-group carbons while the charged amine or carboxyl group can stick out of the dimer. Mutation of these residues to a smaller charged amino acid can introduce a charge to the hydrophobic core that may not be tolerated. This could be the reason for the significant number of non-hydrophobic to non-hydrophobic variants seen at 'a' and 'd' positions.

Residues at 'e' and 'g' positions form intra-dimer salt bridges; mutations at these positions may affect the stability of the dimer. Amino acids present at 'b', 'c' and 'f' positions are responsible for the inter-dimer interactions that hold the filament together. These residues are non-hydrophobic and are exposed to the environment. Mutations at these positions may disrupt inter-dimer bonds, or interactions with non-IF cytoskeletal proteins.

The majority of the non-heptad region mutations are present in the head and tail regions, which contain Serine and Glycine repeats. These Serines get phosphorylated by kinases and are key for disassembly of the filament. We do not see many disease-causing variants at these positions, likely due to the redundancy provided by the large number of Serines present in the heads and tails.

The heads and tails of dimers are positively charged, while the coiled-coil region is negatively charged (Figure 12). Mutations from Lysine (positively charged) to Glutamic Acid (negatively charged) may lead to instability of the dimer due to negative-negative charge repulsion between the head and the coiled-coil domain.

Analysing which interactions are disrupted in these variants may give us insight into the mechanism of various keratinopathies. Although, since we only have a model of the K8K18 filament with the heads, we need to fully build structures of other keratins as well.

It is crucial to note that this dataset does not contain variants that do not show a phenotype or (embryonic) lethal mutants. We also have to take into account the recovery of partial

phenotypes due to presence of other keratins. Therefore, more experimental data is needed to fully understand the pathology of these phenotypes.

Keratins: An evolutionary perspective

Despite their prevalence in metazoans, the structure of IFs is not fully understood in humans or other organisms. Annotated and high-quality sequence data is unavailable for most species. Phylogenetic studies found that, in humans, all type III, IV, V and VI genes appeared earlier in time than keratins (Ho et al., 2022). Type I and II genes appeared later in time but diversified rapidly (Preisner et al., 2018). I wanted to take a broader look and study the diversification of these genes, for which, I studied keratins in different organisms.

I used human keratins as a reference to determine orthologs, or the type of keratin, if that gene is not found in humans. I limited my analysis to 23 chosen organisms due to these practical considerations. I discovered that the proteins could be divided into distinct type I and type II IFs in all these organisms. I also observed that the keratins are well-conserved across organisms. Similarly to previous studies, I noted the high sequence identity between keratin genes in zebrafish and mice. This data suggests that the diversification of these genes occurred before the sea-to-land transition (Ho et al., 2022).

It is known that the coiled-coil domains of keratins are well-conserved in humans. We wanted to study the conservation of different domains across species. Given the lack of domain information, we used multiple sequence alignments with human proteins to infer these data. We found that in keratins across different organisms, the coiled-coil and the head domains are well-conserved.

Given the high degree of conservation of keratins across organisms, we can use human filaments as a template to build a structure of keratins in other organisms as well. This can be used to find structurally conserved regions in filaments, and build interaction networks with other cytoskeletal proteins.

We are aware that some redundancies exist in keratin genes; in the absence of its canonical interaction pair, a keratin may pair up with a different one to recover a partial phenotype. However, we do not know what drives their regulation and interactions. Using the Magnetic webserver, I have found that the keratin pairs are not regulated by the same transcription factors

(Chakraborty et al., 2025). Previous research shows that they may be regulated by the same family of transcription factors (Kalabusheva et al., 2023); it is possible that they have a master regulator.

In order to answer the aforementioned questions, more experimental data is needed, particularly *in vivo* studies and gene localisation data. Studying the structure, mechanism of assembly/disassembly and regulation of these genes is key to understanding the distinct mechanical properties of cells. This research may help in the development of sustainable biomaterials as well.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Astbury, W. T. (1933). *Fundamentals of fibre structure*. Oxford university press, H. Milford.

Bekker, H., Berendsen, H., Dijkstra, E., Achterop, S., Vonddrumen, R., VANDERSPOEL, D., SIJBERS, A., Keegstra, H., & RENARDUS, M. (1993). GROMACS - A PARALLEL COMPUTER FOR MOLECULAR-DYNAMICS SIMULATIONS: 4th International Conference on Computational Physics (PC 92). *PHYSICS COMPUTING '92*, 252–256.

