COMPUTATIONAL MODELING OF 3D STRUCTURE OF KERATIN INTERMEDIATE FILAMENT BY SATISFACTION OF SPATIAL RESTRAINTS

A THESIS
SUBMITTED TO THE DEPARTMENT OF BIOLOGY
OF IISER PUNE
IN PARTIAL FULFILLMENT
FOR
FIVE YEAR BS-MS DUAL DEGREE PROGRAMME
UNDER SUPERVISION OF
DR. M.S. MADHUSUDHAN
BIOINFORMATICS INSTITUTE SINGAPORE

Neelesh Soni
April 2012

I certify that I have read this thesis and that, in my opinion, it is fully adequate in scope and quality as a dissertation for Five Year BS-MS Dual Degree Programme.

Dr.M.S.Madhusudhan Dr. Girish Ratnaparkhi

Approved for the Committee on UNDER-GRADUATE STUDIES IISER pune.

Abstract

Keratin intermediate filaments are the key structural component of epithelia and form cytoskeletal network. Till date, the molecular mechanism governing the keratin assembly are unknown. To get insight into keratin (and in general intermediate filament (IF)) assembly and function, we need to know their three-dimensional structures at atomic resolution.

To do this, Hybrid Modeling method was used. In Hybrid Modeling, experimental data is used as spatial restraints. Making use of molecular modeling and converting the chemical cross linking data into spatial restraints a near atomic resolution model of the Keratin (K5/K14) intermediate filament has been created. For the first time, filament structures are solved with almost all residues (missing head/tail).

The resulting structures reveals the configuration of lower order structures inside the filament. These models can help explain the mechanism of keratin related diseases. They can also be exploited to design experiments to manipulate cytoskeletal structure. Models at atomic level of detail are attractive starting point for the rational drug design. The present approach should be applicable to all proteins of IF family.

Acknowledgement

I would like to express my deep sense of gratitude to Dr. M.S.Madhusudhan, Principal Investigator, Bioinformatics Institute, Singapore for giving me a wonderful opportunity to work on this project. He was a constant source of inspiration and his continuous support and the faith he had in me from the very early stages of the project, helped me overcome all the obstacles and finish my project successfully.

My sincere thanks to Dr. Girish Ratnaparkhi for the invaluable support he gave me. Words are insufficient to express my deep sense of gratitude towards IISER, Pune and KVPY for giving me this wonderful opportunity to have hands on research at such an early stage in my career.

I would also like to thank my lab mates of Bioinformatics Institute, Singapore who were always willing to extend a helping hand in times of need.

I would also take this opportunity to thank Dr. K. N. Ganesh, Director, IISER Pune and Dr. L.S.Shashidhara,Head Biology Department, IISER Pune for providing us with state of the art amenities and Dr. Shivprasad Patil for his constant encouragement and much needed advice during times of need as my faculty advisor.

Last but not the least; I would like to thank my family and friends who were always there for me through my ups and downs throughout my life.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction to Intermediate Filament

Actin microfilaments (MFs), microtubules (MTs) and intermediate filaments (IFs) are the three major filament system of the cytoskeleton of all metazoan cells (Fuchs and Weber, 1994). The integrated networks formed by this filament system are involved in many cell processes and are also responsible for the mechanical integrity of the cell.

Intermediate Filament super family proteins are principle structural elements in nucleus, cytoplasm and epithelial cells, and with essential scaffolding functions in metazoan cells that share common structural and sequence feature. They have an average diameter of 10 nanometers, which is between that of MFs and MTs. In human, $\sim$75 different IF proteins are expressed (Sokolova et al., 2006). However different kinds of IFs have different properties based on the tissue types.

IF proteins are broadly classified in six different types based on the sequence homology (Szeverenyi et al., 2008):

- Type I: (Acidic) eg. Keratins K14, K10 etc.
- type II: (Basic) eg. Keratins K5, K8 etc.
- Type III: (Form homo and heteropolymeric proteins) eg. Desmin, Vimentin
- Type IV: eg. Neurofilaments, α-internexin, NF-H, NF-L

- Type V: Nuclear lamins
- Type VI: eye lens intermediate filaments

The main proposed function of IFs is to provide cells with resistance to mechanical stress caused by external forces or internal processes, e.g., during cell division or migration. The stress buffering functionality of these proteins is designed in their filamentous structure which is yet to be completely understood.

The molecular building blocks of IFs are fibrous proteins. They consist of long, often uninterrupted, segments of alpha helices which adopts a rope-like structure to shield the aqueous environment by forming multi stranded left-handed coils. Electron microscopy study revealed the coiled-coil IF protein dimers as distinct 45-50 nm long rod-like molecules (Aebi et al., 1988). The molecular mechanism of assembly of these IF protein dimers into 10-nm filaments (Figure 1.1), consisting of 24 to 40 monomers (according to mass-per-length measurements by scanning transmission electron microscopy (Fuchs, 1996), is still unclear. One of the main problems in understanding the mechanism is the unavailability of complete molecular structure of IF protein molecules.

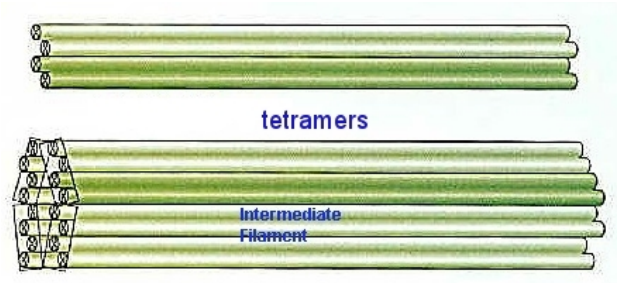

Figure 1.1: Schematic of IFs and lower order structure

The elementary IF dimer includes a central coiled-coil "rod" domain containing ~300 residues and non-alpha-helical N- and C-terminal domains (Strelkov et al., 2002). Rod domains have 4 segments 1A, 1B, 2A and 2B joined by the linker regions (Figure 1.2) . Atomic structures by X-ray crystallography were only solved for the segment 1A [PDB ID: 1GK7] (Strelkov et al., 2002) and two-third of segment 2B [PDB ID: 1GK4] for human vimentin (Strelkov et al., 2002), a type III IF protein, and for human lamin A, segment 2B [PDB ID: 1X8Y] (Strelkov et al., 2004). The C-terminal tail of human lamin A/C assumes immunoglobin fold of globular proteins hence its structure has been reported by crystallography [PDB ID: 1IFR] (Dhe-Paganon et al., 2002) and NMR solution studies [PDB ID: 1IVT]

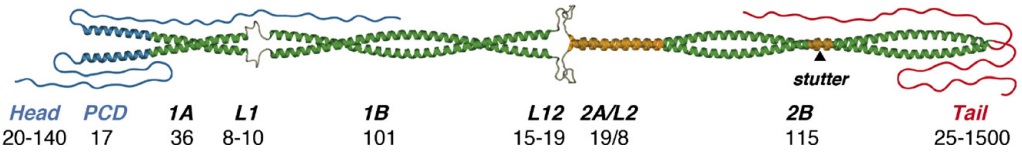

Figure 1.2: K5-K14 dimer are divided into four segments 1A, 1B, 2A, 2B joined by linker regions L1, L12, L2. (Herrmann et al., 2007)

(Krimm et al., 2002). The N- and the C-terminal regions of almost all IFs (except C-term of lamin) are unstructured but assume some specific inter/intra molecular interactions in the tetramer or higher oligomers (Figure 1.3) (Herrmann and Aebi, 2004). However, based on small-angle x-ray scattering experiments supplemented

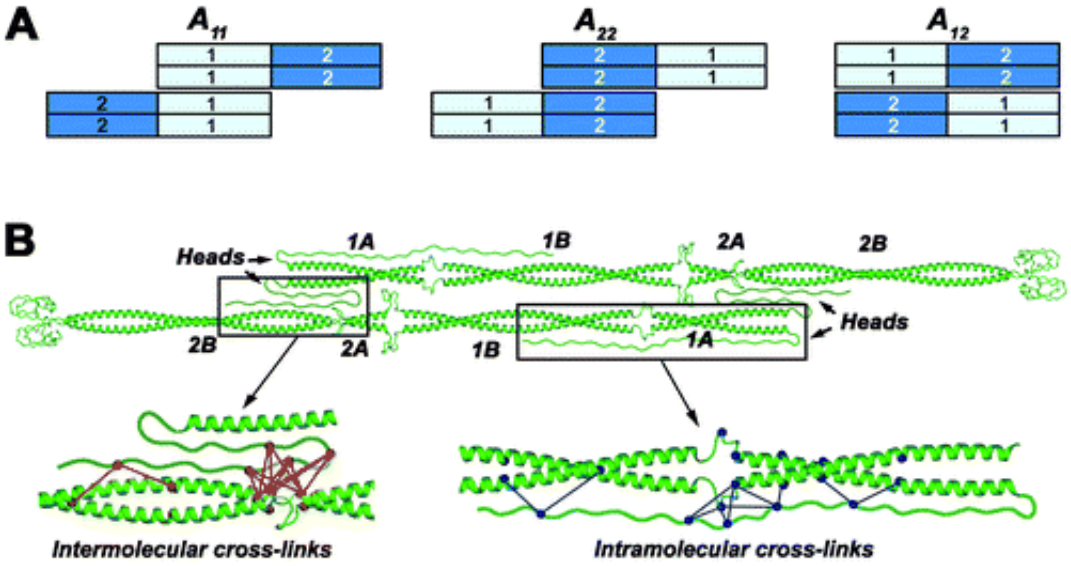

Figure 1.3: (A) Schematic model of dimer-dimer orientation as revealed from chemically cross-linked IFs. (B) Atomic model of a tetramer in A11 configuration with the head domains depicted according to the identification of cross-linking products(Herrmann and Aebi, 2004).

by crystallographic data and additional structural constrains, 3D molecular models of the vimentin tetramer, octamer, and unit-length filaments (ULF) were constructed (Sokolova et al., 2006) (Figure 1.4).

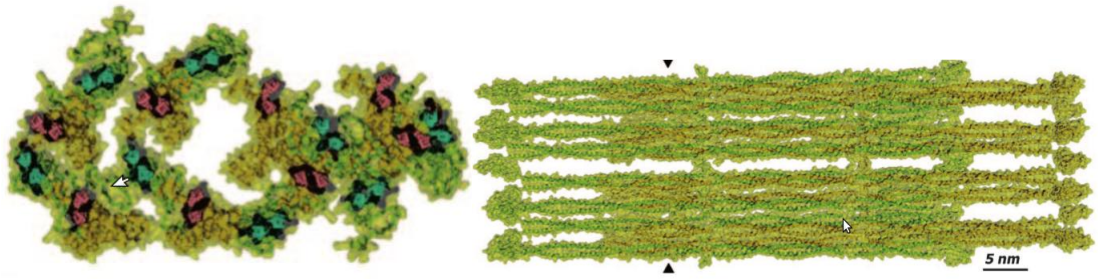

Figure 1.4: Vimentin filament model and its cross-section as given by 3D molecular model (Sokolova et al., 2006).

