COMPUTATIONAL MODELING OF THE 3D STRUCTURE OF INTERMEDIATE FILAMENTS AND MOLECULAR MECHANISM OF ASSOCIATED DISEASES

A thesis

Submitted in partial fulfillment of the requirements

Of the degree of

Doctor of Philosophy

By

Neelesh Soni

Roll No 20133244

Indian Institute of Science Education and Research, Pune

2019

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that violation of the above will be cause for disciplinary action by the Institute and can evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: 7/8/2019

Neelesh Soni
Roll No: 20133244
IISER, Pune

CERTIFICATE

Certified that the work incorporated in the thesis entitled ("Computational Modeling of 3D Structure of Intermediate Filaments and Molecular Mechanism of Associated Diseases") Submitted by Neelesh Soni was carried out by the candidate, under my supervision. The work presented here or any part of it has not been included in any other thesis submitted previously for the award of any degree or diploma from any other University or institution.

Date: 7/8/2019

Dr. M. S. Madhusudhan
Supervisor

IISER, Pune

ACKNOWLEDGEMENTS

This thesis is the culmination of years of research work accompanied by excitement, hard work, brainstorming discussions, persuasions, and writing. The feeling of fulfillment that I am enjoying today would not have been possible without the constant support of many people including my research supervisor, colleagues, friends and family members.

At this moment of accomplishment, I am greatly indebted to my research supervisor, Dr. M. S. Madhusudhan for his invaluable guidance, not just in research topics but also in other academic matters and personal life. Regular discussions with him on various research topics encouraged me to explore areas other than my thesis project. Working on various research projects helped me develop critical thinking, diverse skill set, and essential scientific methods. His zeal for perfection towards science has always inspired me to do more. It has been wonderful knowing him for more than eight years (including my master's degree). I sincerely thank him and grateful for having had the chance to work with him.

I would like to give a special thanks to Dr. Girish Ratnaparkhi and Dr. Sutirth Dey for introducing me to Dr. M. S. Madhusudhan and inspiring me to join him in Bioinformatics Institute, Singapore for my master's thesis. I also express my gratitude toward Dr. L.S. Shashidhara for motivating me to join the PhD program in IISER and to the registrar Col. (Retd.) G. Raja Sekhar and Col. Sanjeevan Sharma for their invaluable support.

I would like to express my deep sense of gratitude to Dr. Girish Ratnaparkhi, Dr. Saikrishnan Kayarat, Dr. Pranay Goel, and Dr. Kundan Sengupta for providing constant guidance, support and suggestions as a part of my research advisory committee.

I would like to acknowledge my present and past labmates in IISER, Pune and in Bioinformatics Institute, Singapore: Minh, Kuan Pern, Binh, Shipra, Parichit, Sagar, Yogendra, Neeladri, Abhilesh, Akash, Sanjana, Tejashree, Golding, Ankit, Gulzar and many others for being wonderful friends and for their helpful discussions on various projects. It is a blessing to work with such helpful, compassionate and friendly teammates.

The thesis would not have been possible without the contributions of the collaborators. I express my utmost thanks to Dr. Birgit Lane (IMB, Singapore), Dr. Thomas Magin, (Leipzig Univ. Germany), Dr. Lars Norlen, (KI, Sweden), Dr. R. Varadarajan, (MBU,

India), Dr. C.S. Verma, (IMB, Singapore), Dr. David Parry, (Massey University, New Zealand), Dr. Ashok Venkitaraman (MRC, Cambridge) and Dr. V. Prasanna (ACTREC, Mumbai).

My sincere thanks to academic, IT and biology department (especially Tushar, Neeta, Mahesh) and many others for their countless support on many different occasions in the pursuit of my doctorate degree. I would like acknowledge Wellcome Trust, DST and Infosys for providing financial support for many international conferences.

I sincerely thank my lifelong friend and fiancée Neha, for providing constant and selfless support during good and bad times. Her company, moral support and assistance in numerous ways ensures I look forward and achieve my dreams. I thank my family and her for their selfless love and for being the part of my life. I thank my parents, sister and brother-in-law for their constant support and understanding throughout the journey.

I would like to thank the company of my friends, Danveer, Ayesha, Shivik, Kunalika, Shrikant, Anushree for giving me delightful life in IISER. I thank all my friends and colleagues whom I have not mentioned here but have always supported me. Last but not the least; I would like to thank IISER community for the memorable twelve years of my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Synopsis

Motivation

The three-dimensional structures of large and dynamic protein complexes could elucidate their functional roles in various molecular/cellular phenomena. Typically, 3D structures are determined by techniques such as X-ray crystallography, NMR spectroscopy, SAXS, etc., that can render structural information from near atomic to nano-meter resolution. In macromolecular complexes especially those with non-globular components, usually a single experimental technique cannot resolve the 3D structure on its own. In such cases, integrative modeling could be used to determine the structures of the complexes [1–4]. The idea is to combine structural data, obtained from experiments at different resolution using computational techniques to construct a three-dimensional model. Such integrative methods have solved many protein complexes that otherwise would not have been possible with conventional techniques [5–10].

Literature review and Problem Statement

This thesis focuses on macro-molecular assembly formed by a family of fibrous proteins known as Intermediate filaments. Intermediate filaments (IFs), microtubules and microfilaments are the three major filament systems of the cytoskeleton of metazoan cells[11]. The network of these integrated filament systems imparts mechanical stability and is involved in several cellular processes such as cell signaling, transport and differentiation, metabolism of cholesterol, supporting/anchoring cell organelles etc.[12–15]. IFs, unlike the other two-cytoskeletal filaments, are different in different cell types. For instance, in humans, there are ~75 different IF proteins expressed in various tissues [11,16]. They are classified into six classes based on sequence homology. In general, Type I and Type II IFs co-expressed to form a heterodimer whereas Type III to Type VI IFs form homodimers[17]. An IF dimer is composed of two alpha-helical chains in parallel orientation forming a coiled-coil structure. Unstructured head and Tail domains flank conserved coiled-coil domain of IFs[18–20]. About 90 different diseases, such as Alexander disease, Dowling degos disease, etc. have been associated with single point mutations in IF genes[21]. Similar point mutations in the Type I-Type II heteromer, also known as Keratins, leads to various “Keratin disorders” such as Epidermolysis Bullosa Simplex (EBS) with an incidence rate of 1 in 3000 [22]. The resulting amino acid change in the gene product leads to the destabilization of the filament, leading to diseased

conditions. Despite the severity of these diseases, little is known about the Keratin IF assembly and the molecular mechanism of the onset of disease conditions. Three-dimensional (3D) structural information of these filaments could give insights into their physical properties and functions.

Objectives

The main objective of this study is to construct the 3D structure of the Keratin IF using Integrative modeling. The modeling protocol uses chemical cross-links and X-ray crystallography data combined with computational methods. The 3D structural model was validated in five different ways. 1) Identify cross-links that were not utilized in the modeling procedure. 2) Comparing model with the low-resolution profiles of SAXS and TEM. 3) Rationalizing disease-causing mutations by using structural rules obtained from the model. 4) Prediction of new filament disrupting substitutions even cases such as hydrophobic to hydrophobic substitutions. 5) Predicting temperature-sensitive mutations for keratin filament.

Other objectives of this study were to use the structural rules obtained from the model in various applications such as designing inhibitors for coiled-coil proteins, designing new stable coiled-coil proteins and predicting the oligomeric state of any coiled-coil assembly.

Methodology

The Keratin intermediate filament structure was determined using Integrative modeling technique described in the following sections.

Dimer Modeling: Initially, the coiled-coil dimer of the heteromeric K5-K14 was constructed using homology modeling. The template chosen was a construct of 17 molecules of leucine zippers (PDB:2ZTA) to span the entire K5-K14 dimer target sequence. Heptad repeats that are inherent to the coiled-coil structures guide the target-template alignment. The keratin K5-K14 dimer model was validated and employed in the tetramer modeling protocol.

Tetramer Modeling: The K5-K14 dimer model was duplicated to satisfy the chemical cross-links computationally. The exhaustive sampling of all possible dimer-dimer configurations was performed to identify all 19 inter-dimer cross-links observed in the experiment. All inter-dimer configurations that parsimoniously satisfy the cross-links were employed in the filament modeling protocol.

Filament Modeling: Inter-dimer configurations for a filament comprising of 16 dimers results in 10^{16} possibilities. Simulating all configurations on a three-dimensional model is an intractable problem. To solve this problem, Disctransformation software was created that converts the 3D object into a quasi-3D object. Each keratin dimer was mapped as a two-dimensional disc by taking a projection along the dimer axis. The software can then sample and score all possible configurations exhaustively. Models were selected in sequential steps using different scoring criterion at each step. In all 699 models were clustered into six topologically different configurations. The cluster with the optimal packing of the dimers was chosen as the final model of keratin filament.

Filament Validations: Disctransformation model was reverse mapped into a 3D structure along with succeeding and preceding filaments to obtain three unit lengths of the whole keratin (K5-K14) filament at near atomic resolution. The stability of this 500,000-atom model was verified using the 50ns Molecular Dynamics (MD) simulation. One thousand snapshots from the MD simulation trajectory were obtained to infer experimental data not utilized during modeling. The validations included identification of cross-linkable residues, comparing SAXS and radial density profile of keratin filaments with the model, fitting the filament bundle model into TEM images and identification of substitutions that could lead to filament disruption.

Molecular determinants of stability: The analysis of the 3D structures of the keratin filaments and coiled-coil proteins provide the molecular basis of their stability and specific pairing. The molecular determinants include coiled-coil core packing density, charge distribution, amino acids pair preferences, hydrogen bonding, and salt bridges. These structural properties were tested thoroughly on various molecular dynamic simulations and subsequently used for the prediction of mutants.

Prediction of Mutations: The molecular determinants derived from the Keratin Intermediate filament structure explain 86% (105/122) of K5-K14 disease causing mutants. The molecular determinants were then used to score novel filament destabilizing mutations obtained through *in-silico* saturated mutagenesis of all IF family members. The validations of the predicted filament disrupting mutants were performed through *in-vivo* cell culture experiments on K5-K14 Intermediate filament system. In all 16 different substitutions were predicted and the experimental outcomes agreed with the predictions. The molecular determinants of the K5-K14 filament provided temperature-sensitive substitutions that would greatly help researchers to study the effects of filament assembly at restrictive temperatures.

Applications: Molecular determinants of the stability of coiled-coil structures were subjected to various applications such as predicting the oligomeric state of a coiled-coil assembly. The method can predict the oligomeric state of any coiled-coil with near perfect accuracy (0.99) and Mathews Correlation Coefficient (0.98).

The determinants were also used to design inhibitors for Nipah Virus Fusion protein (F protein). Nipah F protein has a coiled-coil 6HB domain that exists in two different states. A small helix (inhibitor) was designed that would bind to the 6HB domain. This inhibitor would then interfere with the transition process of the protein, thereby affecting the virus life cycle.

The structural rules were also used to design novel anti-parallel eight helices coiled-coil protein. This coiled-coil topology was not seen in the PDB database. Six variants of anti-parallel octamer were designed and analyzed *in-silico*. These designs are currently under laboratory testing.

Conclusion

This is among the first 3D structure of an IF filament and could be used as a template to model other IFs. This structure could also serve as an attractive starting point for the rational design of therapeutic agents. This work has many potential applications such as designing inhibitors of coiled-coil proteins and predicting temperature-sensitive mutations for IF. Most importantly, the method can be used to predict complementary substitutions that will stabilize the unstable disease-causing IF variants.

Thesis Organization

This dissertation is organized into nine chapters. Following is the list of thesis chapters and their summary.

Chapter 1: Introduction

> Chapter 1 introduces macro-molecular assemblies and techniques for determining their 3D structure. The chapter first introduces integrative modeling technique through the previously published Nuclear-pore structure. It then highlights the advantages of Integrative modeling technique over other methods for constructing large macromolecular complexes such as intermediate filaments.

Chapter 2: Background and Related Work

The focus of this chapter is to introduce the coiled-coil proteins and their structural properties such as heptad repeats. It also describes the existing coiled-coil databases, software and experimental studies related to coiled-coils. The focus then shifts to a coiled-coil family of protein known as Intermediate filaments (IFs). The chapter highlights the Intermediate filaments classification, functions, and the current understanding of the molecular architecture. The last section deals with the introduction to the keratin intermediate filaments and their associated diseases.

Chapter 3: IF Dimerization and Stability Determinants.

Chapter 3 introduces the molecular determinants of Intermediate filament dimerization and stability. All Intermediate filament dimers were constructed and validated for structural correctness. The analysis of the Intermediate filament dimers provides the molecular basis of their stability and specific pairing. The structural properties include coiled-coil core packing density, charge distribution, amino acids volume, and pair preferences. These properties were tested thoroughly using molecular dynamic simulations and subsequently used to rationalize and predict new filament disrupting mutants at heptad positions a/d. The chapter ends with the validations of the predicted filament disrupting mutants through *in-vivo* experiments on K5-K14 Intermediate filament system.

Chapter 4: Integrative Modeling of 3D Structure of Intermediate filament

This chapter describes the three Integrative modeling steps performed to determine the structure of Keratin Intermediate filament. Step 1) Construction of Keratin K5-K14 dimer model using homology modeling. Step 2) Construction of possible tetramer models by computationally satisfying the chemical cross-links. Step 3) Filament construction by simulating all Inter-dimer configurations between 16 dimers. The chapter then introduces the Disctransformation software that simulated and scored ~10^{16} possibilities exhaustively. The Root Mean Square Deviation values cluster the structurally similar models into different sets. The chapter concludes with the reverse mapping of the Disctransformation model to provide the 3D coordinate of the best representative of the Keratin K5-K14 filament model.

Chapter 5: Validation of 3D Structure of Intermediate filament

Chapter 5 describes the validation methodologies performed on whole Keratin (K5-K14) filament model. The stability of the 500,000-atom model was verified using Molecular Dynamics (MD) simulation. The 50ns MD simulation trajectory

of Intermediate filament model was utilized to infer experimental data. The comparison of these deductions with the experimental data not employed during the modeling validates the model. The validations included identification of cross-linkable residues, comparing SAXS and radial density profile of keratin filaments with the model, fitting the filament bundle model into TEM images and identification of substitutions that could lead to filament disruption.

Chapter 6: Intermediate filament Assembly and Stability Determinants

This chapter highlights the methodology involved in obtaining the molecular determinants of intermediate filament stability. Keratin K5-K14 filament model provides the stability parameters such as inter-dimer hydrogen bonding and salt bridges from the heptad positions b/c/e/f/g. These parameters provide a foundation to rationalize and predict new filament destabilizing mutations along with temperature-sensitive mutations in K5-K14 filament.

Chapter 7: Applications of Intermediate filament Structural Properties

The chapter describes the potential applications of this work by subjecting molecular determinants to various research issues. The applications include predicting the oligomeric state of a coiled-coil assembly, designing inhibitors for Nipah Virus Fusion protein (a coiled-coil protein), and designing new coiled-coil proteins. The description of the detailed methodology used in addressing these research issues and their successful applications.

Chapter 8: Discussions and Outlook

Chapter 8 summarizes the critical conclusions from this study and discusses various ways through which the scientific community could use the knowledge gathered in this work.

Chapter 9: Appendices

Appendices contain a detailed description of software and algorithms developed during this study. All these software are open source under LGPL License and available for download from the links provided in their respective sections.

CHAPTER 1
Introduction

SALIENT FEATURES

- ***Background and Motivation***
  - Introduction to Integrative Modeling
  - Nuclear-pore complex structure as an example
  - Advantages of Integrative modeling over conventional approaches
  - Previous studies incorporating Integrative Modeling
  - Software for Integrative Modeling
  - Conclusions, Challenges and Possible solutions
- **Summary**

1.1 Background and Motivations

Three-dimensional structures of large and dynamic protein complexes could elucidate their functional roles in various molecular/cellular phenomena. These functions are mediated through interactions with other molecules inside the cell. Typically, 3D structures are determined by techniques such as X-ray crystallography, NMR spectroscopy, SAXS, etc. that can render structural information from near atomic to nano-meter resolution [23]. A single experimental technique alone cannot resolve the 3D structure of protein complexes, usually those with non-globular components. Although computational methods such as *ab-initio* modeling, iterative threading assembly refinement, and homology modeling can also predict structures [24–31]. However, the accuracy of these methods is usually lower than that of the experimentally determined structures. In such cases, integrative modeling could be used to determine the structures of the complexes [1–4].

Table 1-1 Examples of experimental methods that produce structural information

Structural information obtained from experiments combined with molecular modeling to compute 3D structure. Adapted from [32]

Method	Measured Data	Structural data generated	Example
Small Angle X-Ray Scattering (SAXS)	Scattering intensity as a function of momentum transfer	Pair distribution function; shape envelope	[33,34]
Circular Dichroism (CD)	Mean residue ellipticity as a function of wavelength	Secondary structure content	[35]
Forster Resonance Energy Transfer (FRET)	The yield of fluorescence energy transfer	Distance between donor and acceptor pair	[36]
Electron Paramagnetic Resonance (EPR)	Dipole-dipole coupling between electron spins	Spin label environment and distance between pairs of spin labels	[37,38]
Deuterium Exchange Mass Spectroscopy (DXMS)	Rate constant of H/D exchange	Solvent exposure	[39]
Radical footprinting	Rate constant from the dose-response curve	Solvent exposure	[40]
Chemical crosslink	Mass/charge ratio of joint fragmentation	Upper limit on pair distance between reacted groups	[41–43]
Mutagenesis	Yield from ALISA	Interfaces or binding sites of interacting proteins	[44–46]
Coevolution	Pairwise correlation between mutations	Interacting amino acid pairs	[47,48]

The idea of integrative modeling is to combine structural data obtained from multiple experiments at different resolutions using computational techniques to construct a three-dimensional model. Previously, several studies employed experimental techniques that provide structural data at different levels of resolution (Table 1-1). Integrative methods have successfully combined these structural data to solve many protein complexes that would not have been possible with conventional techniques alone [5–10,27,30,31,49]. Models determined through integrative approaches could be used to design and analyze new hypotheses and experiments. These experiments in turn, enhance the resolution and speed up the process of structure determination. Integrative modeling is akin to solving a sophisticated 3D-jigsaw puzzle where data from various sources integrated to determine the structure.

1.2 Introduction to Integrative Modeling

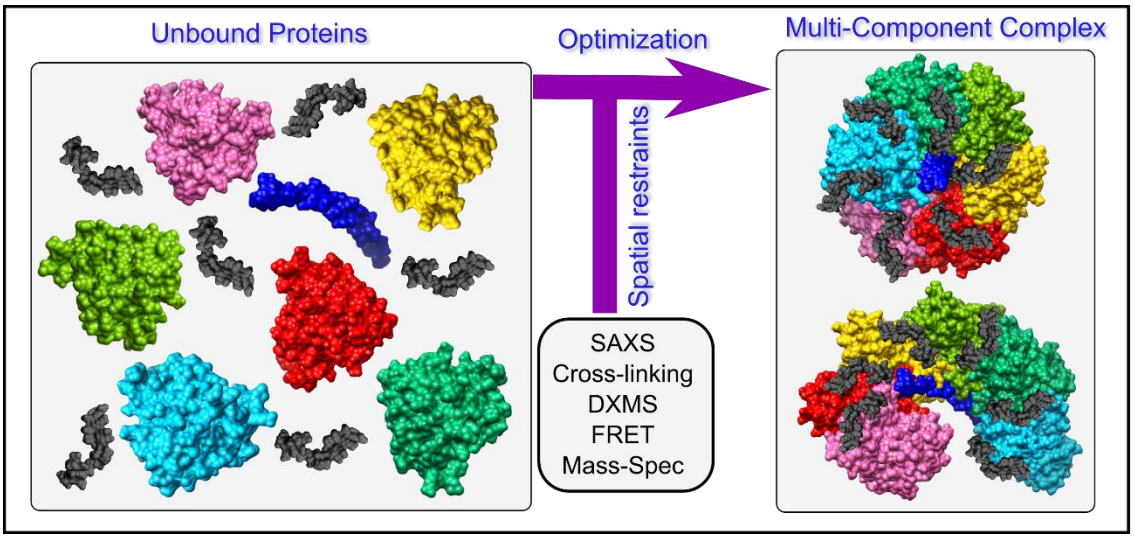


Figure 1-1 Schematic of Integrative modeling protocol of macromolecular assembly. *Individual proteins structures (unbound proteins) were determined through X-ray crystallography, NMR, etc. Interaction between unbound proteins provides spatial restraints among themselves obtained from experimental techniques such as SAXS, cross-linking, etc. The protein entities optimized computationally along with spatial restraints (Purple Arrows) to determine the multi-component macromolecular assembly. The multi-component complex shown in the figure is Rad51 filament assembly in an active state with six Rad51 monomers bound on DNA (blue) along with BRC repeat (gray)* [30]. *Some of the studies that incorporate such integrative modeling protocol to determine their structures involve nuclear pore complex*[5], *26S proteasome*[50], *RyR1 channel*[51], *80S Ribosome*[52], *etc. among others (*Table 9-2*).*

1.2.1 Integrative Modeling Overview

Integrative modeling follows the standard scientific cycle of collecting data, proposing hypotheses, followed by testing and refining those hypotheses. It can utilize computational optimization along with experimental data to build structural models (Figure 1-1). The models are evaluated using data from experiments that were not involved in structure construction. The construction of Nuclear-pore complex describes the integrative modeling protocol in the following section.

1.2.2 An example of Integrative Modeling: The Nuclear-pore complex

The nuclear envelope, along with a double membrane contains many macromolecular assemblies known as the nuclear pore complex (NPC). NPC functions as the mediators in the bidirectional exchange between nucleoplasmic and cytoplasmic compartments [53]. NPC is one of the largest macromolecular assemblies inside the cell. The structure determination of NPC represents a significant challenge to conventional approaches for structure determination. Owing to its massive size along with a high degree of flexibility, no single experimental technique including the computational approach would by themselves resolve the shape of the molecular complex. Proteomics experimental data combined with computational methods were then used to determine the NPC structure [2]. Following section describes the stages of integrative modeling of NPC.

Stage 1: Data Collection.

The first step in integrative modeling is to collect data from various experimental and computational procedures that provide structural information. Sometimes statistical data could supplement experimental observations. The amount and quality of the data can significantly increase the accuracy of the models. Data relevant to the structure of the NPC explained as follows.

1. *NPC component list and stoichiometry.* A single NPC contains about 30 nucleoporins that constitute the assembly with multiple copies of each nucleoporin [54].
2. *Shape and size of each component and symmetry within NPC.* Shapes of each nucleoporin was estimated using sedimentation coefficients determined by ultracentrifugation [2]. Electron microscopy and cryo-electron microscopy provide the overall shape and symmetry of the NPC. Heparin treatment of isolated NPC provided an upper bound diameter of the NPC [55].
3. *The localization of each component in NPC.* The constituent proteins of the NPC have

been localized using the immune localization map generated from immuno-EM method [54].

Stage 2: Representation and evaluation.

Ideally, computational modeling requires spatial restraints from the experiments. Only few experimental methods such as NMR, X-Ray crystallography etc. directly provide spatial restraints for modeling. Otherwise data collected from other experimental methods need to be represented using a scoring function. The representation depends on the type of experimental data collected. For instance, chemical cross-linking may provide maximum possible distances between the cross-linked pair of atoms and thus represented by an upper bound function. Similarly, EM data represented as volume restraints to fit structures onto the density maps. The global optimum of this function corresponds to the native assembly structure [56,57]. An example of such function is a joint probability density function (pdf), $p(C|I)$, given by the product of individual pdf's, $p_f(C|I_f)$, each describing the assembly features such as angles, distances, relative orientations, etc. The scoring function, $F(C)$, would then be defined as the negative logarithm of joint pdf.

$$F(C) = -ln \prod_f p_f\left(C|I_f\right)$$

where C stands for the Cartesian coordinates of assembly given prior information I about the system. The experimental data generated from a proteomics study of NPC provided spatial restraints between protein entities. Spatial restraints include protein shape, stoichiometry, contacts, localization, envelope surface, etc. A probability distribution function defines the scoring scheme, which encodes the structural information of the system. This scoring scheme evaluates the model consistency and the corresponding score determines the model quality. A good scoring model should satisfy all the restraints.

Stage 3: Sampling and Optimizing.

The computational assembly starts with sampling a random configuration. The scoring scheme constructed in stage 2 evaluates the 3D structure/model. The optimizer minimizes the violated restraints in the models. The final score calculated for the optimized models defines the quality of the models by setting a pre-defined cutoff. This optimization protocol was repeated for different initial random configurations to generate an ensemble of valid configurations.

Stage 4: Analyze resulting models.

The structural similarity scores cluster the ensemble of valid configurations generated in the stage 3. Analysis of the obtained clusters provides protein positions, contacts and the localization probability distributions. The localization probability distributions should peaks at a single value with little standard deviation (single narrow maximum). This validates that the input restraints were sufficient to determine the protein positions. The occurrence of relative contact frequency of any two proteins can give the measure of protein proximities. Contact frequency defines the number of times two proteins come close to each other in an ensemble. A well-defined structure must have a single narrow maximum. Thus, the relative contact frequency should be highly conserved in the ensemble. At this stage, data not used in the structure construction/optimization was utilized to validate the models. Analyses such as electrostatic complementarity, EM density map fitting and suggesting wet lab experiments would further validate the models.

Valid Integrative Model

The output of the integrative modeling may lead to many clusters of models that satisfy the restraints. This clustering leads to three possible outcomes.

- *A single cluster satisfies all restraints.*
  Implies that the data is sufficient for determining the unique native structure.
- *Two or more clusters satisfy the restraints.*
  Implies that data is insufficient to resolve a unique native structure or there are multiple conformations of the system.
- *No cluster satisfies the restraints.*
  Implies that either the data were wrong or there was an error in data interpretation.

The structural differences among the models inside a cluster is an estimate of the quality of the models generated through the integrative approach.

An Integrative modeling approach was able to solve the structure of NPC using the diverse proteomics data. The assembled structure was consistent with the unused experimental data during modeling. NPC structure, when viewed as projections of the nucleoporins mass density, resembles the EM maps obtained earlier [58–60]. These EM maps independently validate the NPC structure. The diameter of the transport channel, ~39nm, was also consistent with the diameter established earlier [61]. The structure of the NPC was robust as omitting the 10% of the randomly chosen subset of the interaction

data do not change the localization probability. The structural model of NPC provided abundant insights into the function and evolution of the cell [5]. For instance, how NPC restricts the entry of inert molecules but allows the entry of selected cargo molecules [62,63]. It also provides a reason why certain nucleoporins were crucial in the assembly of NPC [64]. Including the additional high-resolution structural data concerning different stages of a dynamic process may provide NPC's assembly and transport mechanisms [65].

1.2.3 Advantages over conventional approaches.

Several methodologies exists to determine the structure of macromolecular assemblies (Figure 1-2). However, Integrative modeling provides a way to determine structures using information collected from different experiments. This methodology is especially useful when no single experimental technique alone can determine the structure. There are many advantages of the integrative modeling approach over traditional methods briefly described below.

1. Using Diverse and new data.

Integrative modeling provides the flexibility to incorporate the diverse and new information into the models. This incorporation is useful when a single type of structural data was insufficient to describe the system whereas a combination of different types can help assemble the whole system.

2. Increasing accuracy, precision, and completeness.

The models generated through integrative modeling satisfy the variety of restraints derived from the collected data. Thus, final models were more complete, precise and accurate compared to the models built using data from a single source.

3. Acquiring model dynamics and function.

The exhaustive sampling done in this approach generates all possible models that satisfy the given restraints. Thus, multiple models may capture the dynamic functionality of the protein (or any bio-molecule in general) and helps in understanding the protein's function inside the cell.

4. Experimentation planning.

Computationally testing the speculative information available from the generated models could help design further wet lab experiments.

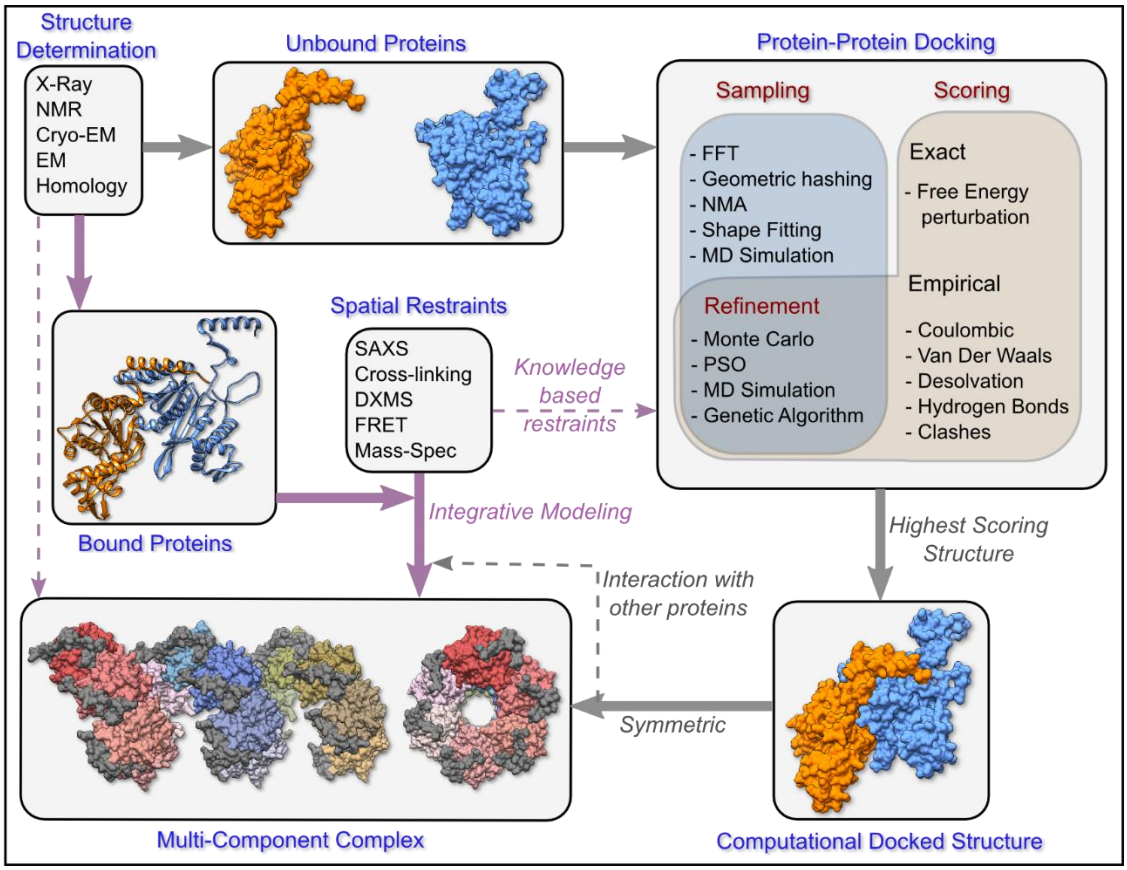


Figure 1-2 A general approach to determine the macromolecular assembly.

This approach uses computational techniques and is necessary when no single experimental method can determine the structure. Various experimental methods provide structure of individual protein molecules. Few experimental methods directly provide the interactions between individual proteins (purple arrows). Otherwise, protein-protein docking exhaustively samples and scores configurations to provide the docked structures (gray arrows). Integrative modeling using unbound proteins (dashed arrow) or bound protein structures along with other spatial restraints determine the macro-molecular complex. Adapted from [1].

1.2.4 Previous studies incorporating Integrative Modeling

Previously, Integrative modeling protocol determined the 3D structures of biologically important macromolecular complexes (Figure 1-3 and Table 1-2). The dimensions of these macromolecular assemblies range from 18nm to 6μm (See Table 9-2 in Appendix for a complete list of Integrative modeling studies).

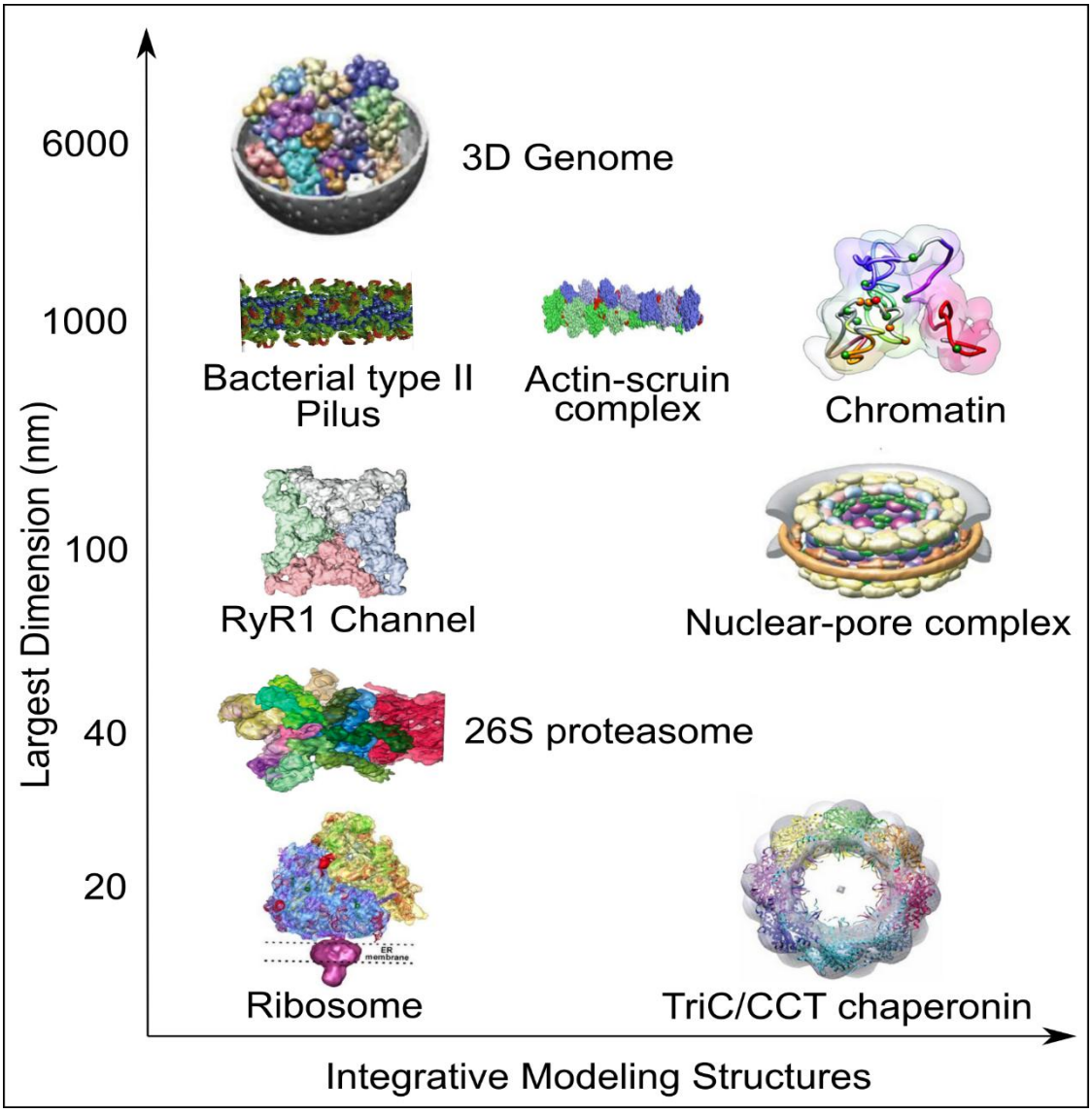


Figure 1-3 Examples of structures determined using Integrative modeling.

Pictorial representation of structures determined using integrative modeling [2,50,51,66–71]. *The ordering of the structures was according to their approximate sizes in nanometer (Y-axis).*

Table 1-2 Selected Integrative Modeling Examples

Table list selected list of models determined through Integrative modeling protocol. See Table 9-2 *in Appendix for a complete list of studies. 1st and 2nd Column lists the solved macromolecular complex and the experimental data used in the construction respectively. 3rd column gives the resolution of the final model.*

Model System	Structural Data	Resolution	Reference

26S proteasome	Cryo-EM, X-Ray, Cross-linking, Mass spectrometry	~8.4 Å	[50]
RyR1 Channel	Cryo-EM, Homology Modeling	~9.6 Å	[51]
TriC/CCT chaperonin	Cryo-EM, Homology Modeling	~15.0 Å	[66]
Bacterial type II pilus	Cryo-EM, EM, NMR, Cross-linking	~12.5 Å	[68]
Actin-scruin complex	X-Ray, Chemical cross-linking,	~9.5 Å	[69]
Chromatin	Cryo-EM, Mass spectrometry	~5-25 Å	[70]
Eukaryotic Ribosome	Cryo-EM, Homology Modeling	~11.7 Å	[72]
Mammalian Ribosome	Cryo-EM, Homology Modeling	~8.7 Å	[73]

1.3 Software for Integrative Modeling.

Integrative Modeling Platform (IMP) software package provides a wide range of modules for integrative modeling applications [74]. The software assists in the development of new model representations, scoring schemes, sampling algorithms, and various analysis methods for macromolecular assemblies. IMP encodes the models as a collection of particles, each of which could represent a part of the system. For instance, a single IMP particle could represent an atom, a molecule, a protein chain, or a complete protein. This feature of IMP helps reduce the complexity of the system.

For any system, IMP generates an atomic-level detail, coarse-grained or a hierarchical representation depending on the data available for modeling. Thus, IMP provides a straightforward method to represent any bio molecular system at various resolution levels. During modeling, IMP provide flexibility to represent different parts of the protein using different entities/particles. Every IMP particle has inbuilt attributes like coordinates, atomic symbol, mass, residue name, etc. It also provide modules for adding new attributes. All models within IMP could be evaluated using scoring functions rendered by the spatial restraints. For any configuration, the number of spatial restraints satisfied determine the overall score. The computational optimization of these configurations provide the high scoring models that satisfy the restraints. There are various sampling and optimization algorithms available in IMP. Optimization algorithms include Monte Carlo [75], conjugate gradients [76], the simplex optimizer from Gnu Scientific Library GSL [77]. Simulation schemes include molecular and Brownian dynamics [78], sampling schemes like Gibbs sampler [79], replica exchange [80] and a divide and conquer sampler named as DOMINO sampler [81]. IMP also provide modules for comparing, clustering

and analyzing the models.

1.4 Conclusions, Challenges and Possible Solutions

The Integrative modeling technique provides a standard scientific methodology to construct and analyze the 3D structures of bio-molecules and bio-molecular complexes. This technique is particularly useful when no single experimental method on its own can construct the structures. Integrative approaches have an advantage over traditional methods as they provide a system to combine vastly different experimental information and then integrate them into a structural model. If the experimental data is not enough to determine the structure, the integrative approach could provide structural models for testing new hypotheses.