Berendsen, H. J. C., Grigera, J. R., & Straatsma, T. P. (1987). The missing term in effective pair potentials. *The Journal of Physical Chemistry*, *91*(24), 6269–6271. https://doi.org/10.1021/j100308a038

Bork, P., Sander, C., & Valencia, A. (1992). An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. *Proceedings of the National Academy of Sciences*, *89*(16), 7290–7294. https://doi.org/10.1073/pnas.89.16.7290

Branden, C. I., & Tooze, J. (2012). *Introduction to Protein Structure* (2nd ed.). Garland Science. https://doi.org/10.1201/9781136969898

Chakraborty, A., Chopde, S., & Madhusudhan, M. S. (2025). Motif distribution in genomes gives insights into gene clustering and co-regulation. *Nucleic Acids Research*, *53*(1), gkae1178. https://doi.org/10.1093/nar/gkae1178

Conway, J. F., & Parry, D. A. D. (1988). Intermediate filament structure: 3. Analysis of sequence homologies. *International Journal of Biological Macromolecules*, *10*(2), 79–98. https://doi.org/10.1016/0141-8130(88)90015-3

Crick, F. H. C. (1952). Is alpha-keratin a coiled coil? *Nature*, *170*(4334), 882–883. https://doi.org/10.1038/170882b0

Crick, F. H. C. (1953). The packing of α-helices: Simple coiled-coils. *Acta Crystallographica*, *6*(8–9), 689–697. https://doi.org/10.1107/S0365110X53001964

Dominguez, R., & Holmes, K. C. (2011). Actin Structure and Function. *Annual Review of Biophysics*, *40*(Volume 40, 2011), 169–186. https://doi.org/10.1146/annurev-biophys-042910-155359

Edgar, R. C. (2004). MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*, 113. https://doi.org/10.1186/1471-2105-5-113

Eibauer, M., Weber, M. S., Kronenberg-Tenga, R., Beales, C. T., Boujemaa-Paterski, R., Turgay, Y., Sivagurunathan, S., Kraxner, J., Köster, S., Goldman, R. D., & Medalia, O. (2024). Vimentin filaments integrate low-complexity domains in a complex helical structure. *Nature Structural & Molecular Biology*, *31*(6), 939–949. https://doi.org/10.1038/s41594-024-01261-2

Francou, A., & Anderson, K. V. (2020). The Epithelial-to-Mesenchymal Transition in Development and Cancer. *Annual Review of Cancer Biology*, *4*(1), 197–220. https://doi.org/10.1146/annurev-cancerbio-030518-055425

Fuchs, E., & Weber, K. (1994). INTERMEDIATE FILAMENTS: Structure, Dynamics, Function and Disease. *Annual Review of Biochemistry*, *63*(Volume 63, 1994), 345–382. https://doi.org/10.1146/annurev.bi.63.070194.002021

Gowers, R. J., Linke, M., Barnoud, J., Reddy, T. J. E., Melo, M. N., Seyler, S. L., Domański, J., Dotson, D. L., Buchoux, S., Kenney, I. M., & Beckstein, O. (2016). MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Scipy*. https://doi.org/10.25080/Majora-629e541a-00e

Gruber, M., Söding, J., & Lupas, A. N. (2006). Comparative analysis of coiled-coil prediction methods. *Journal of Structural Biology*, *155*(2), 140–145. https://doi.org/10.1016/j.jsb.2006.03.009

Ho, M., Thompson, B., Fisk, J. N., Nebert, D. W., Bruford, E. A., Vasiliou, V., & Bunick, C. G. (2022). Update of the keratin gene family: Evolution, tissue-specific expression patterns, and relevance to clinical disorders. *Human Genomics*, *16*(1), 1. https://doi.org/10.1186/s40246-021-00374-9

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., Grubmüller, H., & MacKerell, A. D. (2017). CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nature Methods*, *14*(1), 71–73. https://doi.org/10.1038/nmeth.4067

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD - Visual Molecular Dynamics. *J. Molec. Graphics*, *14*, 33–38.

Ishikawa, H., Bischoff, R., & Holtzer, H. (1968). Mitosis and intermediate-sized filaments in developing skeletal muscle. *The Journal of Cell Biology*, *38*(3), 538–555. https://doi.org/10.1083/jcb.38.3.538

Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L. E., Brookes, D. H., Wilson, L., Chen, J., Liles, K., Chun, M., Li, P., Gohara, D. W., Dolinsky, T., Konecny, R., Koes, D. R., Nielsen, J. E., Head-Gordon, T., Geng, W., … Baker, N. A. (2018). Improvements to the APBS biomolecular solvation software suite. *Protein Science*, *27*(1), 112–128. https://doi.org/10.1002/pro.3280