1.2 Keratin and diseases

Keratins are members of the intermediate filaments family of proteins that lines the periphery of epithelial cell. In humans, more than 54 functional keratin genes exist (Moll et al., 2008). Expression of these genes is highly specific in patterns related to the epithelial type and stage of cellular differentiation. Keratins in the epithelial cytoskeleton play a major functional role in mechanical stability and integrity of the epithelial cells through cell-cell contacts. They are also involved in intracellular signaling pathways, e.g. wound healing, protection from stress, and apoptosis. Consequently, from the single cell to the tissue formation, they are an inherent part of the system. Some mutations to the Keratin genes lead to diseases such as EBS that are hereditary. The importance of the genetic constitution of Keratins has been studied extensively using transgenic mice models (Lane and McLean, 2004).

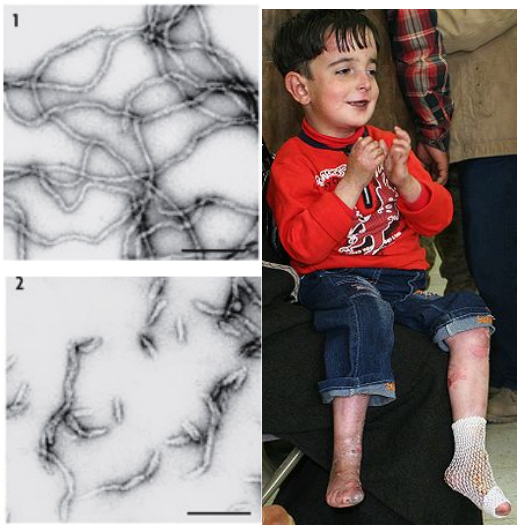

Figure 1.5: EM images of Keratin IFs.(Herrmann et al., 2007) (1)Normal (2)Diseased. (3)Individual with EBS disease.(Milks, 2012)

Approximately 85 distinct diseases have been associated with the intermediate filament gene family. These diseases are generally a result of point mutations in the protein sequence. In diseased individuals, keratins are susceptible to fragmentation. Figure 1.5 clearly shows fragmented keratins.

An accurate, near-atomic level resolution, structural model of Keratin filaments would give insight into the mechanism of filament assembly and how mutations cause disease. The latter would help us to rationally design drugs against disease conditions.

1.3 Keratin structure determination

Most IFs are assembly of dimers. Vimentin are homo-dimeric assembly but keratin are hetero-dimeric assembly. Keratin dimer molecules initially assemble as a heterodimer of type I and type II polypeptides. These keratin molecules are thus the most complex ones out of all IF proteins. Their structural aspects are less clearly understood than other types of IF proteins. Keratins first dimerize forming coil-coiled proteins. Several such dimers further oligomerize to form filaments that are ~10-12nm in thickness (Figure 1.6). Due to filamentous structure, these proteins are hard to crystallize. It is also difficult to obtain the NMR structure due to their size constraint. Thus very little structural information is

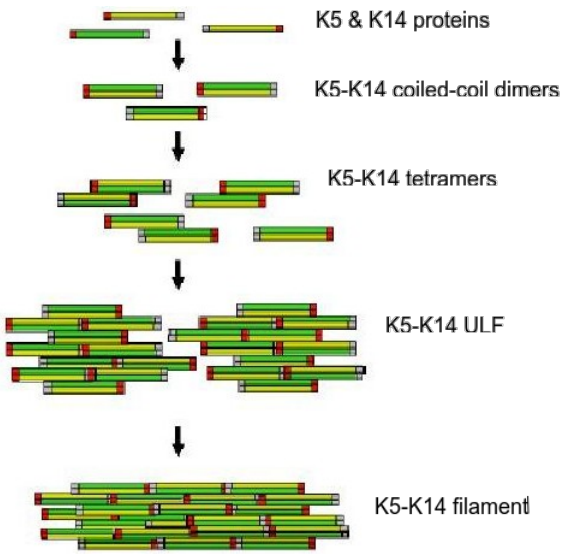

Figure 1.6: Schematic diagram of filament formation. Initially K5 & K14 proteins form coiled-coil dimer which then co-assemble to form higher-order structures.

available even at the dimer level. Key structural information available is that keratin heterodimers have coiled-coil geometry at the central region with unstructured head and tail regions.

The broad aim of this work is to model the unit length filament (ULF) formed by the hetero-dimeric assembly of keratin K5 (Type II) and K14 (Type I) at atomic resolution. To achieve this aim, hybrid modeling method is used which is introduced in the next section.

1.4 Hybrid Modeling

The standard scientific cycle includes the assembling of data, proposing hypotheses, and then gathering more data to test and refine those hypotheses. Hybrid Modeling or Integrative modeling implies a computational conversion of this scientific cycle (figure 1.7). It continues with the cycle of generating & gathering experimental data, design of representation and scoring, structure enumeration and analyzing the structures until the convergent ensemble of structure is found fitting the current information. The hybrid modeling approach has several advantages over conventional methods. It requires encoding modeling efforts into Integrative modeling applications to gain the advantages of hybrid modeling. These applications generally consist of scripts and associated information.

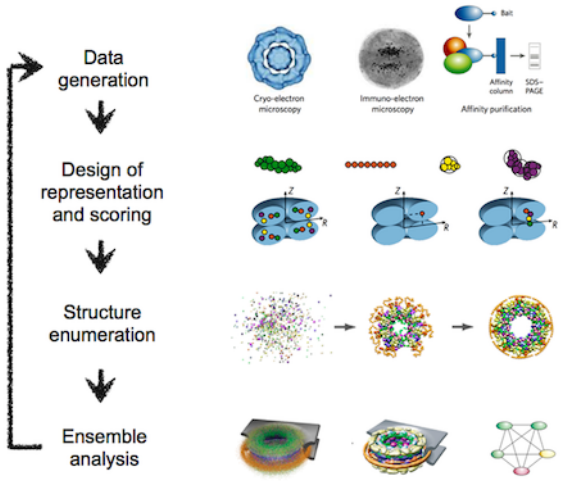

Figure 1.7: Stages of integrative modeling cycle(Alber et al., 2007).

The IMP software package alleviates the composing of these applications (Webb et al., 2011). This includes development of new representations, scoring functions, sampling techniques and the distribution of integrated modeling applications.

In IMP, each piece of the modeling system is encoded as a collection of particles. For example, a protein structure as given by a pdb file is represented as a hierarchy of residue type and atoms in each of the residues. Depending on the information available, representations can be atomic, coarse-grained or hierarchical. Hence a protein can be easily represented at any resolution from rigid bodies, to coarse grained, to full atom models. Each particle in IMP can have any number of attributes like label, coordinates, radius, atom type, rigid body composition etc. If inbuilt attributes are not sufficient more attributes can be added very easily. Models that need to be evaluated will be done using scoring function consisting of restraints. Restraints can be distance restraints, angle restraints, excluded volume restraint etc. Each of them measures how well the model fits the given information for restraints.

In this particular application of hybrid modeling, data from chemical cross-linking, electron microscopy (EM), small angle X-ray scattering (SAXS), and homology modeling are utilized. Each of these sources of data provides information about the system at different levels of atomic/molecular organization. For instance, EM and SAXS give the overall (sectional) shape of the filament, cross-linking identifies distances between residues within the filament, and homology modeling computes the structure of the Keratin dimers. The problem is akin to solving a jigsaw puzzle in 3 dimensions.

1.5 Thesis organization

In this work, we have used the Hybrid modeling technique to model the keratin K5-K14 ULF. A divide and conquer approach is used to reduce the sampling space. Chapter 2 introduces all the methods and approach used to generate the models. Several techniques like homology modeling, electrostatic analysis etc are explained in this section. Chapter 3 explains about the generated models and some deductions made from the models and their analysis. Features of the models that are in accordance with the experimental results are described here. Appendix A gives information about the cross-links observed experimentally.

Chapter 2

Methods

2.1 Comparative Modeling

Three dimensional protein structures are invaluable sources of information and are best determined by experimental methods such as X-Ray crystallography and nuclear magnetic resonance spectroscopy. However these methods have limitations that restrict the numbers of proteins to which they can be applied. In such cases, computational methods are useful for protein structure prediction. Following are some computational methods for protein structure prediction:

- Homology or Comparative Modeling
- Fold Recognition or threading Methods
- Ab initio methods that utilize knowledge-based information
- Ab initio methods without the aid of knowledge-based information

Comparative modeling and threading are based on statistical learning while Ab inito methods are based on the laws of physics. Among these methods, comparative modeling gives the most accurate results. In this work, comparative modeling technique is significantly used to model the protein structure. Hence the following some sections briefly describes the comparative modeling and its application.

2.1.1 Introduction to Comparative Modeling

Comparative modeling, also known as homology modeling of protein, refers to constructing an atomic-resolution model of the protein from its amino acid sequence. It uses the structural information from an experimentally determined three-dimensional structure of a related homologous protein (the "template"). Two conditions have to be met for predicting the protein structure by comparative modeling. First, the target sequence must be homologous to the template sequence of known structure. Second, accurate alignment between the target and the template sequence must be possible.

Comparative Modeling consists of 4 sequential steps: fold assignment & template selection, target–template alignment, model building, and model assessment (Figure 2.1) .