Identifying the amount of data needed to uniquely determine the structure is one of the problems in Integrative modeling. Sometimes available data may not be sufficient to determine the structure uniquely. A diverse range of expertise is needed to combine and derive explicit experimental restraints for generating the missing data. This enhances the resolution of the assembled structure. Techniques such as the incorporation of sub-sequences of amino acids as fluorophores, cross-linkable groups, and EPR labels have to be developed at low cost to simplify the process of deriving experimental restraints. Combining these with the development of an efficient and automated software system for experimental data analysis can produce rapid and low-cost structural information.

Another major problem in this approach is combining the spatial restraints at different levels of resolution. For instance, structural data for a protein will give information about its overall shape as well as about finer atomic details. Combining the data that has different resolution is a time-consuming and challenging task. Efficient and intelligent algorithms to automate this process would ease the modeling protocol. Various techniques like machine learning, computational geometry algorithms, support vector machines, etc. are well developed but have not been used extensively for this kind of applications. A variety of learning algorithms targeted to solve this problem would be useful.

Resolution of a model is an important parameter in determining the specificity in protein interactions, functions and dynamics. It provides specific details with a certain level of confidence. It is often difficult to determine the resolution of the structures generated from the integrated modeling approach. This is a big challenge as different parts of the model have a different resolution. Clustering of the assembled models could provide resolution

of the overall model. Defining the resolution of different parts of the model could also be a possible solution. An automated software that combines the various data at different resolutions and reports the resolution of the assembled models would be useful.

At this stage, focused efforts from diverse experimentalist groups together on a particular problem can generate data at various levels of resolution quickly and efficiently. This type of data will be more oriented towards integrative modeling applications and hence easy to interpret. Combining data from different experiments can overcome the limitations of a single experimental and computational method. Integrative modeling technique is progressing very rapidly, and at this stage, we can build time-resolved models. The experimental data at various time points could help capture the dynamics of the system. Structurally oriented bio-informatics is currently the need for better and more efficient structure determination through integrative modeling. Integrative approaches are a recent innovation, and it is expected that the few existing examples would grow to many more in the immediate future.

CHAPTER 2
Background and Related Work

SALIENT FEATURES

- ***Introduction to Coiled-coil proteins***
    - Coiled-coil protein architecture
    - Amino acid sequence pattern and coiled-coil geometry
    - Databases and Software for Coiled-coil proteins
    - Experimental studies on Coiled-coil proteins
    - Coiled-coil stability and specificity determinants
    - Biological implications of coiled-coils
- ***Introduction to Intermediate Filaments (IFs)***
    - Intermediate filaments and functions
    - Types of IFs, Architecture and Assembly
    - Current status of 3D structure of IF
    - Introduction to Keratin K5-K14 Intermediate filament
    - Keratin K5-K14 associated diseases
    - Structural Instability within Intermediate filaments
- **Summary**

2.1 Introduction to Coiled-coil proteins

2.1.1 Coiled-coil Protein Architecture

A coiled-coil is a frequently occurring structural motif formed by a set of alpha helices that wind around each other to form a supercoil. Approximately 3-5% of all amino acids in proteins are coiled-coils [82,83]. They consists of two to seven alpha helices aligned parallel or antiparallel to one another [84]. Interactions between the alpha helices in a coiled-coil guide their orientation.

2.1.2 Amino acid Sequence Patterns and Coiled-coil Geometry

Coiled-coil protein sequences exhibit a pattern of hydrophobic and hydrophilic amino acids with the smallest unit known as a repeat. There are different types of repeats such as a heptad repeat and a hendecad repeat. The most common of them is a heptad repeat which is a sequence of seven amino acids and positions labeled as a/b/c/d/e/f/g [20,85]. Heptad positions **a** and **d** contains hydrophobic amino acids that form the core of the coiled-coils whereas charged groups occupy heptad positions **e** and **g** and facilitate the formation of inter-helical salt bridges (Figure 2-1). Polar residues that form the outer surface of the coiled-coil occupy heptad positions **b**, **c** and **f**. The amino acids at these heptad positions interact with the solvent molecules. Though these are the most common patterns observed, deviations also exist in many coiled-coil protein sequences.

2.1.3 Databases and Software for Coiled-coil Proteins

There exist a range of databases and software for characterizing and predicting coiled-coil proteins. Software such as PCOIL, Paircoil2, MARCOIL, Multicoil2 and DeepCoil can detect the heptad repeats in protein sequences [82,86–90]. They compare the protein sequence with the sequence of known coiled-coil structures over the sequence window lengths of 14, 21 or 28. Other software such as SOCKET, assign the heptad repeat to the protein sequence by detecting knob into hole packing of side chains [91]. These programs have been used to detect the coiled-coil geometry packing in known PDB structures and housed in CC+ database [92]. Additionally, the coiled-coil structures can be defined mathematically using parametric equations [93]. Despite this simplicity in the structure, the molecular determinants of the coiled-coil stability and specific pairing are still not clear [94–96].

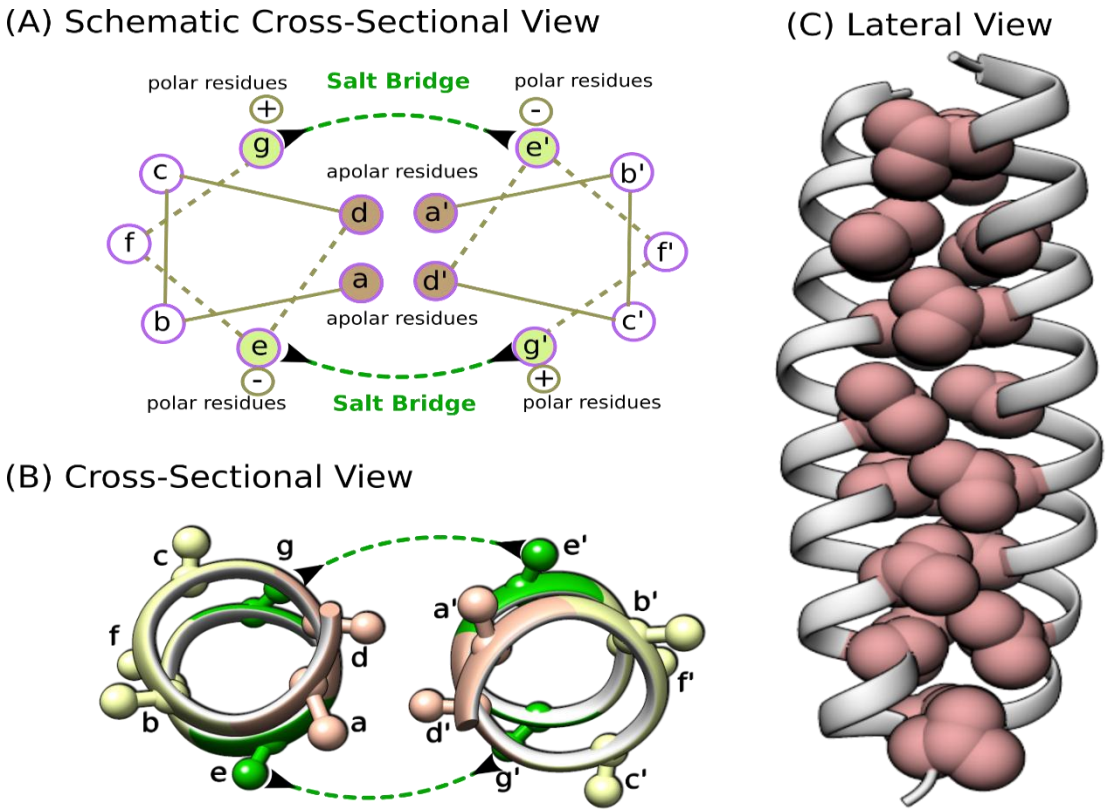


Figure 2-1 Schematic representation of coiled-coil and heptad geometry.

(A) Schematic representation of a 2-helix coiled-coil labeled with heptad repeats ***a-b-c-d-e-f-g*** *(apostrophe indicates partner helix). The circle around heptad repeat represents the residues in the helix connected through gray lines. Green and brown color shows positions of polar and apolar residues favoring possible salt bridges (green dotted line) and hydrophobic core respectively. (B) The cross-sectional view of a 2-helix coiled-coil in ribbon representation labeled with heptad positions. Side chains at heptad position* ***d*** *point towards each other whereas side chains at heptad position* ***a*** *are anti-parallel to each other. (C) A coiled-coil helix (gray ribbon) with amino acids at heptad positions* ***a*** *and* ***d*** *forms the core of the coiled-coil (brown spheres).*

2.1.4 Experimental Studies on Coiled-coil Proteins

Many experimental studies have attempted to determine the stability of coiled-coils by changing one amino acid at a time and then measuring the change in free energy [97–101]. Several other studies attempted to determine the role of buried residues at heptad positions **a**/**d** and the relative orientation of the helices [102–106]. There were few studies that predict the interaction specificity directly from coiled-coil sequence and predict the oligomerization states [87,107,108]. Few studies try to identify the molecular determinants of the oligomerization-state specificity, core packing geometry, inter-helical

electrostatic interactions [109–111]. These studies provide information on the stability and specificity of the coiled-coils. However, the exact molecular mechanisms of coiled-coil stability and the molecular determinants associated with them are not clearly understood.

Experimentally, it is difficult to study all free parameters for the coiled-coil proteins. Parameters include free energy change due to mutations at various heptad positions, addition of polar residues in the hydrophobic core, the distance of mutation from N and C terminus, and the distribution of salt bridges (ion-pairs) along the coiled-coils [112–119]. Studies that determine the frequency of the amino acids at different heptad position gave insight about the pair preferences between the coiled-coil forming helices [120]. Such studies individually gave enormous amount of information about the coiled-coils. However, these studies do not satisfactorily account for all the disease-causing substitutions in the coiled-coils.

2.1.5 Coiled-coil Stability and Specificity

The “Peptide Velcro” (PV) hypothesis outlines the elements involved in the formation of coiled-coils [121]. The hypothesis rationalizes the formation of hetero-dimeric coiled-coils by destabilizing homo-dimeric coiled-coils. It outlines the three structural elements categorized based on the heptad repeat positions. 1) Coiled-coils are stabilized through hydrophobic contacts with heptad positions **a**/**d** occupied by hydrophobic amino acid. 2) Coiled-coils are stabilized (electrostatic attraction) or destabilized (electrostatic repulsion) through charged groups present at heptad positions **e** and **g**. 3) Polar groups at the **b**, **c**, and **f** positions help in stabilizing the coiled-coil through the interaction with the solvent molecules. The variants of these interactions determine the specificity, orientation and oligomerization state of the two participating helices.

2.1.6 Biological Implications of Coiled-coils

Many proteins involved in the cell growth, proliferation, development of bones and cartilage, and some DNA binding proteins such as Fos and Jun transcription factors have coiled-coil domains [122,123]. Coiled-coils have been shown to function as antibody stabilizers, anti-cancer drugs, temperature regulators, purification tags and linker systems [83]. Varied distribution of coiled-coil domains and their simple geometry that can be explained using parametric equations, makes them potential biochemical and therapeutic agent [93,124]. The structural and functional diversity associated with coiled-coil proteins makes them a suitable model system for current studies [125].

2.2 Introduction to Intermediate Filaments (IFs)

2.2.1 Intermediate Filaments and Functions

Actin microfilaments (MFs), microtubules (MTs) and intermediate filaments (IFs) are the three leading filament system of the cytoskeleton of metazoan cells [11]. The network of these integrated filament systems imparts mechanical stability and is involved in several cellular processes such as cell signaling, transport and differentiation, metabolism of cholesterol, supporting/anchoring cell organelles etc. [12–15]. Among these, IFs are one of the most abundantly found proteins in metazoan cells and tissues [126]. They have derived their name from the diameter, about 8-12 nm, that is in between that of MFs and MTs. Among the three classes of filaments, IFs were discovered first by electron microscopy of skeletal muscle cells and analyzed by X-ray crystallography. They are a class of filaments first discovered by electron microscopy of skeletal muscle cells and one of the first to get analyzed by X-Ray crystallography in 1930 [127]. The sequences of Intermediate filaments are highly conserved irrespective of their origin [128–130]. IFs are the principal structural elements in the nucleus, cytoplasm, and epithelial cells [131–134]. They are similar to non-IF like proteins such as myosin, collagen, and fibrinogen but distinctly different when compared to globular proteins [135]. They are highly insoluble and require specific mechanism or proteins such as kinases to solubilize them [136]. The mesh formed by IF inside the cell provides resistance to mechanical stress during cell division or migration, thereby protecting the cell from external and internal processes [136–138]. The stress buffering functionality of the Intermediate filaments protects the cell from various types of stress such as chemical, microbial and physical [139,140]. This stress buffering system due to their filamentous structure is most prominent in the epithelial layer of the skin that are exposed to changing environment. Experimental studies have shown that the Intermediate filaments can stretch up to 3.4 times the original length while exposed to physical stress [141]. These properties of Intermediate filaments is one of the current topic of research [139].

2.2.2 Types of IFs, Architecture and Assembly

IFs, unlike the other two cytoskeletal filaments, are different in different cell types. For instance, in humans, 73 genes code for various IF proteins that get classified into six classes based on the sequence homology and functions [11,16,142] (Table 2-1). IFs of Type-I and Type-II are co-expressed in tissues to form heterodimers whereas Type III to

Type VI IFs in general form homodimers [17]. There are few exceptions such as Vimentin-Desmin that assemble as heterodimers when co-expressed *in-vivo* [143].

The dimers are the basic building blocks of all IFs. Various experiments and sequence analyses have confirmed the presence of 3 domains in the IF dimer, i.e., Head domain, Central domain and tail domain [18–20,144]. Electron Microscopy (EM) study revealed the coiled-coil IF protein dimers as distinct 45-50 nm long rod-like molecules [145]. These dimers consist of long, often uninterrupted, segments of parallel alpha helices that wind around each other to form central coiled-coil rod domain. The 8-17 amino acids long non coiled-coil linker domains (L1, L12, and L2) connects the parallel coiled-coil segments of alpha helices (1A, 1B, 2A, 2B) [11,146]. The central coiled-coil domain contains about 624 amino acids sandwiched by non-alpha helical Head (N-terminal) and Tail (C-terminal) domains [18–20,144,147]. Head/Tail domains are inherently unstructured due to the presence of a large number of glycine and serine amino acids that provides more flexibility to their overall structure [148]. They also make essential molecular interaction with the neighboring dimers for the formation of the tetramer or higher oligomers [16,149].

Table 2-1 Classification of Intermediate filaments.

Classification of different types of Intermediate filaments. Six types of IFs (Row 1), the examples of Intermediate filaments (Row 2), and example for type of tissues in which they are found (Row 3).

IF types	Type I (Acidic)	Type II (Basic)	Type III	Type IV	Type V	Type VI
Example of IF	Keratin K10 to K18	Keratin K1 to K8	Vimentin Desmin	Neurofilament, Peripherin, GFAP, Nestin	Lamin	Phakinin, filensin
Example Location	Simple epithelial, Epidermal, hair, nails		Blood vessels	Nervous system	Nucleus	Eye Lens

The molecular mechanism of assembly of IF protein dimers into 10-nm thick filaments is still unclear [150–152]. The concept of IF assembly is summarized using the schematic representation (Figure 2-2). Cross-section of the IFs consists of 24 to 40 monomers according to mass-per-length measurements by Scanning Transmission Electron Microscopy (STEM) with a peak at about 32 monomers [149]. One of the main problems in understanding the mechanism is the unavailability of the complete molecular structure

of the IF protein dimer. The second biggest challenge is to construct the complete filament structure along with head/tail domains.

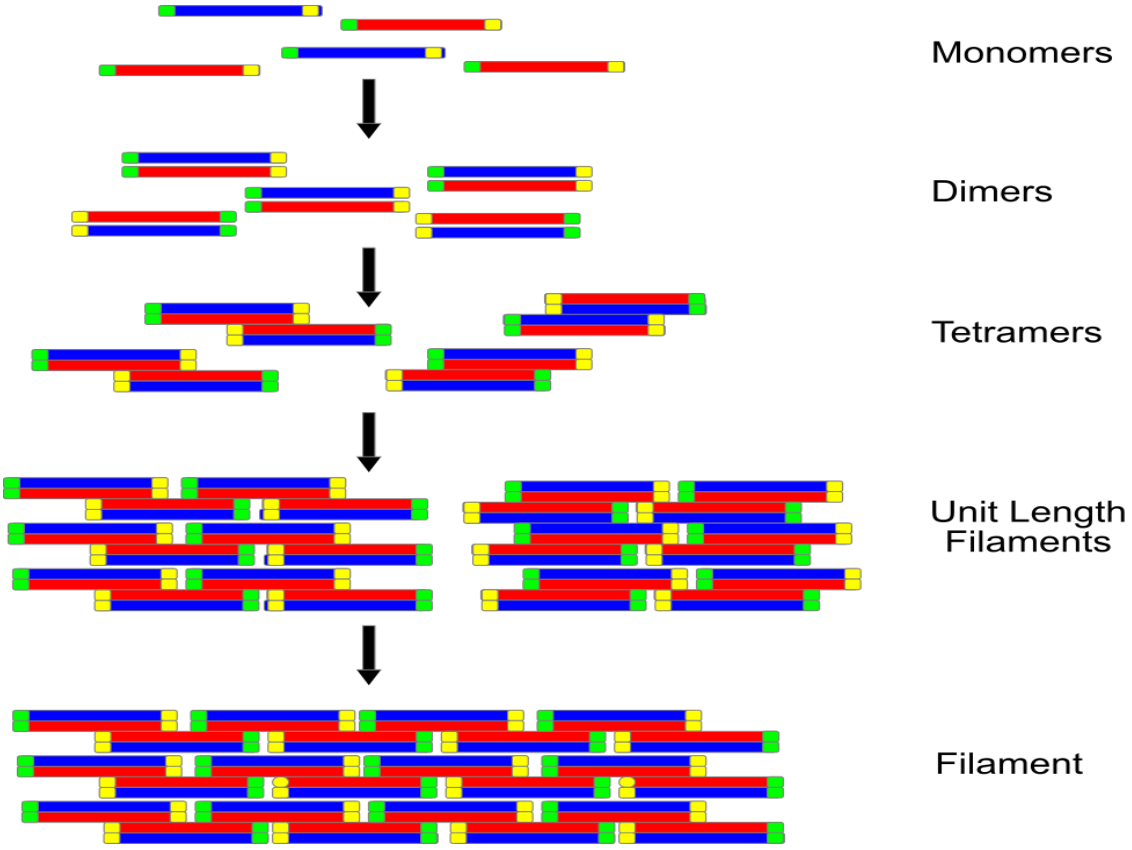


Figure 2-2 Schematic of IF filament architecture and assembly process

The assembly process of Intermediate filament starts with two different monomers with central alpha helical domain (blue and red) flanked by non-alpha helical head/tail domains (green/yellow). Initially, different monomers combine to form parallel coiled-coil heterodimers. The coiled-coil dimers then interact in anti-parallel arrangement to form tetramers and higher order oligomers. Adapted from www.interfil.org [21].

2.2.3 Current Status of 3D Structure of Intermediate Filament

In the past 50 years there were considerable improvements in getting more insights about the structure of Intermediate filaments [135,153–155]. Among Intermediate filaments, the majority of the studies were done on Keratin and Vimentin [156–158]. The technological advances limit our knowledge about the 3D organization of Intermediate filament. Various groups have tried to crystallize the Intermediate filaments dimers. However, with current technology only segments of crystal structures of Intermediate filament dimers (PDB ID: 3S4R [159], 3UF1 [160], 3SWK [159], 3G1E [161], 3TRT [162], 1GK4 [147], 4ZRY [163]) were obtained. These segments were short coiled-coil dimers and have large

deviations from the standard coiled-coil geometry. These deviations, established by large beta factor values, was always near the linker domains. Such coiled-coil segments of Intermediate filament dimers tend to show crystal artifacts and have variable pitch as in the case of Vimentin study [159].

Many groups have modeled the coiled-coil domain of Intermediate filament dimer by joining various crystal structure segments [164,165]. This was done by aligning and stitching together different crystalized segments on an axis [159]. The temperature fluctuations in the crystal segments leads to tilted/bent models. These models deviate from standard coiled-coil geometry leading to packing defects. Complete filament models constructed using such non-straight and improperly packed dimers could lead to incorrect and functionally infeasible models.

Similar study on the Keratin Intermediate filament built the model of Keratin dimer. Along with the central coiled-coil domain of the dimer, a possible head and tail domain was constructed through random placement with prior information. The complete model was then subjected to molecular dynamics simulation for trajectory analysis [19].

A computational study models the Vimentin Intermediate filament with the low-resolution SAXS data. However, the study provides no conclusive model as a randomly arranged dimer model of the filament fits best with the experimental data [166].

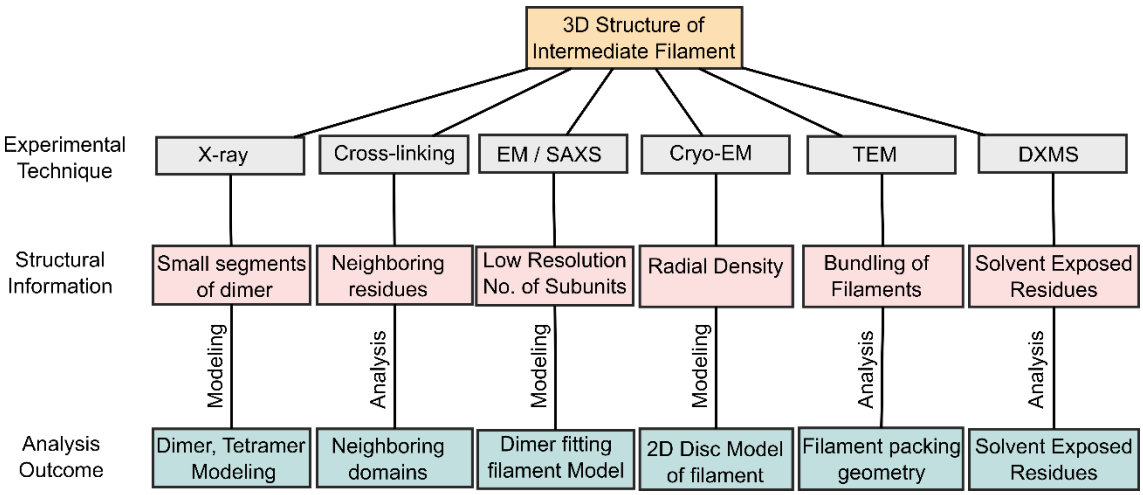


Figure 2-3 Summary of studies relating to structure of Intermediate Filament.

Studies for determining 3D structure of the intermediate filaments (coral box) utilizes several experimental methods such as X-ray crystallography, cross-linking amino acids, electron microscopy etc. (gray boxes). These experimental techniques provide structural/spatial information (pink boxes). Subsequently, the modeling/analysis of the structural information provide insights about the structure of the dimer/filament (blue boxes).

A deuterium exchange experiment provides some knowledge about the accessibility of the amino acids with the solvent. The study provides regions on the filament that were masked by head/tail domain [167]. Deuterium exchange data provide essential interaction information between head/tail domain and coiled-coil domains. However, the absence of 3D structure of head/tail domains remains the bottleneck in modeling the complete filament models.

A significant study that provided the 2D topology of the filament at the proto-fibril level justifies the "8+0" model rather than "7+1" model. The study uses fitting of Intensity transform profile of single dimer to match with that of complete filament. This was done by 2D models considering a single dimer as a disc [168].

2.2.4 Keratin Intermediate Filament

Keratins are the members of the Intermediate Filament family of proteins (Type I and Type II) that line the periphery of the epithelial cell (Section 2.2.2, Table 2-1) [169]. They are also called as cytokeratins and account for most of the Intermediate filaments [19,148]. Keratin encodes for 54 functional genes out of the 73 human IF genes [11,16,136]. Unlike most other IFs that form homodimers, keratin assembles with Type-I and Type II IFs to form heterodimers [17]. Different epithelial cells co-express more than one kind of keratin pair leading to another subdivision of the keratins [138,170,171]. These include 'structural keratins' of trichocyte that form hard appendages such as nails, hairs, horns (K86-K36, K82-K80, K71-K25), 'simple keratin' of one layered epithelium (K8-K18) and 'barrier keratins' of stratified epithelia (K1-K10, K3-K12, K5-K14) [21,148]. Simple and barrier keratins are also known as soft keratin or epithelial keratin whereas 'structural keratin' are known as hard keratin or trichocyte keratin owing to their rigidity due to additional disulfide cross-links within the structure [172,173]. Soft keratin redistributes mechanical forces across the cell membrane, thus, preventing the cell from damage and keeps it in its native shape [131,174]. Soft keratin IF network is one of the major contributing factors of the elastic properties of the skin.

2.2.5 Keratin Intermediate Filament and Associated Diseases

Expression of Keratin genes (or in general IFs) is highly specific in patterns related to the epithelial type and stage of cellular differentiation. Keratins in the epithelial cytoskeleton play a significant functional role in mechanical stability and integrity of the cells through cell-cell contacts. They are also involved in providing rigidity to the nuclear shape [175].

They also play a role in intracellular signaling pathways; for instance, wound healing, protection from stress, and apoptosis. Consequently, they are an inherent part of the system at various levels of cell organization from single cell to tissue formation. Mutations in the keratin genes lead to more than 70 hereditary diseases also known as "Keratin disorders" such as Epidermolysis bullosa simplex and its subtypes, Dowling-Degos disease, and Epidermolytic palmoplantar keratoderma [21]. The incidence rate of keratin disorders are 1 in 3000. In diseased individuals, keratin is susceptible to fragmentation. The importance of the genetic constitution of keratin has been studied extensively using transgenic mice models [22,176]. Approximately 119 distinct diseases such as Alexander disease, EBS etc. have been associated with single point mutations in IF gene [21]. The resulting single amino acid change in the gene product leads to a destabilization of the IF filament or the IF network, leading to diseased conditions. Complete information about the Intermediate filament proteins and their associated diseases could be found in the Intermediate filament database hosted at http://www.interfil.org [21].

2.2.6 Structural Instability within Intermediate Filament dimers

Despite the severity of the Intermediate filament associated diseases, little is known about their molecular mechanism. For instance, disease conditions are attributed to the replacement of one hydrophobic amino acid by another hydrophobic amino acid. Studies on the factors that stabilize coiled-coil structures, throw little or no light on why hydrophobic to hydrophobic substitutions cause structural instability. Few studies have been done that takes in to consideration of pairing preferences of amino acids [108,120]. However, these pairing preferences would not be able to explain structural instability caused by hydrophobic to hydrophobic substitution.

To determine the structural instability, various substitutions have been made for different heptad positions of the coiled-coil proteins [110,177]. Several groups have identified packing preferences that play significant role in the IF dimer stability [97,100,118,177]. A single amino acid change can change the oligomerization state of the monomer. Other rules such as presence of beta branching amino acids at heptad position **d** destabilize the standard coiled-coil dimers due to atomic clashes. However, IF dimers have beta branching amino acids and remains stable. This remains a topic of interest in the field and holds the key to the mechanism of disease causing or filament destabilizing mutations.

CHAPTER 3

IF Dimerization and Stability Determinants

SALIENT FEATURES

- ***Construction of Intermediate filament dimers database***
  - Heptad repeat alignment and conservation across IFs
  - Distribution of disease causing variants
- ***Intermediate filament stability feature determination***
  - Sequence and structural analysis of IF dimers
  - Molecular determinants of stability
- **Construction of Neural Network Schemas**
  - Training and testing dataset and their representation
  - Bootstrapping and performances
- **Using Neural network Schema for classifying Hex pairs**
  - Prediction of novel filament disrupting mutations
  - Experimental validation of novel K5-K14 mutations
- **Homology modeling of naturally occurring and non-native dimers**
  - Structural analysis of energy minimized models
  - Comparison if stability
  - Equilibrium between homo-dimers and hetero-dimers
- **Molecular level insight of disease conditions**
- **Scope for the development of therapeutic agents**
- **Summary**

The previous chapter discusses about the preferential binding of two Intermediate Filaments (IFs) monomers to form a coiled-coil dimer (Section 2.2.2). IFs of Type-I and Type-II are co-expressed in tissues to form heterodimers whereas Type III to Type VI IFs in general form homodimers [17]. The molecular interactions within a coiled-coil geometry govern the stability of IFs dimers. This chapter describes the methodology involved in determining the Intermediate filament dimerization principles and the molecular determinants governing the stability and specificity of any coiled-coil (including IF) dimer.

3.1 Constructing the Database of IF Sequences for Stability Analysis

To determine the Intermediate filament dimerization principles, 21 human intermediate filament sequences were downloaded from www.interfil.org [21]. Total 35 IF sequence pairs (dimers) were constructed and equally divided into five sets under two categories (1) Naturally occurring or native dimers comprising of co-expressed heterodimers of type I-II, homodimers of type III to type VI, and (2) synthetic or non-native dimers consisting of non-co-expressed type I-II heterodimers, homodimers of type I and type II. The dimer types along with the examples is shown in Table 3-1. The molecular determinants of IF dimers pair-preferences were extracted from the naturally occurring dimers while the non-native dimers served as a negative control. From here on, the "IF Dataset" will represent the pairwise sequences of native and non-native dimers described in Table 3-1, unless mentioned otherwise. The following sections describes the sequence and structural analysis of the coiled-coil domains from all the dimers in the dataset (Refer Section 2.2.2 for details on coiled-coil domains). The analysis of the IF dataset leads to various molecular determinants that provide stability and specificity to the coiled-coil dimers. The next few sections briefly describes the methodology used to align the coiled-coil domains of IFs.

Table 3-1 Data set of native and non-native IF Dimers.

First column indicates whether the IF dimers are naturally occurring or non-native dimers. Second and third column provides IF dimer type and their lists respectively. All Intermediate filament sequences were obtained from www.interfil.org [21].

	IF Type	IF Dimers
Naturally Occurring dimer	Type I-II Hetero-dimer (co-expressed)	Keratins: K1-K10, K3-K12, K4-K13, K5-K14, K6a-K16, K6b-K17, K8-K18
	Type III-VI Homo-dimer	Light (NFL) and Medium (NFM) Neurofilament, Peripherin, Vimentin, α-internexin (ANX), Desmin, Glial fibrillary acidic protein (GFAP)
Non-native Dimer	Type I Homo-dimer	Keratins: K10-K10, K12-K12, K13-K13, K14-K14, K16-K16, K17-K17, K18-K18
	Type II Homo-dimer	Keratins: K1-K1, K3-K3, K4-K4, K5-K5, K6a-K6a, K6b-K6b, K8-K8
	Type I-II Hetero-dimer (non-co-expressed)	Keratins: K1-K18, K3-K10, K4-K12, K5-K13, K6a-K14, K6b-K16, K8-K17

3.2 Sequence alignment and Conservation across IF dimers

3.2.1 Heptad repeat based alignment

The Intermediate filament coiled-coil domains are about 41% identical to one another (Table 3-2). Within different classes of IFs, Type II IF is the least diverging with 65% identity when compared to other IFs (Table 3-2). The alignment of all Intermediate Filament sequences was performed by aligning the coiled-coil domains guided by PCOIL predicted heptad repeats [87]. Heptad repeats of the linker domains (see Section 2.2.2) were interpolated from the heptad repeats of the neighboring coiled-coil segments. The resulting alignment (refer to file: ALL_IF_alignment.txt in Supplemental Material) was used for collecting statistics such as frequency of different amino acids, pairwise volumes, atomic density, polar and charge interactions, etc. for various heptad positions. Being a coiled-coil structure, the hydrophobic core (formed by heptad positions **a**/**d**) stabilizes the

IF dimers. This chapter will discuss the stability and specificity of the IF dimers inferred based on the heptad positions **a/d**.

Table 3-2 Intermediate filament sequence identities and similarities.

In addition to IFs described in Table 3-1, f*ollowing IFs were also utilized to calculate identity and similarity. Keratin K2, K6c, K7, K9, K15, K19, K20, K23-28, K31-40, K71-86. Following amino acid groups were considered for similarity calculations: Aromatic: F, Y, W. Aliphatic: V, I, L. Positive charged: R, K, H. Negative charged: D, E. Small: S, T. Polar: N, Q.*

Intermediate Filament	Average Sequence Identity	Standard Deviation (Identity)	Average Sequence Similarity	Standard Deviation (Similarity)
Type I	57.2%	11.9%	67.2%	9.7%
Type II	65.0%	11.2%	74.1%	8.3%
Type III-VI	54.8%	7.3%	67.8%	7.1%
ALL Sequences	41.3%	17.2%	54.7%	14.5%

3.2.2 Distribution of Disease Causing Mutations in IF dimers

All native dimers (Table 3-1) were analyzed for 438 disease causing single point mutations at different heptad positions along the length of coiled-coil domain (amino acid numbers from 1 to 325). The analysis suggest (a) 220 (about 50%) of all the disease-causing mutations were present at the ends of the coiled-coil domains (Figure 3-1). (b) 153 mutations of the total 438 (~35% of mutations) were observed to occur at the heptad positions **a**/**d** (Table 3-3). (c) The 86 heptad position **a**/**d** mutations out of 153 (~56% of all **a**/**d** mutations) were present at the ends of the coiled-coil domains (Figure 3-1). This suggests that IF coiled-coil dimer is susceptible for destabilization at the hydrophobic core formed by amino acids at heptad positions **a**/**d**. This is an important parameter for IF dimer coiled-coil stability. However, the occurrence of mildly destabilizing disease mutations such as hydrophobic to hydrophobic substitutions cannot be explained. Thus, a more robust analysis at the molecular level is required to explain the onset of diseases.

Table 3-3 Intermediate filament disease causing mutations at heptad a/d

All disease-causing mutations for Intermediate filaments described in Table 3-1 were obtained from www.interfil.org [21].

IF	Mutant	IF	Mutant	IF	Mutant	IF	Mutant	IF	Mutant
Desmin	L274P	K1	F191I	K12	Y429C	K16	L124P	K5	D328E
Desmin	L274R	K1	F191C	K12	L433R	K16	L124H	K5	A428T
Desmin	L338R	K1	F191L	K14	M119I	K16	L128Q	K5	A428V
Desmin	N342D	K1	L208P	K14	M119V	K16	K354N	K5	L463P
Desmin	L345P	K1	F231L	K14	M119T	K16	L421P	K5	I467L
Desmin	I367F	K1	D340V	K14	L122F	K17	M88T	K5	I467T
Desmin	L370P	K1	D340G	K14	Y129D	K17	M88K	K5	I467M
GFAP	M73R	K1	L437P	K14	Y129C	K17	L91P	K5	Y470H
GFAP	M73K	K1	I479F	K14	V133L	K17	L95Q	K5	E477K
GFAP	L76F	K1	I479T	K14	V133M	K17	L95P	K5	E477G
GFAP	L76V	K1	Y482C	K14	V133A	K17	Y98D	K5	E477D
GFAP	Y83H	K1	L486P	K14	L136Q	K17	V102M	K6a	I167N
GFAP	V87L	K1	L486R	K14	L136P	K17	N109D	K6a	I167S
GFAP	V87G	K1	E489K	K14	N140S	K17	L388P	K6a	L170F
GFAP	L90P	K10	M150R	K14	L143P	K17	L388R	K6a	F174V
GFAP	L97P	K10	M150T	K14	I176M	K18	I150V	K6a	F174S
GFAP	L101P	K10	L153V	K14	R211P	K18	Q285R	K6a	F174C
GFAP	E207K	K10	Y160D	K14	D273G	K3	E509K	K6a	I462S
GFAP	E207Q	K10	Y160S	K14	I377N	K4	E520K	K6a	I462N
GFAP	L235P	K10	K439E	K14	I377T	K5	I172V	K6a	Y465H
GFAP	Y242D	K10	L442Q	K14	L384P	K5	L175F	K6a	Y465C
GFAP	Y257C	K10	I446T	K14	L401P	K5	F179S	K6a	L469R
GFAP	L331P	K10	Y449D	K14	L408M	K5	V186L	K6a	L469P
GFAP	L352P	K10	Y449C	K14	I412F	K5	V186M	K6a	E472K
GFAP	L359V	K10	L453P	K14	I412N	K5	V186E	K6a	E472D
GFAP	L359P	K12	M129V	K14	Y415H	K5	N193K	K6b	E472K
GFAP	Y366H	K12	M129T	K14	Y415C	K5	R265P	NFL	L94P
GFAP	Y366C	K12	L132P	K14	L419Q	K5	V324D	NFL	I213M
GFAP	E373K	K12	V143L	K14	E422K	K5	V324A	NFL	L268P
GFAP	E373Q	K12	I426V	K16	M121T	K5	D328H	NFL	L333P
GFAP	E373D	K12	I426S	K16	M121K	K5	D328V	NFL	E396K
K1	L187F	K12	Y429D	K16	L124R	K5	D328G	PRPH	D141Y

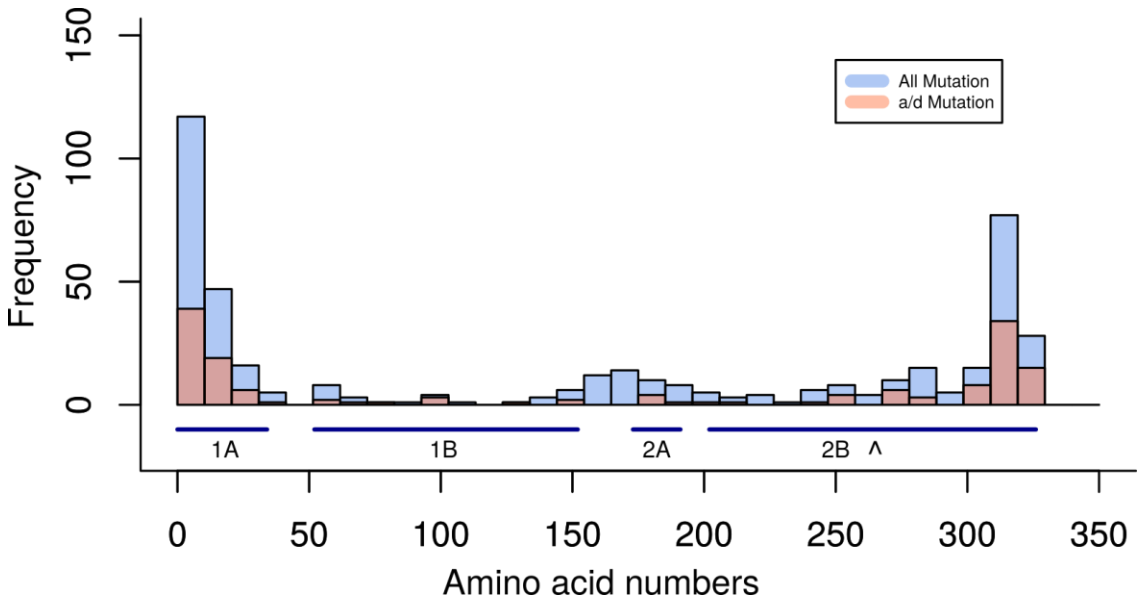


Figure 3-1 Distribution of Disease Causing Mutations in IFs.