Kalabusheva, E. P., Shtompel, A. S., Rippa, A. L., Ulianov, S. V., Razin, S. V., & Vorotelyak, E. A. (2023). A Kaleidoscope of Keratin Gene Expression and the Mosaic of Its Regulatory Mechanisms. *International Journal of Molecular Sciences*, *24*(6), Article 6. https://doi.org/10.3390/ijms24065603

Kapli, P., Yang, Z., & Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, *21*(7), 428–444. https://doi.org/10.1038/s41576-020-0233-0

Le, S. Q., & Gascuel, O. (2008). An Improved General Amino Acid Replacement Matrix. *Molecular Biology and Evolution*, *25*(7), 1307–1320. https://doi.org/10.1093/molbev/msn067

Letunic, I., & Bork, P. (2024). Interactive Tree of Life (iTOL) v6: Recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research*, *52*(W1), W78–W82. https://doi.org/10.1093/nar/gkae268

Liu, Y., van den Ent, F., & Löwe, J. (2024). Filament structure and subcellular organization of the bacterial intermediate filament–like protein crescentin. *Proceedings of the National Academy of Sciences*, *121*(7), e2309984121. https://doi.org/10.1073/pnas.2309984121

Ludwiczak, J., Winski, A., Szczepaniak, K., Alva, V., & Dunin-Horkawicz, S. (2019). DeepCoil-a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics (Oxford, England)*, *35*(16), 2790–2795. https://doi.org/10.1093/bioinformatics/bty1062

McKean, P. G., Vaughan, S., & Gull, K. (2001). The extended tubulin superfamily. *Journal of Cell Science*, *114*(15), 2723–2733. https://doi.org/10.1242/jcs.114.15.2723

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mitchison, T., & Kirschner, M. (1984). Dynamic instability of microtubule growth. *Nature*, *312*(5991), 237–242. https://doi.org/10.1038/312237a0

Mücke, N., Kreplak, L., Kirmse, R., Wedig, T., Herrmann, H., Aebi, U., & Langowski, J. (2004). Assessing the Flexibility of Intermediate Filaments by Atomic Force

Microscopy. *Journal of Molecular Biology*, *335*(5), 1241–1250. https://doi.org/10.1016/j.jmb.2003.11.038

Omary, M. B., Ku, N.-O., Tao, G.-Z., Toivola, D. M., & Liao, J. (2006). 'Heads and tails' of intermediate filament phosphorylation: Multiple sites and functional insights. *Trends in Biochemical Sciences*, *31*(7), 383–394. https://doi.org/10.1016/j.tibs.2006.05.008

Pauling, L., & Corey, R. B. (1953). Compound helical configurations of polypeptide chains: Structure of proteins of the alpha-keratin type. *Nature*, *171*(4341), 59–61. https://doi.org/10.1038/171059a0

Pollard, T. D., & Goldman, R. D. (2018). Overview of the Cytoskeleton from an Evolutionary Perspective. *Cold Spring Harbor Perspectives in Biology*, *10*(7), a030288. https://doi.org/10.1101/cshperspect.a030288

Preisner, H., Habicht, J., Garg, S. G., & Gould, S. B. (2018). Intermediate filament protein evolution and protists. *Cytoskeleton*, *75*(6), 231–243. https://doi.org/10.1002/cm.21443

RayChaudhuri, D., & Park, J. T. (1992). Escherichia coli cell-division gene ftsZ encodes a novel GTP-binding protein. *Nature*, *359*(6392), 251–254. https://doi.org/10.1038/359251a0

Rölleke, U., Kumari, P., Meyer, R., & Köster, S. (2023). The unique biomechanics of intermediate filaments – From single filaments to cells and tissues. *Current Opinion in Cell Biology*, *85*, 102263. https://doi.org/10.1016/j.ceb.2023.102263

Sapra, K. T., & Medalia, O. (2021). Bend, Push, Stretch: Remarkable Structure and Mechanics of Single Intermediate Filaments and Meshworks. *Cells*, *10*(8). https://doi.org/10.3390/cells10081960

Schrödinger, LLC. (2015). *The PyMOL Molecular Graphics System, Version 1.8*.