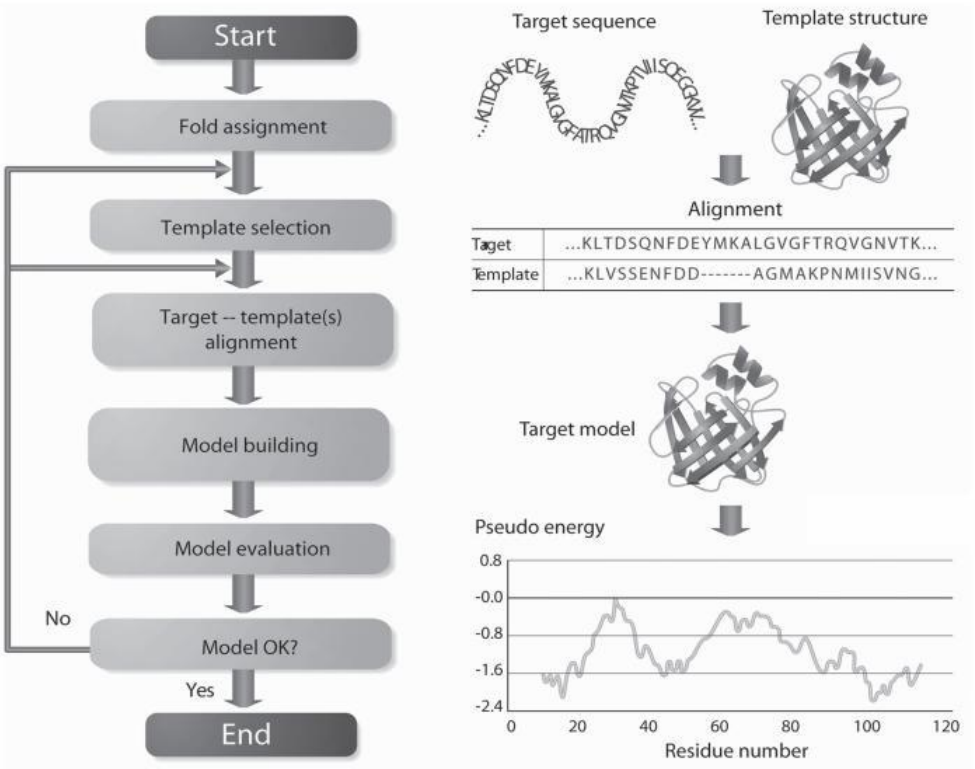

Figure 2.1: A flowchart of steps involved in Comparative Modeling (Eswar et al., 2006)

2.1.1.1 Fold Assignment and Template selection

This is the initial step in the comparative modeling, assigning the likely fold of the target sequence. Template selection can be achieved by scanning the sequence and structure databases, such as Protein Data Bank (PDB) (Westbrook et al., 2002) structural classification of proteins (SCOP) (Lo Conte et al., 2002) etc. This can be done by using the programs like BLAST(Altschul et al., 1990) and FASTA (Pearson, 1990), which aligns the target sequence with all the sequences in the databases of known structures.

2.1.1.2 Target-Template alignment

The next crucial step in comparative modeling is to align the target sequence to the template sequence. Even though template recognition methods produce a target-template alignment, there is sometimes a need to use a specialized alignment method to realign the sequences because the template-identification step is often optimized to identify distance relationship at the cost of alignment accuracy. This step of comparative modeling is an important step in the model building process because a wrong alignment will lead to incorrect model.

2.1.1.3 Model Building

The mapping of the target sequence on the template structure is generated by target-template alignment. This mapping is thus utilized to build the 3-D model of the target protein. Various model building methods exists for building the 3-D models as listed below:

- Modeling by Assembly of Rigid Bodies. Eg. COMPOSER (Sutcliffe et al., 1987) software programe.
- Modeling by Segment Matching or Coordinate Reconstruction. Eg. SegMod software program (Levitt, 1992).
- Modeling by Satisfaction of Spatial Restraints. Eg. Modeller software program (Sali and Blundell, 1993).

2.2 Comparative Modeling of K5-K14 Dimer

2.2.1 Template Selection

No suitable template was found to model the entire coiled-coil region of K5-K14 dimer. Different parts of the dimer can be modeled based on the crystal structures of fragments of another intermediate filament protein vimentin (1GK4, 1GK6, 1GK7) with sequence similarity of about 29%. However, it was difficult getting continuous coiled-coil conformation from fragment templates due to uncertainty in the linker regions. In order to construct the coiled-coil dimer, an appropriate template is needed, which should have regular coiled-coil geometry with two parallel strands. Hence, Leucine Zipper (PDB code: 2ZTA) was found to be of such proto-type, having regular coiled-coil geometry. Considering this, Leucine Zipper was selected as our template structure for K5-K14 dimer modeling (Figure 2.2 A) . Leucine Zipper,

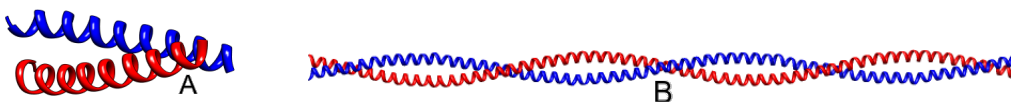

Figure 2.2: (A)Leucine zipper, (B)Generated long leucine zipper

2ZTA, was limited by its short stretch of amino acid sequences, only 31 residues in each chain. On the other hand, K5-K14 dimer has 313 residues in K5 chain and 311 residues in K14 chain (excluding the head and tail domains). Therefore, a long leucine zipper molecule (Figure 2.2 B) was generated by superimposing 16 2ZTA molecules using program CLICK(Nguyen et al., 2011). 10 residues at N-Terminal of one 2ZTA molecule are superimposed on 10 other residues at C-terminal of duplicate 2ZTA. The residues found redundant after superimposition were removed to make one continuous molecule with two chains winding around each other as coiled-coil dimer. Thus the molecule becomes a sufficiently long straight coiled-coil dimer with 692 residues to serve as a template for K5-K14 dimer without head and tail domains. Our models do not contain head and tail domains because these domains contain repeats of Serine and Glycine which are unstructured.

2.2.2 Heptad Repeat and Target-template alignment

An alpha-helical coiled-coil structure is easily recognized by a heptad repeat sequence (a-b-c-d-e-f-g), where positions a and d are commonly occupied by apolar residues, and positions e and g by charged and hydrophilic residues (figure 2.3) . The apolar residues form a left-handed seam that gradually winds round the axis of each of the constituent right-handed-helices (Parry, 2006). Positions “a” and “d” form the hydrophobic core of coiled-coil protein. Hydrophobic forces drive the assembly of the chains into supercoils to form multi-stranded ropes (classically these are left-handed), thereby shielding the apolar stripes from the aqueous environment. Electrostatic interactions, primarily involving residues in the e and g positions, play a particular role in specifying the alignment of the chains.

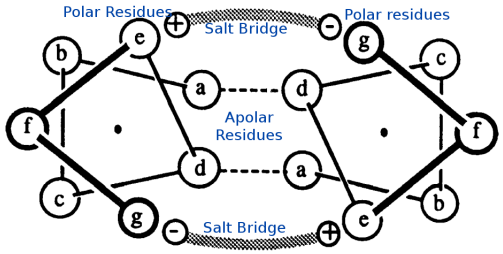

Figure 2.3: Position of Hydrophilic and Hydrophobic Residues. Positions “a” and “d” are occupied by Apolar residues while postions “g” and “e” are occupied Hydrophilic residues.

In our case, since target-template alignment is not homologous, alignment is done manually by matching the heptad position of cross-linked Lysine residues. Figure 2.4 shows the target-template alignment.

During this alignment, it is made sure that apolar residues occupy "a & d" positions while hydrophilic residues occupy "e & g" positions. Also final alignment is constructed such that the cross-linked lysine should not be buried.

Figure 2.4: Target(K5-K14 sequence)-Template(long leucine zipper sequence) Alignment with color code for different segments. Blue color represents 1A region, Green represents 1B region, Red represents 2A region and Orange color represents 2B. Highlighted sequence with yellow color are the linker regions. Chain break are represented using "\"

2.2.3 Model Building using Modeler

Using the target-template alignment, the model of K5-K14 is generated by using the MODELER suite of programs (Sali and Blundell, 1993). MODELLER calculates a 3D model of the target completely automatically, using its automodel class. Total 5 models are generated and the best one is selected based on the objective

function score, as the final model for K5-K14 dimer. Models are build such that the dimer remains straight. These models are based on the “long leucine zipper” as template structure and target-template alignment as constructed earlier.

2.2.4 Model Assessment

In this case of comparative modeling, the keratin k5-k14 dimer is built such that hydrophobic residues are buried inside, i.e., between k5 and k14 protein. Similarly, hydrophilic residues should not be buried. Figure 2.4 gives the schematic positions of the hydrophilic and hydrophobic residues. This model was constructed to have

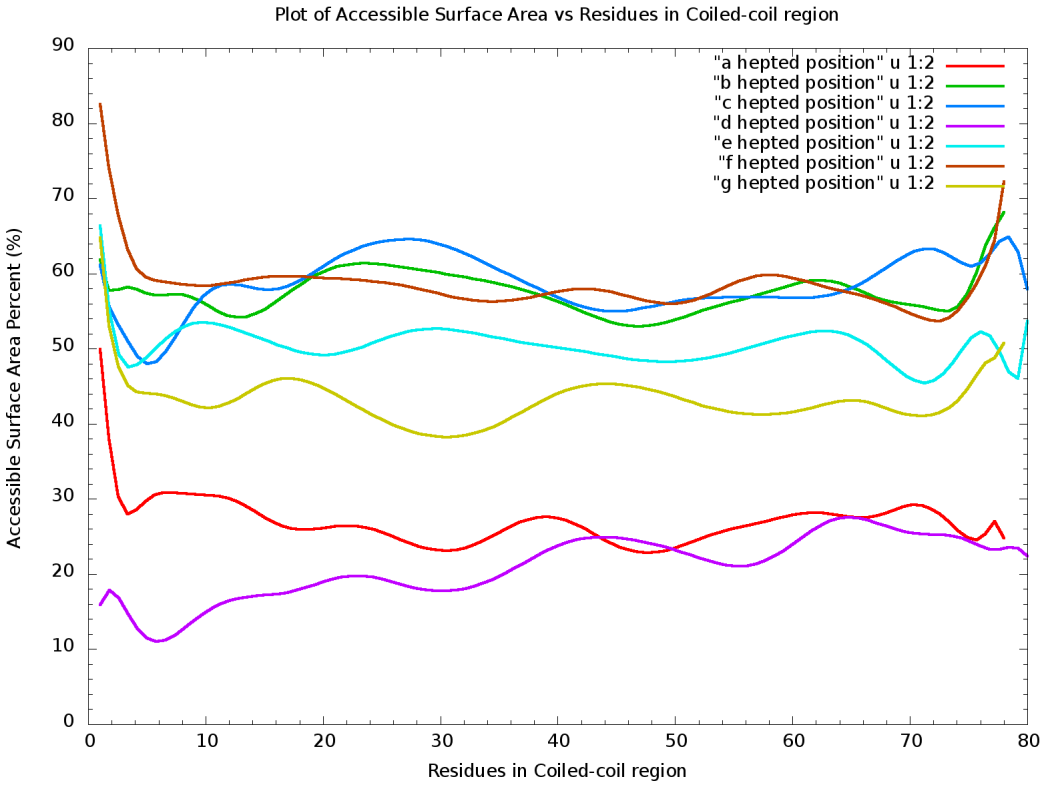

Figure 2.5: Plot (smooth) of Solvent Accessible Surface Area vs Residues for each heptad position on a coiled-coil dimer.

the coiled-coil geometry. Hence there is no need to do the separate assessment of the models. Figure 2.5 shows the Solvent Accessible Surface Area (SASA) of

residues in coiled-coil region with their heptad position in the coiled-coil geometry. Figure clearly shows the heptad positions “a” and “d” have the least SASA for all residues. Similarly, heptad positions “b”, “c” and “f” have higher SASA compared to others. Hence, k5-k14 dimer model appears to be a good model.