The frequency distribution (y axis) of disease causing mutations along the length of coiled-coil dimer (x axis: amino acid residue position). Mutations at all heptad positions (blue bars) were plotted along with mutations at only heptad positions ***a/d*** *(red bars). Residue numbers (1 to 325) starts from the coiled-coil domain.*

3.3 Intermediate Filament Stability Feature determination

The stability analysis of any protein at the molecular level requires determining its molecular interactions. As coiled-coil proteins follow heptad repeats, its amino acid sequence was mapped to a standard geometry to infer information about the amino acid positions (Figure 3-2 A, C). The mapped amino acids then provide interactions within a coiled-coil structure. For instance, in a coiled-coil protein sequence, the predicted heptad repeat sequence maps ASN amino acid at heptad position **a**. Since heptad **a** pairs with heptad **a** of partner helix, a hydrogen bond is assumed to be formed between ASN amino acids of the two helices in a coiled-coil structure (Figure 3-2 B). Similarly, electrostatic repulsion was assumed if two similarly charged amino acids were mapped at both **a**/**d** heptad positions (Figure 3-2 B, C). Only heptad positions **a/d/e/g** of the two helices directly interact with one another. Such interactions determine the stability and specificity of coiled-coil structures. In the following section, the methodology adopted to extract various features important for a coiled-coil structure stability is discussed. The features include molecular determinants such as amino acid hydrophobicity, branching, charge position, volume and heuristic (Table 3-6).

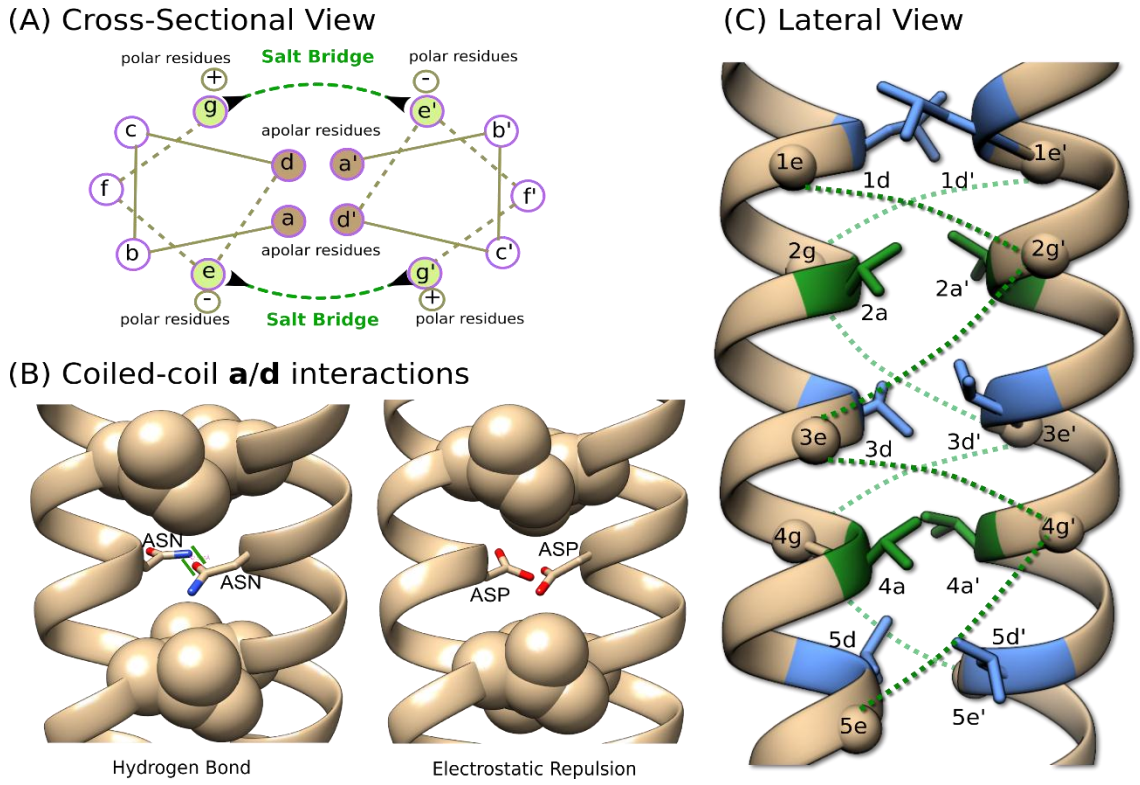


Figure 3-2 Elements of a coiled-coil heptad repeat.

*(A) Schematic represents heptad positions **a-b-c-d-e-f-g** and **a'-b'-c'-d'-e'-f'-g'** as labeled circles on two different helices of a coiled-coil in a cross-section view. Gray lines connect the positions, solid and dotted lines indicate different planes. Polar (green) and apolar (brown) residues form salt bridges (green dotted line) and hydrophobic core respectively. (B) The ribbon representation of coiled coil indicates the interactions at heptad positions **a** and **d**. Sphere represents hydrophobic core formed by side chains at heptad position **a/d**. The interaction types at heptad positions **a/d** is represented as: (Left) Hydrogen bond formed by ASN-ASN and (Right) Electrostatic repulsion from charged amino acids (ASP-ASP). (C) Representation of a coiled-coil helix (gray ribbon) in a lateral view with amino acids at heptad positions **a** (green) and **d** (blue). Salt bridges between **e-g' and e'-g** are shown in green color dotted lines.*

3.3.1 Pairwise Distribution of Amino acid

Heptad positions **a**/**d** forms the hydrophobic core of a coiled-coil dimers with amino acids at heptad pairs **a-a'** and **d-d'**. The hydrophobic interaction at **a-a'** and **d-d'** heptad positions plays the dominant role in keeping the two helices together, whereas electrostatic interactions provide specificity [83,100,104,121,177]. Native IF dimers from the "IF Data set" (Table 3-1) and non-intermediate filament coiled-coil dimer

sequences were used to calculate amino acid pair preferences. Non-intermediate filament coiled-coil dimer sequences were downloaded from CC+ database [92]. All two helices parallel homodimers and heterodimers sequences were considered from the database that contained at least 28 amino acids per helix. The analysis of dimer sequences combined with heptad positions provided the frequency of amino acid pairs at heptad positions **a-a'** and **d-d'**. This identified the differences in amino acid pair preferences between coiled-coils and IF proteins.

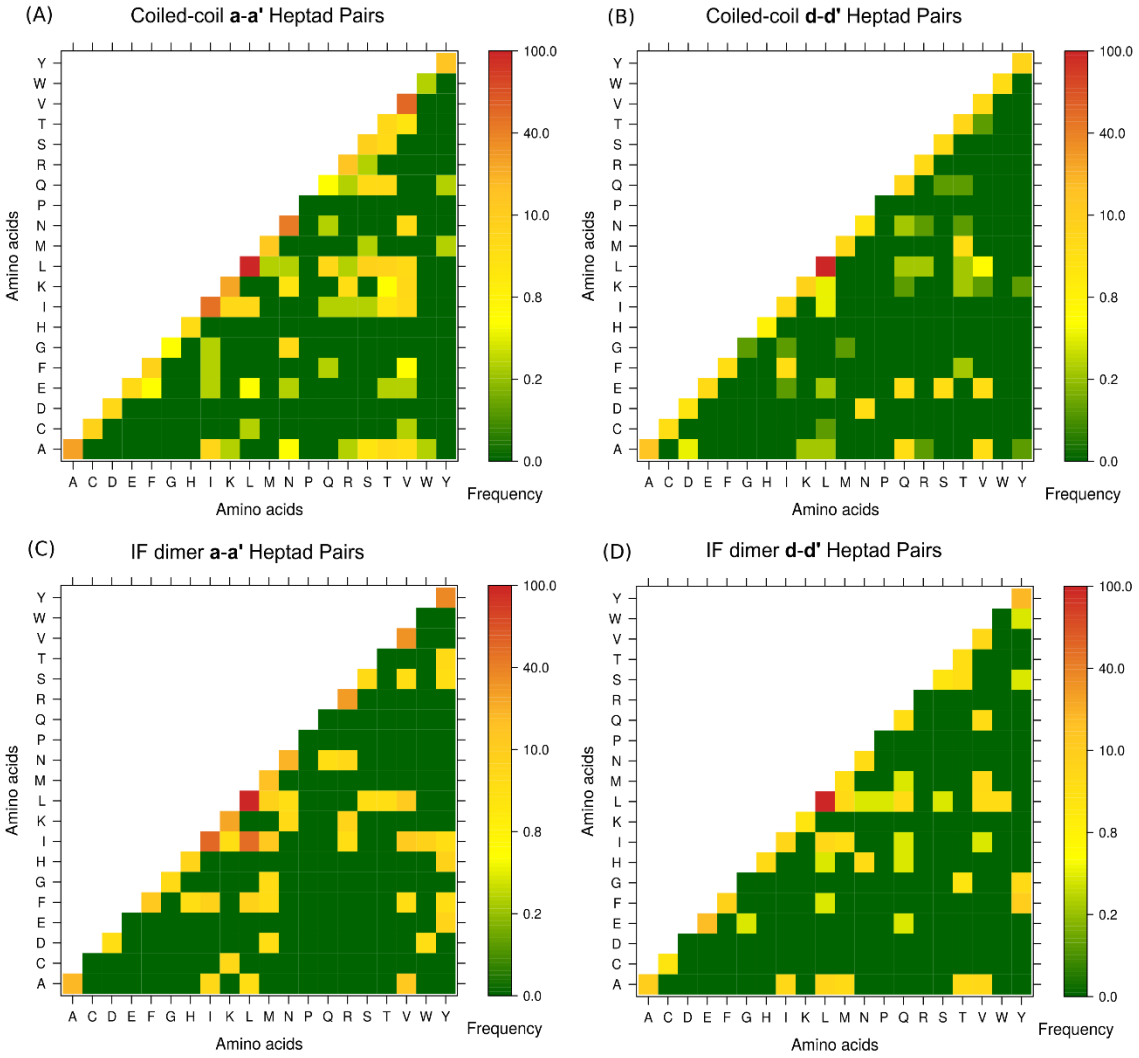


Figure 3-3 Pairwise Residue frequency of Coiled-coil and IF dimers.

*Frequency of pairwise occurrence of amino acids at heptad positions is represented as a heatmap for coiled-coil in (A and B) and for IF in (C and D). Pairwise occurrence in the coiled-coil dimers at heptad positions of (A) **a-a'** and (B) **d-d'** and in IF dimers at heptad positions of (C) **a-a'** and (D) **d-d'**. Color bar represents normalized frequency 0 (green) to 100 (red) of amino acid pairs. Amino acid residues are represented on X and Y axis.*

This analysis provides dimerization rules by comparing the naturally occurring intermediate filament dimer with the non-native dimers. To make these rules more general and applicable to all coiled-coil proteins, we have collected the statistics of amino acid pairs of all coiled-coil and intermediate filament proteins separately. Heptad position **d-d'** shows less variability as compared to heptad positions **a-a'** in both coiled-coils and intermediate filaments (Figure 3-3). At heptad position **d-d**, Leu-Leu amino acid pairs were the most predominant pair in both coiled coil and intermediate filaments. At heptad position **a-a'**, Leu-Leu, Ile-Ile, Val-Val and Asn-Asn amino acid pairs were the most dominant pair in coiled-coil whereas Leu-Leu, Ile-Ile, Leu-Ile amino acid pairs were dominant in intermediate filaments. In general, the distribution of amino acids in both Intermediate filament dimers and coiled-coil dimers are similar. Hence, the obtained rules apply to coiled-coil dimers as well.

3.3.2 Inter-helical hydrogen bonds, electrostatic repulsions, and attractions.

Among the interactions formed by coiled-coils, inter-helical hydrogen bonds and inter-helical repulsions determine the helix pairing specificity [83,121]. However, there is no knowledge about the number of interactions required to quantitate whether a dimer stable or unstable. To determine the significance of inter-helical interactions, native or naturally occurring IF dimers were compared with the non-native or artificial IF dimers (Table 3-1). Thus, the non-native dimers acts as a negative control. The comparison between native and non-native IF dimers (Table 3-1) involve the following interactions:

1. Inter-helical ASN-ASN hydrogen bonds at heptad position **a** and **a'**.
2. Inter-helical electrostatic repulsion between **a-a'**, **d-d'**, **a-g'** and **d-e'** heptad pairs. Similarly charged residues could be tolerated at **e-g'** and **e'-g** pairs [110].
3. Inter helical electrostatic attraction between **a-a'**, **d-d'**, **a-g'**, **e-g'**, **g-e'** and **d-e'** heptad pairs.

Pairwise interactions between amino acids were assumed if they were spatially proximal to one another. These interactions provide a measure of the stability and specificity of the IF dimers. The total number of all possible destabilizing (electrostatic repulsions) and stabilizing (electrostatic attraction, hydrogen bonds) inter-helical interactions were computed for all native and non-native IF dimers (Table 3-4). The non-native dimers on average are expected to have more destabilizing interactions when compared to the native IF dimers.

Table 3-4 Inter-helical polar interactions in native and non-native dimers.

Average number of possible electrostatic repulsions, attractions and hydrogen bonds in native and non-native IF dimers. These numbers accounts for possible interactions between a-a', d-d', a-d', a-g' and d-e' heptad pairs.

	Native IF dimer		Non-native IF dimer		
	Type I-II Hetero-dimer (co-expressed)	Type III-VI Homo-dimer	Type I Homo-dimer	Type II Homo-dimer	Type I-II Hetero-dimer (non co-expressed)
Repulsions	9.85	13.28	9.42	13.42	10
Attractions	4	6.85	4.28	4.28	4
Hydrogen bonds	1	1.85	2.14	1.14	1

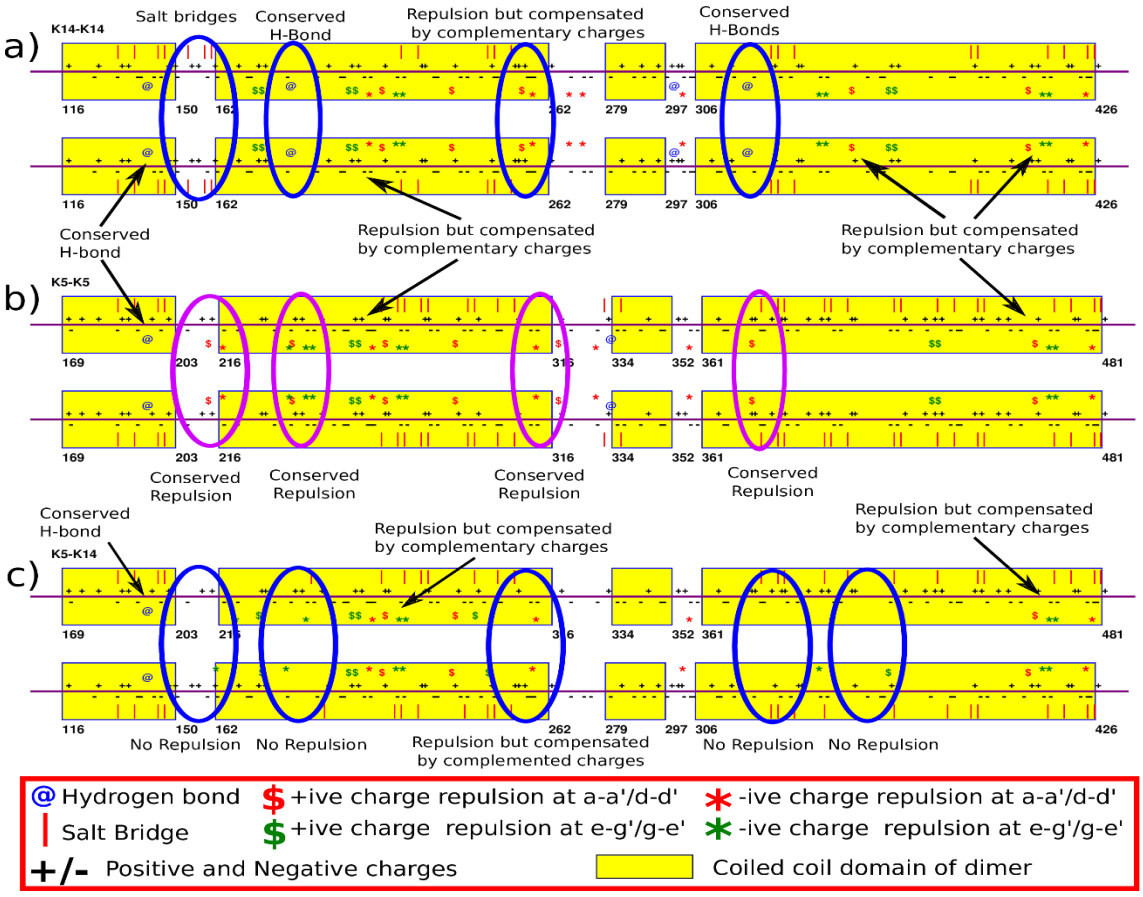


Figure 3-4 Visualization of critical heptad Interactions within IF dimers.

(A) Type I-I homo-dimer (K14-K14), (B) Type II-II homo-dimer (K5-K5), (C) Type I-II heterodimer (K5-K14). Purple encircled regions show conserved repulsion sites whereas blue encircled regions highlight sites without repulsion in comparison to K5-K5 homo-dimer.

A higher number of favorable electrostatic attractions and hydrogen bonds complemented the electrostatic repulsions of native Type III-VI homo-dimers. However, in non-native Type II homo-dimer, the average electrostatic repulsions were more in comparison to all other dimers. Thus, Type II homo-dimers are least stable compared to all other dimers (Figure 3-4). Native Type I-II heterodimers show similar trends for repulsions and attractions as compared to non-native Type I-II heterodimers and non-native Type I homodimers. Thus, Inter-helical interactions (hydrogen bonds and electrostatic repulsions) at **a-a'** and **d-d'** heptad pairs provides pairing specificity through destabilization between different IF pairs as also previously observed [121]. A single conserved hydrogen bond in native IF dimers provides orientation to the coiled-coil dimer (parallel or anti-parallel) at the cost of destabilizing the hydrophobic core. Additional hydrogen bonds may destabilize the coiled-coil dimer due to the loss of hydrophobic core.

3.3.3 Hydrophobic Core Atomic Density

The coiled-coil forming helices interact through their spatially proximal amino acids at heptad positions **a**, **d**, **e,** and **g**. These interactions play a crucial role in stabilizing and the specific pairing of the IF dimers [83,97,100,104,121,177]. Heptad pairs **a-a'** and **d-d'** interactions constitute the hydrophobic core of the dimer. The apostrophe indicates the heptad position of the partner helix. The amino acids at heptad pairs **a-a'** forms the periphery of the hydrophobic core whereas amino acids at the heptad pairs **d-d'** occupy a more central location (Figure 2-1 B). This is sometimes also referred to as knob into hole packing [20]. The following section discusses packing of atoms in two ways. 1) Pairwise interactions (**a-a'** and **d-d'**) and 2) Extended pairwise interaction (**ada** and **dad**).

1. Atomic Density Distribution at a-a' and d-d' pairs

To determine the favorable packing of atoms at heptad pairs **a-a'** and **d-d'**, the optimal number of atoms that can be packed for each position were computed for the entire dimer length. Every amino acid was assigned a score computed as the total number of non-hydrogen atoms at β and γ positions (Table 3-5). Amino acids having the following atoms were given extra scores as indicated; C^{γ} sp^2 hybridized carbon = +1, S^{γ} = +1.5, S^{δ} = +0.5, C^{δ} sp^3 hybridized carbon = +0.25. These extra scores accounted for the bulkiness/size (such as an S atom being bulkier than C) and degrees of freedom (for sp^2 hybridization, planer molecule). All other atoms at C^{δ} and beyond were ignored, as they do not contribute to the hydrophobic core of the coiled-coil dimer.

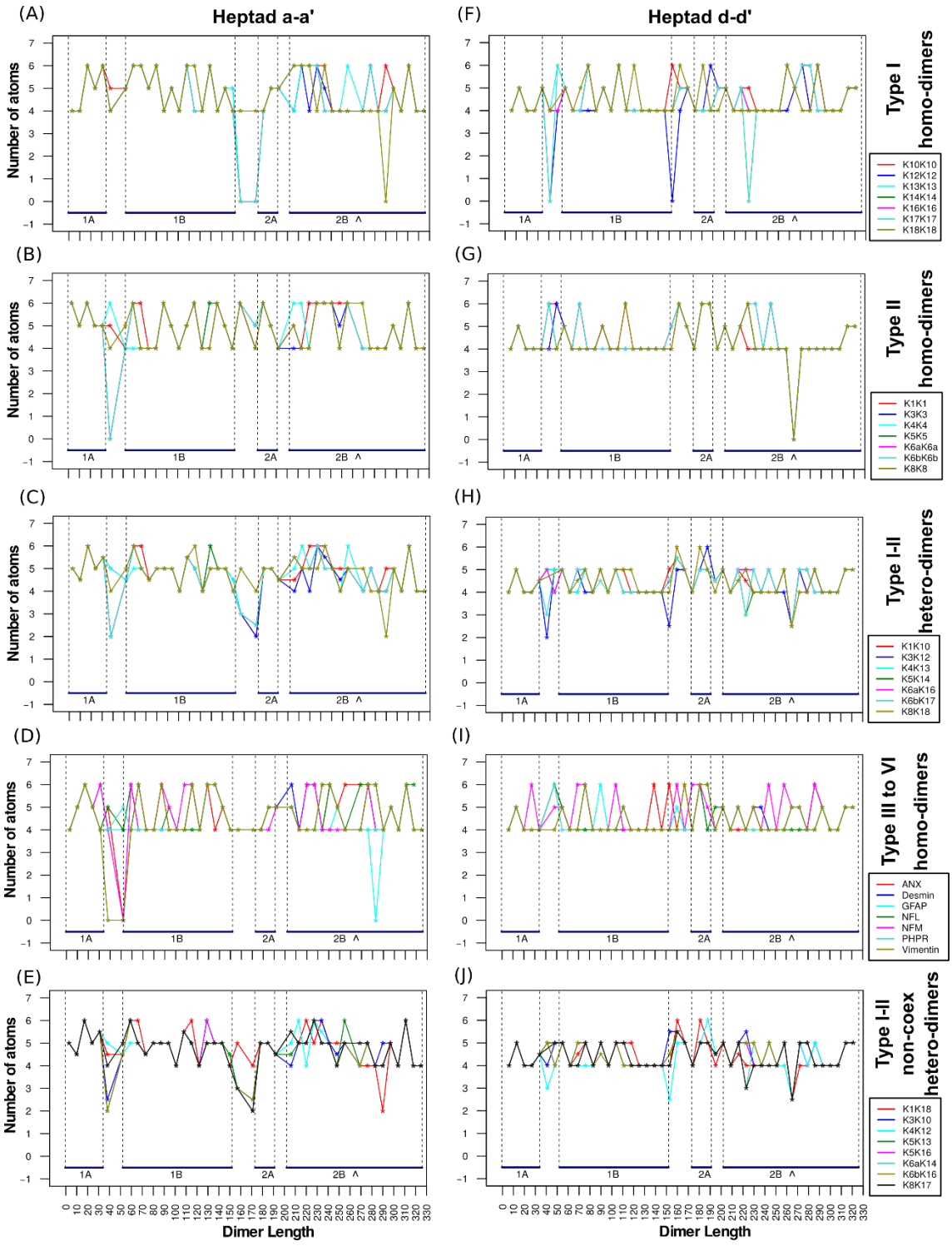


Figure 3-5 Density scores of a-a' and d-d' residue pairs.

(A, F) Type I homodimers, (B, G) Type II homodimers, (C, H) Type I-II heterodimers, (D, I) Type III-VI homodimers and (E, J) Type I-II heterodimers non co-expressed. Density scores of ***a-a'*** *and* ***d-d'*** *heptad pairs calculated from* C^{β} *and* C^{γ} *atoms of amino acid. Color indicate different dimers. The blue horizontal line represents the coiled-coil domain of the dimer. Caret (^) symbol shows the location of the stutter region.*

Table 3-5 Molecular determinants score values for amino acid.

	Hydrophobicity	Branching of Amino acids	Charged Score	Stacking	Volume (in $Å^3$)	Proline Absence
A	1.8	1	0	0	88.6	1
C	2.5	3.5	0	0	108.5	1
D	-3.5	3	-2.666	0	111.1	1
E	-3.5	2	-2	0	138.4	1
F	2.8	3	0	1	189.9	1
G	-0.4	0	0	0	60.1	1
H	-3.2	3	2	0	153.2	1
I	4.5	3.25	0	0	166.7	1
K	-3.9	2	1.6	0	168.6	1
L	3.8	2.5	0	0	166.7	1
M	1.9	2.5	0	0	162.9	1
N	-3.5	3	0	0	114.1	1
P	-1.6	4	0	0	112.7	0
Q	-3.5	2	0	0	143.8	1
R	-4.5	2	1.32	0	173.4	1
S	-0.8	2	0	0	89	1
T	-0.7	3	0	0	116.1	1
V	4.2	3	0	0	140	1
W	-0.9	3	0	1	227.8	1
Y	-1.3	3	0	1	193.6	1

Total score of every **a-a'** or **d-d'** amino acid pair was determined by adding the individual scores of amino acids at these positions. The total scores of all **a-a'** and **d-d'** heptad positions were calculated and compared with the scores of the non-native dimer sets. The non-native dimers were expected to have large variability in the packing (less stable). Several studies have determined the amino acid preferences across various heptad positions in coiled-coil dimers [100,118,177]. However, no study has reported pairwise preference in the coiled-coil dimers. In an ideal coiled-coil dimer, heptad pairs **a-a'** and **d-d'** should complement each other resulting in an evenly distributed atomic density along the length of coiled coil. An uneven density (low/high) in coiled-coils either

destabilizes dimer or favors the formation of trimer and higher order structures [104,178,179]. However, intermediate filaments and other coiled-coils often have high and low-density regions. This is possible only through re-distribution of the atomic density in the adjacent heptad pairs so that the overall distribution remains uniform. Beta-branching amino acids at heptad position **d** increases the atomic density in the hydrophobic core that should be complemented by the low atomic density partner at heptad **d'** and sandwiching heptad **a-a'** pairs (Figure 3-5). The re-distribution of atomic density was accomplished by locally changing the pitch of the coiled-coil dimer. The change in pitch creates enough space to accommodate extra atoms and thereby maintaining the uniform density along the coiled-coil.

2. Atomic Density Distribution at ada and dad hex pairs

Atomic density distributions of heptad pairs **a-a'** and **d-d'** were extended by appending the preceding and succeeding **a-a'** and **d-d'** pairs density. This accounts for cases where a high-density heptad pair is surrounded by low-density heptad pairs or vice versa. This arrangement of atoms is represented here as **ada** and **dad** hex pairs. A **dad** hex pair is a set of twelve amino acids at heptad position **a-a',** sandwiched by **d-d'** and **e-e'/g-g'** pairs. Similarly, **ada** hex pair is a set of twelve amino acids at heptad position **d-d',** sandwiched by **a-a'** and **e-e'/g-g'** pairs. For instance, a **dad** pair consists of heptad positions **1d**, **1d'**, **2a**, **2a'**, **3d**, **3d'**, **1e**, **1e'**, **2g**, **2g'**, **3e** and **3e'** where the number before heptad signifies the position (from start to end) along the length of a coiled-coil dimer. Atomic distribution scores for **ada** and **dad** hex pairs were calculated as follows.

$$Score(\boldsymbol{ada_3}) = \frac{Score(\boldsymbol{a_2} - \boldsymbol{a'}_2) + Score(\boldsymbol{a_4} - \boldsymbol{a'}_4) - 2 * Score(\boldsymbol{d_3} - \boldsymbol{d'}_3)}{2}$$

$$Score(\boldsymbol{dad_3}) = \frac{Score(\boldsymbol{d_2} - \boldsymbol{d}_2) + Score(\boldsymbol{d_4} - \boldsymbol{d'}_4) - 2 * Score(\boldsymbol{a_3} - \boldsymbol{a'}_3)}{2}$$

Score of every **a-a'** and **d-d'** pair was calculated as explained in the previous section (See Section 3.3.3-1 for more details). The subscript indicates the position of **ada** or **dad** hex pairs.

A low-density region at the heptad pair **a-a'** allows invasion from solvent molecules. The solvent molecules could easily penetrate the hydrophobic core, thus destabilizing the coiled-coil dimer. A stable coiled-coil should have complemented high-density heptad pairs around low-density heptad pairs (Figure 3-6). In such cases, the neighboring ionic interactions protect the exposed low-density hydrophobic cores through hydrophobic methylene groups of the participating LYS and ARG amino acids.

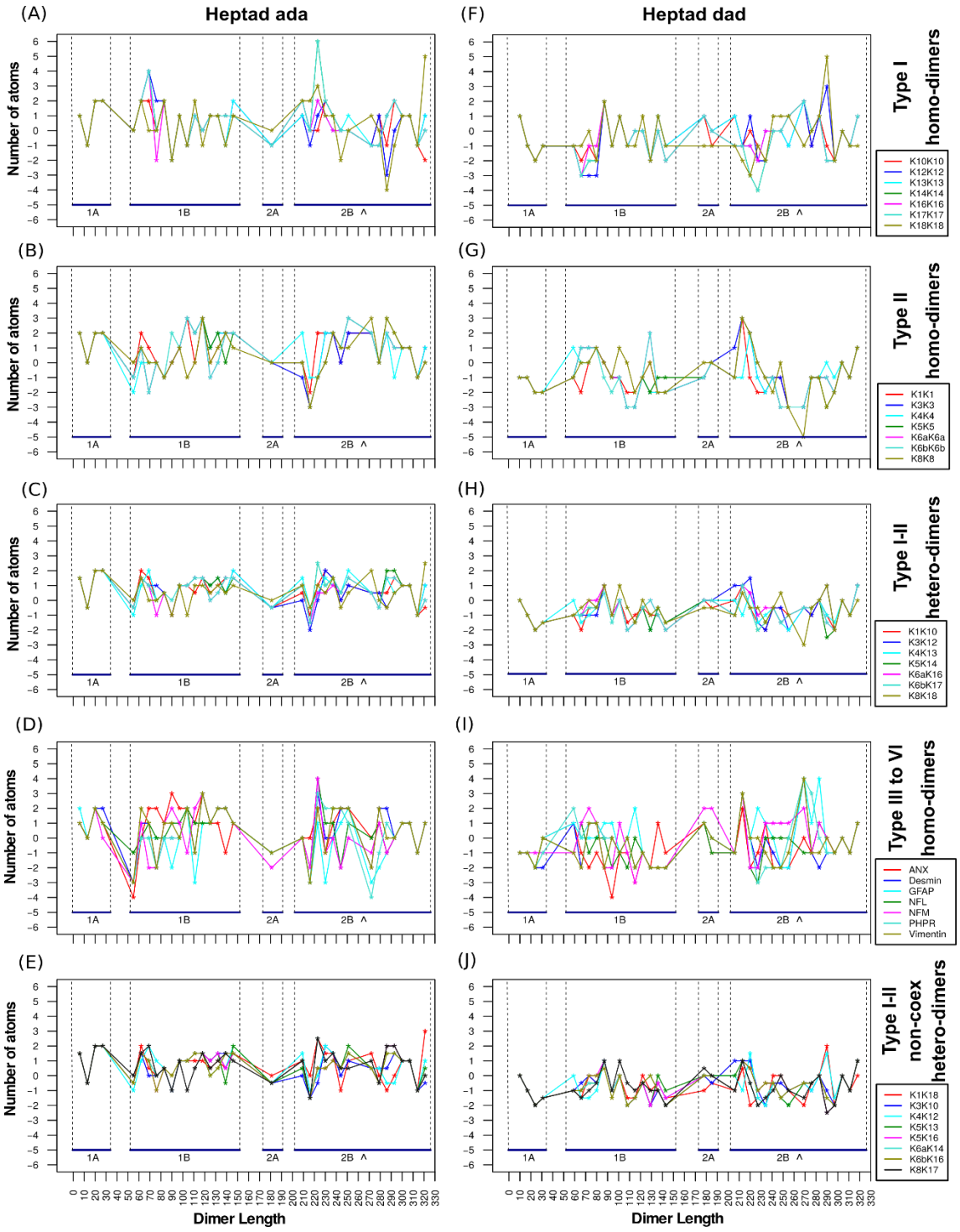


Figure 3-6 Density scores of ada and dad residue pairs.

(A, F) Type I homodimers, (B, G) Type II homodimers, (C, H) Type I-II heterodimers, (D, I) Type III-VI homodimers and (E, J) Type I-II heterodimers non co-expressed. Density scores of ***ada*** *and* ***dad*** *heptad pairs calculated from C^{β} and C^{γ} atoms of amino acid. Color indicate different dimers. The blue horizontal line represents the coiled-coil domain of the dimer. Caret (^) symbol shows the location of the stutter region.*

The hydrophobic core atomic density scores suggests that a uniform density is stable whereas the non-uniform density is unstable. For instance, a substitution that decreases the atomic density of hydrophobic core could destabilize the dimer due to invasion of water molecules [180]. The invasion that fills the hydrophobic core with water molecules were seen in several coiled-coil crystal structures (PDBID: 1ZIM, 4DME, 1IJ2, 1MG1, 1Y4M, 1MOF, 2EBO). Similarly, a substitution that increases the atomic density of the hydrophobic core would favor the formation of trimer or higher-order oligomer structures but not dimer [181].

3.3.4 Volume distribution of Amino acids at a-a' and d-d' positions

The total volume of amino acids at **a-a'** and **d-d'** heptad pairs were calculated for all dimers in the "IF Data set" (Table 3-1) to account for the bulkiness of the amino acids. The total volume of each amino acid pair was calculated by adding the individual volume of amino acids [182]. The volume of the amino acids also plays a significant role in the formation of the dimer. Only amino acid pairs with appropriate volume can form a well-packed structure. Large protrusion of amino acids from the hydrophobic core could be destabilizing especially in the context of the filament. In this section, the volume preferences of the native IF dimers were compared with non-native IF dimers. Non-native IF homo-dimers of Type I and Type II have many irregular volume pairs at heptad position **a-a'** and **d-d'** (Figure 3-7). However, native IF hetero-dimers of Type I-II have a narrow band of allowed pair volume suggesting the importance of uniformity in the volume of hydrophobic core.

Apart from that, a lower hydrophobic core volume in native IF dimers was always compensated by a hydrophobic methylene groups of LYS/ARG amino acids while forming ionic interactions. Non-native IF homo-dimers of Type I and Type II do not show these trends. Native IF homo-dimers of Type III-VI, however, show few fluctuations but were always protected by methylene groups of salt bridge forming amino acids. Type I-II non-native heterodimers also have a narrow band suggesting their capability to form stable heterodimers. The volume analysis suggests that the volume of pairing amino acid is an essential factor in dimerization.

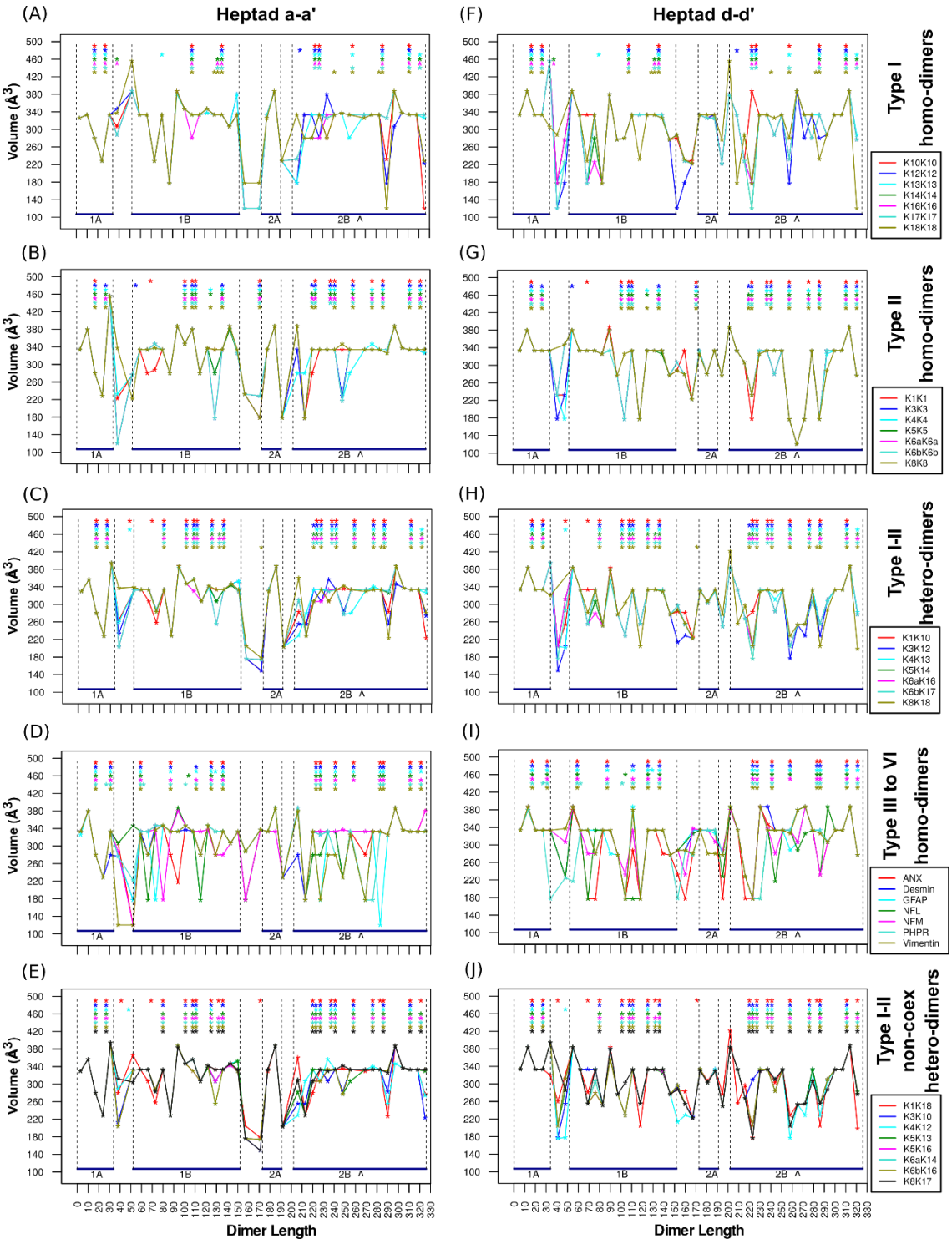


Figure 3-7 Volume of a-a' and d-d' residue pairs.