Soni, N. (2019). *Computational modelling of the 3D structure of intermediate filaments and molecular mechanism of associated diseases*.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., & Wu, C. H. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, *31*(6), 926–932. https://doi.org/10.1093/bioinformatics/btu739

Szeverenyi, I., Cassidy, A. J., Chung, C. W., Lee, B. T. K., Common, J. E. A., Ogg, S. C., Chen, H., Sim, S. Y., Goh, W. L. P., Ng, K. W., Simpson, J. A., Chee, L. L., Eng, G. H., Li, B., Lunny, D. P., Chuon, D., Venkatesh, A., Khoo, K. H., McLean, W. H. I., … Lane, E. B. (2008). The Human Intermediate Filament Database: Comprehensive information on a gene family involved in many human diseases. *Human Mutation*, *29*(3), 351–360. https://doi.org/10.1002/humu.20652

Turner, P. (2005). XMGRACE, Version 5.1. 19. *Center for Coastal and Land-Margin Research, Oregon Graduate Institute of Science and Technology, Beaverton, OR*, *2*(5), 19.

Webb, B., & Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics*, *54*(1), 5.6.1-5.6.37. https://doi.org/10.1002/cpbi.3

Yamada, S., Wirtz, D., & Coulombe, P. A. (2002). Pairwise Assembly Determines the Intrinsic Potential for Self-Organization and Mechanical Properties of Keratin Filaments. *Molecular Biology of the Cell*, *13*(1), 382–391. https://doi.org/10.1091/mbc.01-10-0522

Appendix

Classification of Amino Acids

Acidic: D, E

Basic: R, H, K

Polar: S, T, N, Q, C, Y

Hydrophobic: A, V, L, I, M, F, W, P, G

Size of different domains of human keratins

Table 12: Size of domains in human keratins.

Keratin	Head	1A	L1	1B	L12	2A	L2	2B	Tail
KRT1	180	35	12	101	17	19	8	121	150
KRT2	178	35	12	101	17	19	8	121	201
KRT3	198	35	14	101	17	19	8	121	114
KRT4	137	35	12	101	17	19	8	121	69
KRT5	168	35	12	101	17	19	8	121	108
KRT6A	163	35	12	101	17	19	8	121	87
KRT6B	163	35	12	101	17	19	8	121	87
KRT6C	163	35	12	101	17	19	8	121	87
KRT7	91	35	11	101	17	19	8	121	65
KRT8	91	35	10	101	17	19	8	121	80
KRT71	130	35	12	101	17	19	8	121	79
KRT72	125	35	12	101	17	19	8	120	72
KRT73	132	35	12	101	17	19	8	121	94
KRT74	140	35	12	101	17	19	8	121	75
KRT75	149	35	12	101	17	19	8	121	88
KRT76	183	35	12	101	17	19	8	121	141
KRT77	164	35	12	101	17	19	8	121	100
KRT78	111	35	12	101	17	19	8	121	95
KRT79	142	35	14	101	17	19	8	121	77
KRT80	83	35	10	101	17	19	8	121	57
KRT81	106	35	10	101	17	19	8	121	87
KRT82	120	35	10	101	17	19	8	121	81
KRT83	111	35	10	101	17	19	8	121	70
KRT84	165	35	10	101	17	19	8	121	123
KRT85	123	35	10	101	17	19	8	121	72

KRT86	106	35	10	101	17	19	8	121	68
KRT9	153	35	12	101	16	19	8	121	157
KRT10	146	35	14	101	16	19	8	121	123
KRT12	125	35	15	101	16	19	8	121	53
KRT13	103	36	19	92	23				45
KRT14	115	35	11	101	16	19	8	121	45
KRT15	105	35	12	101	16	19	8	121	38
KRT16	117	35	11	101	16	19	8	121	44
KRT17	84	35	11	101	16	19	8	121	36
KRT18	80	35	10	101	17	19	8	122	37
KRT19	80	35	11	101	16	19	8	121	8
KRT20	70	35	11	101	16	19	8	121	42
KRT23	72	35	11	101	16	19	8	120	39
KRT24	140	35	16	101	16	19	8	121	68
KRT25	79	35	15	101	16	19	8	121	55
KRT26	83	35	15	101	16	19	8	121	69
KRT27	84	35	15	101	16	19	8	121	59
KRT28	86	35	15	101	16	19	8	121	62
KRT31	56	35	11	101	16	19	8	121	48
KRT32	96	35	11	101	16	19	8	121	40
KRT33a	56	35	11	101	16	19	8	121	36
KRT33b	56	35	11	101	16	19	8	121	36
KRT34	98	35	11	101	16	19	8	121	26
KRT35	97	35	11	101	16	19	8	121	46
KRT36	93	35	11	101	16	19	8	121	62
KRT37	104	35	11	101	16	19	8	121	33
KRT38	104	35	11	101	16	19	8	121	40
KRT39	96	35	11	101	16	19	8	121	83
KRT40	89	35	11	101	16	19	8	121	30