2.3 Hybrid Modeling of K5-K14 filament

2.3.1 Sources of Experimental Information

2.3.1.1 Chemical Cross linking

To obtain the adjacency of the molecules in the keratin filament, cross-linking experiments were performed on them. In two separate studies (Steinert et al., 1993b; Parry et al., 2001), they found 14 and 9 different intermolecular cross-links respectively. These two experiments were done in similar conditions except using higher resolution of HPLC and better reducing agent in the later experiment. The molar ratio of the cross-links found in (Steinert et al., 1993b) are in range $0.1 - 0.62$ mol/mol, but molar ratio of the cross-links found in (Parry et al., 2001) are not significant, $0.02 - 0.001$ mol/mol. Details of cross-links are in Appendix A.

The experiment was performed with intact k5-K14 filament using the cleavable bifunctional reagent disulfosuccinimidyl tartrate (DST) (Bragg and Hou, 1980)(figure 2.6). All reactions were performed for 30 min, and excess reagent was quenched with 0.1 M NH_4HCO_3. Several DST-cross-linked products were then digested to completion with trypsin enzyme. The peptides were then resolved by reverse-phase HPLC using a 60-min gradient of $0 - 40\%$ acetonitrile in 0.1% trifluoroacetate(Steinert et al., 1993a). Comparisons of two HPLC profiles before and after cleavage of the glycol DST bond with 0.1 M sodium periodate revealed several potential cross-linked species, which were then recovered.

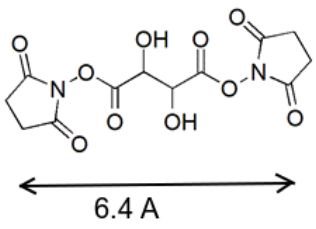

Figure 2.6: DST cross-linker with spacer arm length 6.4Å

Original cross-linking study led to the presentation of the two-dimensional Surface lattice model (Steinert et al., 1993b).

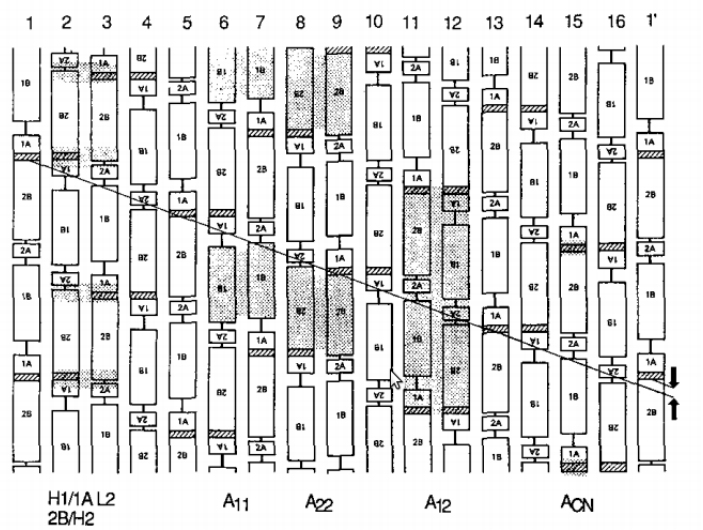

Figure 2.7: Two-Dimensional Surface Lattice Model was introduced in (Steinert et al., 1993b). Model is an arrangement of alternate anti-parallel dimers.

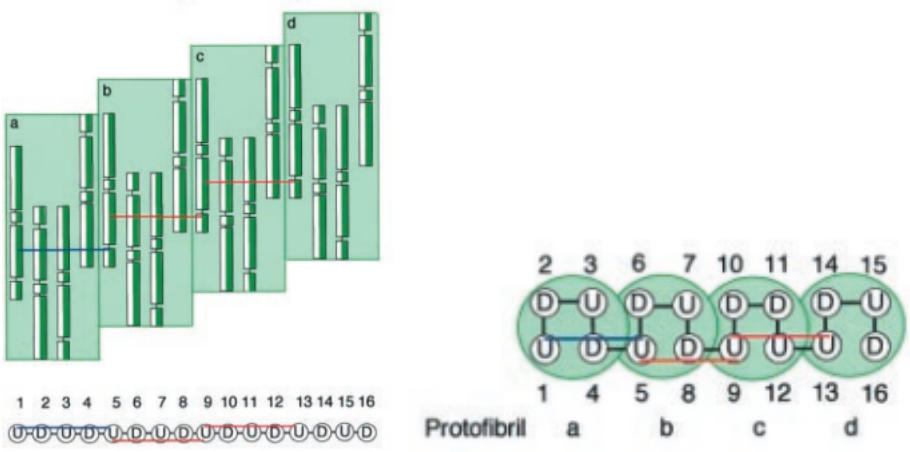

Figure 2.8: Tetramer arrangement in ULF. Cross-links observed in (Parry et al., 2001) was deduced to occur between 1^{st} and 5^{th} dimers in the Surface lattice model. Figure shows the association of of parallel dimers within the filament.

This model is an arrangement of alternate anti-parallel modes of association of dimers (figure 2.7). No explicit information about the three-dimensional structure

was provided in this study. Based on the cross-links observed in (Parry et al., 2001), parallel association of dimers were introduced. The 2-D Surface Lattice model were extended by showing additional cross-links between 1^{st} and 5^{th} dimer in the model (figure 2.8). Details of cross-links are given in appendix A.

2.3.1.2 Number of dimers in the filament

From electron-microscopy, the cross-section of keratin filament was found to be $10nm$ (Aebi et al., 1983). The number of the dimers to be fitted in filament cross-section has been calculated in two different ways, based on the cross-sectional area ratio and volume ratio of filament to the dimer. Volume of a dimer is calculated based on atomic volumes. Volume of unit length filament is calculated from the cylindrical geometry, where cross-sectional radius is from EM study and the unit length is the length of model dimer. Both approaches have shown that 16 dimers can fit into the filament cross-section of $10nm$ width.

2.3.1.3 Lower Order Organization

To simplify the modeling of the filament, it may be useful to first construct the sub-structures of the filament and then their assembly. This will definetly reduces the sampling of the possible configurations by orders of magnitude. Recent experimental validation in (Lichtenstern et al., 2012) confirmed the tetramer existence in the filament. This experimental validation of filamentous nature of keratin filament makes the modeling of ULF easier. Figure 2.9 shows a schematic representation of the filament. This work is started by building the dimer model which is then followed by the tetramer. Finally a filament model was created using tetramers.

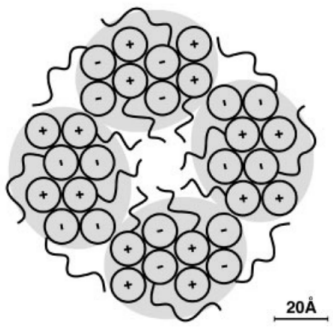

Figure 2.9: Schematic of the filament model (Strelkov et al., 2003). Four dimers associate into tetramers (shaded oval), and four tetramers constitute the filament.

2.3.2 A Divide and Conquer approach

2.3.2.1 Sampling dimer-dimer configuration

To reduce the search space during optimization, divide and conquer approach is used. A cross-link can only form between two dimers. Using this fact, all possible arrangements of two dimers are explored that satisfy the cross-links. This is done by creating an exact replica of the dimer at a distance 18.5Å parallel/anti-parallel to the original dimer and then translating one dimer with respect to the other. In this work, longitudinal translation of the dimer is referred as "shear". For each shear value, both the dimers are rotated along their longitudinal axis (Figure 2.10). This samples all possible cross-linked dimer-dimer configurations. A cross-links is recorded based if $C^{\alpha} - C^{\alpha}$ distance is less than 19.4Å. The cross-links are categorized as true positives and false positives. True positives are those cross-links that were observed in any of the two cross-linking experiments. All other recorded cross-links are false positives. The configurations that satisfy the most number of cross-links are termed as modes of alignment. This is done for parallel and anti-parallel association of two dimers.

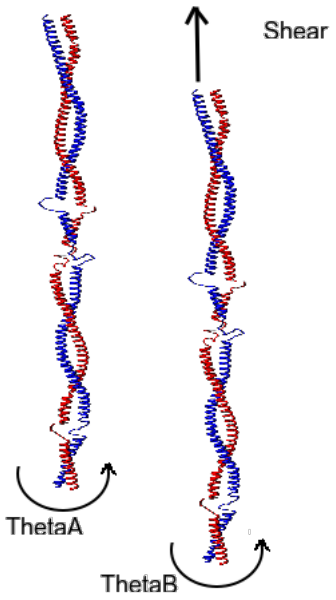

Figure 2.10: Transformation on dimer-dimer system. Dimers are rotated and sheared along the longitudinal axis as shown by the arrows.

A database of all lysine-lysine distances encompassing both true and false positive cross-links was created. Table 2.1 shows a sample table of database.

This configuration search results in large database set. This database is queried through MySQL. This table contains all information about the pairs of lysine that are cross-linked at different configurations.

Shear	ThetaA	ThetaB	Distance between dimers	Cross-link	True/ false +ive	Distance of cross-link
190	40	160	18.5	10.A-495.B	T	16.973
190	40	160	18.5	10.A-493.B	F	15.421
190	40	160	18.5	96.A-108.B	F	12.356
...	...	...	18.5	...	...	...
190	40	180	18.5	10.A-495.B	T	18.953
...	...	...	...	...	...	...

Table 2.1: Cross-link Database table

2.3.2.2 Modes of association in anti parallel arrangement

For anti-parallel dimer-dimer configuration, only three modes of association are found with respect to shear or longitudinal movements that satisfy cross-links. Those are named as A11, A12 and A22. Mode A12 is found in two different configurations with respect to the rotation of dimers, and are named as A12_1 and A12_2. All these modes satisfy all the cross-links observed in (Steinert et al., 1993b). The following are the values of longitudinal shear ("shear"), rotation of dimer A ("θA"), rotation of dimer B ("θB") for each mode of association.