*(A, F) Type I homodimers, (B, G) Type II homodimers, (C, H) Type I-II heterodimers, (D, I) Type III-VI homodimers and (E, J) Type I-II heterodimers non co-expressed. Color indicate different dimers. The blue horizontal line represents the coiled-coil domain of the dimer. The asterisk (*) symbol provides the position of salt bridge formed by the neighboring **e-g'** pair. Caret (^) symbol shows the location of the stutter region.*

3.3.5 Distribution of Hydrophobic and Hydrophilic Amino acids

To understand the distribution of hydrophobic amino acids across various heptad repeats, amino acids were categorized into hydrophobic and non-hydrophobic residues. Amino acids Ala, Ile, Leu, Met, Phe, Trp, Tyr, and Val were considered hydrophobic, while other 12 amino acids were considered non-hydrophobic. For all native and non-native dimers, the percentage of hydrophobic and non-hydrophobic amino acids were recorded for all heptad pairs.

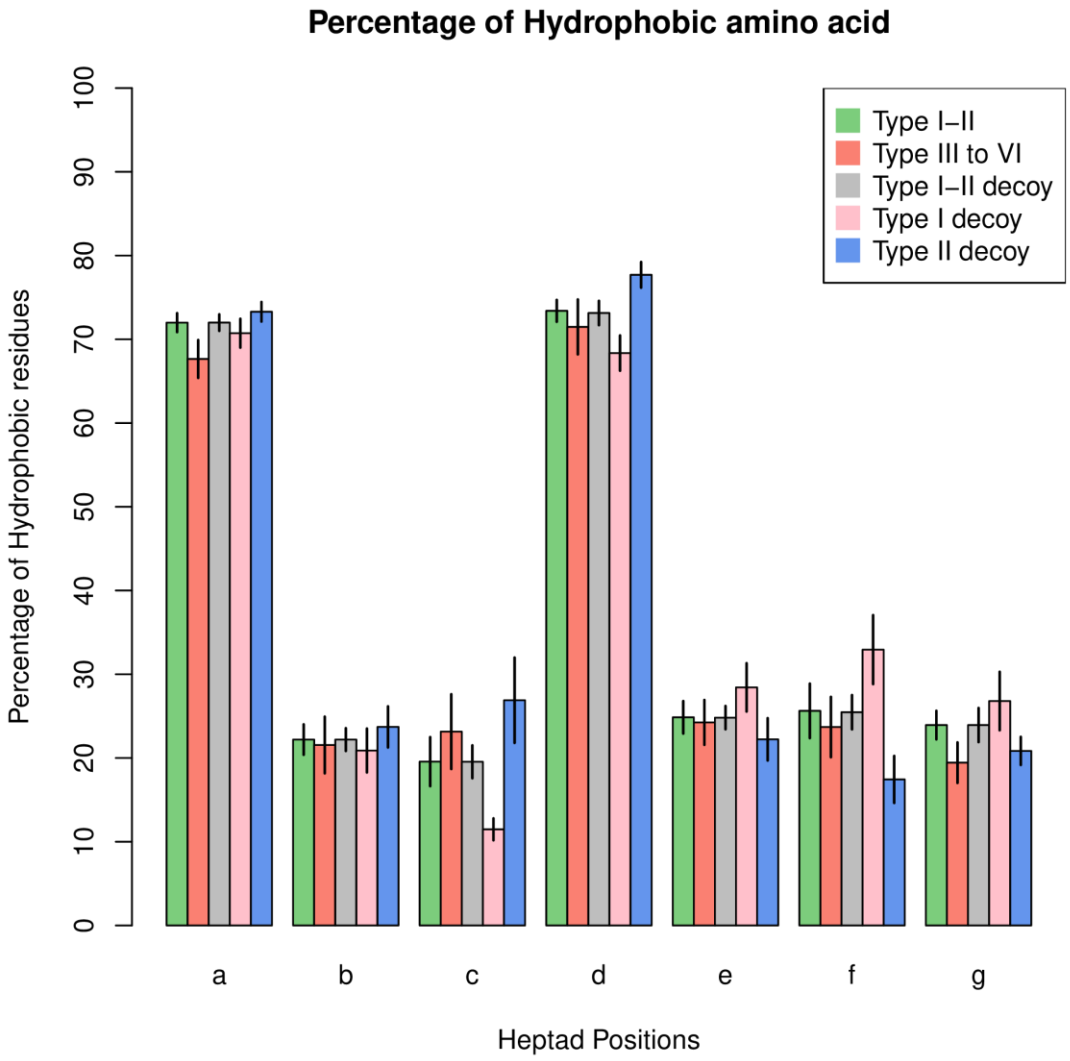


Figure 3-8 Distribution of hydrophobic amino acids across heptad positions.

The percentage of hydrophobic amino acids at various heptad positions along the x axis. Error bar indicates standard deviation. The bars are grouped according to different types of intermediate filament dimers i.e. native and non-native dimers (decoys).

Coiled-coil dimers in heptad positions **a** and **d** are known to have hydrophobic amino acids whereas polar amino acids occupy heptad positions **b**, **c,** and **f**. However, heptad positions **c** and **f** in non-native dimers deviate from the standard distribution when compared against native dimers (Figure 3-8). The hydrophobic residue percentage for Type I non-native homodimers are ~12 and ~33, for Type II non-native homodimers are ~28 and ~17 whereas for naturally occurring dimers the percentage is ~20 and ~25. The difference is about 8% (13 amino acids out of 157) for heptad position **c** and **f** when compared with naturally occurring dimers. Similarly, the difference for heptad position **d** is about 5% (4 amino acids out of 79) when compared to the native dimers.

Previously, Intermediate filaments were shown to have hydrophobic patches on the dimer surface (heptad positions **b**, **c,** and **f**) that stabilize the dimer-dimer contact [159]. The differences shown here are significant enough to cause destabilization because of unbalanced hydrophobic amino acid patches along the length of non-native homodimers. The analysis provides critical positions for dimer-dimer contact as well as the fact that the deviations in the distribution of hydrophobic amino acids play a crucial role in the intermediate filament stabilization.

3.3.6 Heuristic Features

Other than the molecular determinants obtained from the sequence analysis of native and non-native IF dimers, it has been observed that the presence of Proline amino acid induces a kink in the coiled-coil helix [183]. The kink in the helix leads to the loss in intra-helical hydrogen bonds, thereby affecting the coiled-coil stability. Thus, the presence of Proline amino acid in the coiled-coil dimer was used as another molecular determinant. The stability of coiled-coil dimers also depends on other factors such as the neighboring amino acids, stacking interaction of aromatic residues etc. Disrupting such interactions could cause the functional breakdown of the intermediate filaments.

3.3.7 Summary of structural features

The previous sections discussed the methodology adopted for determining the structural features obtained for the Intermediate filament dimers. The features were obtained by comparing the molecular determinants of the native IF dimers with the non-native IF dimers. In all six features provided maximum contrast between native and non-native IF dimers. Table 3-6 highlights a brief summary of structural features, their physical significance along with the calculation methodology. These structural features in their numerical form (Table 3-5) were used for classifying the hex pair stability. The next

section will discuss the construction of neural network schema that would use these structural features to classify the hex pair stabilities.

Table 3-6 Physical significance of features and their calculation methodology.

Table list the features and their physical significance that was used for predicting the coiled-coil stabilities and the calculation methodology to obtain the corresponding numerical values.

Feature	Physical significance	Calculation method
Hydrophobicity of residues	Heptad positions **a** and **d** are generally hydrophobic.	Hydrophobic index
Branching of residues	Heptad positions **a** and **d** have different preference of branching residues.	Total number of atoms at C^{β} and C^{γ} positions.
Position of charged atom	This rule provides a measure of how far is the charge from the hydrophobic core.	Charge on amino acid divided by number of atoms between the charged atom and C^{α} atom in amino acids.
Stacking of residue	Residues at heptad positions **a** and **d** appear to have stacking interaction with the partner residues	Stacking interaction is assumed if aromatic rings were present at heptad position **a** and **d'** and vice versa.
Volume of residues	This rule provides a measure of bulkiness of the residues	Total volume of amino acid pair.
Addition of Proline	Proline is well known to kink the alpha helix. Disease causing mutations also shows large number of Proline mutants.	Proline substitution has higher value.

3.4 Construction of Neural Network Schemas

3.4.1 Obtaining training and testing data

When single point mutants cause disease, it was conjectured that this is because of the destabilization of the IF filament and/or IF dimer. In this study, a multi-layer perceptron (MLP) network was used to characterize and predict the effects of point mutations on the stability of the IF dimer. The multi-layer perceptron network learns the weights of the network from the input hex pairs (see Section 3.3.3-2) obtained from naturally occurring dimers and their corresponding disease causing variants.

3.4.2 Representing and scoring hex pairs

A hex pair (Figure 3-9 A) was represented as a column vector of twelve amino acids ($V_{12\ X\ 1}$) (Figure 3-9 B). Each amino acid in the vector $V_{12\ X\ 1}$ was mapped to six scores given by its molecular determinants (Table 3-5), thereby generating a score matrix ($M_{12\ X\ 6}$) (Figure 3-9 C). All elements of the score matrix (in row major order) were normalized across features that feeds in to the input layer of the neural network schema (Figure 3-9) (Figure 3-9 D). The hidden layer of the schema has logistic activation function. For every hex pair the output layer provides a probability value in range 0 to 1 (0 stable hex pair, 1.0 unstable hex pair).

3.4.3 Bootstrapping and Performance

The molecular determinants numerical values for heptad pairs **a-a'** and **d-d'** were different for native dimers when compared to non-native dimers (Figure 3-5, Figure 3-6, Figure 3-7). Owing to this difference, two independent neural network schemas were constructed. One of the schema was trained on **ada** hex pairs whereas other was trained on **dad** hex pairs. To ensure that there was no over-fitting, 50 bootstrap iterations were performed for both schemas by dividing their corresponding input into training and testing set in ratio 0.7 and 0.3 respectively. Training and testing data are of size 186 and 82 for **dad** pairs schema whereas 196 and 85 for **ada** pairs schema respectively.

The method had an average accuracy of about 0.944 and 0.938 for heptad position **dad** and **ada** respectively (Table 3-7). The specificity of the method is about 0.982 and 0.974 for heptad position **dad** and **ada** respectively (Table 3-7). The method is nearly perfect in determining non-disease causing substitutions. Overall the trained model has Mathew correlation coefficient (MCC) 0.757 and 0.722 for heptad position **dad** and **ada** respectively.

3.4.4 Neural Network Implementation parameters

Neural network schemas (multi-layer perceptron) were developed using scikit-learn library with the following parameters; optimization algorithm = 'l-bfgs', regularization rate = 0.05, number of neurons in hidden layer = 10, activation function of each neuron = 'logistic', learning rate = 0.3, tolerance for optimization = 0.000001, number of iterations = 50000 [184].

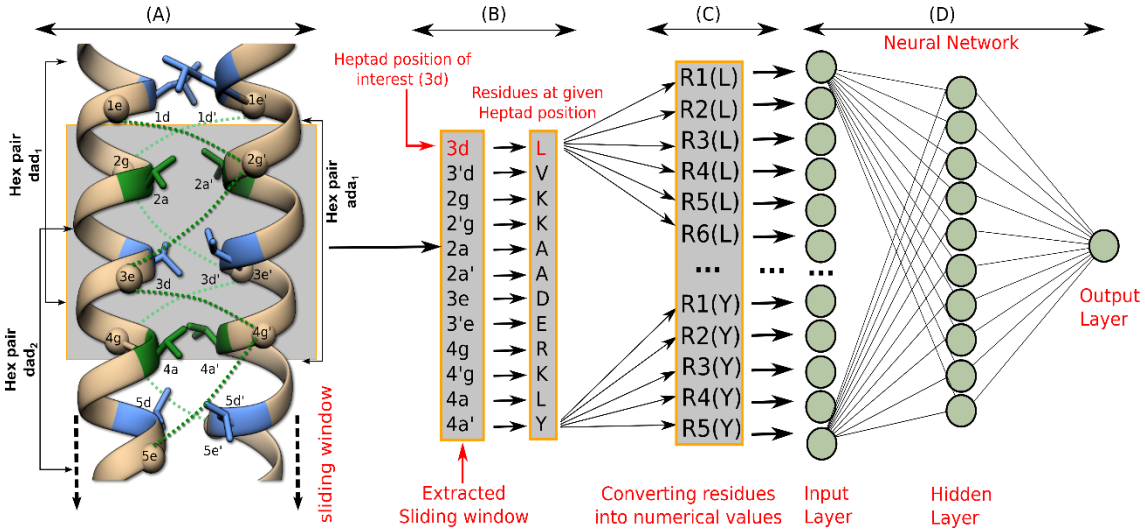


Figure 3-9 Schematic showing conversion of a hex pair to neural network input.

*(A) A hex pair (**ada** and **dad**) extracted from a coiled-coil protein. A sliding window (gray box) around the helix shows an ada hex pair. (B) Extracted heptad positions and corresponding amino acids. (C) Conversion of amino acids to numerical values according to six molecular determinants represented in the figure as R1 to R6 (see* Table 3-5 *for details about molecular determinants). (D) Neural network with input layer, hidden layer and output layer. All input layer nodes were connected to all other hidden layer nodes.*

Table 3-7 Performance of trained neural network model.

Rows indicate heptad position whereas Column indicate various statistical measures such as Accuracy, Sensitivity, Specificity, and Mathew correlation coefficient (MCC).

	Accuracy	Sensitivity	Specificity	Precision	MCC
Schema **dad**	0.944	0.711	0.982	0.877	0.756
Schema **ada**	0.938	0.706	0.974	0.816	0.722

3.5 Using Neural Network Schema for classifying hex pairs

Once the neural networks schemas were trained to distinguish between stable and unstable (disease causing) **dad** and **ada** hex pairs, they could be used to predict the effects of substitutions at heptad positions **a** and **d** respectively. Following sections describes the usage of trained models for classifying hex pairs as stable or unstable.

3.5.1 Prediction of Novel filament disruptive substitutions in all IF dimers

The trained neural network models were utilized to predict new filament destabilizing substitutions for all Intermediate filament family menbers. Every heptad position **a** and **d** in the intermediate filament dimers was substituted to the other 19 amino acids. The stability of the hex pairs generated through *in-silico* saturated mutagenesis was classified using trained neural network schemas (see Section 3.4 for details). The trained models provide a probability value (representative of stability) for every hex pairs. A hex pair with a probability value greater than 0.9 was considered unstable whereas probability less than 0.1 was considered stable. Hex pairs with probability values in range 0.1 to 0.9 cannot be classified as stable/unstable with confidence, hence were marked as "Not Confident". Table 3-8 list the total number of filament disrupting substitutions and control mutations (little or no effect on the stability) for all intermediate filament family. No restrictions were applied (such as hydrophobic to polar substitutions) on the type of filament disrupting mutations. The control mutations for all Intermediate filaments were low in number compared to filament disrupting mutations. This suggests that the heptad positions **a/d** provide little space for variability and hence explains the high conservation of Intermediate filaments coiled-coil domain sequences.

3.5.2 Experimental validation of novel K5-K14 mutations

To validate the *in-silico* predictions of filament disrupting substitutions, keratin K5-K14 was selected as a model system. The neural network models provided around 635 filament-disrupting substitutions along the length of dimer (Table 3-8). Majority of these predicted mutants were hydrophobic to polar/charged amino acids. These extreme substitutions were likely to be filament destabilizing and hence do not need experimental validation. However, all hydrophobic to hydrophobic predicted mutants have similar chemical nature and hence, are not trivially validated. Thus, all hydrophobic to hydrophobic residue mutants along with their predicted controls were experimentally validated (Table 3-9). GFP-tagged K14 plasmids for all mutants were generated through site directed mutagenesis. These plasmids were co-transfected with DsRed-tagged K5

wild-type plasmid into PtK2 cells [185]. The cells were then analyzed for keratin aggregates using light microscopy. Sixteen different substitutions at six positions were validated and agree with the computational predictions. For every position, the filament disrupting substitutions were predicted along with a possible control substitution. Most of these substitutions (Table 3-9) disrupt the filament without the need of hypo-osmotic stress. Two representative images of L126 position substitutions (control and filament disrupting) show the Keratin K5-K14 aggregates (Figure 3-10). However, mutant I169G was filament destabilizing only when the cell was subjected to hypo-osmotic stress with 150mM Urea followed by 1 hour recovery (Figure 3-11).

Table 3-8 Filament disrupting substitutions for all Intermediate Filaments

Column 1 lists native IFs, Column 2 lists number of filament disrupting mutations predicted in the coiled-coil domain through neural network schema. Column 3 list number of control substitutions predicted for each type of Intermediate filament. Control substitutions would have little or no effect on the stability of the Intermediate filaments.

Intermediate Filament	Filament disrupting substitutions	Control substitutions (no effect)
K1-K10	619	57
K3-K12	628	63
K4-K13	641	64
K5-K14	635	67
K6a-K16	639	63
K6b-K17	644	62
K8-K18	616	53
Desmin	210	34
GFAP	197	31
NFL	198	30
NFM	207	32
Peripherin	196	29
Vimentin	211	30

Table 3-9 List of all in-vivo validated list of substitutions.

*Column 1 provide six different position along the length of K5-K14 Intermediate filament dimer coiled-coil domain. Column 2 provide the predicted filament disruptive substitution details with wild type amino acid with position and corresponding substitution. Column 3 and 4 provide average probability predicted from 50 bootstrap iterations and the corresponding predicted effect (Intact and Disruption). Column 5 provides effects observed in the in-vivo experiments. * indicates filament aggregates observed in some cells.*

Position	Substitution	Average Probability	Predicted Effect	Experimentally Observed Effect
L126	L126F	0.01	Intact	Intact
	L126I	0.92	Disruption	Disruption
	L126A	1.00	Disruption	Disruption
	L126G	1.00	Disruption	Disruption
Y162	Y162I	0.17	Intact	Intact
	Y162M	0.96	Disruption	Disruption
I169	I169V	0.00	Intact	Intact
	I169G	0.94	Disruption	Disruption on stress
V186	V186F	0.28	Intact	Not confident*
	V186M	0.98	Disruption	Disruptive
I239	I239V	0.00	Intact	Intact
	I239M	0.92	Disruption	Not confident*
	I239G	1.00	Disruption	Disruptive
I412	I412W	0.99	Disruption	Disruption
	I412M	1.00	Disruption	Intact
	I412L	1.00	Disruption	Intact

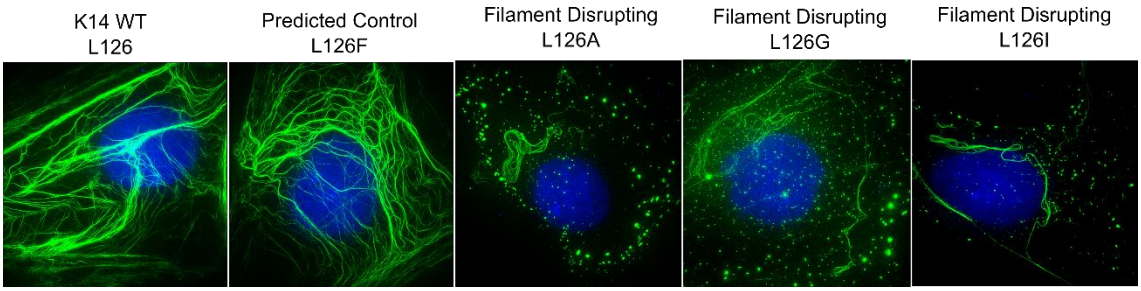


Figure 3-10 Experimental validations of K5-K14 novel K14-126 mutant.

Representative images of K5-K14 keratin filaments and aggregates in WT and mutants. (First column:) Keratin K14 filaments in the wild type Ptk2 cells. (Second column:) Keratin K14 filaments in the predicted control mutant and (third to fifth columns:) Keratin K14 aggregates in the predicted filament disrupting mutants. Mutants were co-transfected with DsRed-tagged K5, and GFP-tagged K14 in K5, K14 null PtK2 cells and residues substituted at the position K14-126 for L126A, L126G, L126I. (Blue: DAPI labelled nucleus, Green: K14 Keratin). Experiments: Dr. Birgit Lane's group, IMB Singapore.

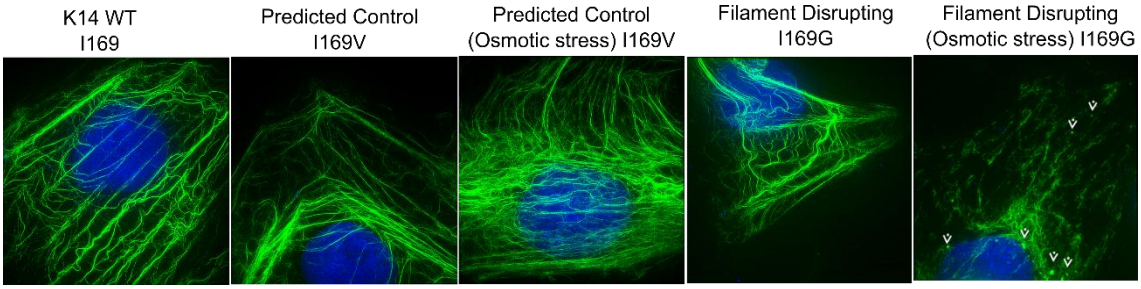


Figure 3-11 Experimental validations of K5-K14 novel K14-169 mutant.

Representative images of K5-K14 keratin filaments and aggregates in WT and mutants. (First column:) Keratin K14 filaments in the wild type Ptk2 cells. (Second column:) Keratin K14 filaments in the predicted control mutant. (Fourth column:) Keratin K14 aggregates in the predicted filament disrupting mutants. (Third and fifth column:) K14 filament mutants subjected to hypo-osmotic stress followed by 1-hour recovery Mutants were co-transfected with DsRed-tagged K5, and GFP-tagged K14 in K5, K14 null PtK2 cells and residues substituted at the position K14-169 for I169V, I169G. (Blue: DAPI labelled nucleus, Green: K14 Keratin). Experiments: Dr. Birgit Lane's group, IMB Singapore.

3.5.3 Determining unstable positions in native and non-native dimers

Molecular determinants (listed here Table 3-6) decide the stability of any given position across the IF coiled-coil dimers. These determinants were used to construct neural network model that scores any hex-pair and classify them as stable or unstable (Figure 3-9 Section 3.4 Section 3.5). Every molecular determinant provided contrasting difference between native and non-native dimers without considering the effects of other molecular determinants. Thus, their combined effect on the dimer stability should independently classify them as native or non-native dimers.

The trained neural network model verifies the stability of all unique hex-pairs from native and non-native dimers. The non-native dimers were expected to have at least one or more filament destabilizing hex-pairs whereas native dimers should not have any unstable hex-pairs. The hex-pairs with probability greater than 0.9 were considered unstable (as previously done in section 3.5.1). All non-native homo-dimers of type II have an unstable hex-pair at a specific position, whereas non-native homo-dimers of type I do not have such unstable hex pairs (Table 3-10). This suggests that type I homo-dimers could form stable dimers whereas type II homo-dimers are unstable. Similarly, no native dimers (type I-II hetero-dimer and type III-VI homo-dimers) had any filament destabilizing hex pairs. An important thing to note here is that non-native "non co-expressed type I-II hetero-dimer" also do not show any filament destabilizing hex pair. This agrees with the previously observed non co-expressed pairing of Intermediate filaments. See section 3.7 for more details on stability analysis of native and non-native dimers. In the hex-pair stability analysis, Asn-Asn hydrogen bonding pairs were excluded as they were functionally important and favor the formation of parallel dimer at the cost of hydrophobic core disruption [102,113,186].

Table 3-10 Native And Non-native Dimers Independent Validation

Independent testing of filament destabilizing substitutions using trained neural network. Column 1 represents the different category of dimer ("IF Data set"), Column 2 represents IF type, Column 3 gives substitution position and probability, Column 4 states presence of filament destabilizing hex-pair (probability cutoff >0.9) and Column 5 provides reason for disruption.

	IF Type	Position & (Probability)	Unstable hex-pair	Reason for disruption

Type I homodimers	K10-K10	None detected	No	Not applicable
	K12-K12	207 (0.89)	No	Not applicable
	K13-K13	32 (0.97), 322 (0.97)	Yes	High density core followed by a repulsive pair E-E
	K14-K14	280 (0.89)	No	Not applicable
	K16-K16	77 (1.0)	Yes	Proline pair followed by a repulsive pair E-E
	K17-K17	Not detected	No	Not applicable
	K18-K18	Not detected	No	Not applicable
Type II homodimers	K1-K1	53 (1.0)	Yes	A Repulsive pair E-E followed by low density
	K3-K3	53 (1.0), 88 (0.9)	Yes	Same as above
	K4-K4	53 (0.99)	Yes	Same as above
	K5-K5	53 (1.0)	Yes	Same as above
	K6a-K6a	53 (1.0)	Yes	Same as above
	K6b-K6b	53 (1.0)	Yes	Same as above
	K8-K8	53 (1.0)	Yes	Same as above
Type I-II hetero-dimer Non co-expressed	K1-K18	Not detected	No	Not applicable
	K3-K10	Not detected	No	Not applicable
	K4-K12	Not detected	No	Not applicable
	K5-K13	Not detected	No	Not applicable
	K6a-K14	Not detected	No	Not applicable
	K6b-K16	Not detected	No	Not applicable
	K8-K17	Not detected	No	Not applicable
Type I-II hetero-dimer	All dimers	Not detected	No	Not applicable
Type III to VI homo-dimer	All dimers	Not detected	No	Not applicable

3.6 Homology modeling of native and non-native dimers

To get insight into the structural properties of intermediate filament dimers, the coiled coil domain structures of native and non-native dimers were constructed using multi-template homology modeling [28].

The first template for modeling was constructed through end to end linking of different segments (1A, 1B, 2A, 2B) of crystal structures of vimentin (PDB ID: 3S4R [159], 3UF1 [160], 3TRT [162], 1GK4 [147]). The linking was done by superimposing the overlapping residues using CLICK [187,188]. The redundant residues were removed after superimposition to make one continuous molecule. The obtained model was not straight owing to large atomic displacements towards the ends of the crystal structures. Analysis of vimentin segments showed that the structures have large B-factor values especially at the ends of the segments (Figure 3-12). The vimentin segments also have large deviations in pitch values. The pitch of the structures were calculated using an in-house script that works by considering C^{α} atoms [189]. The outliers (values outside the range 100 Å to 200 Å) were removed during pitch calculation. The average pitch of the coiled-coil domain was ~140 Å. Thus, along with this template, a second straight template having pitch of about ~140 Å was required for multi-template modeling.

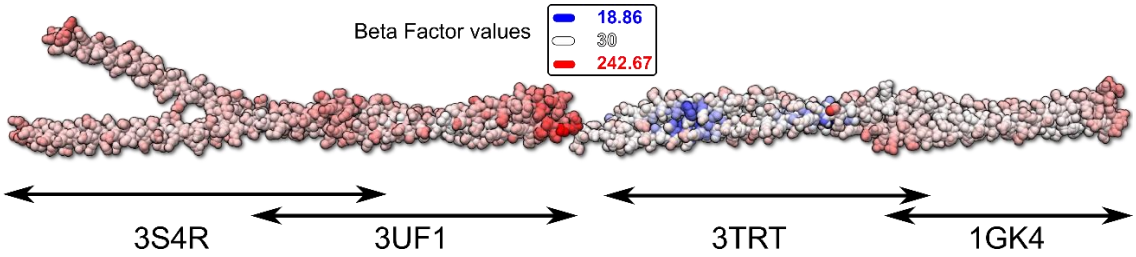


Figure 3-12 Vimentin dimer model with temperature factor.

Vimentin dimer model constructed using crystal structure segments (sphere representation). The spheres were colored according beta factor values ranging from 18.86 (blue) to 30 (white) to 246.67 (red). Coiled coil dimer near linker region L12 shows high beta factor values suggesting uncertainty in the position of coiled coils.

The second template was built using the Leucine zipper (~40 Å; 31 residues in each chain; PDB id: 2ZTA), a coiled coil dimer with a pitch of 143 Å similar to Vimentin. A straight long coiled coil model was constructed by linking together 17 molecules of leucine zippers using the program CLICK [187,190]. Ten residues at C-terminal of one 2ZTA molecule were superimposed with the ten residues at N-terminal of another 2ZTA molecule. The residues found redundant after superimposition were removed to make one

continuous molecule with two chains winding around each other forming an unbroken coiled-coil dimer of 714 residues. The resulting template had a coiled-coil cylindrical geometry (Figure 3-13).

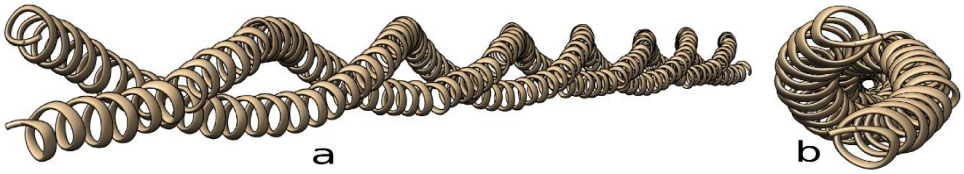


Figure 3-13 Template for homology modeling of IF Dimer.

Ribbon representation of Leucine zipper, a 17-mer molecule (PDB id: 2zta). a) Side view of the template b) Cross-sectional view of the template. The model have two right-handed alpha helices wrapped around each other to form left-handed super coil structure.

This straight long coiled coil dimer was selected as another template for homology modeling. By considering the above two templates, multi-template homology modeling of all native and non-native dimers from "IF Data set" (Table 3-1) was performed using the Modeller suite of programs with default parameters [26,191].

3.6.1 Structural analysis of Energy minimized models

Structural models of native and non-native IF dimers from the "IF Data set" were prepared for energy minimization. The system was placed in the triclinic water box around the dimers using GROMACS software package [192]. The water box contains explicit SPCE water model with the boundary extended by 25 Å around the dimer. The total charge of the system was neutralized by adding counter ions equal to the total charge (Table 3-11). The energy minimization was performed for 5,000,000 steepest gradient steps. Energy minimized models were then used for further structural analysis.

Table 3-11 Number of counter Ions added for every IF dimer

IF dimer	Ions	IF dimer	Ions	IF dimer	Ions	IF dimer	Ions	IF dimer	Ions
K10-K10	48	K1-K1	30	K1-K10	39	K1-K18	30	ANX	32
K12-K12	46	K3-K3	30	K3-K12	38	K3-K10	39	Desmin	38
K13-K13	42	K4-K4	24	K4-K13	33	K4-K12	35	GFP	40
K14-K14	40	K5-K5	24	K5-K14	32	K5-K13	33	Peripherin	40
K16-K16	40	K6A-K6A	22	K6A-K16	31	K6A-K14	31	Vimentin	48
K17-K17	42	K6B-K6B	22	K6B-K17	32	K6B-K16	31		
K18-K18	34	K8-K8	30	K8-K18	32	K8-K17	36		

3.6.2 Dimer stability assessment using Molecular Dynamic (MD) Simulations

Section 3.3.3, section 3.3.7, and section 3.5 establishes that non-uniform atomic density destabilizes the coiled-coil dimer. However, the dynamics involved in the destabilization

of the hydrophobic core was not clear. This section first provides the structural analysis of small coiled-coil dimers and presents the dynamics involved at the molecular level. The next part of this section provides the structural analysis of IF dimers taking the cues from small coiled-coil dimers.

1. Analysis of Leucine Zipper and its Variants

To study the factors that influence the robustness of coiled-coil structures, eight variants of leucine zipper structure (PDB id: 2ZTA) were generated by mutating residues using the Chimera swapaa command (Table 3-12). Case 1 was the actual crystal structure of leucine zipper that act as a control for verifying the robustness of the coiled-coil structures. Case 2 to case 8 provide variants of leucine zipper representing various possibilities for the hydrophobic core density and corresponding salt bridges (Table 3-12, Figure 3-14 D). In Figure 3-14 D, case 2/3/4 present uniform densities of hydrophobic core, whereas case 5/6 present high-density core followed by low-density core. Case 7/8 provide low and high-density core at single position respectively. In some of these cases, salt bridges were incorporated to negate their effect in the coiled-coil stability. The leucine zipper structure and its variants were energy minimized as done for the IF Data set (see section 0). All energy-minimized models were equilibrated for 1 ns in NVT ensemble followed by 1 ns in NPT ensemble. MD simulations were performed on the equilibrated systems having 16064 atoms for 100 ns using GROMACS[192–194]. Snapshots of the trajectories were extracted at intervals of 0.1 ns.

Table 3-12 Variants of Leucine zipper structure

Amino acid substitutions in variants (Case 2 to Case 8) of leucine zipper sequence (Case 1:2ZTA) are shown in bold.

Variant name	Sequence
Heptad Sequence	gabcdefgabcdefgabcdefgabcdefgab
Case 1: 2ZTA	RMKQLEDKVEELLSKNYHLENEVARLKKLVG
Case 2	RMKQLEDKVEEL**E**SK**V**YHLENEVARLKKLVG
Case 3	RMKQLEDKVEE**IE**SK**L**YH**I**ENEVARLKKLVG
Case 4	RMKQLEDKVEE**IE**S**SL**YH**IS**NEVARLKKLVG
Case 5	RMKQLEDKVEE**IE**S**SI**YH**AS**N**SA**ARL**S**KLVG
Case 6	RMKQLEDKVEE**IE**S**SI**YH**AK**NE**A**ARL**S**KLVG
Case 7	RMKQLEDKVEEL**E**S**SV**YH**AS**N**SA**ARL**S**KLVG
Case 8	RMKQLEDKVEE**IE**S**SI**YHL**S**N**S**VARL**S**KLVG

An independent confirmation of the stability through the redistribution of atomic density can be seen in the molecular dynamics simulations trajectory of Leucine zipper (PDB ID: 2zta) and its variants (Table 3-12, Figure 3-14). Case 1 was the crystal structure (PDB id 2zta) and case 2 was N16V variant to remove the bias due to (16N-16N) inter helical hydrogen bond. Both of these structures were control simulations and did not show any unusual fluctuation in RMSD/RMSF. Case 3 and case 4 both have high-density core due to 12I, 19I. Even though case 3 has more salt bridges, it shows mild increase in distance between two helices. The increase was due to inter helical repulsion from 20E and 22E amino acid pairs along with neighboring large density core. This phenomenon agrees with the theory that repulsion inversely governs the stability whereas salt bridges give specificity [121]. Case 5 has a high-density region when followed by low-density region that destabilizes/break the coiled-coil dimer completely (Figure 3-14 B, D, Figure 3-15) and shows gradual increase of distance between two helices. In case 6, two salt bridges were included to protect the low density hydrophobic core from the methylene groups of lysine. However, due to extremely low-density core, the fluctuation occurred after 30 ns and did not stabilize the structure completely. This was an extreme case of destabilization simulated to differentiate the stable and unstable density distribution. Case 7 and case 8 were controls to show that only consecutive high and low density core destabilizes the dimer.

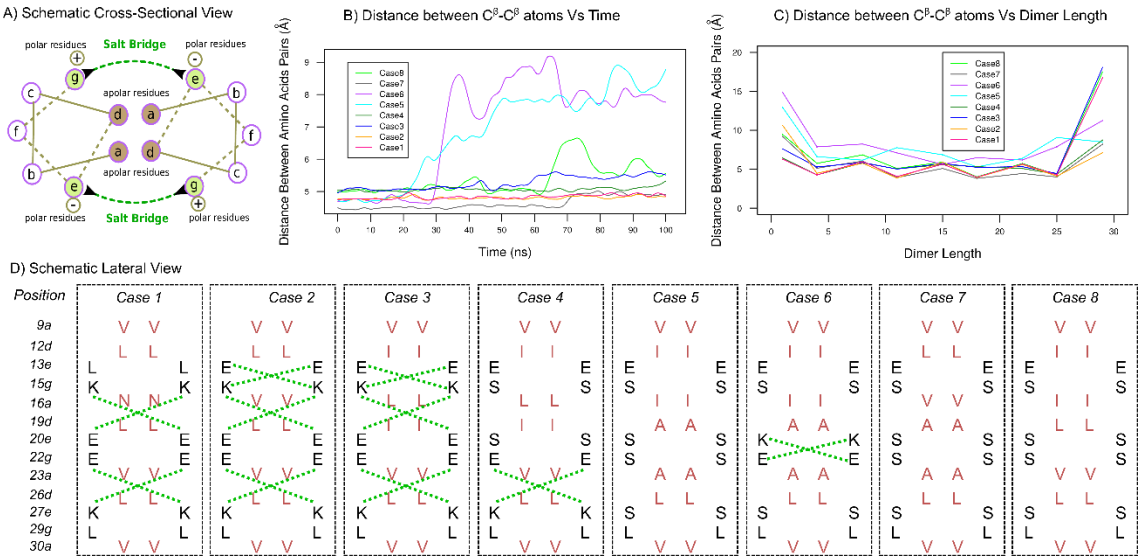


Figure 3-14 MD Simulations of coiled-coils with variable atomic density.

*(A) Schematic of standard coiled-coil with heptad **a/d** forming the hydrophobic core and heptad **e/g** forming the salt bridges. (B) Distance between C^{β}-C^{β} atoms over 100 ns. (C) Distance between C^{β}-C^{β} atoms as a function of dimer length. (D) Schematic lateral view*

*of the coiled-coil dimers sequence from N to C terminus (top to bottom). First column presents the position number of only heptad repeats **a/d/e/g**. Case 1 presents the sequence of Leucine zipper whereas case 2 onwards provides different variants. Amino acids at heptad positions **a/d** are shown in red color, whereas amino acids at heptad positions **e/g** are shown in black color that forms the salt bridges (green dotted lines).*

Previous section describes a set of 100 ns Molecular Dynamics (MD) simulation of standard coiled-coil protein leucine zipper and its variants. In these variants, the atomic density either remains constant, increases or decreases along the coiled-coil. The MD simulation trajectory of constant density structures (crystal structure and controls) remains stable. The non-uniform density structures leads to larger fluctuation in the coiled-coil dimer and eventually breaks the structure. A high atomic density region followed by low-density region breaks the coiled-coil immediately. The salt bridges at **e** and **g** heptad positions could partially protect a less dense core through hydrophobic methylene groups of Lysine and Arginine amino acid side chains. Similarly, high atomic density at consecutive heptad **a-a'** and **d-d'** pairs temporary destabilizes the dimer. The exposed hydrophobic core of high-density region favors the formation of trimers and higher order oligomers.

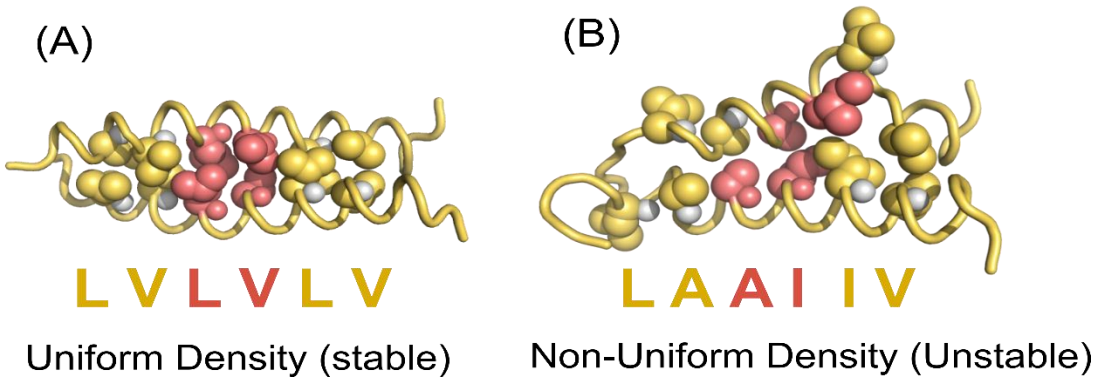


Figure 3-15 Snapshots of two coiled-coil system after 100 ns MD simulation.

A) Snapshots corresponds to Case 2. (B) Snapshot corresponds to Case 5. For a movie of complete 100 ns MD simulation trajectory, refer to the following files: C2C2_movie.mpg and C5C5_movie.mpg in Supplemental Material.

2. Analysis of IF dimers (K5-K5, K5-K14, K14-K14)

The energy minimized models (see section 0) of Type I homo-dimer (K14-K14), Type II homo-dimer (K5-K5) and Type I-II hetero-dimer (K5-K14) were equilibrated for 1 ns in NVT ensemble followed by 1 ns in NPT ensemble. MD simulations were performed on

the equilibrated systems having ~300,000 atoms for 100 ns using GROMACS[192–194]. Snapshots of the trajectories were extracted at intervals of 1 ns for IFs. The following analyses were performed on the MD simulation trajectories:

- *RMSD of atoms at **a-a'** and **d-d'** heptad positions.*
  Distances between C^{α}-C^{α} and C^{β}-C^{β} atoms of amino acids at heptad positions **a-a'** and **d-d'** along the length of dimer averaged over all time intervals. The analysis identified the unstable or fluctuating atomic positions along the length of the dimer.
- *RMSF of atoms at **a-a'** and **d-d'** heptad positions.*
  Distances between C^{α}-C^{α} and C^{β}-C^{β} atoms of amino acids at heptad positions **a-a'** and **d-d'** across all time intervals averaged over dimer length. The average distance between the monomers were calculated by summing up distances of all **a-a'/d-d'** amino acid pairs divided by the total number of pairs. The analysis identified the fluctuating atomic positions over time intervals.

The above results (redistribution of atomic density) were also independently seen in the molecular dynamics simulations trajectory of naturally occurring K5-K14 hetero-dimer, K5-K5 and K14-K14 non-native homo-dimers. An overly dense region shows RMSD fluctuation compared to an averagely dense region in K14-K14 homo-dimer (Figure 3-16 A). This region correspond to the beta branch amino acids pairs 186V-186V and 190I-190I at heptad position **d**/**a** respectively followed by low-density core 193A and 197A. This was similar to the case 5/6 in Figure 3-14.

A similar fluctuation is also seen in K5-K5 homo-dimer in heptad position **d** with 261E-261E pair and in heptad position **a** with 265R-265R pair. These residues were consistent across Type II intermediate filament. Due to repulsion, the atomic density reduces locally and water could penetrate the hydrophobic core. This repulsion-induced invasion of water molecules destabilizes the dimer and prevents the formation of Type II intermediate filament dimers/filaments.

Above described variations in non-native homo-dimers also leads to increase in RMSF over time as compared to the native heterodimers showing Type I and type II non-native homodimers were transiently stable and unstable respectively (Figure 3-16 B).

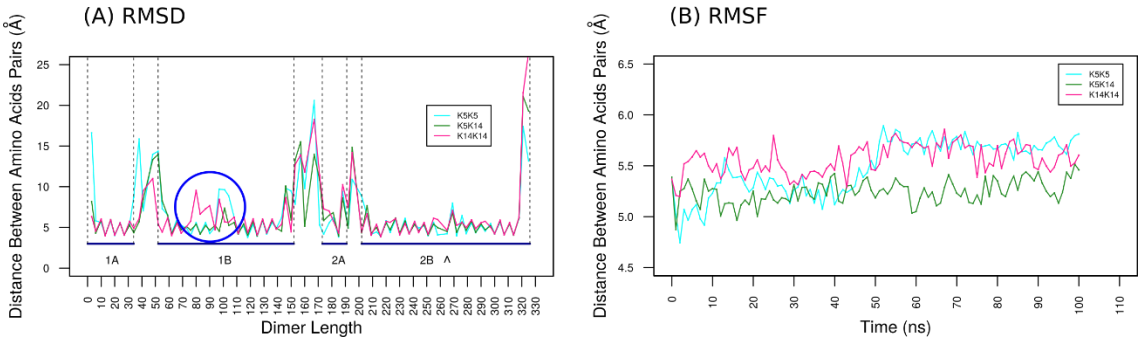


Figure 3-16 RMSD and RMSF of amino acid pairs at a-a' and d-d' heptad positions. *(A) Average distance between C^{β}-C^{β} along the length of dimer. The distances are averaged over 100 ns simulation and are only accounted for amino acids at heptad positions **a-a'** and **d-d'**. The highlighted region with blue circle show the fluctuation in the distances with respect to K5-K14 dimer. Caret (^) symbol shows the location of the stutter region. (B) Average distance between C^{β}-C^{β} atoms over 100 ns MD simulation. The distances are averaged for all **a-a'** and **d-d'** residue pairs.*

3.7 Equilibrium between homodimers and heterodimers

To explain the dimerization preference of Intermediate filaments, the simulation of the reaction system of homogeneously mixed intermediate filament monomers was performed describing the kinetics. In the reaction system, monomers reversibly react with one another to form homo-dimers and hetero-dimers (see below equations).

$$M_I + M_I \underset{k_2}{\overset{k_1}{\rightleftharpoons}} D_{I-I}$$ Monomers of Type I interact to form Type I homo-dimer

$$M_{II} + M_{II} \underset{k_4}{\overset{k_3}{\rightleftharpoons}} D_{II-II}$$ Monomers of Type II interact to form Type II homo-dimer

$$M_I + M_{II} \underset{k_6}{\overset{k_5}{\rightleftharpoons}} D_{I-II}$$ Monomers of Type I and II interact to form Type I-II hetero-dimer

Where *M* and *D* stands for Monomer and Dimers of Intermediate filaments. Subscript indicates the Type of Intermediate filament (Type I, Type II). k_i indicates rate constants for the forward and reverse reactions.

A Kinetic Monte Carlo simulation using the Gillespie algorithm was used to numerically solve for this system [195]. The system consists of six equations (above equations) with initial concentration of Monomers equal to 100 (in arbitrary units). Rate constants of the reactions were chosen to meet the following criteria:

- Type II homo-dimer is unstable due to many unfavorable interactions. Hence, the

rate constant for the formation of the Type II homo-dimer should be less than its dissociation rate.

- The rate constant of the stable reaction should be greater than the unstable reaction. The values for the rate constants were chosen from the kinetic studies done on coiled-coil FOS-JUN proteins [123,196]. The chosen values were sampled exhaustively in the range of 10^4 to 10^7 for stable reactions and $6x10^{-2}$ for unstable reaction. Choosing a single value for unstable reaction do not change the outcome, as the system was analyzed only for the qualitative behavior (Refer to file: Analytical_Solution_IF_Equilibrium.pdf in Supplemental Material). The results only highlight the stability and not the actual concentrations in the equilibrium mixture.

A simulation run was performed until the system attained equilibrium (i.e. there was no change in the concentration) over the period. The next section discusses the insights obtained from the Kinetic Monte Carlo simulation on Intermediate filament assembly.

3.7.1 Type I intermediate filaments are a meta-stable species

For an optimal packing, the high-density residue pair 186V-186V at heptad position **d** in Type I intermediate filament dimers should be surrounded by low-density residue pairs at position 183 and 190. However, the neighboring pair 190I-190I also has high atomic density (see section 3.3.3 and section 3.6.2-2 for details of atomic density). Moreover, the other neighboring pair 183N-183N forms a strong hydrogen bond, thereby restricting the inter-helical movement to accommodate high density of atoms. This asparagine-asparagine hydrogen bond is also consistent in all Type I intermediate filaments. The redistribution of atomic density at the cost of breaking a stable hydrogen bond make the Type I intermediate filaments homo-dimers a meta-stable species as shown in the Monte-Carlo simulations (Figure 3-17). The meta-stability of Type I Intermediate filaments was previously been seen in the assembly experiments [197]. Type I could theoretically form homo-dimeric filaments in the absence of appropriate Type II intermediate filament just like K16, K17 and K18 are already known to form filaments [198,199]. The formation of type I-II heterodimers along with trace quantities of type I and type II homo-dimers were seen in the equilibrium mixture as also observed during in-vitro assembly experiments.

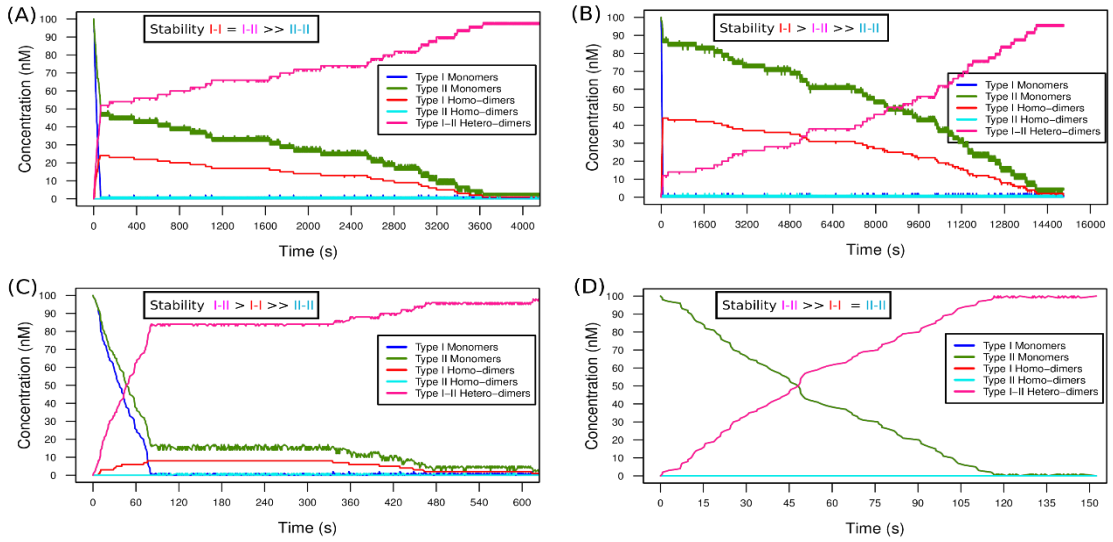


Figure 3-17 Equilibrium concentration of different IF dimers over time.

(A to D) Concentration (Y axis) of reactants and products (Type I and II monomers, Type I-I and Type II-II homodimers and Type I-II heterodimers) as a function of time (x axis). Range of rate constant values were chosen proportional to stability parameter (box at the top of the plots). Irrespective of the rate constant values, hetero-dimers (pink curve) dominate the equilibrium mixture. Plots (A to C) shows Type I homo-dimer (red curve) was meta-stable as compared to Type II homo-dimer (cyan curve).

3.8 Summary

In this chapter, the molecular basis of Intermediate filament diseases caused due to point mutations has been determined. Analysis of the structures and sequences of coiled-coil proteins, provides the molecular determinants of their stability and specific pairing. The numerical form of these molecular determinants was used to train neural network schema that explains 35% (153/438) of disease mutants with 94% accuracy, accounting for most mutations in the heptad positions **a** and **d**. The method was nearly perfect in determining the filament destabilizing substitution. The trained models were then used to score new filament destabilizing mutations obtained through *in-silico* saturated mutagenesis of all IF family member. To validate these findings, K5-K14 IF system mutations were validated through *in-vivo* tissue culture experiments. The stability determinants also establish the meta-stability of Type I homo-dimers and unstable nature of Type II homo-dimers. This leads to rationalize preferential pairing of IFs. i.e. why type III to type VI intermediate filaments form homodimers whereas type I and type II form heterodimers using kinetic Monte Carlo simulations.

CHAPTER 4

Integrative Modeling of 3D structure of IFs

SALIENT FEATURES

- ***Estimating the number of dimers in the filament***
  - Using mass measurements
  - Using volume measurements
- ***Homology Modeling of K5-K14 dimer***
  - Incorporation of Head/Tail domains
  - Assessment of K5-K14 dimer
- **Modeling K5-K14 tetramer configurations**
  - Exhaustive sampling of dimer-dimer configurations
  - Sub-division of interacting configurations
  - Assessment of K5-K14 tetramer
- **Modeling K5-K14 filament**
  - Conversion of 3D models to quasi 3D models
  - Sampling longitudinal and cross-sectional space
  - Model clustering and selection
  - Mapping 2D models to 3D full atom models
  - Model statistics
- **Summary**

4.1 Experimental Information utilized for construction of filament.

Experimental information related to Intermediate filament structure exists in bits and pieces (Figure 2-3). As these experiments provide information at various levels of assembly, information from a single experiment alone cannot be used to construct the structure. This chapter provides the methodology developed/adopted to construct the full atom models based on the experimental data. It describes the structural information obtained from experiments such as K5-K14 filament cross-links, small X-ray segments of IF dimers, heptad repeat geometry from bioinformatics analysis etc. Other experimental information such as K1-K10 filament cross-links, EM, Cryo-EM, TEM etc. were utilized to validate the K5-K14 filament model (see CHAPTER 5 for validation).

4.1.1 Geometry of individual dimer

Keratin filament or any Intermediate filament is an assembly of individual IF dimers [200–202]. These coiled-coils IF dimers are the basic building blocks of the filament. A coiled-coil dimer is a stable protein assembly and their structure could be described in the form of heptad repeats [20,93] (see Section 2.1.1 and Section 2.2.2). Thus, utilizing their sequence information and mapping them into structural coordinates through techniques such as homology modeling, energy minimization etc. are core to building the individual dimer. The coiled-coil dimers are structurally conserved even with low sequence identities. Thus, smaller variations (due to side chain packing) during homology modeling would make the Keratin dimer model close to near atomic level resolution.

4.1.2 Cross-linking Experiments

Cross-linking experiments [203–205] were done earlier on intact and reassembled Keratin Intermediate filament with cleavable bi-functional reagent Disuccinimidyl Tartrate (DST) [206]. All reactions were performed for 30 minutes and then quenched with 0.1M NH_4HCO_3. Cross-linked Lysine amino acids were then digested using trypsin enzyme. They were resolved using reverse phase HPLC column with 60 minute gradient of 0.1% trifluoroacetate and 0-40% acetonitrile. The cross-linking experiment had shown that 10% of proteins undergo modification and 50% of available lysines get cross-linked. In this experiment, two lysine amino acids cross-linked at their N^{ζ} positions with a cross-linker that has an arm length of 6.4 Å (Table 4-1). Cross-links identified in the experiments were used as spatial distance restraints between cross-linked lysine residues. It provide a LYS C^{α}-C^{α} upper bound distance restraint of 19.4 Å obtained by adding the length of the

stretched lysine side chains and the cross-linker arm length. These distance restraints were used for determining the orientation of the dimer-dimer interaction within the filament. There were 23 inter-molecular cross-linking restraints, out of which 19 were inter-dimer cross-links. Other than these restraints, the cross-links of homologous K1-K10 Keratin Intermediate filaments could also be used during the modeling procedure [204]. This includes LYS-LYS and CYS-CYS cross-links. In this study, K1-K10 cross-links were used to validate the models (see Section 5.4.7 for validation protocol).

Table 4-1 List of all cross-linked lysine amino acids

First column provides the interacting mode of cross-link sets and the oligomer (dimer) size of protein. Second and third columns provide the cross-linked lysine identities (LYS1, LYS2) and modified identities according to coiled-coil domain respectively. First amino acid number in the coiled-coil domain starts from 1 to 313 for K5 and 14 to 624 for K14.

Interacting Modes Type	**www.Interfil.org numbering K5:169-481, K14:429-739**		**Coiled-coil Numbering K5:1-313, K14:314-624**	
	LYS 1	**LYS 2**	**LYS 1:heptad**	**LYS 2:heptad**
Mode A Oligomer 2/3/F	K5:209	K5:443	41:c	275:e
	K5:264	K14:641	96:g	526:e
	K5:178	K5:472	10:g	304:f
	K14:556	K14:610	441:e	495:g
	K14:480	K5:431	365:f	263:g
Mode A Oligomer 2/3	K5:264	K5:383	96:g	215:e
	K5:426	K14:480	258:b	365:f
	K5:343	K14:564	175:e	449:f
Mode B1 Oligomer 2/3/F	K5:178	K14:610	10:g	495:g
	K5:264	K5:264	96:g	96:g
	K5:255	K5:276	87:e	108:e
	K14:480	K14:563	365:f	448:e
	K14:480	K14:564	365:f	449:f
	K14:488	K14:556	373:g	441:e
	K5:185	K5:343	17:g	175:e
Mode B1 Oligomer 2/3	K5:276	K5:257	108:e	89:g
	K5:276	K14:516	108:e	401:g
	K5:304	K14:480	136:e	365:f
Mode B2 Oligomer 2/3/F	K5:404	K5:443	236:e	275:e
	K5:404	K5:441	236:e	273:c
Mode B2 Oligomer 2/3	K5:383	K14:718	215:e	603:a
	K5:405	K5:443	237:f	275:e
Intra-dimer Oligomer 1/2/3/F	K5:185	K14:445	17:g	330:g
	K5:257	K14:516	89:g	401:g
	K5:383	K14:641	215:e	526:e
	K5:364	K14:622	196:g	507:g
	K5:304	K14:563	136:e	448:e
	K5:304	K14:564	136:e	449:f

Cross-linking studies also yield relative frequency of each cross-links. The relative frequency of cross-links provides an additional restraint. However, due to poor resolution given by the HPLC at the time of the experiment (year 1993), provides little confidence in the relative frequencies of the cross-links.

4.2 Estimating the number of dimers in the filament.

The estimation of the number of dimers in the unit length filament was done in two ways that leads to approximately same result.

4.2.1 Mass of Keratin filament and dimer

To estimate the number of dimers in the cross-section, the mean mass of the Vimentin (a homologue of Keratin) unit length filament (1.7 MDa) as determined by SAXS [142] was divided by the mean mass of a single dimer (107 KDa).

$$Number\ of\ Dimers = \frac{1700}{107} = 15.88 = \sim 16\ dimers$$

4.2.2 Volume of Keratin filament and dimer

The volume of a K5-K14 dimer was computed by summing the volumes [182] of individual residues.

$$Volume\ of\ K5\ Keartin\ chain = 73.84\ nm^3$$

$$Volume\ of\ K14\ Keratin\ chain = 60.99\ nm^3$$

$$Volume\ of\ dimer = Volume\ of\ K5 + Volume\ of\ K14 = 134.83\ nm^3$$

$$Length\ of\ a\ dimer = 46\ nm \text{ [205]}$$

$$Radius\ of\ the\ dimer = \sqrt[2]{Volume\ of\ dimer/46 \times 3.14} = \sqrt[2]{134.83/46 * 3.14}\ nm$$

$$= 0.966\ nm = \sim 1\ nm$$

$$Volume\ of\ dimer\ per\ unit\ length = 3.14 \times (1)^2 = 3.14\ nm^2$$

$$Average\ Filament\ Diameter = \sim 10nm \text{ [16,207,208]}$$

Assuming the filament as cylinder along the filament axis.

$$Volume\ of\ filament\ per\ unit\ length = 3.14 \times (10/2)^2 = 78.5\ nm^2$$

$$Number\ of\ dimers\ in\ the\ filament = \frac{Volume\ of\ Filament}{Volume\ of\ dimer} \times packing\ density$$

$$= \frac{78.5}{3.14} \times 0.71 = 17.75$$

The packing density considered here was for the interior of a globular protein domains [209]. In the Keratin filament, packing density of the dimers would be lower (than 17.75)

as it is the packing of already folded domains (dimers). Hence, the number of dimers (from 17.75) was rounded to 16 as it contains an inherent symmetry consideration. The number 16 was also in agreement to mass measurements (see Section 4.2.1). This computation estimates the number of dimers in the cross-section to be 16 which is also consistent with the number of dimers in the Vimentin Intermediate filaments as measured using STEM [210].

The 3D structure of the Keratin dimer was not known in its entirety and its structure was needed to determine their arrangement through possible interactions within the Keratin filament. Homology modeling was used to construct the Keratin dimer as described in the next section.

4.3 Homology Modeling of K5-K14 dimer

The structure of the coiled-coil dimer was constructed using homology modeling [26–28] (see Section 3.6). A good template was mandatory to minimize the errors during homology modeling. Since, no crystal structure cover the entire length of the keratin dimer. Thus, a straight modeling template was constructed using coiled-coil leucine zipper protein (see Section 3.6 for modeling details). The resulting template had a coiled-coil cylindrical geometry (Figure 3-13).

To further increase the accuracy of the side chains packing, the Vimentin dimer model was built from multiple available fragments (see Section 3.6 for modeling details). Vimentin dimer model was utilized as another template in the multi-template homology modeling (Figure 3-12).

Along with a good template structure, a sequence alignment of the target and template was also required for homology modeling. In general, a sequence alignment done by well-established methods was used for this purpose. Here template was chosen by considering structural similarity rather than sequence similarity. Therefore, the alignment was done by considering the heptad repeat geometry. Heptad positions of the vimentin structure and the computationally constructed long 2ZTA molecule were extracted from their 3D structure. Similarly, heptad geometry of the keratin K5 and K14 sequences was predicted using PCOIL Server [87,120]. The heptad repeats guide the alignment of the target (K5-K14 sequence) and templates for model building by matching the heptad positions.

Keratin dimer models were then built using modeller by considering template structure and the target-template alignment [26,211]. Five models were constructed and the model with the lowest “Discrete Optimized Protein Energy” (DOPE) score [57] was considered

for further analysis. DOPE score is an atomic distance-dependent statistical potential derived from native structures. All other parameters of the Modeller were set to their default values. The selected model was energy minimized and then optimized through Molecular Dynamics simulation for 10 ns using GROMACS package [192]. This was to ensure that the models were in low energy state and packing of the atoms were efficient. The final K5-K14 dimer model was a straight long coiled-coil molecule of length approximately 465 Å and about 45 heptad repeats (Figure 4-1). The dimer model contains 624 amino acids with 313 amino acids from K5 keratin and 311 from K14 keratin. As there was no template available for modeling the unstructured head/tail domains, the filament modeling methodology chosen such that it remain in-sensitive to head/tail domains. Only coiled-coil domain of the dimer model was utilized for integrative filament modeling.

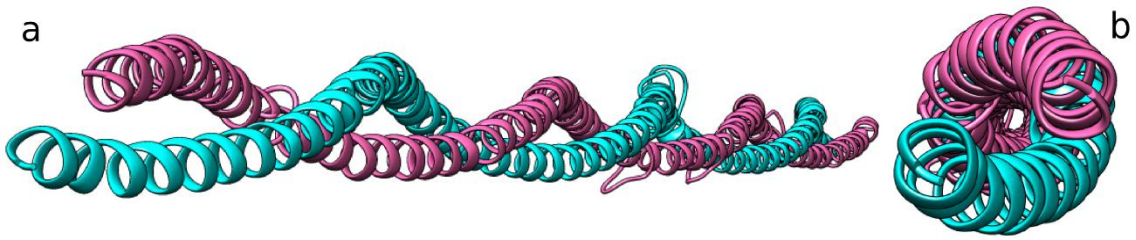


Figure 4-1 Homology model of coiled-coil domain of K5-K14 dimer.
Ribbon representation of Keratin K5(cyan)-K14(pink) dimer model constructed through homology modeling. a) Side view of the keratin dimer b) Cross-sectional view of keratin dimer. The straight coiled-coil domain with K5 monomer helix wound around K14 monomer helix to form a super-coil structure.

4.4 Incorporation of Head and Tail domain in K5-K14 dimer

The K5-K14 filament modeling strategy was in-sensitive to the presence of unstructured Head/Tail domain of K5-K14 dimer. However, some of the filament validations such as dimensions, comparison of radial density etc. required head/tail domains (see Section 5.2.2 and Section 5.3). This section describes the methodology utilized to build the models of head/tail domains. Previously, head/tail domain of the K1-K10 dimer model was constructed computationally by minimal sampling that satisfies a small set of cross-links [19]. The study provided plausible models for K1-K10 head/tail domains. The Intermediate filament head/tail domains were known to be unstructured and construction of the models with limited sampling (25 samples) could provide completely wrong

models. Thus, the accuracy of the K1-K10 dimer models could not be ascertained. However, as there was no template for modeling, the K1-K10 head/tail domain model was used to build the head/tail domain of K5-K14 dimer.

The homology modeling of complete K5-K14 dimer model involves alignment of head/coiled-coil/tail domains with respective template structures. The K1-K10 head/tail domains acts as template for K5-K14 head/tail domains whereas K5-K14 coiled-coil domain model (see Section 4.3) acts as template for itself (Figure 4-2). The sequences were aligned using Modeller align2d module (Refer file: K1K10_K5K14_alignment.ali in Supplemental Material) [26,211]. Homology modeling of complete K5-K14 dimer was performed using Modeller suite of programs with "refine_level" parameter set to "very slow", "dynamic_coulomb" and "dynamic_lennard" set to True [26,191]. Modeling produced a long straight K5-K14 dimer model with globular head/tail domains (Figure 4-2).

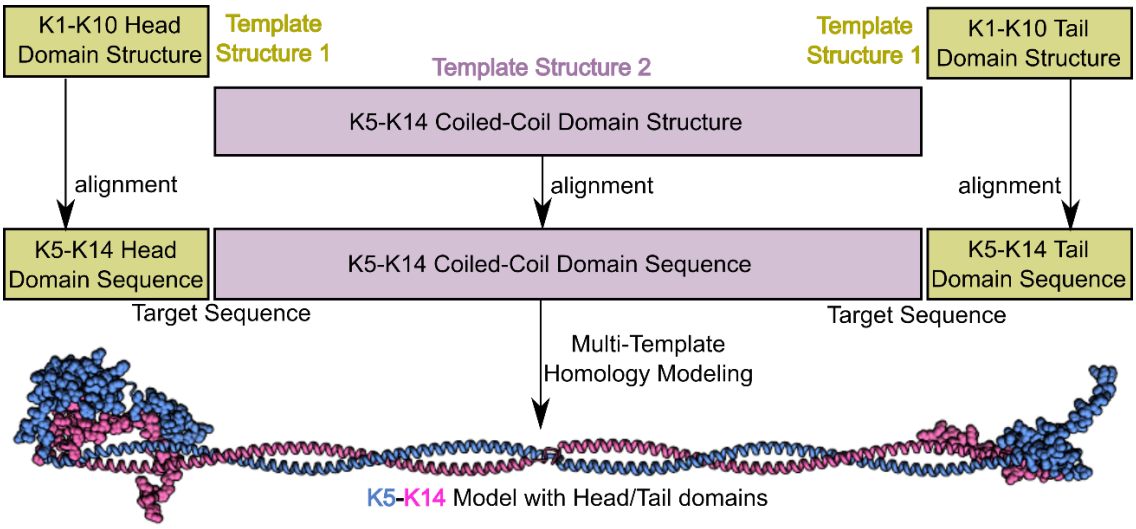


Figure 4-2 Homology modeling of K5-K14 dimer with Head/Tail domains

Construction of K5-K14 dimer model. Schematic representation of the templates used for the model construction. Keratin K1-K10 head/tail domain structures acts as template for K5-K14 head/tail domain sequences (Gray boxes). K5-K14 coiled-coil domain structure is a template for itself with identical self-alignment (purple box). After multi-template homology modeling, keratin K5-K14 dimer model (K5:blue and K14:pink) with globular head/tail domain (spheres) and straight long coiled-coil domain (ribbon) is represented.

The K5-K14 head/tail domain model was based on a template (K1-K10 model) that was in-turn dependent on the cross-linking information. This raises uncertainty in the atomic

positions of most (except cross-linked atoms) head/tail domain atoms. However, the overall position of the entire head/tail domain with respect to the coiled-coil domain was certain. Thus, the entire head/tail domains were replaced with equivalent unstructured atomic density of carbon atoms.

The number of carbon atoms in the head/tail domains were calculated by dividing the mass of head (26365.7 kDa) and tail (14628.7 kDa) with molecular weight of carbon atom (12 Da). For head domain, number of carbon atoms (N_h) were 2195 whereas for tail domain (N_t) they were 1217. The shape of head/tail domains were assumed to be ellipsoid with protein packing density (d) as 0.7 [212–214]. The volume of a carbon atom (V_c) was calculated by considering the Van der Waals radii as 1.7Å.

$$V_c = \frac{4}{3}\pi\ 1.7^3 = 20.56\ \text{Å}^3$$

The volume of the ellipsoid with radius (r_x, r_y, r_z) was given by the following equations.

$$V_{ellipsoid} = \frac{4}{3}\pi\ r_x r_y r_z = \ (\ N_{h,t} * V_c\ )/d$$

For elongated ellipsoid in x direction, radius of head/tail ($r_{h,t}$): $2r_{h,t} = r_x = 2r_y = \ 2r_z$

Substituting values for r_x, r_y, r_z, radius of head/tail ($r_{h,t}$) was given as

$$\frac{4}{3}\pi\ 2r_{h,t}^3 = \ (\ N_{h,t} * V_c\ )/d$$

$$r_{h,t} = \sqrt[3]{\left(\frac{3}{8\pi d}(\ N_{h,t} * V_c\ )\right)} = \sqrt[3]{\left(\frac{3}{0.56\pi}(\ N_{h,t} * 20.56\ )\right)}$$

For a given number of head/tail atoms, the above relation was used to determine the radius of head/tail domain next to coiled-coil domain (Figure 4-3). The obtained K5-K14 dimer model was utilized to build the complete Keratin model. The following few sections describe the validation protocol for the K5-K14 dimer model.

Figure 4-3 K5-K14 dimer with unstructured atomic density of head/tail domains.
The representation of the unstructured head/tail domains (shaded circles) as ellipsoids of the same atomic density (yellow spheres). Coiled-coil dimer shown in ribbon (blue and pink) and sphere representation (cyan and pink).

4.5 Assessment of K5-K14 Dimer

4.5.1 Hydrophobic core Analysis

To validate the Keratin dimer model, the amino acids in the heptad positions [**a-g**] should follow the standard coiled-coil guiding principles. A stable super-coiled structure should have a dominant hydrophobic core formed by residues in heptad positions **a** and **d**. Amino acids in the dimer model were categorized as hydrophobic and non-hydrophobic (see Section 3.3.5 for classification). The frequencies of hydrophobic and non-hydrophobic amino acids were recorded for every heptad position in the coiled-coil dimer. The heptad positions **a** and **d** were predominantly (more than 70%) hydrophobic (Figure 4-4 A). The rest 30% of amino acids at heptad position **a/d** that were not hydrophobic were mostly present in the non-coiled-coil domains (linker domains).

4.5.2 Solvent accessible surface area analysis

Due to coiled-coil geometry, the amino acids at heptad position **a** and **d** remain inaccessible to the solvent. The Solvent Accessible Surface Area (SASA) of amino acids, provides a measure to determine the solvent accessibility. To validate the inaccessibility to the solvent of amino acids at heptad **a/d**, SASA for all amino acids were computed across heptad positions for the K5-K14 dimer model. All atom solvent accessible surface area of amino acids at heptad positions **a** and **d** were below 30%, whereas for all other positions it was above 50% (Figure 4-4 B). Heptad position **g** showed values around 40% due to the presence of partially buried side chains of charged amino acids such as lysine and arginine.

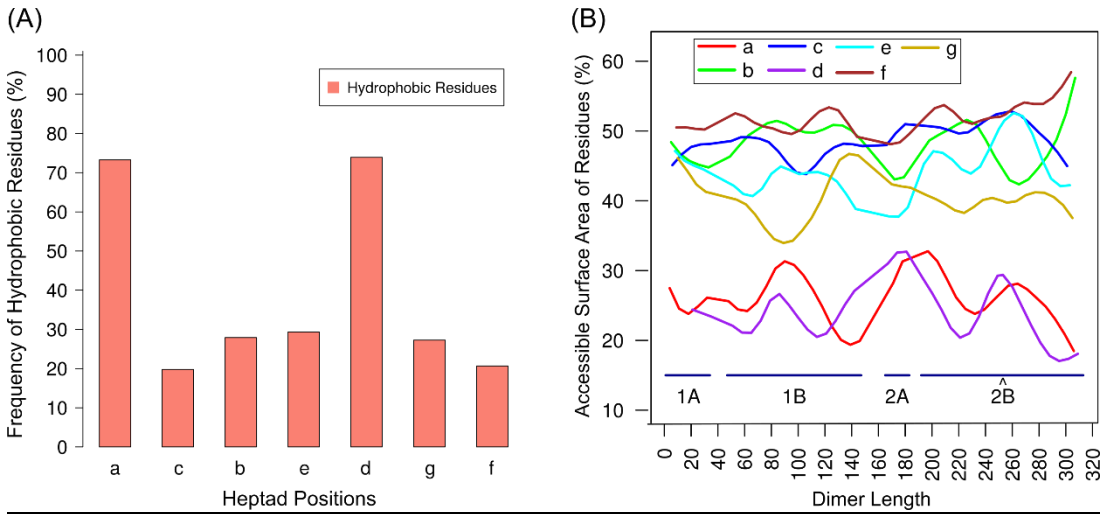


Figure 4-4 Distribution of hydrophobic amino acids and SASA across heptad.

*(A) Frequency of hydrophobic residues (red bars) across heptad positions (x axis). Heptad positions **a/d** have more than 70% of hydrophobic residues in contrast to other heptad positions. (B) All atom smooth Solvent Accessible Surface Area (SASA) in percentage (y axis) was plotted against the dimer length (x axis) for residues at heptad positions **a** to **g** (different colors). SASA of residues at heptad position **a/d** was consistently low as compared to residues at other heptad positions.*

4.5.3 Comparison with the crystal structure

The Keratin dimer model was compared with a recently crystalized 2B segment (PDB code 3TNU) [175] of the Keratin dimer. The Keratin dimer model had a Root Mean Square Deviation (RMSD) value of 0.9 Å (Figure 4-5), the fraying at the ends notwithstanding. This implies that our model is accurate to within atomic resolution.

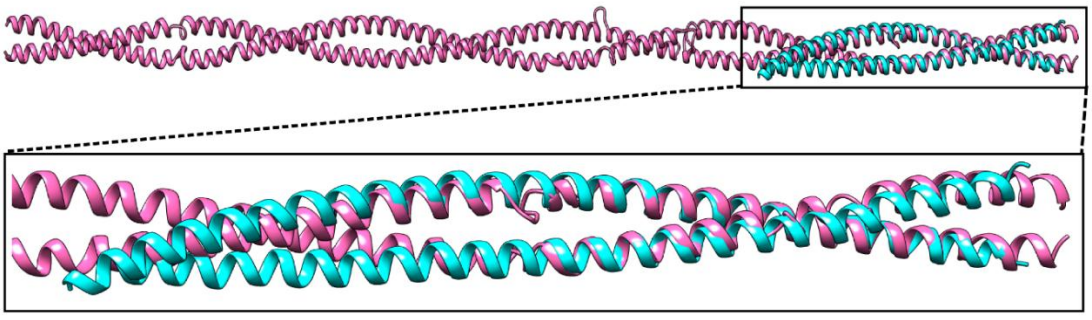


Figure 4-5 Comparison of Model with crystal structures.

Comparison of Keratin 2B segment crystal structure (PDB id: 3TNU) with Keratin dimer model. Superimposition of dimer model (pink) with crystal structure (cyan). Inset shows larger view of superimposition.

4.5.4 Electrostatic Analysis

Electrostatic calculations were performed to determine the charge distribution in the Keratin dimer model. The charge distribution is helpful for determining the interaction sites of the Keratin dimer. At pH 7.0, coiled-coil domain have net negative charge -32 units whereas the head and tail domain has net positive charge +12 and +7 units respectively. The charges were localized in various positions of the coiled-coil domain (Figure 4-6). The negative charge was present in different pockets along the coiled–coil dimer. Some of these sites could be the binding sites for head and tail domains. All electrostatic calculations were performed using Adaptive Poisson Boltzmann Solver (APBS) [215] and Poisson Boltzmann Equations solver (PBEQ-Solver) online server [216] and visualized in Pymol (http://pymol.sourceforge.net/) and Chimera [191].

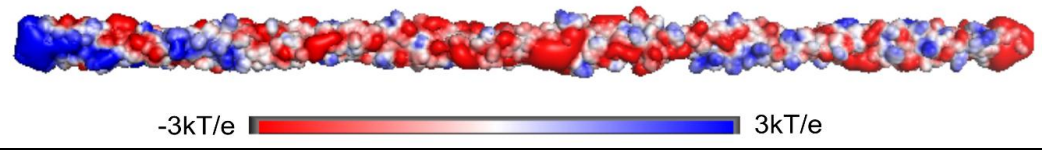


Figure 4-6 Electrostatic potential of Keratin dimer model.

Keratin dimer model (surface representation) showing rendered electrostatic potential in dimensionless units of kT/e. Blue and red colors represent positive and negative charges respectively.

4.5.5 Intra-dimer cross-links analysis

Along with the inter-dimer cross-links, the cross-linking experiments [203,204] obtained six intra-dimer cross-links (cross-links within a dimer) (Table 4-1). All intra-dimer cross-linked residues were at **g-g'** and **e-e'** heptad positions. Due to coiled-coil geometry, both heptad pairs were at two opposite ends of the coiled-coil dimer. Theoretically, cross-linker needs to pass through the hydrophobic core to form an intra-dimer cross-link (Figure 4-7 A). Since the K5-K14 dimer model was consistent with the X-ray crystal structure (Figure 4-5), there were no ambiguities in the heptad repeat pairs **g-g'** and **e-e'.** Thus, the only way to form cross-link was to disrupt the hydrophobic core, which also leads to protein modification. The cross-linking experiment had shown 10% of proteins having a modification and 50% of available lysines getting cross-linked [204].