Mode A11:
shear=190;θA=340;θB=160; satisfies 7 cross-links, $1-7$ in Appendix D

Mode A12_1:
shear=0;θA=340;θB=260; satisfies 3 cross-links, $8-10$ in Appendix D

Mode A12_2:
shear=0;θA=240;θB=120; satisfies 2 cross-links, $11-12$ in Appendix D

Mode A22:
shear=-300;θA=340;θB=160; satisfies 2 cross-links, $13-14$ in Appendix D

Figure 2.11 shows various anti-parallel modes.

2.3.2.3 Modes of association in parallel orientation

For parallel configuration, only two modes of association are found with respect to shear or longitudinal movement. These are named as P1 and P2. All these modes

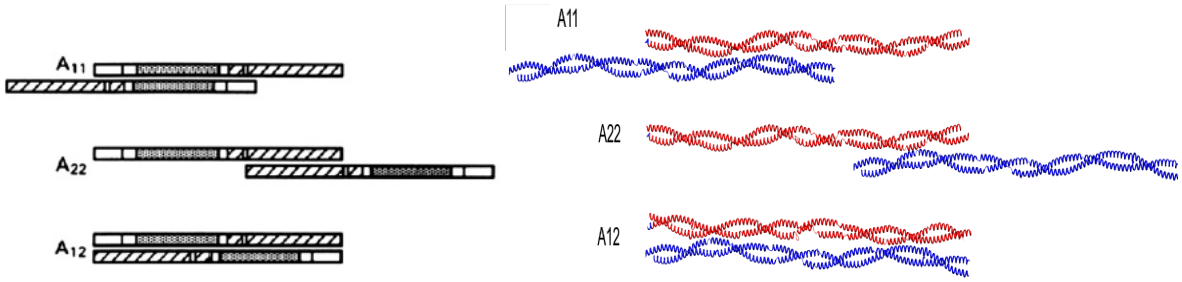

Figure 2.11: Schematic modes of association,(Parry et al., 2007) and Generated anti-parallel modes of alignment between dimers.

satisfy all the cross-links observed in (Parry et al., 2001). Following are the values of shear ("shear"), rotation of dimer A ("θA"), rotation of dimer B ("θB") for each mode of association.

Mode P1:

shear=116;θA=340;θB=120; satisfies 7 cross-links, $15-21$ in Appendix D

Mode P2:

shear=-374;θA=340;θB=120; satisfies 2 cross-links, $22-23$ in Appendix D

Figure 2.12 shows P1 and P2 modes with exact measurement.

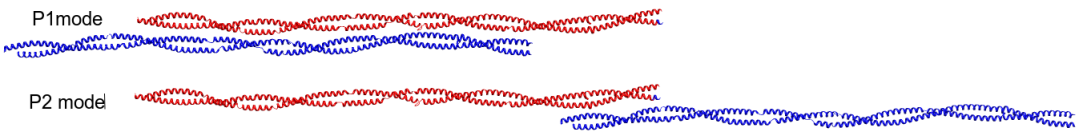

Figure 2.12: Generated parallel modes of alignment (P1 & P2 mode) between dimer with values as reported in section 2.3.2.3

2.3.3 Model generation and evaluation

2.3.3.1 Tetramer models generation

A two-dimensional surface lattice models (figure 2.7 and 2.8) helps explain the parallel and anti-parallel modes. Consider the figure 2.13 as an expected model, intra-tetramer cross-links have a higher relative frequency of occurrence compared to the inter-tetramer cross-links. For instance, this structure has 12 possible sites

for intra-tetramer cross-links (A1-A2, A2-A3, A3-A4, B1-B2, ..., D2-D3, D3-D4), which is three times more compared to inter-tetramer cross-links (A4-B2, B4-C2, C4-D2, D4-A2). Also, the Solvent Accessible Surface Area (SASA), at the interface of two tetramers is less compared to the other dimers in the tetramers.

Hence, among the parallel (P1,P2) and anti-parallel (A11, A12_1, A12_2, A22) modes of association, only anti-parallel modes are used to build the tetramer models and parallel modes are used to join tetramers to form a filament.

All the modes are dimer of dimers, arranged in a specific value of shear, thetaA, thetaB. Total 24 tetramer models are generated as explained in figure 2.14.

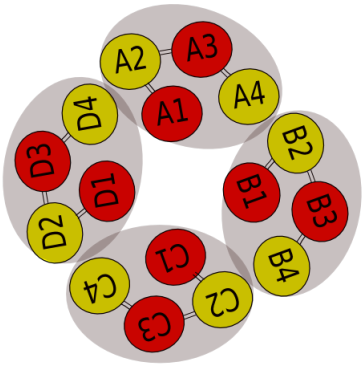

Figure 2.13: Expected schematic filament structure. Red dimers coming out of the plane and yellow dimer are going into the plane.

Among the anti-parallel modes of arrangement, A11 and A22 modes are shown in figure 2.15). Two dimers shown in blue are cross-linked to the single red dimer. Dimer A is in arrangement of A11 mode with dimer B. Also, dimer A is in arrangement of A22 mode with dimer C. Dimer B and C are parallel to one another. Dimer C is just a longitudinal translation of the dimer B with a gap between two dimers as 36Å (Figure 2.15). This is how the filament would elongate longitudinally. In addition to A11 and A22 mode, P1 and P2 mode are also similar (Figure:2.16). Hence, this is a way to join succeeding and preceding tetramer. A22 mode is used to join the three tetramers to create a succeeding and preceding tetramer structure.

Succeeding and preceding tetramer structures generation are automated. Total 24 structures were created. Valid structures are selected based on the absence of serious steric clashes. Selected models were optimized using IMP. The variables for optimization are cross-links and excluded volume restraint. Excluded volume restraint is used to avoid clashes during satisfaction of cross-links by the optimizer. Here in IMP, only Conjugate Gradient optimizer is used because the dimers are

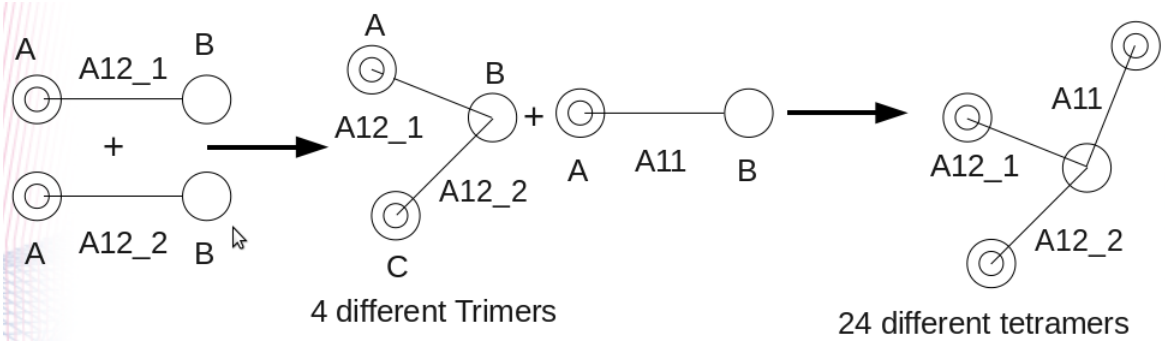

Figure 2.14: Schematic for tetramer generation. Here a dimer of dimer is represented as two circle connected with a line. If the mode is anti-parallel, then one of the circle is concentric. Tetramer creation is a 2 step procedure. First, a single dimer of dimer-dimer system is superimposed with a single dimer of another Dimer-dimer system using CLICK (Nguyen et al., 2011). This step generates the trimer (A-B-C) with 4 different configurations. Secondly, a single dimer of another dimer-dimer system is superimposed with a single dimer of trimer(A-B-C). This step gives 6 possible tetramers for each trimer. Total $6 * 4 = 24$ tetramers are possible.

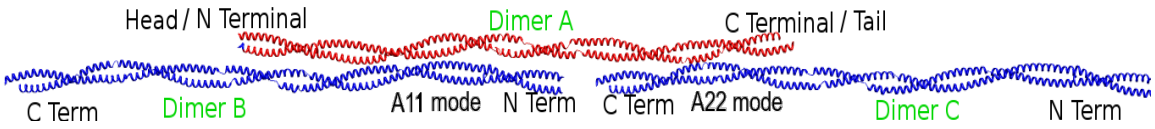

Figure 2.15: A11 and A22 mode with common dimer and 36Å gap between head and tail of blue dimers.

already placed at their appropriate positions to satisfy cross-links and all possible models are optimized. Out of 24 succeeding and preceding tetramers only 4 models satisfy the cross-links without any clashes after optimization. Hence, these 4 models are used to generate the filament models.

2.3.3.2 Filament models generation

Using the succeeding and preceding tetramers, filaments models are created using P1 modes of alignment between two tetramers. Consider the filament model explained in figure 2.17, , dimers A1 and A2 are joined by A12_1 mode, dimers A2 and A3 are joined by A12_2 mode, and dimers A3 and A4 are joined by A11

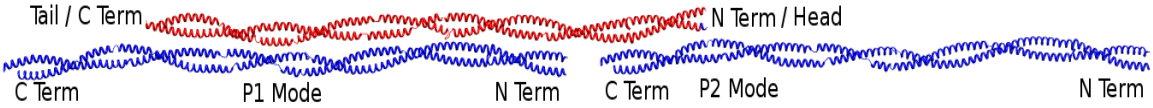

Figure 2.16: P1 and P2 mode with common dimer 36Å gap between head and tail of blue dimers.

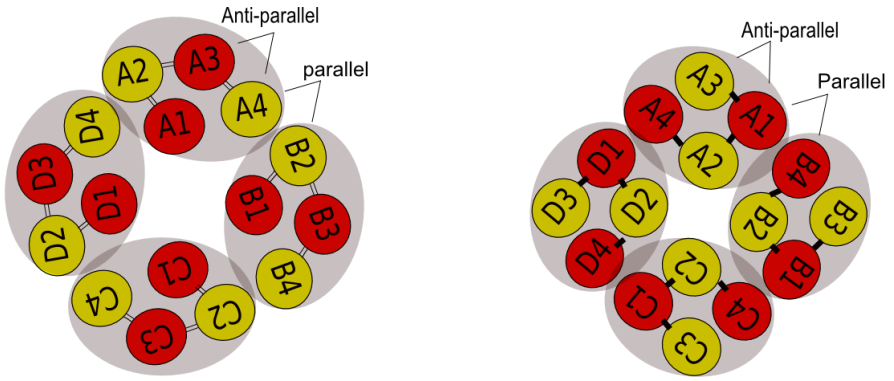

Figure 2.17: Possible Dimer Orientation in filament. Red colored dimer are coming out (tail to head vector coming outward) & yellow colored dimer are going down(tail to head vector going inward).

mode. All the four dimers in each tetramer are alternate anti-parallel arrangement. Using this information, all the possible models of the filament, for each tetramer models are created. This is done by joining tetramer models using P1 mode of arrangement (figure: 2.18).

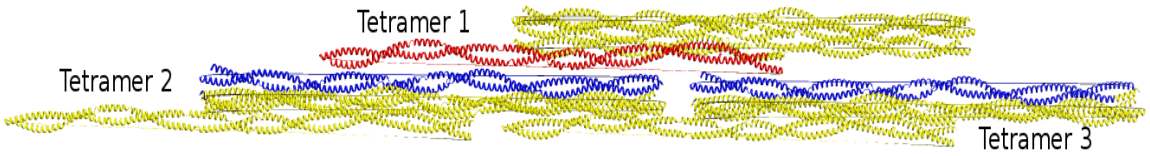

Figure 2.18: Here Tetramer 1 and Tetarmer 2 are joined by P1 mode. Tetarmer 3 is the succeeding tetramer of Tetramer 1. Because, P2 mode is just a translational shear of P1, Tetramer 3 forms P2 mode with Tetarmer 1.

Of the tetramer models generated, best models are selected based on the number of steric clashes.

2.3.4 Full Atom Optimization of the Filament

Initially, at each step of creation of lower order structures, models were optimized by considering side-chains as rigid bodies. Now the filament model is ready to perform the full atom optimization by MODELER. At the time of the thesis submission, only one filament model is generated. More refinement is needed to obtain the final filament models.

Chapter 3

Results and Discussion

3.1 Creation of Coiled-coil Dimer

As described in section 2.2, coiled-coil dimer is generated by comparative modeling. Figure 3.1 show a single K5-K14 coiled coil dimer. The length of the dimer is

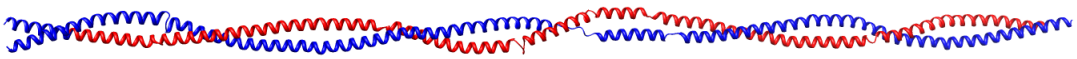

Figure 3.1: Coiled-coil Dimer

454Å. Positions “a” and “d” in the heptad repeats of a coiled-coil dimer should be buried. To confirm this in our models, SASA is calculated for each heptad position in the dimer. Figure 2.5 shows the SASA values of all residues. Clearly “a” and “d” positions in the graph have the least SASA values. To obtain the electrostatic properties of the individual dimer, electrostatic calculations are performed on the dimer using APBS software package (Baker et al., 2001). Figure 3.2 shows the Electrostatic potential surface of the dimer.

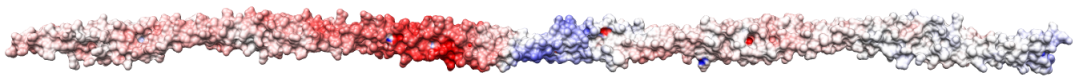

Figure 3.2: Electrostatic Potential Surface

3.2 Description of Dimer of Dimers

As described in section 2.3.2.1, the sampling of dimer-dimer configurations gives certain configurations that satisfy the cross-links. In case of A11 and A12 mode, three to four different configurations satisfy the same cross-links. All these different configurations are variants of each other. For instance, the cross-links of A11 mode is satisfied by the following three different configurations.

1. shear=190,θA=340,θB=140
2. shear=190,θA=340,θB=160
3. shear=190,θA=340,θB=180

To choose among the following configurations, cross-linking distances are obtained from table 2.1. The configurations which give the least value of cross-linking distance are selected.

3.2.1 Electrostatic complementarity of modes

All anti-parallel modes of association are favorable in terms of electrostatic energy. These are confirmed by doing electrostatic analysis of modes using APBS software package.

Anti-parallel arrangement of modes (A11, A12, A22) are confirmed with electrostatic complementarity of the dimer-dimer interaction. Figure 3.3 shows the electrostatics of these modes of association. All these modes are qualitatively given

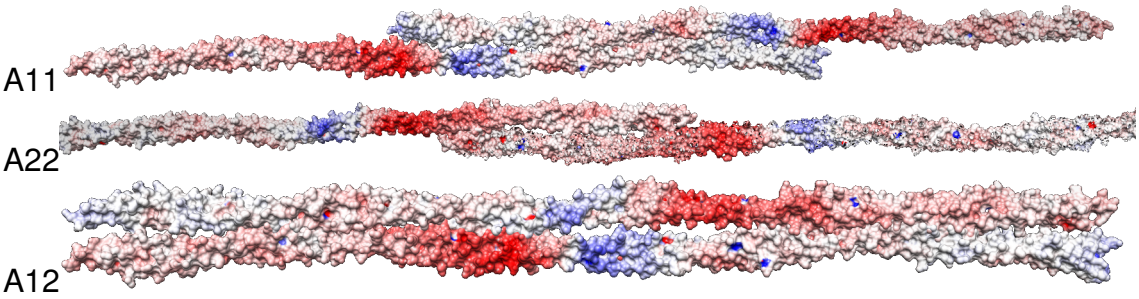

Figure 3.3: Electrostatic potential surface of anti-parallel modes A11, A12 & A22

in (Steinert et al., 1993b) without specifying the exact value of orientation. This is an independent confirmation about the modes of association between dimers with exact measurements.

3.3 Creation of Tetramers

As described in figure 2.14, 24 tetramer models are created. These models are created by superimposition of A11, A12_1 and A12_2 modes. All these models are extended by adding succeeding & preceding tetramers to the original tetramer. Figure 3.4 show a sample tetramer and a succeeding & preceding tetramers. Figure 3.5 shows the electrostatic potential surface of a tetramer and a succeeding & preceding tetramer.

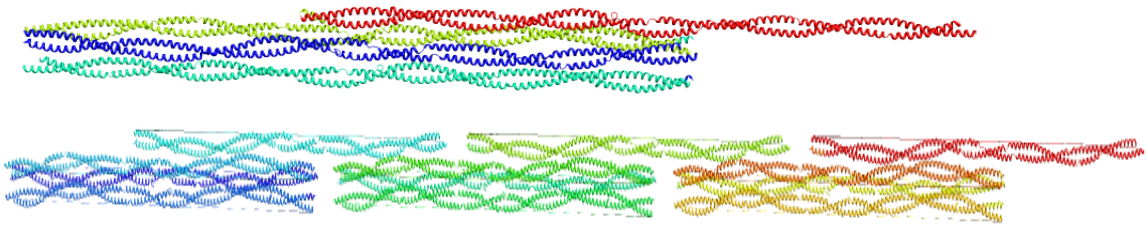

Figure 3.4: (A) Tetramer (B) Succeeding & Preceding Tetramer

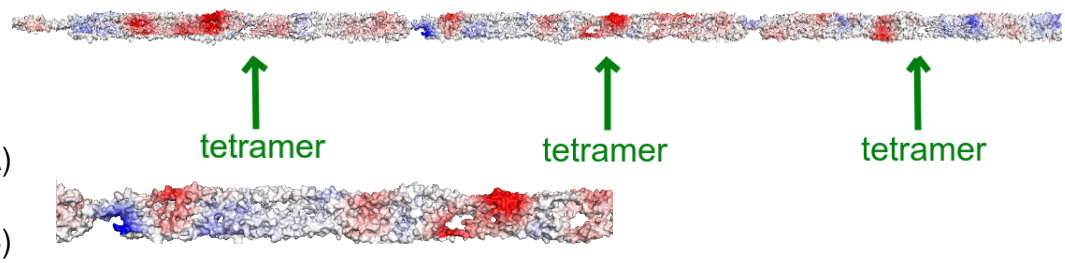

Figure 3.5: Electrostatic surface of Tetramers. (A)Succeeding and Preceding Tetramer. (B)Central Tetramer

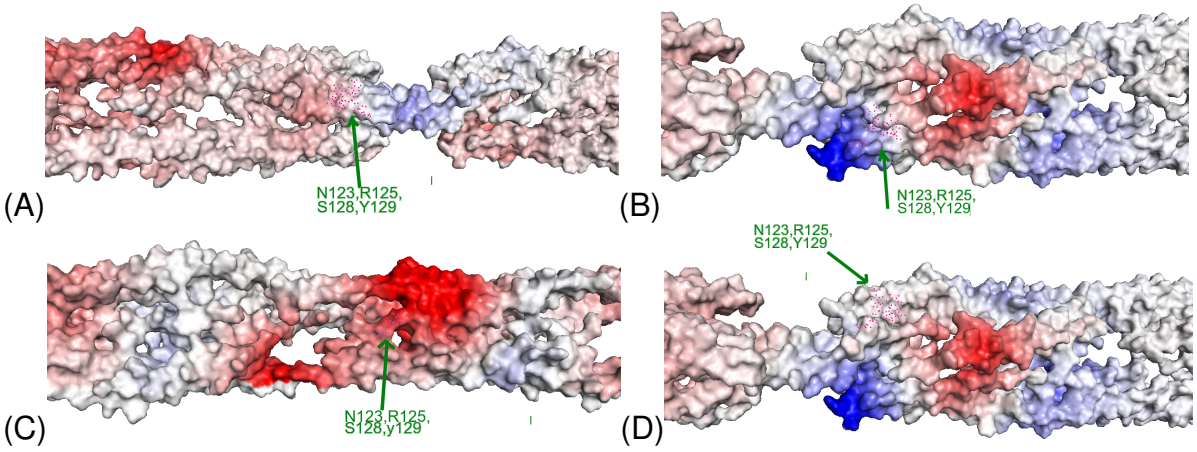

Figure 3.6: (A),(B),(C) & (D) are the closeup of the four different positions of the labeled residues in a single tetramer. Mutating these residues to residues of different polarity results in filament fragmentation.

3.4 Creation of Octamers

Using the valid tetramer structures, possible octamer structures are created. This is done by joining two same copy tetramers using P1 mode of association. Figure 3.5 shows a possible octamer structure. Figure 3.6 shows the cross-section of two different octamers.

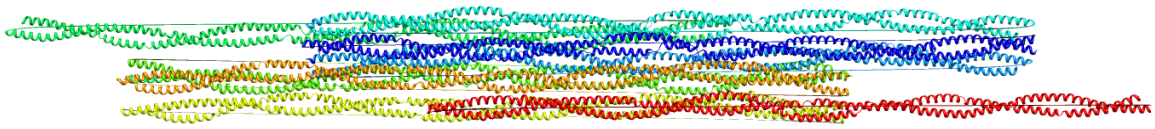

Figure 3.7: Possible Octamers created by joining two tetramers with P1 mode.