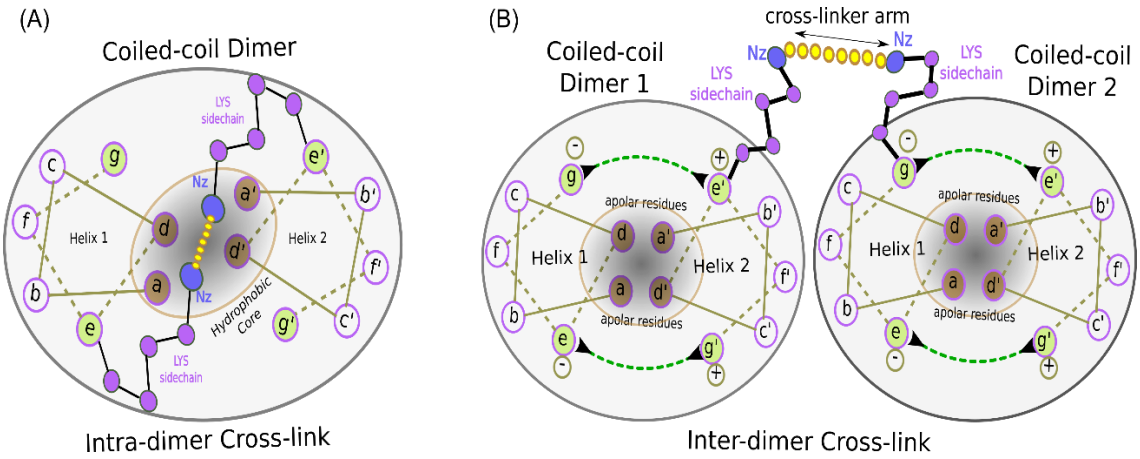


Figure 4-7 Schematic of Intra-dimer and Inter-dimer cross-link.

*(A) Theoretically possible Intra dimer cross-links at **g-g'** and **e-e'** positions. Straight lines (turquoise) between heptad positions represents bond between residues, lysine side-chain atoms (purple circles) cross-linked via N^{ζ} atom (blue circle), cross-linker (yellow circles) and the hydrophobic core (shaded region). (B) Schematic of Inter-dimer cross-links between dimer 1 and dimer 2.*

4.6 Modeling K5-K14 Tetramer (dimer-dimer) Configurations

4.6.1 Getting Restraints between Neighboring Dimers (Inter-dimer restraints)

To deduce the possible dimer-dimer (tetramer) configurations, two dimers (individual dimer construction details in section 4.3) were constructed parallel and anti-parallel to one another separated by an inter-axial distance of 18.5 Å. This was the closest possible distance between the two dimers without atomic clashes. The dimer axes (determined from the first principal component of dimer Cartesian coordinates) were made parallel/anti-parallel with appropriate spatial transformation to prevent atomic collisions during configuration search. All required transformations were determined through standard 3D geometry formulas.

All possible dimer-dimer configurations were exhaustively sampled by translating and rotating one dimer with respect to the other dimer (Figure 4-8). The step size of longitudinal shear along the dimer axis was 5 Å, whereas the step size of rotation of first and second dimer was 5 degrees. Since the dimer length was about 470 Å, the total number of possible dimer-dimer configurations were 974592.

In all these configurations, lysine amino acids whose side chains were within 6.4 Å of one another were likely to form cross-links. As our modeling was not sensitive to relative side chain orientations, C^{α}-C^{α} distances were considered for distance restraints. The upper

bound distance between C^{α}-C^{α} atoms was the sum of length of stretched lysine side chains and the cross-linker arm length. Thus the upper bound distance was estimated as 19.4 Å (Figure 4-7 B). The cross-links (lysine C^{α}-C^{α} distances) were recorded in each of these configurations. Similarly, all possible configurations for anti-parallel dimers (two parallel dimers pointing in opposite directions) were also explored. All possible configurations (interacting modes) were stored in a MySQL database of size approximately ~100 Giga bytes.

Overall, in the MySQL database of dimer-dimer configurations (total size 974592), 128 configurations satisfy Mode A cross-links and 296 configurations satisfy Mode B cross-links. Since these configurations were based on lysine C^{α}-C^{α} distances, the dimer-dimer configurations was further reduced using the availability of lysine side chains. Full atom models of all configurations (128 + 296) were modeled again with N^{ζ} atom of cross-linked lysine restrained to 6.4 Å distance (cross-linker arm length). Configurations where lysine N^{ζ} atom could not come within the cross-linking distance were discarded. After lysine side chain modeling, Mode A had 106 possible configurations whereas Mode B had 164 possible configurations. All these configurations were utilized for filament construction (see Section 4.8).

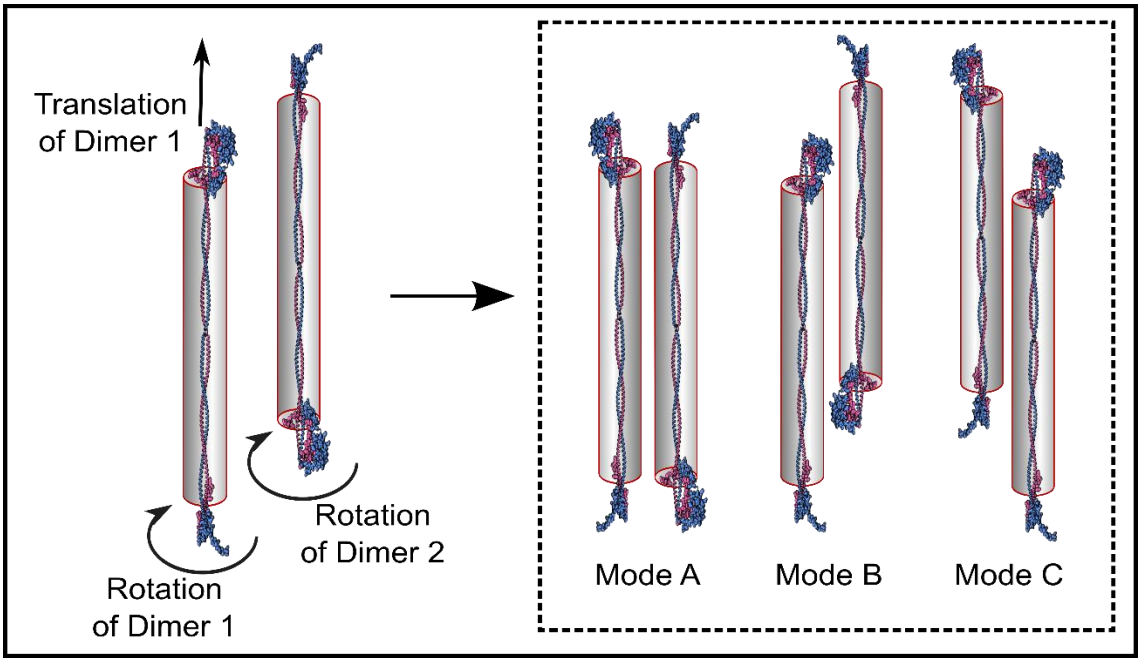


Figure 4-8 Configuration Search.

Sampling all possible configurations between two dimers. θA and θB are the rotation of dimer A and dimer B around the dimer longitudinal axis. Three modes of association were determined to parsimoniously satisfy the crosslinks. Mode A and Mode B are anti-parallel modes whereas Mode C is parallel mode.

4.6.2 Sub-division of Interacting configurations

Sampling of the dimer-dimer configurations was performed to determine the inter-dimer configurations that could form cross-links. Three anti-parallel modes and two parallel modes parsimoniously satisfied all cross-links. Choosing configurations parsimoniously allows minimum number of configurations to satisfy all cross-links. The chosen configurations also accounts for lysine amino acid pairs that were within the cross-linking distance but were not cross-linked in the experiments. These residue pairs could not form cross-links either due to masking by head/tail domains or due to steric clash with other atoms. Such residue pairs were marked as false positives and would be used later to filter out the incorrect models. One of the anti-parallel modes (Mode B1) was just the axial translation of one dimer (equivalent to the dimer length) to form other anti-parallel mode (Mode B2) (Figure 4-9). Thus, Mode B1 and Mode B2 were combined into one mode labeled as Mode B. Similarly, one of the parallel mode (Mode C1) was combined with another parallel mode (Mode C2) to form Mode C (Figure 4-9). Together these interaction modes (Mode A, Mode B, Mode C) between any two dimers (a tetramer) should satisfy all the cross-links observed in the experiment. The assumption here was that the filament contains multiple self-similar building blocks rather than unique tetramers as seen in many studies [17,217–219].

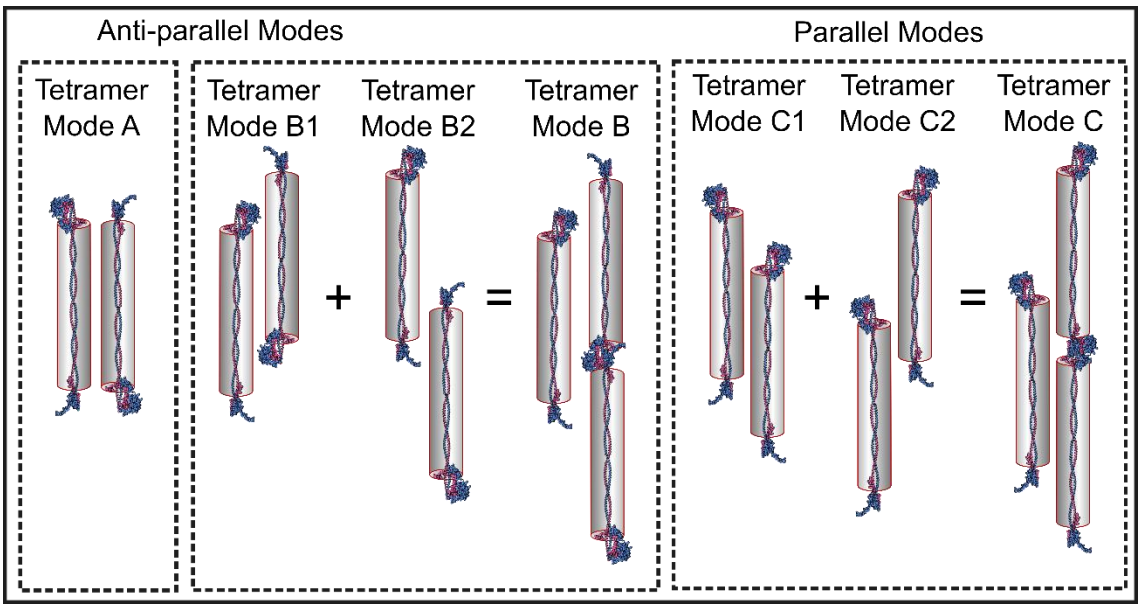


Figure 4-9 Keratin dimer-dimer configuration.

Tetramer Mode A, Mode B1, Mode B2 were anti-parallel modes whereas tetramer Mode C1 and Mode C2 were parallel modes. Tetramer mode B1 and B2 were combined to get tetramer Mode B whereas tetramer mode C1 and C2 combined to get mode C. Each of these interacting modes have a range of values for axial shear and rotations.

4.7 Assessment of K5-K14 Tetramer Configurations

4.7.1 Tetramer models electrostatics

One of the possible tetramer configuration (see Section 4.6.1) was analyzed with APBS generated electrostatics of individual dimers [216]. Head domain of each dimer was positively charged whereas tail domain was negatively charged. Head domain in all the interaction modes (Mode A, Mode B, Mode C) were close to a negatively charged region in an adjacent dimer (Figure 4-10). This charge complementarity was present in all other dimer-dimer configurations due to limited longitudinal shear values. This implies that all the selected dimer-dimer configurations were energetically favorable.

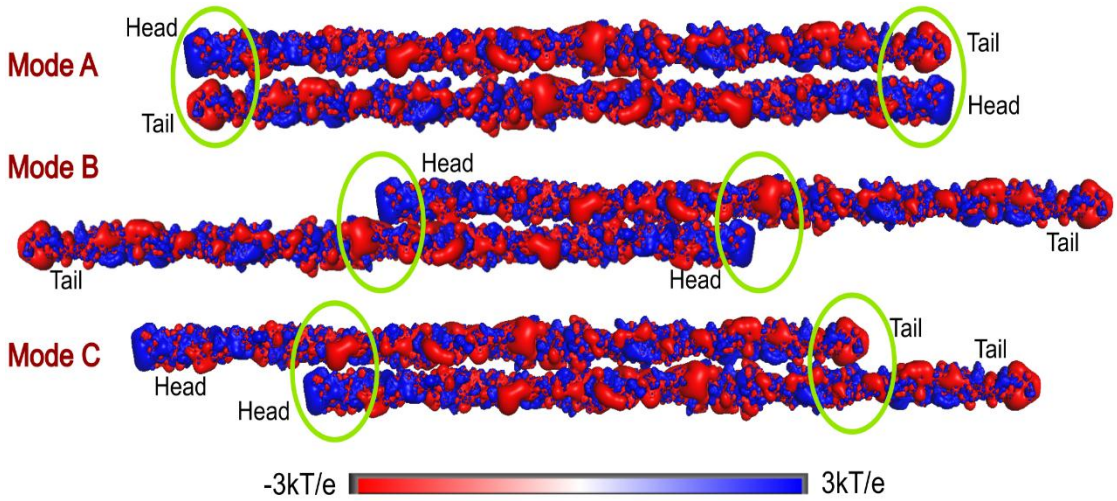


Figure 4-10 Electrostatic complementarity of tetramers.

Each tetramer has two dimers with head and tail domains. The encircled regions show electrostatic complementarity. Scale bar -3kT/e (red) to 3kT/e (blue) represents electrostatic potential in the dimensionless units of kT/e.

4.7.2 Disease causing mutations and their implications on Tetramers

More than 442 different disease-causing mutations were known to fragment intermediate filaments (Section 2.2.5), out of which 143 were in keratin K5-K14 filament (Table 4-2). Since the basic building blocks (Mode A, Mode B and Mode C) were tetramers (Section 4.6.2), the disease mutants at heptad positions **b/c/e/f/g** should affect their stability. More than 442 different disease causing mutations exist in Intermediate filaments, out of which 143 were in keratin K5-K14 intermediate filament (Table 4-2). Qualitative analysis of few mutants show less stability of dimer-dimer modes (tetramers) compared to the wild type (Figure 4-11). A more accurate quantitative analysis needed for all mutants. This

could be done by comparing free energy of mutants and the wild type tetramers. The structure of keratin filament and its analysis could directly be implied to other IF's structure and diseases.

Table 4-2 Disease causing mutations in K5-K14 Intermediate filament

K14 Heptad a/d Mutations									
M119I	M119V	M119T	V133L	V133M	V133A	N140S	R211P	I377N	I377T
L384P	I412F	I412N	L419Q	L122F	Y129D	Y129C	L136Q	L136P	L143P
D273G	M294T	L401P	L408M	Y415H	Y415C	Y415S	E422K		
K5 Heptad a/d Mutations									
I172V	F179S	V186L	V186M	V186E	N193K	R265P	I467L	I467T	I467M
L175F	L203M	L203P	V324D	V324A	D328H	D328V	D328G	D328E	A428T
A428V	L463P	Y470H	E477K	E477G	E477D				
K14 Heptad b/c/e/f/g Mutations									
Q120R	Q120P	R134C	R134P	R148C	A413T	A413P	S128P	V268D	T319P
K116E	K116N	N123S	N123K	L130P	E144A	V270M	V270A	A274D	R288L
E381K	R388C	R388G	R388H	R388P	R416P	R417P	R125C	R125G	R125S
R125H	R125P	R125L	M272R	M272T	E411K	L418V	L418Q		
K5 Heptad b/c/e/f/g Mutations									
K173N	A180P	A180D	R187P	E475K	E475G	S181P	L202Q	V323A	V323G
T469P	G476D	R169G	R169P	N176S	I183F	I183V	I183M	E190K	E190D
L311R	L325F	L325P	N329S	N329K	R352S	A424P	A438D	E466D	R480G
K404E	K443E	K443N	R471C	R471H	E478K	E170K	E170G	N177S	Q191P
T198S	K199T	K199R	K199M	K199N	M327T	M327K	R331C	R331G	R331S
R331H									

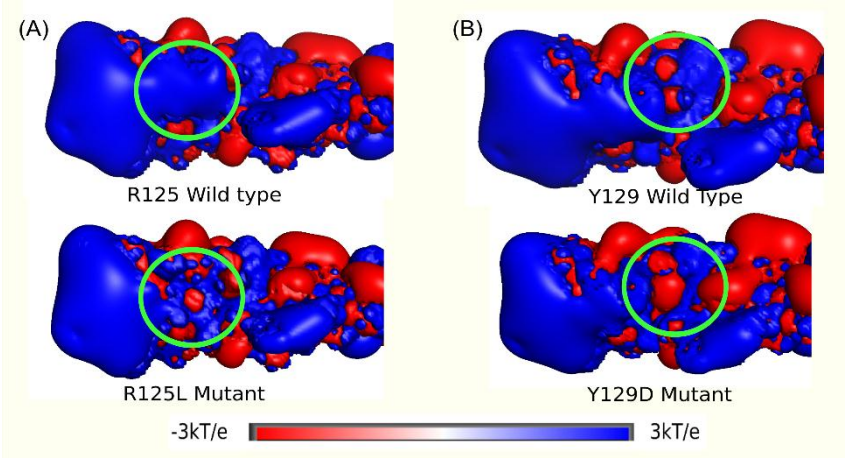


Figure 4-11 Comparison of electrostatic potential of mutants and wild types.

(A) Comparison of wild type R125 and R125L mutant. (B) Comparison of wild type Y129 and Y129D mutant. Scale bar -3kT/e (red) to 3kT/e (blue) represents electrostatic potential in the dimensionless units of kT/e. Encircled regions show position of the wild type and mutant.

4.8 Modeling the Filament

The filament was assembled from the dimers using the possible interaction modes through "divide and conquer" approach. In this approach, tetramer of dimers (octamers) were assembled first to reduce the sampling space. Considering octamers in the next level of hierarchy was an implicit symmetry consideration. All the assembled octamers having clashes were removed from the sampling space. Non-clashed octamer configurations were then used to construct filament models. However, the number of possible configurations were extremely large for complete full atom model simulations. To solve this problem the Disctransformation software was constructed (See Appendix 9.1 for more details). The Disctransformation software convert the 3D structure to a quasi-3D structure, thereby reducing the sampling space. Following section describes the methodology adopted to convert the 3D structure to a quasi-3D structure.

4.8.1 Converting 3D models to quasi-3D models.

The 3D model of keratin dimer was a straight long rod like structure. The projection of the dimer on a plane perpendicular to the dimer axis would be a 2D circular disc. Thus, the dimer was approximated as a circular disc with radius given by the radius of the dimer (Figure 4-12).

Each keratin dimer was sandwiched between globular head/tail domains that also provides the orientation of the dimer with respect to the other dimer. Thus, different colored discs (red/green) represents the orientation of a dimer. If the position of head domain is up and tail domain down, the disc would be red in color. Similarly, if tail is up and head is down, then disc would be green in color (Figure 4-12).

The rotation of the disc around its center provides the rotation of the dimer along the dimer axis (Figure 4-12). Thus, a disc model with appropriate color and rotation completely represents the keratin dimer.

Any dimer-dimer (tetramer) configuration was specified by rotation of dimer 1 ($\Theta 1$), rotation of dimer 2 ($\Theta 2$), relative longitudinal shear (S_l) and their axial distance (A_d). A tetramer configuration in disc models was represented by two discs separated with distance (A_d) and were arranged appropriately using their corresponding rotations values ($\Theta 1$ and $\Theta 2$). A small arc on both the connecting discs represents the point of contact of the two cross-linked dimers. The length of the arc gives the conformation flexibility between the discs/dimers (Figure 4-12). During the conversion of the dimer-dimer model

to the two disc model, the relative shear value between dimers was mapped in the discs and hence retaining the 3D information of the models.

Disctransformation representation described above provide the exact conversion of 3D models to quasi-3D models without any loss of the structural information. The biggest advantage of this representation was the coarse graining of a 3D dimer to a single disc. Since, the filament has many self-similar units (dimers), the assumption that all the units in the filament have similar structure doesn't affect the overall structure of the filament. The next section describes the sampling and scoring methodology used in constructing the filament models.

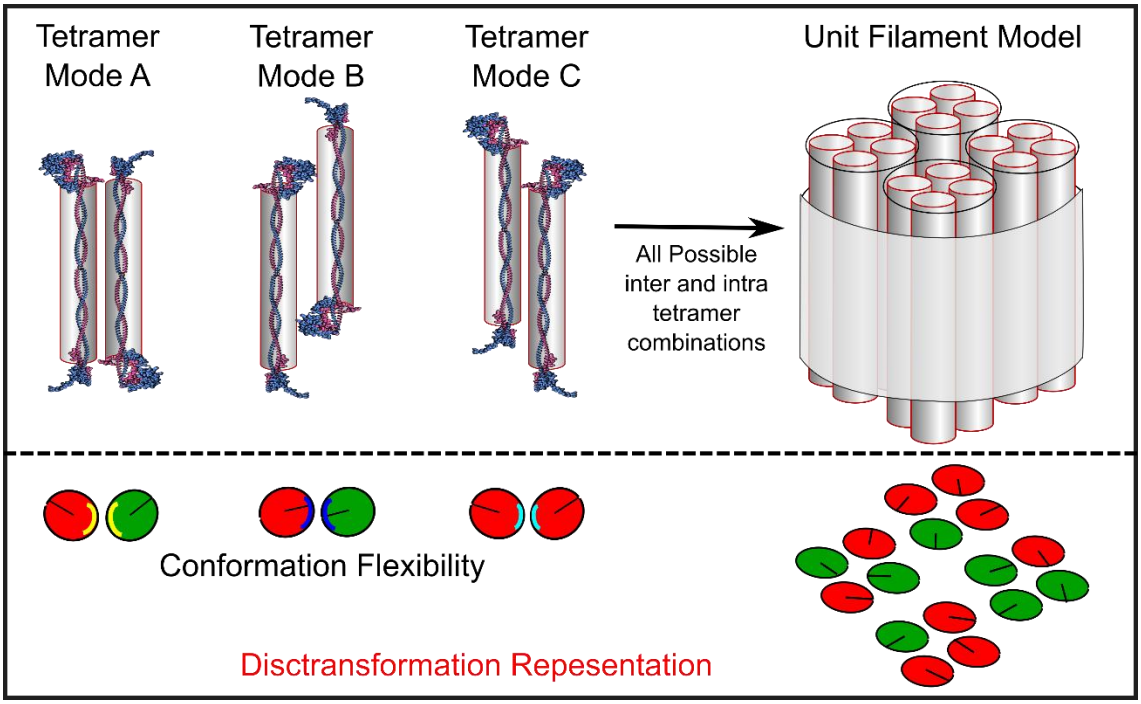


Figure 4-12 Conversion of Full atom models to Disctransformation models.

Disctransformation space consists of red or green colored disc representing a single coiled-coil dimer with head upwards or downwards. The two discs connect to each other (forms cross-link) at a specified angle given by dimer-dimer interaction mode. Black line on the discs represent zero degree position on the dimer. Yellow, blue and cyan arcs on the disc represents all valid angles in a particular dimer-dimer configuration (Mode A, B, C) respectively. Combination of 16 discs (dimers) represent a unit length filament in Disctransformation space.

4.8.2 Sampling longitudinal space

1. Representation of filament topology in Disctransformation

The first step for exhaustive sampling of filament models was to enumerate all possible filament topologies. A filament topology was defined by the set of all dimer-dimer configuration types (Mode A, Mode B) in the filament. The topology do not account for the actual numerical values of dimer-dimer configurations (S_l, $\Theta 1$, $\Theta 2$). A convenient mathematical way to represent a filament was via the square adjacency matrix (Figure 4-13). Rows and columns of the matrix represents the dimers in the filament. The value of the matrix for a particular row (say dimer d1) and column (say dimer d2) provides the interacting mode (mode A or mode B). If the two dimers were not connected (no cross-link) via an interacting mode, its value in the matrix would be zero (Figure 4-13). Thus, a complete filament model topology was represented with this adjacency matrix. To enumerate all possible filament topologies, all dimer-dimer associations in a filament were accounted in the adjacency matrix.

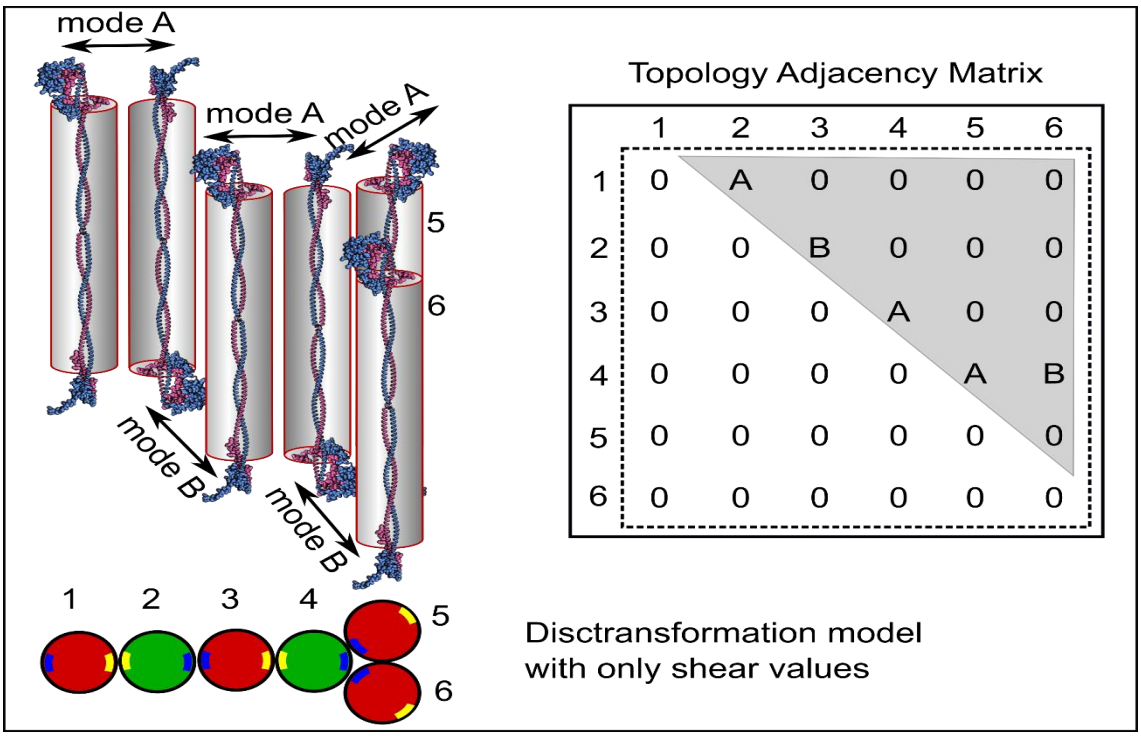


Figure 4-13 Schematic of filament topology through adjacency matrix

Schematic representation of six-dimer model (1-6) with connections described in the upper diagonal adjacency matrix (Mode A or Mode B). The adjacency matrix provide an effective representation even in cases of non-linear connections such as dimer 4 (4th disc) was connected to both dimers 5 and 6 (5th and 6th disc).

2. Filament construction in longitudinal space

A filament model consists of 16 dimers that was sub-grouped into four "tetramers of dimers" or four Octamers. The longitudinal shear values of each dimer, connected through an interacting mode were obtained from the cross-linking analysis of dimer-dimer configurations. For sampling the longitudinal space of the filament, four octamers ("tetramers of dimers") were constructed along with an additional octamer. This additional octamer should overlap with the first octamer in the filament. In the overlapping octamers, at least one dimer from each tetramer should have identical shear (see Figure 4-16 for schematic representation). This overlap restraint ensures that the 2D model remains closed with respect to longitudinal shear (Figure 4-14). All possible models that satisfy longitudinal shear restraint were considered for sampling cross-sectional space. Next section describes the filament sampling in longitudinal space.

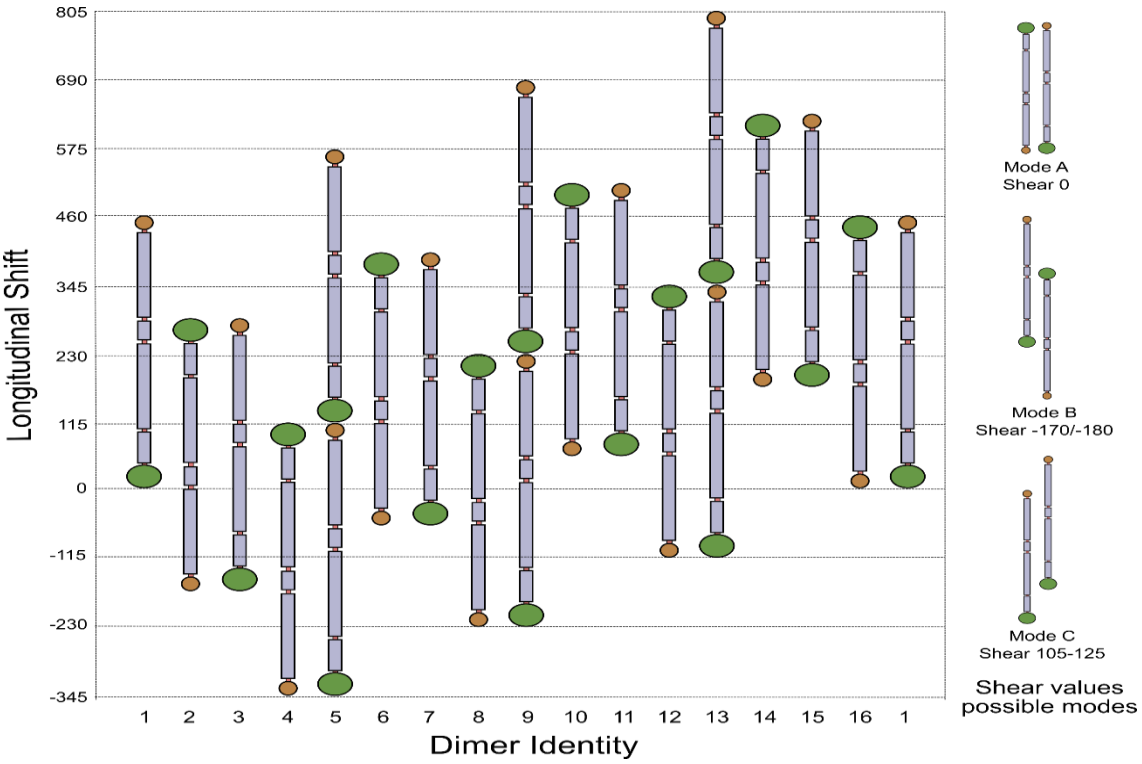


Figure 4-14 Sampling shear values in 2D filament matrix.

Schematic representation of filament model considering only the longitudinal shear values in 2D. A coiled-coil dimer (gray bar) with head (green) and tail (red) domain were aligned parallel to filament axis. Every dimer (x axis) has an absolute longitudinal shear value (y axis) given by the interacting mode. In the schematic, dimer 1 is connected to dimer 2, dimer 2 connected to dimer 3 and so on. A filament model starts from dimer 1 and ends at dimer 16. The 17th dimer show identical shear value with respect to the first dimer. Legend shows the possible shear values of the interacting modes.

4.8.3 Sampling cross-sectional space

In the previous section, all filament models were constructed by connecting dimers through an interacting mode. Each interacting mode has its relative longitudinal shear (S_l) along with rotation of dimer 1 (Θ1) and dimer 2 (Θ2). Relative longitudinal shear values were considered for longitudinal space sampling. For cross-sectional sampling, angle values of interacting mode (Θ1 and Θ2) were accounted. Thus, every sample in the sampling space has a 2D structure given by the coordinates of the discs. All models that satisfies the longitudinal restraint (described in previous Section 4.8.2) were considered for cross-sectional sampling. Every model in the cross-sectional samples were scored with various scoring functions. The following paragraphs describes the scoring functions utilized for building filament models.

- ***Clash score function***:

A clash score was a quantitative measure of the dimer-dimer clash. The distances between every dimer/disc with other dimers/discs were calculated within a model. A clash score has a negative value if the two discs overlap each other otherwise zero (Figure 4-15).

$$"clash\ score" \equiv f_1(x) = -k \times \sum_{i=1}^{n} xi^2$$

where n is the total number of clashes, xi is the length of the overlapping region between disc pairs, k was taken as 1 which could be adjusted according to the weightage needed for this score.

- ***Compactness score function***:

A compactness score was a binary score (values 0 and 1), that represents whether a model was a compact model or not. Models that were confined in a given region were considered compact (Figure 4-15).

$$"compactness\ score" \equiv f_2(x) = \begin{cases} 1, & x < cutoff \\ 0, & x \geq cutoff \end{cases}$$

where x was the distance between the two farthest discs in the model, cutoff value was decided based on the maximum distance allowed between two discs. Maximum distance covered by n number of discs is (n-1)*diameter. For instance, for 16 discs model, cutoff was calculated using eight discs as explained below.

$$cutoff = (8-1) \times Diameter = 7 \times 20 = 140\ .$$

“Compactness score” was a binary score which will either accept the model (score=1) or reject the model (score=0).

- **Hard Clash Count Score function**

A clash count score counts the number of hard clashes between the discs. At most one disc was allowed to clash with other discs. The cutoff value of the hard clash was calculated by determining the radius of dimer excluding side chain atoms. Backbone radius of keratin dimer is ~8.0Å.

$$"Hard\ Clash\ Score" \equiv f_3(x) = \begin{cases} 1, & x < 2 \\ 0, & x \geq 2 \end{cases}$$

where x was the total number of discs clashing with overlap greater than ~8.0Å.

- **Parallel mode cross-link function**

As the relative frequency of the parallel mode (Mode C) cross-links were minuscule, this mode was not considered for determining dimer-dimer interactions. However, Mode C was used as a restraint during sampling/scoring of the models. "Parallel mode score" accounts for presence of at least one parallel mode between the interacting dimers (Figure 4-14, dimer 1 and dimer 5). In this score, angle values were excluded from the scoring function due to low relative frequency and ambiguity in angle values.

$$"Parallel\ Mode\ Crosslink\ Score" \equiv f_4(S_l) = \begin{cases} 1, & 125 \geq S_l \geq 105 \\ 0, & 105 \geq S_l \geq 125 \end{cases}$$

where S_l was the longitudinal relative distance between the two dimer.

- **Closed-ness score function**

Sampling methodology provides large number of models that would have gaps in the 3D structure. A correct 3D structure would be closed structure such that the last "tetramer of dimers" (Octamers) should be interacting with the first octamer. The interaction specified by the dimer-dimer configuration (relative shear, rotation of dimer 1, rotation of dimer 2) that could be from any of the possible modes of associations. An example of closed-ness was illustrated in (Figure 4-16).

$$"Closedness\ Score" \equiv f_5(m) = \begin{cases} 1, & m \in \{mode\ A, mode\ B\} \\ 0, & m \notin \{mode\ A, mode\ B\} \end{cases}$$

where m is a set of $(S_l, \Theta 1, \Theta 2)$ defining an interacting mode.

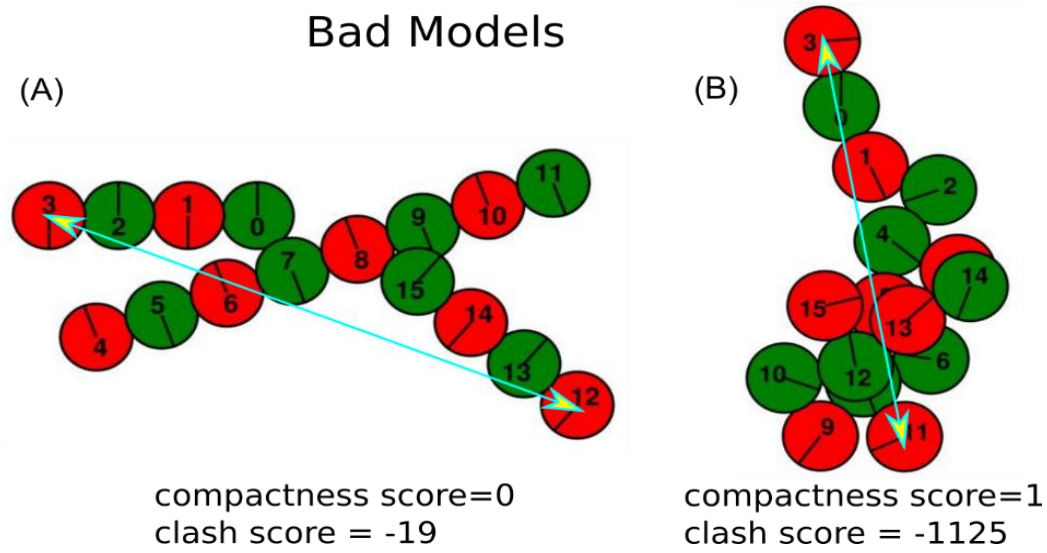


Figure 4-15 Schematic of clash and compactness scoring function.

A dimer is represented as a disc with head domain above (green) and below (red) the plane of paper. Numbers on the discs represents the dimer identity. Arrows indicate the farthest discs. Each model was assigned a score using various scoring functions. Examples of bad models (low scoring models). A) Clash free but not compact model. B) Compact model but have clashes.

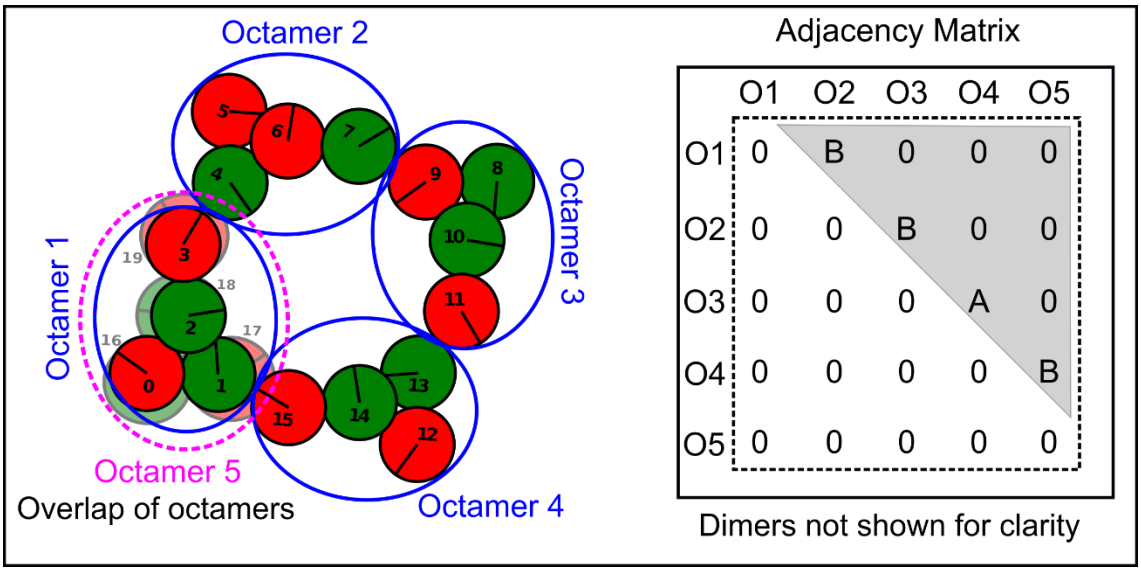


Figure 4-16 Schematic of closed-ness scoring function

A 16 dimer model sub-grouped into four octamers (O1 to O4) along with an additional octamer 5 (O5) that should overlap on top of octamer 1. A sample would be valid model if at least one disc from the octamer 1 and 5 overlap.