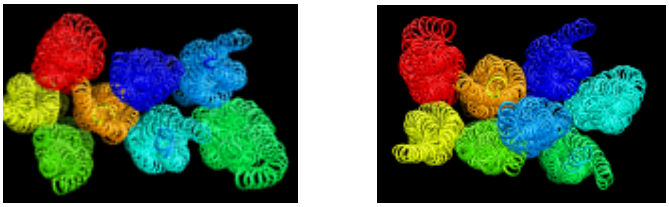

Figure 3.8: Cross-section of two different octamers

3.5 Analysis of the models

3.5.1 Electrostatic Analysis

As described in the introduction section, certain point mutations can lead to filament fragmentation. Some of these mutations are described in figure 3.7. Figure 3.8 shows the electrostatic potential surface in the tetramer. Figure 3.9 and 3.10 illustrate the location of these mutations in the A11 and A12 mode respectively.

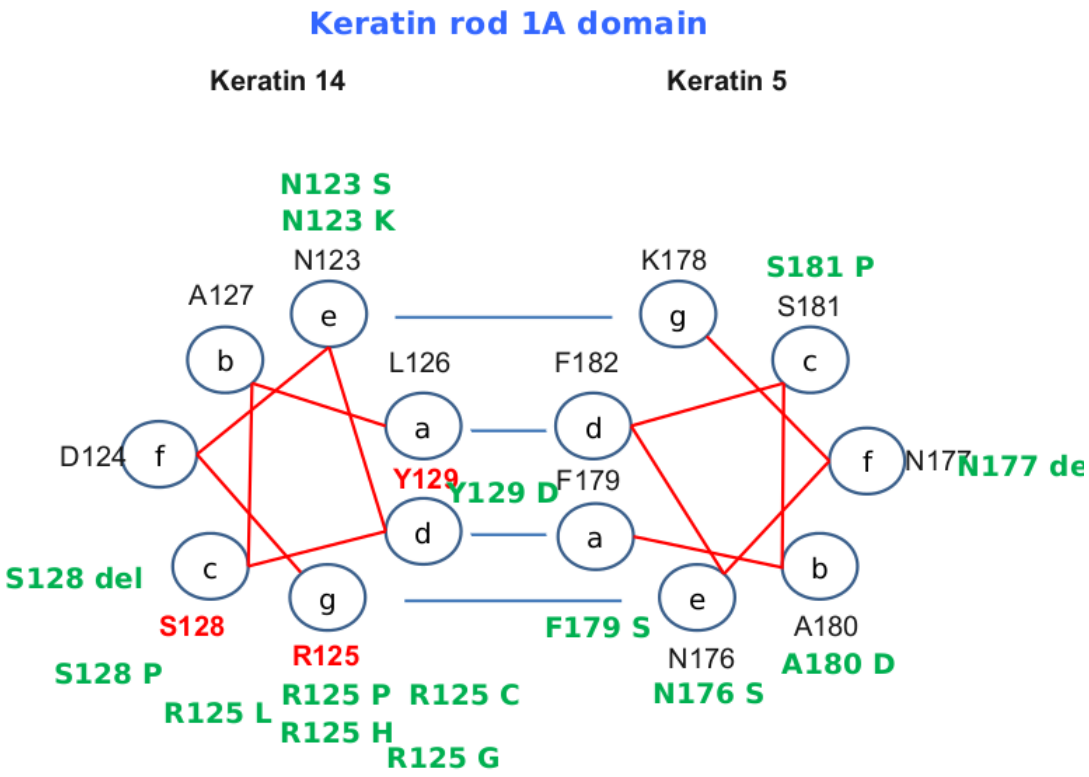

Figure 3.9: Private communication with Birgit Lane and co-workers provided us certain point mutations that leads filament fragmentation. Residues that are red in color affect the filament formation when mutated.

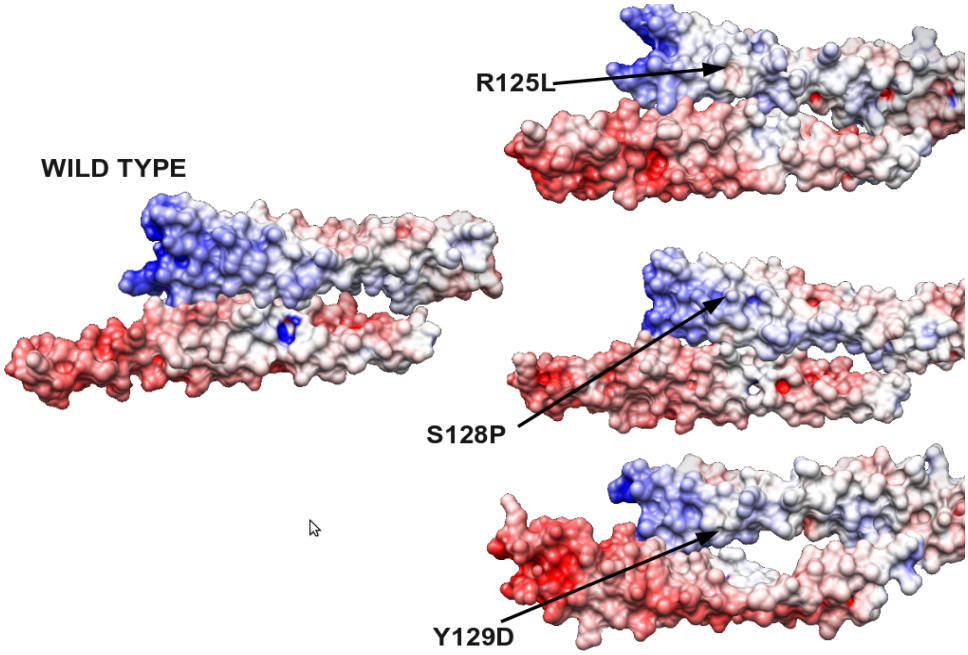

Figure 3.10: Mode A11 with Point mutations

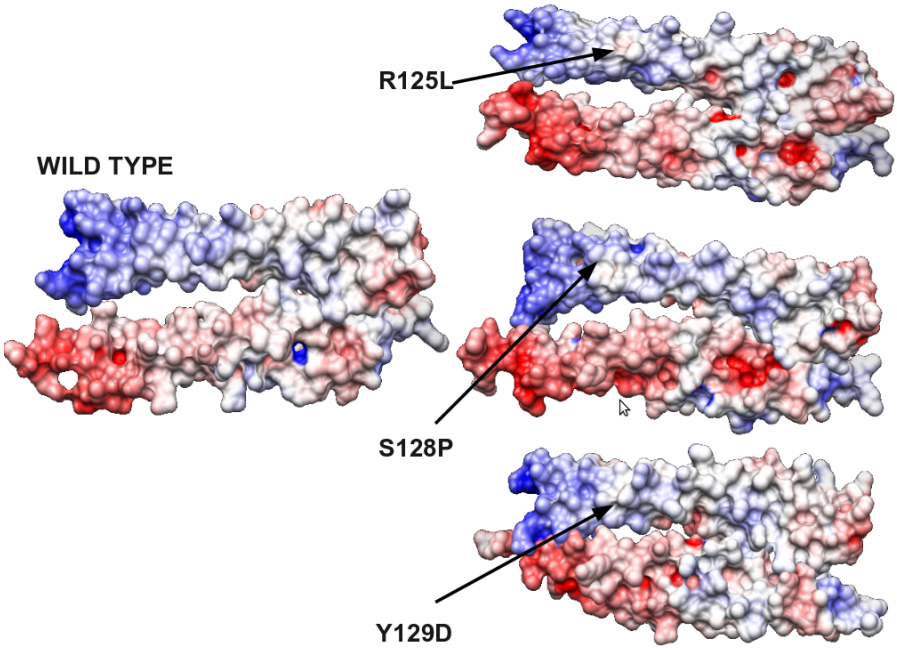

Figure 3.11: Mode A12 with point mutations

From figure 2.4 and 3.3, it is clear that, heptad positions “g” and “e” form salt bridge, which is not possible if either of the position’s polarity is changed. This is the case with R125. Also Y129 position should always be a hydrophobic because it is at position “d”. Mutating these residues affects the electrostatic potential of the surface and reduces the charge complimentarity between the dimers.

3.6 Features of the models that are in accordance with experimental results

3.6.1 Prediction and Comparison of Cross-links Relative Frequency

All atom optimized models are used to predict the relative frequency of the cross-links. This is done by counting the cross-links of each type that are on the surface of the filament and are more likely to bond with the cross-linker. Solvent Accessible Surface Area is used for this purpose. These results are not yet completed but we hope to get an explanation of why certain pairs of lysine get cross-linked while others cannot.

3.6.2 Lower Order Structure Diameter

The diameters of all the four lower order structures (tetramers) as obtained in section 2.3.3.1 are measured. Their values lies between 46.5Å $- 55.0$Å. This is in agreement with the low resolution value of 45.0Å as measured in(Aebi et al., 1983)

3.6.3 Reason for Insignificant Amount of Certain Cross-links

Experimentally, cross-links in P1 and P2 modes have insignificant relative frequency compared to the anti-parallel modes of arrangement. In this mode of association, distance between $C^{\alpha} - C^{\alpha}$

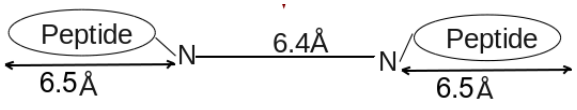

Figure 3.12: Schematic Cross-link

atoms of cross-linked lysine are measured as described in section 2.3.2.1. From table 2.1 their values are in the range of 17.5Å to 19.39Å. The maximum length of the completely stretched lysine side chain is 6.5Å, resulting the maximum distance between C_α atoms as $2 * 6.5 + 6.4\,(cross - linker\,arm\,length) = 19.4$Å. Figure 3.11 show the schematic of cross-link. The cross-links in P1 and P2 modes are mostly strained and hence are less likely to form cross-links with stretched side chains.

Chapter 4

Conclusion and Future Direction

We have successfully built the atomic level resolution model of the K5-K14 filament (Figure 4.1). Filament model needs some refine-

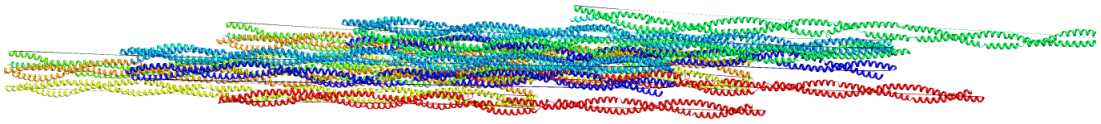

Figure 4.1: Generated K5-K14 Filament Model

ment to be done. All the preliminary results explained in section 3.6 are in agreement with the earlier studies done on the filament. Till date, this is the largest atomic level resolution model created of any protein system. Other large models do exist but they are coarse grained.

Making use of molecular modeling and converting the chemical cross linking data into spatial restraints a near atomic resolution model of the Keratin (K5/K14) intermediate filament has been created.But know we know about the structure with atomic level resolution.

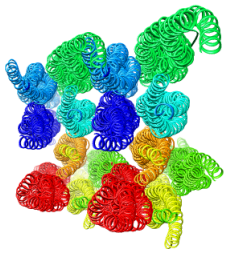

Figure 4.2: Cross-section of K5-K14 Filament Model

The keratin model was built using the K5-K14 dimers as the building blocks. The dimer model was built from K5-K14 sequence using the general information about the heptad repeats and coiled-coil geometry. As the number of cross-links are not sufficient to give the overall structural information, a divide and conquer approach was used. In this approach, cross-links were satisfied by sampling all possible configurations between the dimers. This bottom-up approach reduces the search space for the optimizer, leaving out all but few modes of association between dimers. These modes of association are used to create lower order assembly (tetramers), which is joined to form the filament.

The filament model may provide insight about the dynamics and the network formation of the IF. This model may help explain why certain point mutations lead to various keratin related diseases. This in turn can help design drugs to cure these diseases.

If the models were correct, then the experiments performed on the keratin filaments should be predicted. We can suggest the mutants that can disrupt the filament. Fitting the EM density map of the keratin filament to the generated models will be a direct validation. It will now help understand the IF protein assembly. This method is general and can be applied to all IF proteins.

Appendix A

Cross-linked Lysine Residues

	1st Lysine Residue Number	2nd Lysine Residue Number	Relative Frequency mol/mol	Year of Publish Steinert et al. (1993b); Parry et al. (2001)
1	10	495	0.23	1993
2	96	96	0.51	1993
3	87	108	0.19	1993
4	365	448	0.62	1993
5	365	449	0.62	1993
6	17	175	0.36	1993
7	373	441	0.21	1993
8	41	275	0.22	1993
9	96	526	0.37	1993
10	10	304	0.2	1993
11	441	495	0.41	1993
12	365	263	0.11	1993
13	236	273	0.10	1993
14	236	275	0.25	1993
15	31	108	0.02-0.001	2001
16	330	258	0.02-0.001	2001
17	10	258	0.02-0.001	2001
18	196	275	0.02-0.001	2001
19	109	507	0.02-0.001	2001
20	31	109	0.02-0.001	2001
21	109	196	0.02-0.001	2001
22	175	258	0.02-0.001	2001
23	196	273	0.02-0.001	2001

Table A.1: Cross-linked Lysines with Relative Frequency

K5-K14 dimer is a coiled coil dimer with K5 having residues numbered from 1 to 313 and K14 having residues numbered from 314 to 624. Hence, a lysine residue number 315 is the 2^{nd} residue of K14. T he table records the cross-links with their relative frequency of occurence.

Bibliography

Aebi, U., Fowler, W. E., Rew, P. and Sun, T. T. (1983). The fibrillar substructure of keratin filaments unraveled. J Cell Biol *97*, 1131–43. 2.3.1.2, 3.6.2

Aebi, U., Häner, M., Troncoso, J., Eichner, R. and Engel, A. (1988). Unifying principles in intermediate filament (IF) structure and assembly. Protoplasma *145*, 73–81. 1.1

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Rout, M. P. and Sali, A. (2007). Determining the architectures of macromolecular assemblies. Nature *450*, 683–94. 1.7

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol *215*, 403–10. 2.1.1.1

Baker, N. A., Sept, D., Joseph, S., Holst, M. J. and McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci U S A *98*, 10037–41. 3.1

Bragg, P. D. and Hou, C. (1980). A cross-linking study of the Ca2+, Mg2+-activated adenosine triphosphatase of Escherichia coli. Eur J Biochem *106*, 495–503. 2.3.1.1

Dhe-Paganon, S., Werner, E. D., Chi, Y. I. and Shoelson, S. E. (2002). Structure of the globular tail of nuclear lamin. J Biol Chem *277*, 17381–4. 1.1

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U. and Sali, A. (2006). Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics *Chapter 5*, Unit 5.6. 2.1

Fuchs, E. (1996). The cytoskeleton and disease: genetic disorders of intermediate filaments. Annu Rev Genet *30*, 197–231. 1.1

Fuchs, E. and Weber, K. (1994). Intermediate filaments: structure, dynamics, function, and disease. Annu Rev Biochem *63*, 345–82. 1.1

Herrmann, H. and Aebi, U. (2004). Intermediate filaments: molecular structure, assembly mechanism, and integration into functionally distinct intracellular Scaffolds. Annu Rev Biochem *73*, 749–89. 1.1, 1.3

Herrmann, H., Bar, H., Kreplak, L., Strelkov, S. V. and Aebi, U. (2007). Intermediate filaments: from cell architecture to nanomechanics. Nat Rev Mol Cell Biol *8*, 562–73. 1.2, 1.5

Krimm, I., Ostlund, C., Gilquin, B., Couprie, J., Hossenlopp, P., Mornon, J. P., Bonne, G., Courvalin, J. C., Worman, H. J. and Zinn-Justin, S. (2002). The Ig-like structure of the C-terminal domain of lamin A/C, mutated in muscular dystrophies, cardiomyopathy, and partial lipodystrophy. Structure *10*, 811–23. 1.1

Lane, E. B. and McLean, W. H. (2004). Keratins and skin disorders. J Pathol *204*, 355–66. 1.2

Levitt, M. (1992). Accurate modeling of protein conformation by automatic segment matching. J Mol Biol *226*, 507–33. 2.1.1.3

Lichtenstern, T., Mucke, N., Aebi, U., Mauermann, M. and Herrmann, H. (2012). Complex formation and kinetics of filament assembly exhibited by the simple epithelial keratins K8 and K18. J Struct Biol *177*, 54–62. 2.3.1.3

Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C. and Murzin, A. G. (2002). SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res *30*, 264–7. 2.1.1.1

Milks, M. S. K. A. (2012). Epidermolysis bullosa. 1.5

Moll, R., Divo, M. and Langbein, L. (2008). The human keratins: biology and pathology. Histochem Cell Biol *129*, 705–33. 1.2

Nguyen, M. N., Tan, K. P. and Madhusudhan, M. S. (2011). CLICK–topology-independent comparison of biomolecular 3D structures. Nucleic Acids Res *39*, W24–8. 2.2.1, 2.14

Parry, D. A. (2006). Hendecad repeat in segment 2A and linker L2 of intermediate filament chains implies the possibility of a right-handed coiled-coil structure. J Struct Biol *155*, 370–4. 2.2.2

Parry, D. A., Marekov, L. N. and Steinert, P. M. (2001). Subfilamentous protofibril structures in fibrous proteins: cross-linking evidence for protofibrils in intermediate filaments. J Biol Chem *276*, 39253–8. 2.3.1.1, 2.3.1.1, 2.8, 2.3.1.1, 2.3.2.3, A

Parry, D. A., Strelkov, S. V., Burkhard, P., Aebi, U. and Herrmann, H. (2007). Towards a molecular description of intermediate filament structure and assembly. Exp Cell Res *313*, 2204–16. 2.11

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol *183*, 63–98. 2.1.1.1

Sali, A. and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol *234*, 779–815. 2.1.1.3, 2.2.3

Sokolova, A. V., Kreplak, L., Wedig, T., Mucke, N., Svergun, D. I., Herrmann, H., Aebi, U. and Strelkov, S. V. (2006). Monitoring intermediate filament assembly by small-angle x-ray scattering reveals the molecular architecture of assembly intermediates. Proc Natl Acad Sci U S A *103*, 16206–11. 1.1, 1.1, 1.4

Steinert, P. M., Marekov, L. N., Fraser, R. D. and Parry, D. A. (1993a). Keratin intermediate filament structure. Crosslinking studies yield quantitative information on molecular dimensions and mechanism of assembly. J Mol Biol *230*, 436–52. 2.3.1.1

Steinert, P. M., Marekov, L. N. and Parry, D. A. (1993b). Conservation of the structure of keratin intermediate filaments: molecular mechanism by which different keratin molecules integrate into preexisting keratin intermediate filaments during differentiation. Biochemistry *32*, 10046–56. 2.3.1.1, 2.3.1.1, 2.7, 2.3.2.2, 3.2.1, A

Strelkov, S. V., Herrmann, H. and Aebi, U. (2003). Molecular architecture of intermediate filaments. Bioessays *25*, 243–51. 2.9

Strelkov, S. V., Herrmann, H., Geisler, N., Wedig, T., Zimbelmann, R., Aebi, U. and Burkhard, P. (2002). Conserved segments 1A and 2B of the intermediate filament dimer: their atomic structures and role in filament assembly. EMBO J *21*, 1255–66. 1.1, 1.1

Strelkov, S. V., Schumacher, J., Burkhard, P., Aebi, U. and Herrmann, H. (2004). Crystal structure of the human lamin A coil 2B dimer: implications for the head-to-tail association of nuclear lamins. J Mol Biol *343*, 1067–80. 1.1

Sutcliffe, M. J., Haneef, I., Carney, D. and Blundell, T. L. (1987). Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Eng *1*, 377–84. 2.1.1.3

Szeverenyi, I., Cassidy, A. J., Chung, C. W., Lee, B. T., Common, J. E., Ogg, S. C., Chen, H., Sim, S. Y., Goh, W. L., 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., Lim, Y. P. and Lane, E. B. (2008). The Human Intermediate Filament Database: comprehensive information on a gene family involved in many human diseases. Hum Mutat *29*, 351–60. 1.1

Webb, B., Lasker, K., Schneidman-Duhovny, D., Tjioe, E., Phillips, J., Kim, S. J., Velazquez-Muriel, J., Russel, D. and Sali, A. (2011). Modeling of proteins and their assemblies with the integrative modeling platform. Methods Mol Biol *781*, 377–97. 1.4

Westbrook, J., Feng, Z., Jain, S., Bhat, T. N., Thanki, N., Ravichandran, V., Gilliland, G. L., Bluhm, W., Weissig, H., Greer, D. S., Bourne, P. E. and Berman, H. M. (2002). The Protein Data Bank: unifying the archive. Nucleic Acids Res *30*, 245–8. 2.1.1.1