Final score of the given model is the product of all the scores.

$$F_{model} = f_1\, f_2\, f_3\, f_4\, f_5$$

Since all the scoring functions except f_1 provide binary values (either accept or reject a model), the scoring function were applied to the models one by one in the increasing order of time complexity. This reduces the execution time of the simulation and prevent sampling of invalid models. The optimization was done by using Disctransformation that generated 699 models satisfying all the scoring functions (summarized in Table 4-3).

Table 4-3 Scoring functions for sampling and scoring

Column 1 list the scoring function name. Column 2 provides a brief description of the scoring functions. Column 3 lists the number of models obtained after implementing the corresponding scoring function.

Scoring Function	Description	No of models
Soft Clash	Partial overlap of discs	2060883
Compact-ness	Maximum distance between two discs	441114
Hard Clash count	At most 1 hard clash allowed between discs	174895
Parallel mode	At least 1 disc pair have parallel mode shear	70724
Closed-ness	First and last octamer should interact with mode	699
Clustering	Hierarchical clustering in RMSE space.	6 clusters

4.8.4 Clustering of models

Every disc in the model has the coordinates of the disc center and the corresponding disc rotation. To differentiate the amount of rotation between the two overlapping disc, a fixed point on the disc edge at an angle of zero degrees was obtained. Thus, every disc in the model was defined using two coordinates (Disc center and a fixed point on the disc edge). For a 16 disc models, total 32 coordinates were utilized to determine Root Mean Square Error (RMSE). Assignment of coordinates for RMSE calculations were done based on the disc labels. All 699 models were hierarchically clustered in RMSE space obtained by all verses all RMSE calculations (Figure 4-17). The cutoff RMSE value (300) to determine the number of clusters were obtained by visual inspection of clusters. Models that had identical topology were placed in a single cluster.

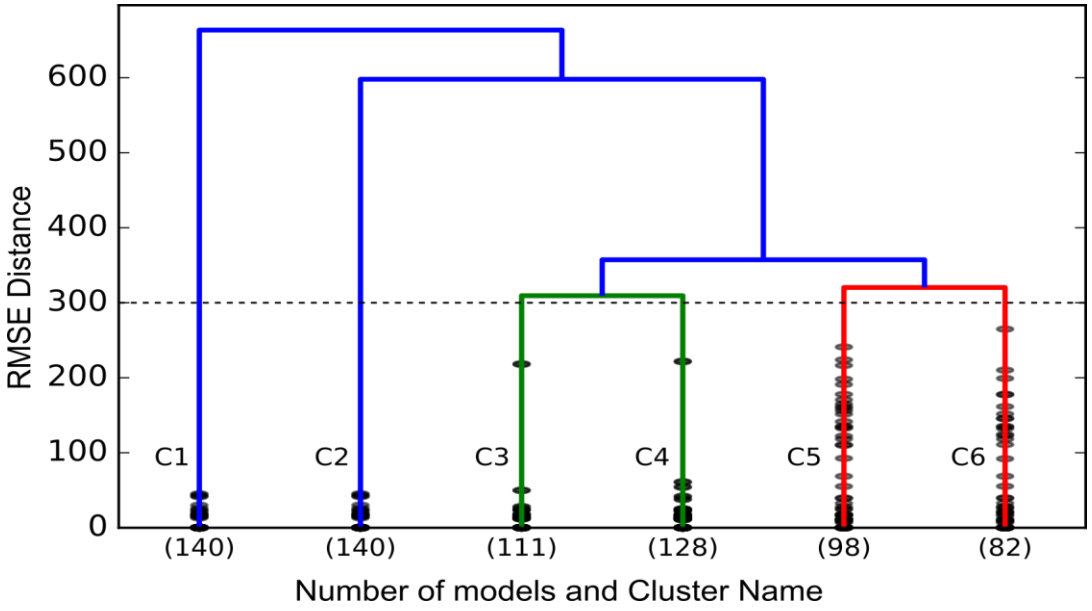


Figure 4-17 Hierarchical Clustering of models based on RMSE distance.

RMSE space distance (y axis) between the models were calculated using UPGMA method that provided six clusters (C1 to C6). The number of models in each cluster shown within brackets (x axis). Dotted line shows the cutoff value for deciding the number of clusters. All models below the dotted line were topologically identical.

4.8.5 Model Selection

The 699 models were clustered according to topological similarity (RMSE space). Total six clusters were obtained that satisfy all scoring restraints (see Section 4.8.2 and Section 4.8.3). Cluster 4 having 128 models (Figure 4-17 x-axis) was selected as a representative model of the keratin filament because of the distribution of head/tail domains (Figure 4-18). Since head/tail domains have net positive charge, a close proximity of similarly charged domains would be destabilizing. The shear values of models from cluster 4 positions head/tail domain spatially separated by around 115Å. Whereas in all other clusters head/tail domains were spatially proximal. Thus, cluster 4 provides the best arrangement of head/tail domains without having repulsions from similarly charged residues of head/tail domains (Figure 4-14).

Along with this, the cluster 4, cluster 5 and cluster 6 had a low-density region in the center of the filament as seen in experiments [172,220]. However, cluster 5 and 6 were ruled out due to the presence of crooked dimers (discs 0, 1 in C5 and disc 2, 3, 14, 15 in C6) that prevents the formation of regular filament bundles (see Section 5.1.3) [207,221]. Thus, only the models in the cluster 4 were chosen as appropriate for further validation.

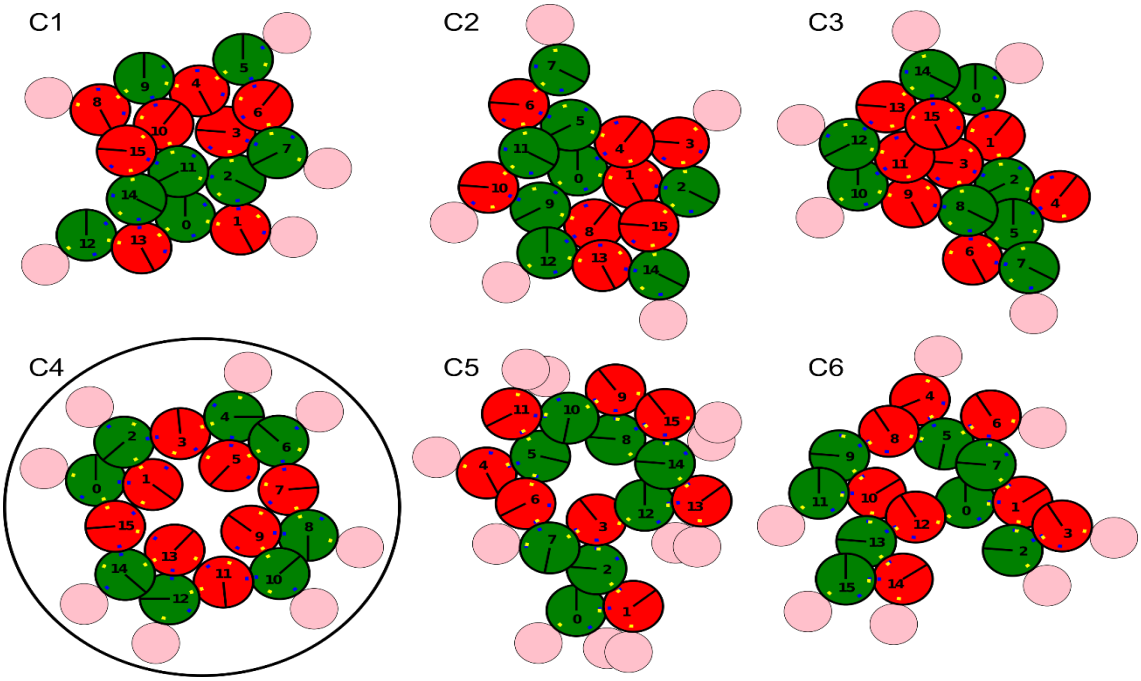


Figure 4-18 Clusters of models obtained after Disctransformation simulation.

Disctransformation simulation clusters topologically different models (green and red discs) into different clusters (C1 to C6). For all models, their possible associations with extra dimers are highlighted (pink discs). These extra dimers provide seed positions for filaments bundles to form. Encircled model cluster 4 (C4) was chosen as the representative model.

4.8.6 Mapping 2D disc models to full atom models

The top scoring cluster (cluster 4) obtained from the Disctransformation simulation was used to represent the filament model (Figure 4-18). This 2D model was converted to full atom models through reverse mapping of the Disc models to full atom models. The disc model provide relative rotations of discs along with relative shear values obtained through dimer-dimer associations. Using a full atom dimers and their interacting modes (relative shears and rotations), a Unit Length Filament (ULF) corresponding to disc model was constructed (Figure 4-19). Since in a ULF there are 16 dimers, 1st and the 17th dimer should overlap (see Section 4.8.2). A mapping of cross-section and longitudinal section provide the arrangement of dimers in a ULF. Six ULF models were constructed by longitudinally shifting the ULF model with the length of dimer. This shifting was also equivalent to interacting mode B2 (Figure 4-20). A cross-section of 3 unit length filaments was extracted to generate a 3ULF model having central ULF succeeded and preceded by ULFs (Figure 4-20).

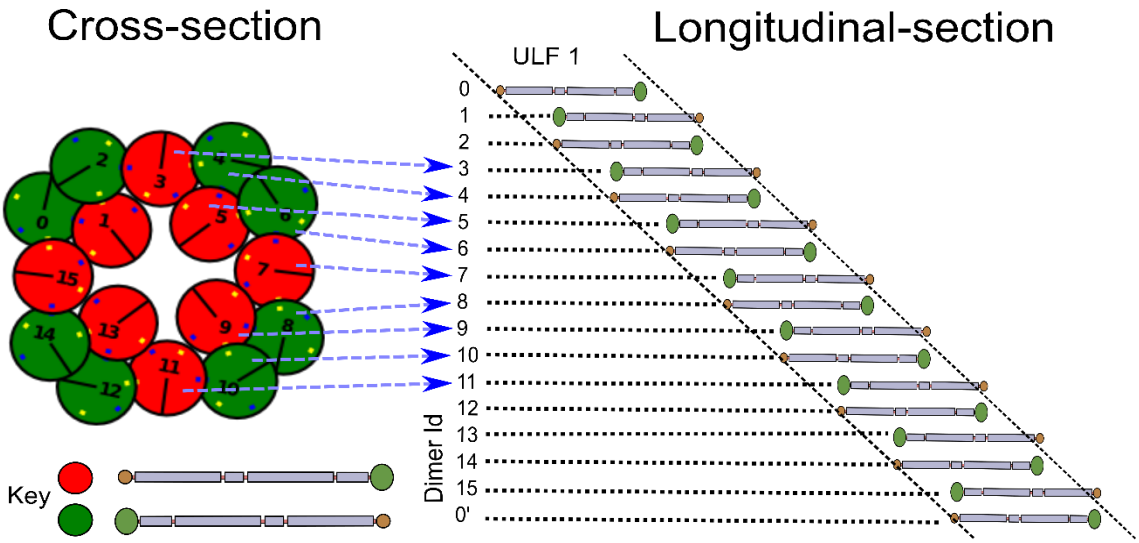


Figure 4-19 Reverse mapping of 2D Disc model to ULF model

Disctransformation model (red and green discs) provide cross-section arrangement of dimers. Shear values obtained from dimer-dimer associations provide longitudinal arrangement of dimers. Cross-section and Longitudinal sections were mapped (blue arrows) using dimer labels. A combination of the two sections provide a ULF model.

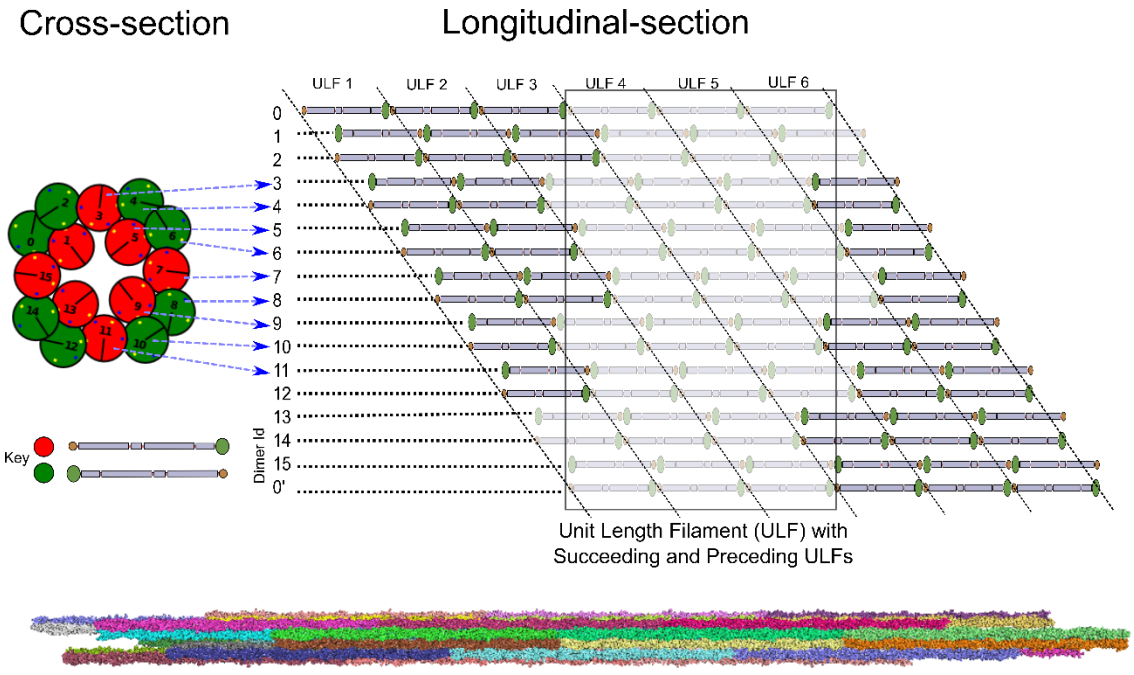


Figure 4-20 Reverse mapping of 2D Disc model to complete full atom models.

Disctransformation model mapped to longitudinal section as explained in Figure 4-19. *Longitudinal section was extended by using tetramer mode B2 (equivalent of end to end joining of dimer) to generate ULF2 to ULF 6. A section comprising of 3ULFs (rectangular box) were extracted (multi-color picture at bottom). This section has filament model with succeeding and preceding filament.*

4.8.7 Integrative Modeling Summary and Statistics

In this chapter, the Integrative modeling (see Chapter 1 for Introduction) of Keratin Intermediate filament was discussed. The different experimental and computational data related to Keratin filament provided restraints at different levels of resolution. Starting with the keratin K5 and K14 sequence, a K5-K14 dimer model was constructed from the bioinformatics data including homology modeling. Volume and mass measurements (obtained from SAXS data and sequence), provided the number of copies of K5-K14 dimers in the Keratin filament. Cross-linking data provided the information about how the neighboring dimers were arranged in the 3D space (interaction modes). Valid dimer-dimer interaction modes were chosen for further analysis and were analyzed for electrostatic complementarity. To get the filament model, the sampling of the 16 Keratin K5-K14 dimers was performed along with interaction modes such that all the available restraints were satisfied. A brief summary of Integrative modeling protocol applied for Intermediate filament was discussed (Figure 4-21).

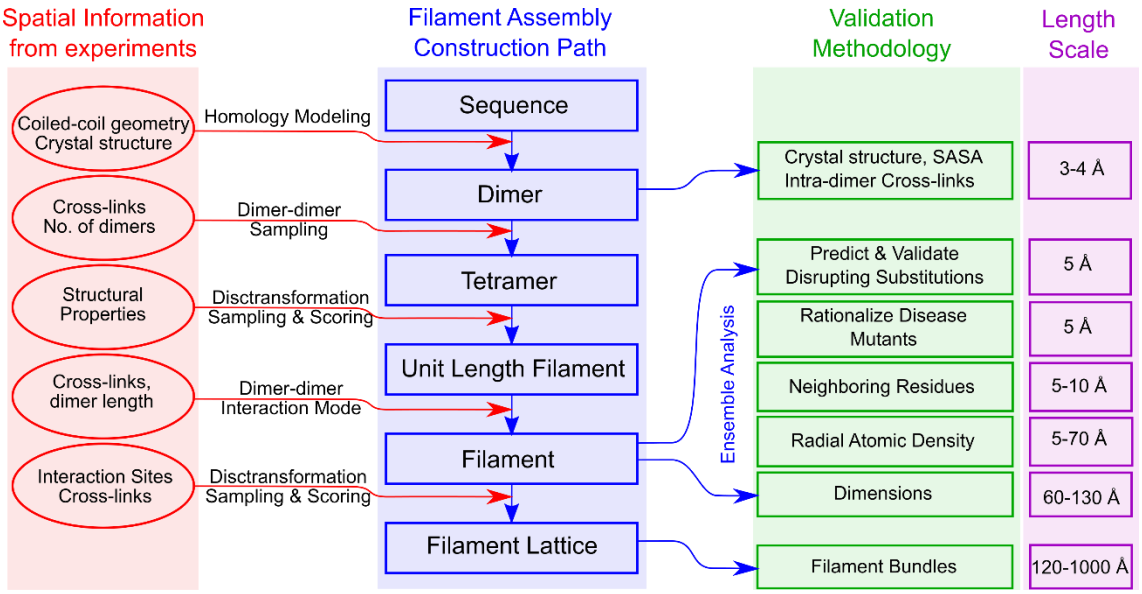


Figure 4-21 Summary of Integrative modeling protocol of Keratin Filament

Spatial information extracted from experimental techniques performed on Keratin and its homologs (red box). Spatial information incorporated with sampling and optimization techniques (red arrows) at different positions in the filament assembly. The obtained structural models were then used for the construction of filament at various length scales (blue panel). Every structure constructed through optimization was validated using different structural data (green box) and at different length scales (purple box). The filament structure was validated on an ensemble of models (blue arrows).

The model built in this work was among the largest near atomic resolution model ever created for Intermediate filament coiled-coil domains. The model also has low-resolution head/tail domains. The filament model consists of three Unit Length Filament (ULF) along the filament axis, a central ULF with preceding and succeeding ULF. Each ULF has 16 dimers arranged according to the interaction modes. A single K5-K14 dimer has 624 residues and ~10,000 atoms (with head/tail domains 1062 residues and ~15,000 atoms). A full atom keratin intermediate filament model has ~30,000 residues and ~500,000 atoms (with head/tail domains ~51,000 residues and ~720,000 atoms). The final full atom model was validated using the experimental data not used during the modeling protocol (see CHAPTER 5 for validation protocol).

4.9 Summary

The chapter discusses about the Integrative modeling protocol applied for determining the structure of Intermediate filament. Initially, the heteromeric K5-K14 coiled-coil dimer was constructed by homology modeling. The model was then validated using various structural properties and compared with available segments of crystal structure. Two computationally duplicated K5-K14 dimer models provide configurations that parsimoniously satisfy dimer-dimer cross-links observed in the experiment. Each keratin dimer was converted to a 2D disc by taking a projection on a plane. Disctransformation software then exhaustively sampled and scored all possible configurations (10^{16} possibilities). In all, 699 valid models were clustered into six topologically different configurations. The cluster with the optimal packing of the dimers was chosen as the final model of keratin filament for further validation.

CHAPTER 5

Validation of 3D structure of IF

SALIENT FEATURES

- ***Bundling of Filament***
  - Bundle Construction & Hexagonal Lattice
  - Comparison with TEM data
  - Improved filament-lipid model
- ***Dimension and Packing***
  - Dimensions of lower and Higher order structures
  - Packing of head/tail domains
- ***Radial Atomic Density***
  - Calculation of radial density distribution
  - Comparison with Cryo-EM data
- ***Cross-link Analysis***
  - MD simulation and Ensemble generation
  - Analysis of K5-K14 cross-links
  - Parallel mode cross-links relative frequency
  - Validation using homolog K1-K10 crosslinks
- ***Model Properties using Voronoi Volumes***
  - Classification of protein into chemical groups
  - Volume statistics of known protein structures
  - Comparison with volume statistics of K5-K14 filament model
- ***Summary***

The previous chapter provided the details of the Keratin K5-K14 filament model obtained through Integrative Modeling. However, the model needs to be validated extensively, which is an essential step of the Integrative Modeling. Since different experiments provide spatial information at various levels of details, individual components of the models have different resolutions. Thus, validation of the models at different length scales is an indicator of robustness. For instance, a range of experimental data providing information at different length scales (1000Å to 5Å) could validate keratin filament model (Figure 5-1). This chapter provides the details of the five major validation protocols. The filament model and filament bundles (length scale 1000Å to 120Å) were validated using Transmission Electron Micrographs (TEM) data, thereby providing details of the higher order organization of intermediate filaments. The overall model dimensions (length scale 130Å to 60Å) and the relative position of head/tail domains were validated with the experimentally observed values. The high-resolution details of the models such as atomic density distribution (length scale 80Å to 5Å) and neighboring residue pairs (length scale 10Å to 5Å) were validated using Cryo Electron Microscopy (Cryo-EM) data and cross-linking data (Figure 5-1). Each of these experimental data validates the filament model at various length scales, thereby indicating the robustness of the obtained models. The next few sections provides the validation protocols applied to the keratin filament.

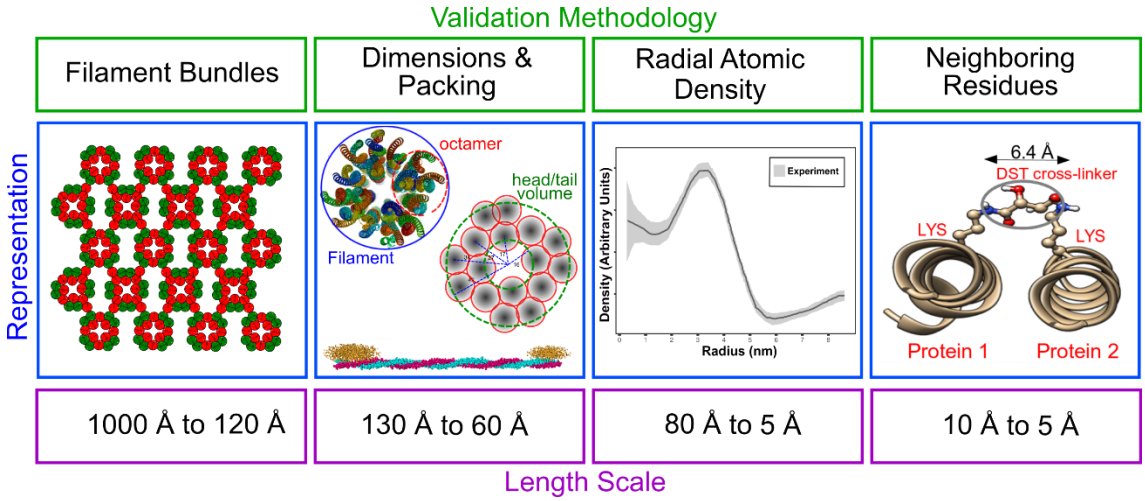


Figure 5-1 Validation performed across different length scales.

Keratin K5-K14 Intermediate filament model was validated using spatial information (green and blue boxes) obtained from the experiments performed at different length scales.

5.1 Bundling of filaments

5.1.1 Bundle construction

A single filament model was constructed by joining interacting dimer pairs in all possible ways (see Section 4.8). However, peripheral dimers of the model have exposed interacting sites for other dimers to bind (Section 4.8.3 and Figure 4-12). These interactions were essential for the formation of higher-order structures by providing seed positions for inter-filament interactions. To validate the filament model, a filament lattice was constructed through its exposed interacting sites (Figure 5-2 A). All the exposed interacting sites were connected to dimers through Mode A or Mode B (Figure 5-2 B). These exposed dimers were then connected to another copy of the filament model. The procedure was repeated for all exposed sites. The filament bundle models that have inter-filament clashes were removed from the possible solutions. Models having a regular geometry were chosen, as it provides an effective packing of filaments (Figure 5-2 C).

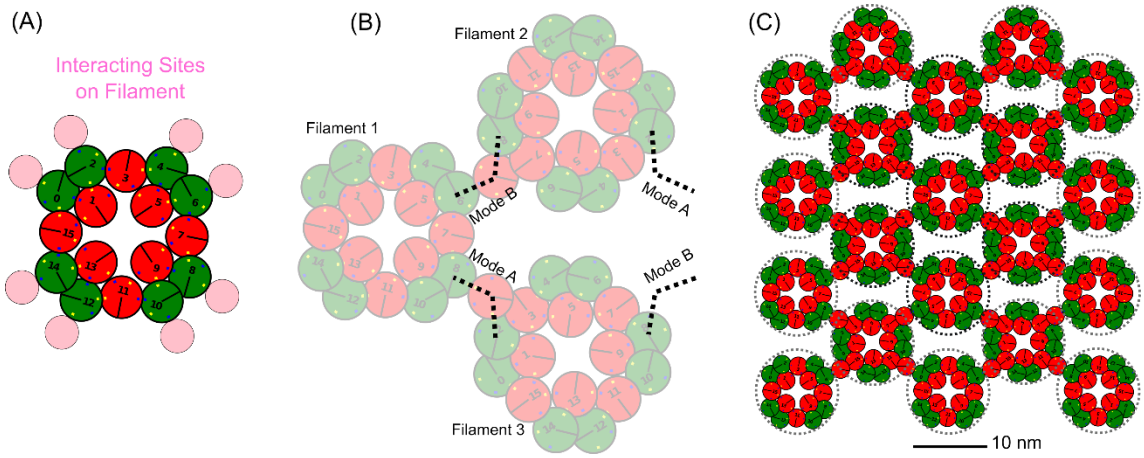


Figure 5-2 Filament lattice construction methodology

(A) Keratin filament model (red/green discs) shown along with possible dimer positions representing filament interacting sites (pink circles) on the periphery of the filament. (B) Interacting modes (Mode A and Mode B) between three keratin filaments through the exposed interacting sites. (C) Filament bundle model constructed that have a regular geometry of filaments (dotted circle).

5.1.2 Hexagonal lattice

The filament lattice model constructed in the previous step were analyzed for packing properties. The lattice vectors of a hexagonal lattice should follow standard geometry rules (Figure 5-3 A). Lattice vectors *a* and *b* should be equal and together define the

hexagon plane. Whereas vector *c* defines the distance between two parallel hexagon planes. The angle γ formed by vectors *a* and *b* should be 120° whereas angles α and β should be 90° (Figure 5-3 A). The filament bundle follows the hexagonal geometry with keratin filaments occupying vector *c* (Figure 5-3 B). The vectors *a*/*b* define the hexagonal plane (formed by the filament cross-section or 2D models) and their magnitude provides the inter-filament distance in the filament bundle lattice (Figure 5-3 C). This distance along with the hexagonal packing were used to compare with the TEM experiment provided measurements of keratin bundles.

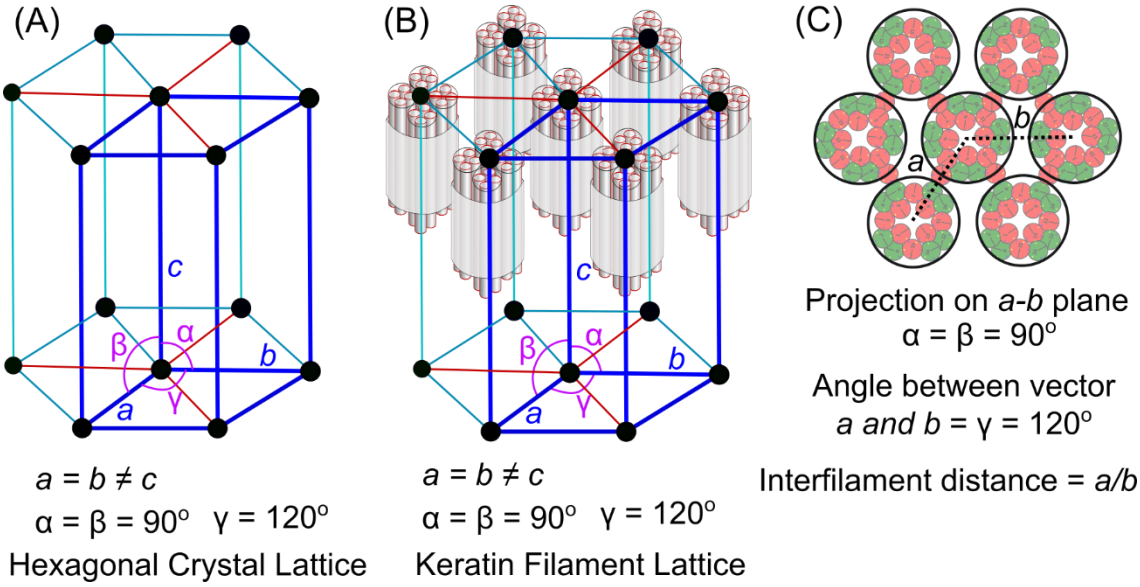


Figure 5-3 Hexagonal lattice of keratin intermediate filament bundle model.

(A) A hexagonal lattice showing lattice vectors and angles between them. (B) Arrangement of keratin filaments in hexagonal arrangement obtained from bundle model. (C) A 2D projection of hexagonal filament bundle lattice along the vector c on a-b plane with lattice parameters.

5.1.3 Comparison with TEM data

The keratin filament bundle obtained as a consequence of interacting filaments had a hexagonal packing (Figure 5-3). In the models, the filament center-to-center distance was measured as ~10 nm. This was in agreement with the Transmission Electron Microscopy (TEM) images of keratin intermediate filaments obtained from native vitreous epidermis. The filament appear to be packed in hexagonal arrangement with filament center to center distance in the range of 9-13 nm with median as ~11 nm calculated over 10 samples (Figure 5-4) [207,221]. The variability in the measurements could be attributed to

experimental procedures (chemical fixation methods, resolution etc.) and filament interaction with other molecules (amino acids, water, lipids etc.).

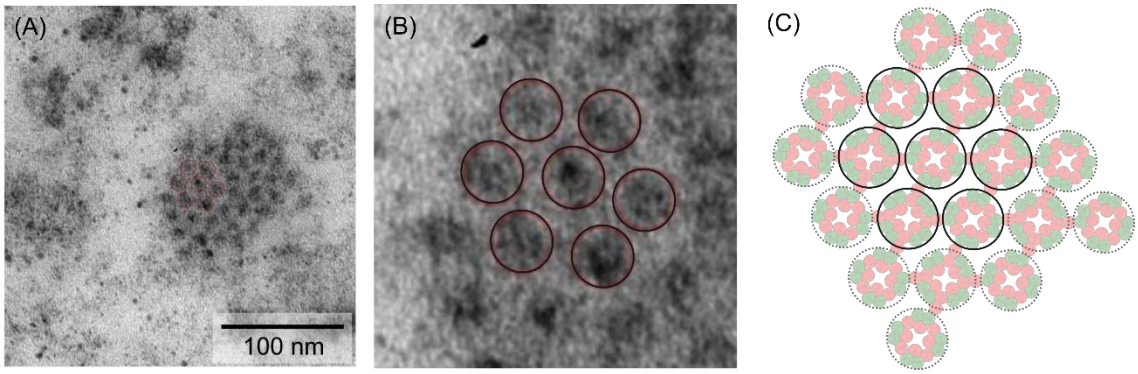


Figure 5-4 Comparison of filament bundle models packing with TEM images.
(A) TEM images of Keratin intermediate filaments obtained from native vitreous epidermis. (B) A zoomed image of the bundle highlighting each filament (red circles). Packing of the filaments shows hexagonal arrangement. (C) Filament bundle model highlighting the hexagonal packing (black circles). (A) and (B) adapted from [221].

5.1.4 Improved filament-lipid model

Keratin filaments were known to have close association with lipid molecules/membranes [207,222,223]. Specifically, filament head domain has a strong binding affinity towards lipid molecules/membranes [224]. Cryo-transmission electron microscopy images of keratin intermediate filaments provide significant insights into the higher order organization of keratin filaments. A filament-lipid model was suggested where cornocyte keratin filaments were arranged in hexagonal lattice surrounded by lipid membrane [207]. These filament bundles upon prolonged exposure to water molecules lead to ~25% increase in the thickness of filament bundles. However, in dehydrated filament bundles, the distance between filament edges reduced to a minimum [225]. Existing filament-lipid model explains most of the observations relating to keratin filament bundles, however, it cannot explain the ~25% swelling of bundles upon prolonged exposure to water (known as swelling behavior).

The keratin bundle model constructed in this study was in agreement with filament-lipid model. The model mimics the partially dehydrated state and provides enough space (~2.5 nm for bi-lipid non-polar core) for lipid molecules to interact with the filaments (Figure 5-5 A, B). The interaction with the lipid molecules were only possible with the peripheral

dimers of the filaments (Figure 5-5-B gray colored region). The filament anchor points (dimers connecting two filaments) in the bundle remain inaccessible to lipid molecules. Thus, inter-filament distances (~10nm) remain unchanged with only lipid molecules. However, with prolonged exposure to water molecules, interactions of the filament anchor points would be replaced by water mediated interaction (charge/hydrogen bonds). Lipid molecules would eventually replace these water-mediated interactions and form a single lipid layer. This lipid layer being in the proximity to other lipid layers around the filaments forms a bi-lipid layer cage in the filament bundles (as seen in Cryo-EM images). The addition of bi-lipid layer hydrophobic core of length ~2.5nm at the filament anchor points increases the inter-filament distance from ~10nm to ~12.5 nm (25% increase). The phenomenon thus explains the ~25% swelling of the filament bundles upon prolonged exposure to water and lipid molecules.

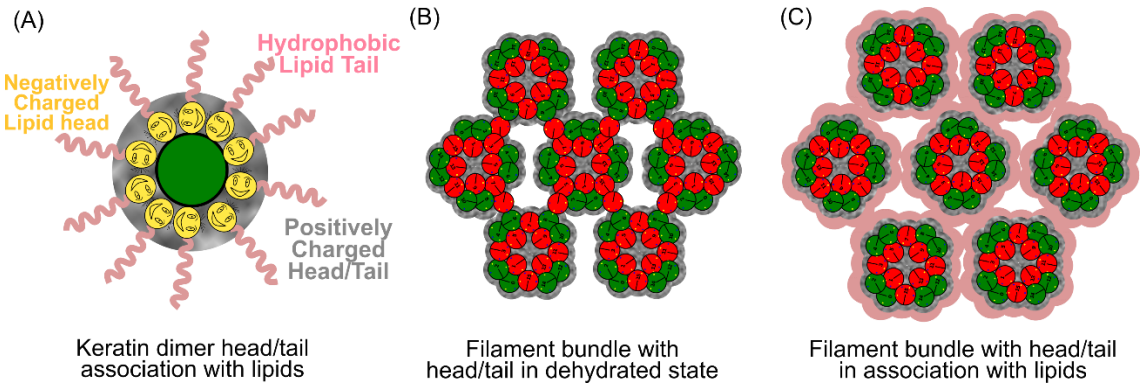


Figure 5-5 Schematic of filament-lipid model with bi-lipid layer around filament.

(A) Schematic diagram showing the interaction of positively charged head/tail domain (gray circle) of a keratin dimer (green/red disc) with negatively charged lipid heads (yellow circle). (B) Filament bundle along with head/tail domains (gray) in the dehydrated state. (C) Schematic of the expanded filament bundle upon exposure to lipid molecules. Hydrophobic tail of lipid molecules (pink) around the filament forms bi-lipid layer with lipid molecules bound to the neighboring filament.

The cryogenic electron micrographs (Cryo-EM) provide a low-resolution information of keratin filaments higher order organization (filament bundles). However, it provides valuable information for the validation of filament models. Since the bundles were constructed directly from dimer-dimer interactions, their agreement with the experimental data validates the model.

5.2 Dimensions & Packing

5.2.1 Dimensions

Previous section validates the filament model at length scale of filament bundles (see Figure 5-1 for length scales). One of the most important validation was the comparison of dimensions of the filament and lower order structures (tetramers and octamers). Due to low-resolution data/images provided by various experimental techniques, it was difficult to measure the exact dimensions of the filaments and sub-structures. However, rough estimates from the experimental data provided dimensions of filament and lower order structures (Table 5-1).

Table 5-1 Comparison of keratin filament and sub-structures dimensions

Comparison of the octamer/filament dimensions obtained from the experiment and models (with and without head/tail domains). Experimental values were obtained from [226].

	Experiment	Model	Model with Head/Tail
Octamer	~6 nm	~5.56 nm	~6.50 nm
Filament	9.4 ± 2.7 nm	~9.22 nm	~12.50 nm

None of the dimensions mentioned in Table 5-1 were considered explicitly during the intermediate filament modeling. However, the dimensions of the selected model agree to the observed values. Other possible models in the clusters (other than C4) obtained during the integrative modeling would have provided dimensions that would not be in agreement with the experimentally determined data (see Figure 4-18 and Section 4.8.5). Thus, choosing the cluster 4 (C4) as the selected model was appropriate and best choice.

5.2.2 Dimension and packing of head/tail domains

Keratin filament models constructed using Integrative modeling approach did not involve positioning of head/tail domains with respect to the dimer orientation. Head/tail domains were constructed by considering them as diffused density of atoms (see Section 4.4). However, the analysis of the full atom models (and 2D models) provides us with valuable information about the relative positioning of head/tail domains. In the filament models, the dimers occupying the central positions (blue outline discs) were surrounded by four peripheral dimers (Figure 5-6 A). The head/tail domains of these buried dimers could only

fit in the central cavity (dashed blue circle). To validate this, the volume of the cavity was calculated and compared with the volume required for head/tail domains of four buried dimers. The following paragraph outlines, the steps for volume calculations (refer to Figure 5-6 B).

Since, the length (perpendicular to the plane of paper) of a single dimer was about 47nm.

$$Volume\ of\ dashed\ cylinder\ V_{cyd} = \pi(1.7^2)47 = 426.9\ nm^3$$

$$Volume\ of\ a\ dimer\ V_d = \pi(0.95^2)47 = 133.3\ nm^3$$

$$Volume\ of\ a\ shaded\ region\ V_s \cong V_d/2$$

$$Volume\ of\ cavity\ V_{cav} = V_{cyd} - 4V_s = 426.9 - 4(133.3/2) = 160.3\ nm^3$$

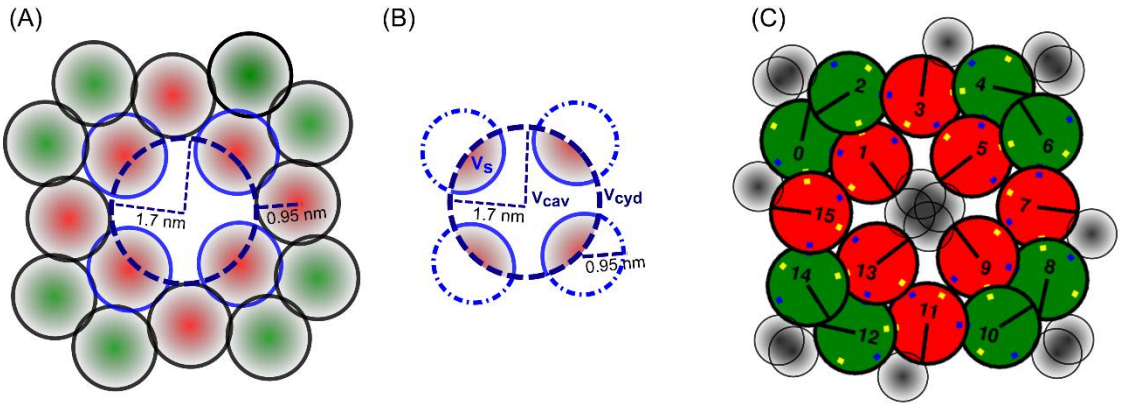


Figure 5-6 Relative position of head/tail domains in keratin filament model.

(A) Filament model without head/tail domains had central cavity (dashed circle) size of approximately 1.7nm radius. (B) Measurements for determination of cavity volume V_{cav} using volume of the overlapping dimers (V_s) and volume of the dashed cylinder (V_{cyd}). (C) Predicted positions of head/tail domains (shaded circles) in the filament models. Overlapping circles (head/tail domains) in 2D arrangement were axially separated by the relative shear between the discs.

The molecular weights of K5-K14 head/tail domains were calculated from the sequence. K5_Head = 16081.6 Daltons, K5_Tail = 9563.2 Daltons, K14_Head = 10284.1 Daltons, K14_Tail=5065.5 Daltons.

Assuming the volume of globular proteins as 1.01 $Å^3$ per Dalton [168,214].

The Volume K5_Head = 16.24 nm^3, Volume K5_Tail = 9.66 nm^3, Volume K14_Head = 10.39 nm^3, Volume K14_Tail = 5.12 nm^3.

Total volume of head/tail domain = 41.41 nm^3. Total volume of head/tail domain of four dimers = 165.64 nm^3.

The volume of the cavity (160.3 nm^3) matches closely with the volume of the head/tail domains (165.64 nm^3) of four internal dimers (blue outlines dimers Figure 5-6 A) as predicted by the model. A close inspection of the internal dimers orientation reveals that all four dimers have same orientation relative to the cavity (Discs 1, 5, 9, 13 Figure 5-6 C). Thus, a favorable position of head/tail domain was near the end of line of discs/dimers (Figure 5-6 C). For peripheral dimers, identical relative position of head/tail domains was considered to build the full filament models for further analysis.

Other arrangements such as all head/tail domains (volume 662.56 nm^3) occupying central cavity were ruled out based on the volume of the cavity (volume 160.3 nm^3) available for head/tail domains. Only the arrangement discussed in Figure 5-6 C favors the optimal packing of head/tail domains.

5.3 Radial Atomic Density

5.3.1 Calculation of radial density distribution

Radial atomic density organization of intermediate filaments provides a comprehensive knowledge of atomic packing across radial axis. Given that the intermediate filaments are made up of multiple copies of dimers, a radial atomic density profile would provide an indirect measurement of dimer packing along the axis perpendicular to filament axis. Since the model built using integrative modeling protocol do not have head/tail domains, an unstructured density of atoms equivalent to head/tail domains for every dimer were constructed (see section 4.4 and section 5.2.2 for details) (Figure 5-7 A). The obtained filament model with head/tail domains was then projected on a plane perpendicular to the filament axis (Figure 5-7 B). The plane was divided with concentric rings (centered on the mean coordinates of filament) of radius 0 to 8 nm and width 0.1 nm. The molecular atomic density for each ring was calculated by adding the molecular weight of all the atoms present in the ring divided by the area of that ring. The following section describes the comparison of this atomic density with the experimental data of trichocyte keratin filaments [172,220].

5.3.2 Comparison with the known experimental data

Radial density of trichocyte keratins were calculated from cryo-EM micrographs by averaging over various segments along the filament axis [220]. The radial density profile

was then compared with the profiles calculated for the K5-K14 intermediate filament model with head/tail domains (Figure 5-7 C). The radial density of the models was in agreement with the experimentally calculated profile. The two profiles however deviate near the center of the filament (radius 0-2 nm) which was occupied by the head/tail domains (Figure 5-6). The experimental profile near the center of the filament also shows the larger error bar indicating the diffusive/flexible head/tail domains. Since head/tail domains were represented as unstructured density, an accurate head/tail domain models in the context of filament structure would provide better estimates of radial density.

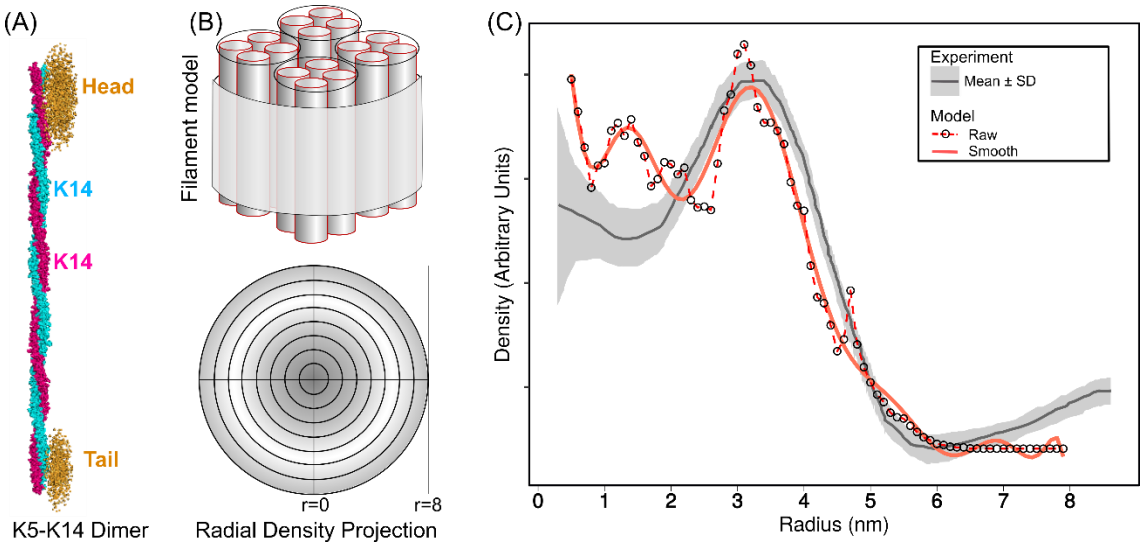


Figure 5-7 Comparison of radial atomic distribution of K5-K14 filament.

(A) Keratin K5-K14 coiled-coil dimer (cyan and pink) with head and tail domains (dark yellow) as diffusive atomic density. (B) Schematic of filament model (gray bundle of dimers) showing radial density projection (shaded concentric rings). (C) Comparison of radial density profile of K5-K14 filament with the experimental data obtained from [172,220].

The comparison made here was between trichocyte keratin (hard keratins) and K5-K14 keratin model (soft keratin). Hard keratin was known to have better packing of head/tails domains compared to soft keratin. The head/tail domains of keratin dimers interact with succeeding and preceding dimers shown by the cross-linking experiment [205,227]. In hard keratin filaments, the absence of water increases the electrostatic attraction between positively charged head/tail domains with negatively charged coiled-coil domains. The stronger electrostatic attraction between the two domains leads to closer packing of head/tail domains with the neighboring dimers. This leads to unavailability of head/tail domains of internal dimers in the core of the filament (Figure 5-6 C). The redistribution

of the head/tail domains of hard keratin results in the lower atomic density in the core compared to soft keratins (Figure 5-7 C). Similar redistribution of head/tail domains of peripheral dimers leads to the lesser radius of the hard keratin filaments as compared to soft keratins (Table 5-1) [220]. The keratin K5-K14 filament model was thus consistent with the previously proposed idea that coiled-coil domains were surrounded by the head/tail domains [228].

5.4 Cross-link analysis

Chemical cross-linking with an appropriate cross-linker could provide spatial information at high resolutions. The modeling of the keratin K5-K14 filament model was dependent upon the cross-linking of lysine amino acids through DST cross-linker [203,204,206]. Fourteen cross-links (Mode A and Mode B) obtained on intact keratin filaments were used to construct the filament models (Table 4-1). However, cross-links that were observed in tetramer and octamers structure (Mode C) were not used during modeling procedure. All these K5-K14 cross-links (including intact filament cross-links) had to be satisfied in the model. Apart from these, nineteen cross-links obtained in a similar cross-linking experiment done on homologous K1-K10 keratin filament should also be satisfied through mapping of amino acids [151]. Next few sections discuss the methodology used to validate the models using cross-linking data.

5.4.1 MD Simulation for generating ensemble

To ensure that the cross-links satisfied in a range of configurations, an ensemble of K5-K14 keratin filament models were generated through molecular dynamics simulation. A three unit length filament (3ULF) model without head/tail domains was used for generating ensembles. The MD simulations were carried out using GROMACS [192–194] with the Gromos54a7 force field [229,230]. A rhombohedral water box was generated around the filament with explicit SPCE water model having boundary extended by 10 Å in each dimension. The system was neutralized by replacing solvent molecules with 32 sodium ions. Particle mesh Ewald sum method [231] was used for electrostatic interactions and LINCS [232] was used to constraint hydrogen bond lengths. The whole system was minimized for 1 ns with integration time step of 2 fs using steepest gradient algorithm. The system was then equilibrated with NVT ensemble simulation followed by NPT simulation at 300K for 1 ns. The production run was simulated for a 50 ns and the trajectory was recorded at every 100 ps. Total ~800 GB of trajectory generated during simulation containing ~3.2 million atoms with 312864 protein atoms. Coordinates of the

filament at every 200 ps were extracted to get 100 snapshots of the filament in the PDB format.

5.4.2 MD Simulation trajectory analysis

Residue wise temperature factor are representatives of thermal fluctuations during the simulation. Average temperature factor of the 3ULF model and central ULF was determined by superimposing C^{α} atoms of all three ULFs and then measure the fluctuations for all snapshots. This procedure was repeated for central ULF2 independently (Figure 5-8). The amino acids in the model shown in sphere representation were colored according to temperature factor (blue to red). End domains of dimers that do not have proper interactions due to missing head/tail were more likely to fluctuate (red) compared to central domain (blue) (Figure 5-8).

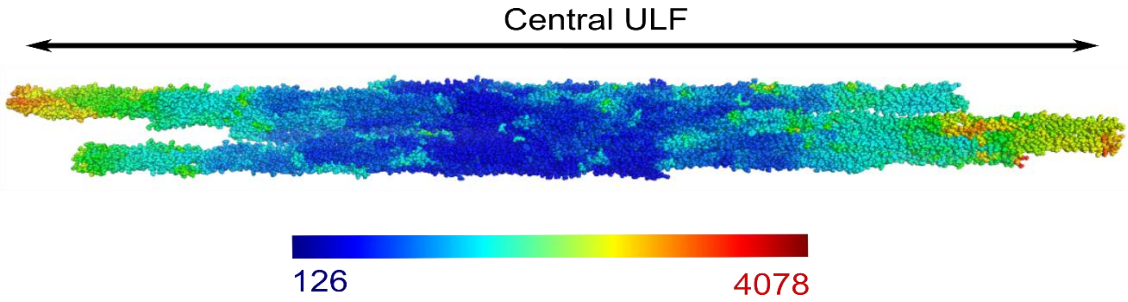


Figure 5-8 Temperature factor of central unit length filament (ULF).

Central unit length filament (ULF) colored according to average temperature factor (blue to red) of amino acids obtained during the simulations.

To determine the deviation of the filament during the simulation, root mean square deviation (RMSD) was measured with respect to the initial structure. RMSD calculations were done for the filament (3ULFs) and the central ULF2 independently. Both RMSD calculations show sudden increase in RMSD values at around ~2.0 ns and then gradual decrease with the saturated RMSD values of about ~1.5/2.0 nm (Figure 5-9). The sudden increase in the RMSD values were due to the unfolding of terminal amino acids in the coiled-coil dimers. This unfolding pushes succeeding and preceding dimers longitudinally along the filament axis leading to increase in filament length (Figure 5-10 at 2ns). However, the increase in length was reversed due to N terminus – C terminus dipole interaction of coiled-coil dimers (Figure 5-10 from 5ns onwards). The movement of filament backbone diminished after 30ns, indicating the stability of the overall structure. Since the missing positively charged head/tail domains were known to interact with negatively charged coiled-coil domains, their addition would have further increased

the stability of the filaments [205,227]. Hundred snapshots from the trajectory at equal intervals were extracted to determine lysine-lysine cross-links.

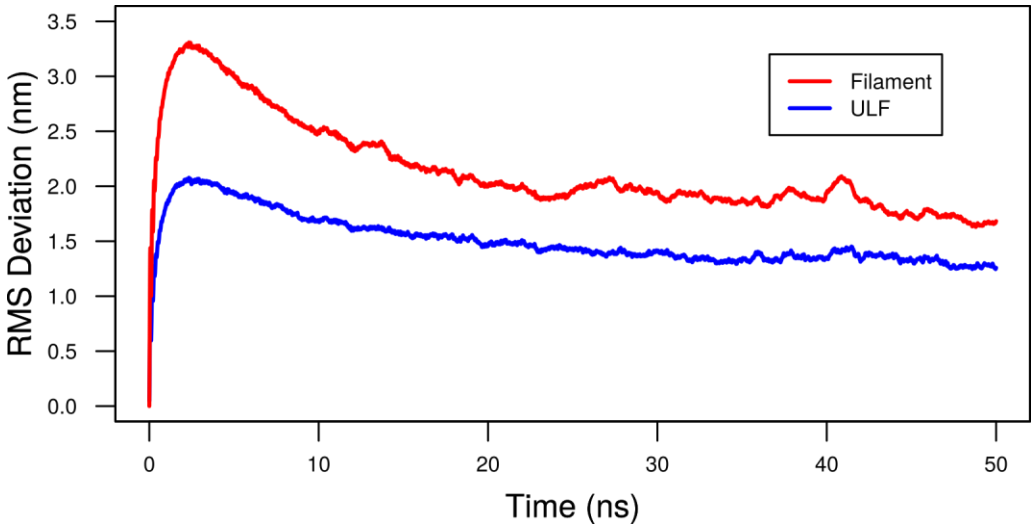


Figure 5-9 RMSD plot of Filament and central ULF.

Root Mean Square Deviation (RMSD) (y axis) of filament model having 3ULFs (red line) and RMSD of central ULF (ULF2) (blue line) during the course of simulation (x axis). Both curves shows increase in RMSD at around 2-3 ns and then gradual decrease.

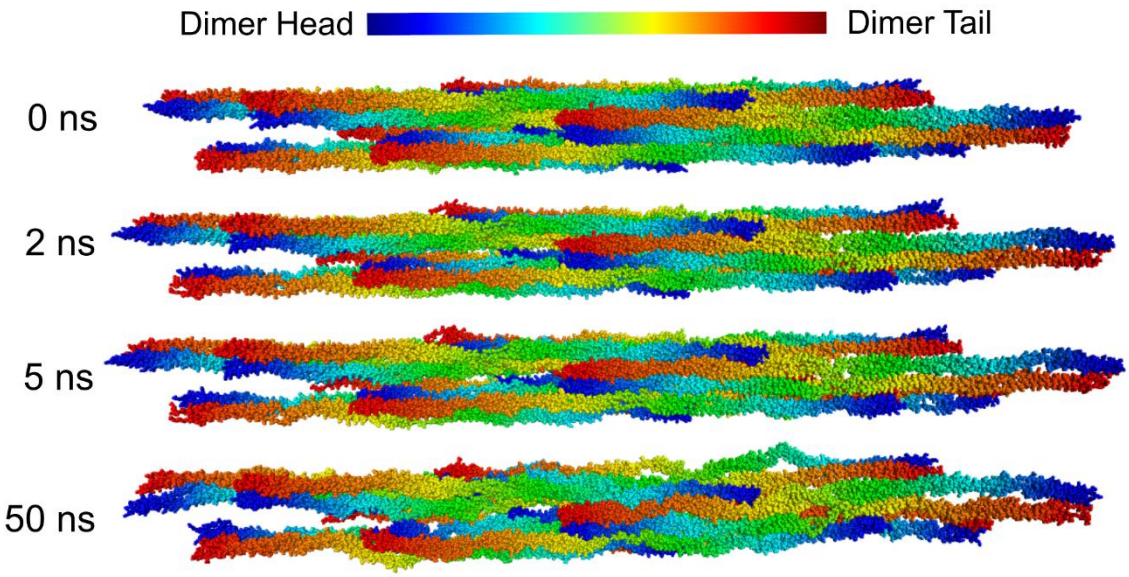


Figure 5-10 Snapshots of MD simulation trajectory at various time steps.

At various time steps (0, 2, 5, 50 ns) a snapshot of central ULF with each dimer colored from blue (dimer head) to red (dimer tail). For a movie of complete 50 ns MD simulation trajectory, refer to the file: IF_MD_Movie.mp4 in Supplemental Material.

The following section discusses about the cross-links used to construct filament model and the validation using all other unused cross-links (including the K1-K10 cross-links). This would further help validate the filament model at amino acid level.

5.4.3 Analysis of K5-K14 cross-links

As discussed in the previous sections (section 4.6.2 and Figure 4-9), K5-K14 cross-links were sub-grouped according to dimer-dimer configurations as mode A, mode B (mode B1 + mode B2) and mode C (mode C1 + mode C2). These dimer-dimer configurations were based on the lysine-lysine N^{ζ} distances restrained to 6.4Å (cross-linker arm length). The maximum C^{α}-C^{α} distance of two cross-linked lysine was ~19.4 Å (see Section 4.6.1 and Figure 4-7 B). Since the lateral distance between the two dimer (each one having a cross-linked lysine) in the filament was unknown, they were laterally separated with the minimum possible distance of 18.5 Å (see Section 4.6.1). Approximately 2Å relaxation was given for distance measurements. Lysine being a positively charged amino acid, two lysines repel each other during the simulation. Thus, considering N^{ζ}-N^{ζ} distance for determining the possibility of a cross-link would not be appropriate. Hence, C^{α}-C^{α} distance with upper limit of 22Å and C^{β}-C^{β} distance with upper limit of 19Å was chosen as cross-linking criteria. Any two lysines following the above criteria were assumed to be cross-linked.

During the simulation all possible lysine-lysine distances following the distance criteria were recorded. As a first check, the mode A and mode B cross-links that were used to construct the filament models should be satisfied according to distance criteria. The distribution of distances between two cross-linked lysine of mode A and mode B would indicate the extent of cross-link formation during the simulation (Figure 5-11). All mode A and mode B cross-links that were used in the construction of the filament satisfied the distance criteria. All mode A and mode B cross-linked lysines were checked for solvent accessibility using DEPTH [233–235]. The lysine should be exposed to the solvent in the filament model to covalently bind the DST cross-linker. A depth value cutoff of 6.0Å was used to determine solvent accessibility of the cross-linked lysine. Depth distribution of mode A and mode B cross-links shows that a significant number of cross-linked lysine pairs were exposed to the solvent (Figure 5-12). Thus, all mode A and mode B lysine pairs were likely to get cross-linked.

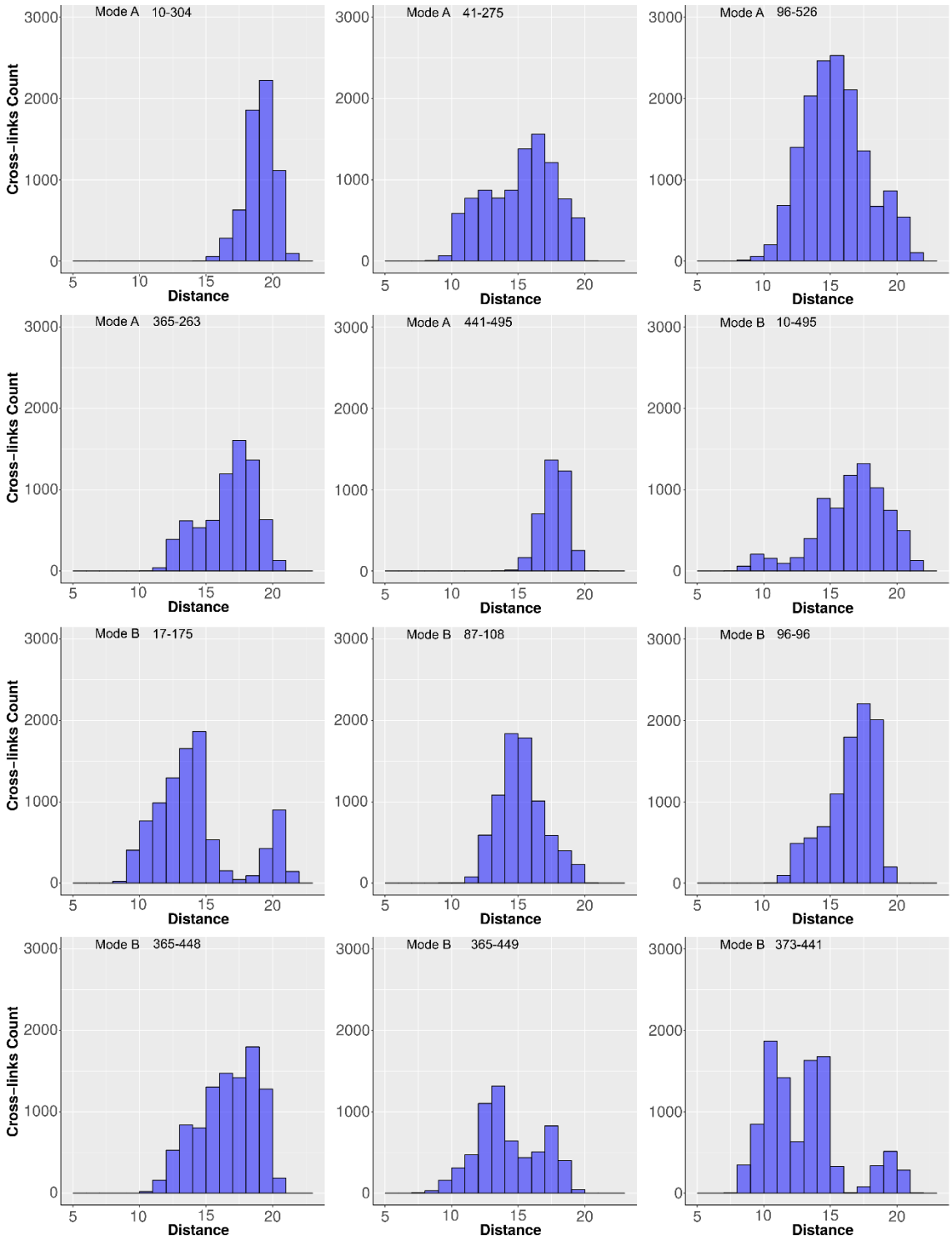


Figure 5-11 Cross-link distance distribution in Mode A and Mode B

Distribution of number of cross-link (y axis) with respect to C_α- C_α distance (x axis) between the cross-linked lysine amino acids. Mode and cross-linked lysines labeled at the top of each sub-figure. See Table 4-1 *for details of cross-linked lysines.*

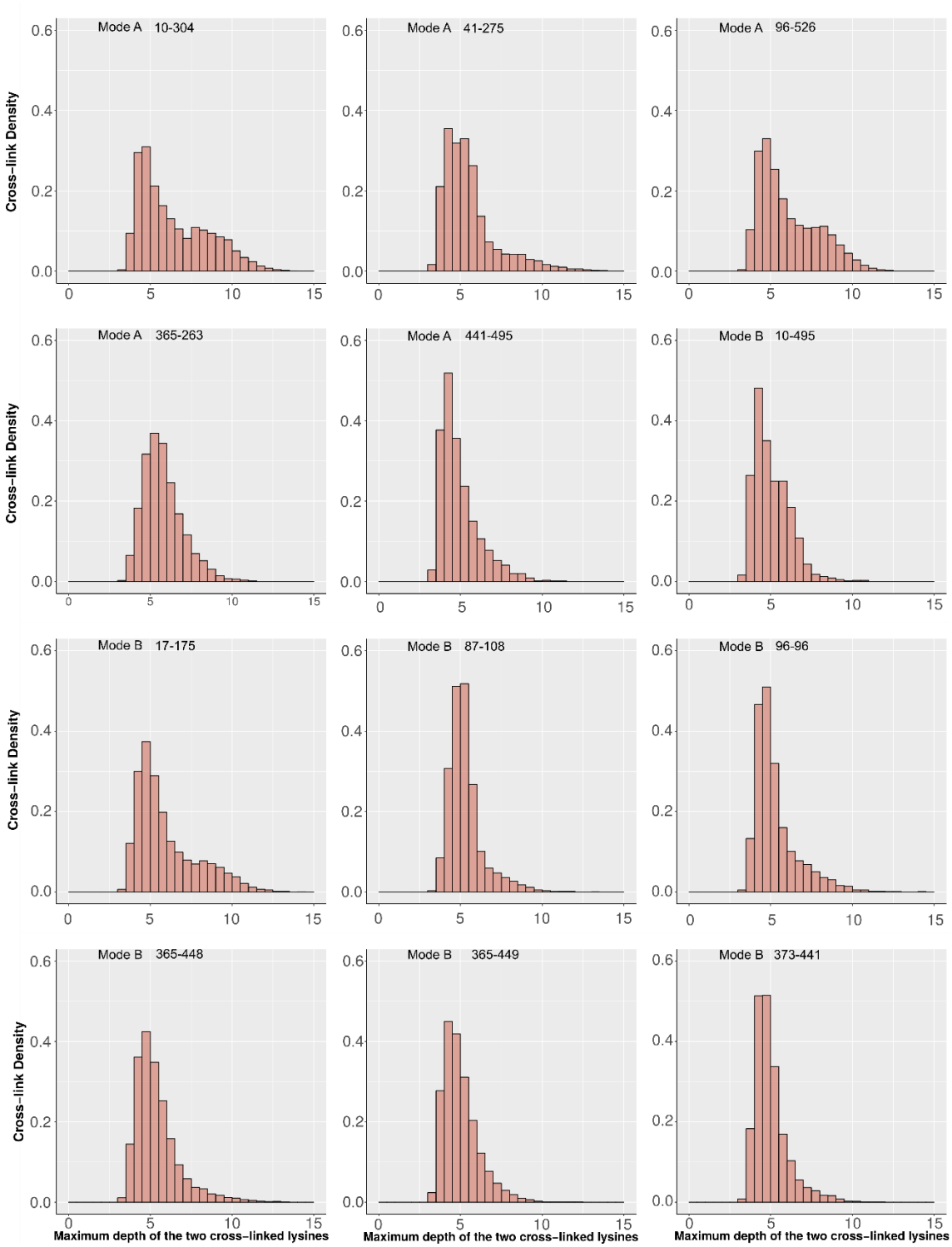


Figure 5-12 Cross-links depth distribution in Mode A and Mode B

Distribution of cross-linking density (y axis) with respect to maximum depth among two cross-linked lysine amino acids (x axis). Mode and cross-linked lysines labeled at the top of each sub-figure. See Table 4-1 *for details of cross-linked lysines.*

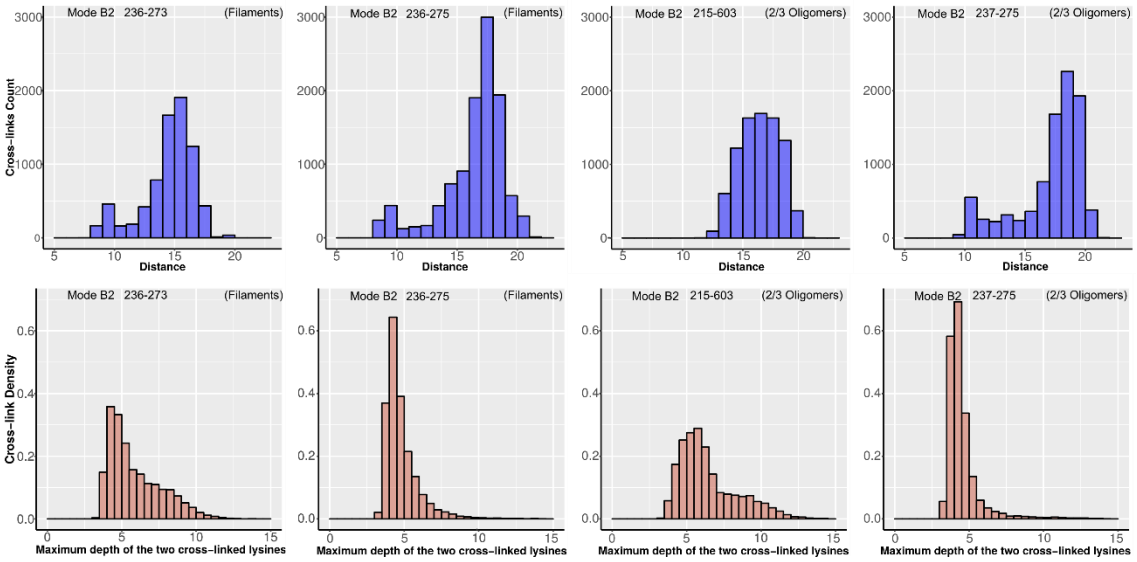


Figure 5-13 Distance and depth distribution in Mode B2 cross-links.

Distance (blue histograms) and depth distribution (red histograms) of mode B2 cross-links formed in the 2/3 oligomer and in the intact filaments.

5.4.4 Analysis of K5-K14 cross-links not used for filament construction

Mode A and mode B (excluding mode B2) cross-links were used to construct the filament model. However, there were more cross-links observed in the cross-link experiment that were not used in the modeling procedure [204].

During the experiment, mode B2 cross-links were obtained in tetramers, octamers and filament (See Table 4-1). In the simulation trajectory, mode B2 showed significant number of cross-linked pairs after applying both the distance and depth cutoff (Figure 5-13). Since mode B2 was just the longitudinal translation (equal to dimer length) of a dimer in mode B (Figure 4-9), the satisfaction of mode B2 cross-links validates the longitudinal separation between two dimers.

Other than mode B2 cross-links, cross-links formed in tetramers and octamers were also obtained (sub-grouped in mode A/B) in the experiment [204]. These cross-linked lysine pairs were not utilized in the filament construction. However, they were analyzed after applying the distance and depth cutoff as done for mode A and mode B filament cross-links. Mode A and mode B tetramer/octamer cross-links were also satisfied according to distance and depth criteria (Figure 5-14). It was hypothesized that though these lysine pairs satisfy the cross-link criteria, they could not form cross-links in the experiment due to their interaction with head/tail domains.

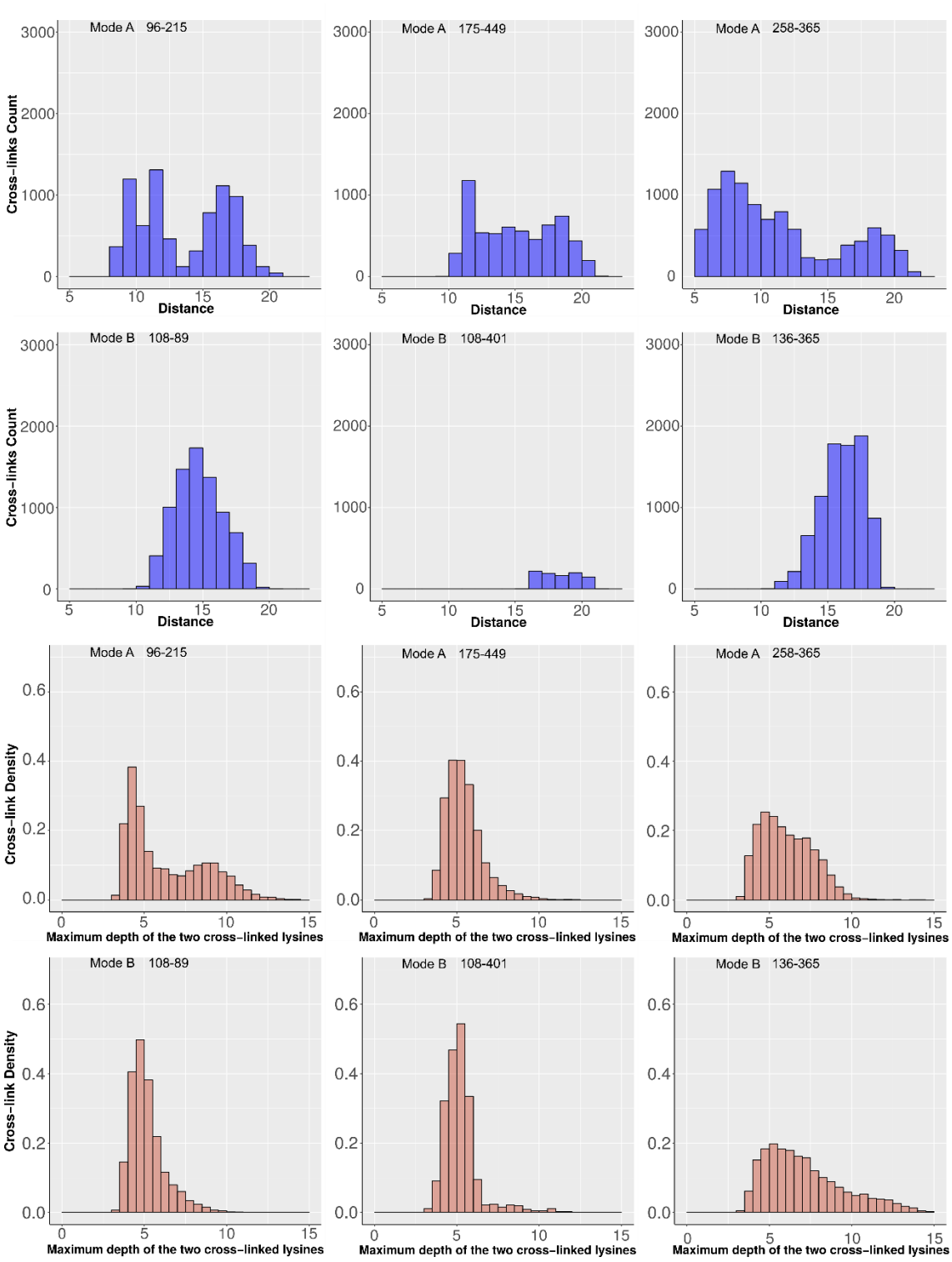


Figure 5-14 Mode A and Mode B cross-links in oligomers not in filament.

Distance (blue histograms) and depth distribution (red histograms) of mode A and mode B cross-links formed in the 2/3 oligomer but not in the intact filaments.

5.4.5 Difficulties in comparing relative frequency of cross-links

The number of cross-links obtained from the model was correlated with the relative frequency of the cross-links obtained in the chemical cross-link experiments. This would also provide a robust way to validate/improve the models. The next few paragraphs discuss the technical difficulties in comparing cross-links with relative frequency.

Counting Methodology

One of the difficulties was to accurately count the valid cross-links. Counting the cross-links through the C^{α}-C^{α} and C^{β}-C^{β} distance along with residue depth cutoff could sometimes be misleading. The distance between lysine pairs could be within the cutoff distance but having a covalently bound cross-linker could be physically impossible due to steric clash from neighboring molecules. Due to the large number of atoms in the model, it was difficult to model the cross-linker in the filament. One of the ways to surpass this is to look at the atomic density between the two cross-linked lysines and decide whether two lysines can cross-link or not.

Multiple Partners

Another important factor that decides the relative frequency of cross-linked lysine pairs was the presence of multiple partners for cross-link. A specific partner may be preferred depending upon several factors such as accessibility to solvent, distance between lysine pairs, temperature factors, local residue packing and stability. Since these governing factors are dynamic in nature, a large number of trials at various time points were attempted to determine the relative frequency of crosslinks. However, no good match was obtained when compared to the relative frequency as observed in the experiments. A more robust algorithm needs to be implemented considering various factors and large number of atoms in the model.

Interaction with Head/Tail domains

The availability of the lysine for cross-link was also dependent on its interaction with the head/tail domains. Head/tail domains are known to interact with coiled-coil domains that could mask some lysines and thus prevent them from forming cross-links.

Resolution of the experimental technique

The resolution of the experimental technique used to determine the relative frequency plays a major role. The K5-K14 and K1-K10 cross-links used in this study were obbtained by HPLC column, thus the accuracy of the relative frequency values could not be ascertained. A similar chemical cross-linking experiment followed by mass spectrometry (instead of HPLC column) would further increase the confidence in the model.

5.4.6 Parallel mode cross-link relative frequency

Though there were technical difficulties in quantitative comparison of relative frequency values, they could be compared qualitatively. Parallel mode cross-links (sub-categorized as mode C) were obtained only after using stronger reducing agent compared to Anti-parallel mode crosslinks (sub-categorized as mode A and mode B) [203]. Many non-reproducible cross-links were also obtained in the experiment [203]. The relative frequency value of parallel mode cross-links were miniscule equivalent to the lowest possible resolution of the experiment. This indicates the presence of few lysine pairs within cross-linking distance, or the lysines were not exposed to the solvent and cross-link only with protein modifications.

The keratin K5-K14 filament model was analyzed for parallel mode cross-links. Example of parallel mode cross-link distribution shows fewer cross-links compared to anti-parallel modes (Figure 5-15). In most of these cross-links lysine side chains had to be completely stretched (C^{α}-C^{α} distance around 19.5 Å) to form covalent bond with cross-linker. The stretched side chains of lysine makes it difficult to form a cross-link. The depth values (representative of solvent accessibility) of cross-links shows many lysine pairs were buried (depth greater than 6.0 Å). Thus, making it difficult for the cross-linker to reach lysines. The models could explain the low relative frequency values of parallel mode cross-links.

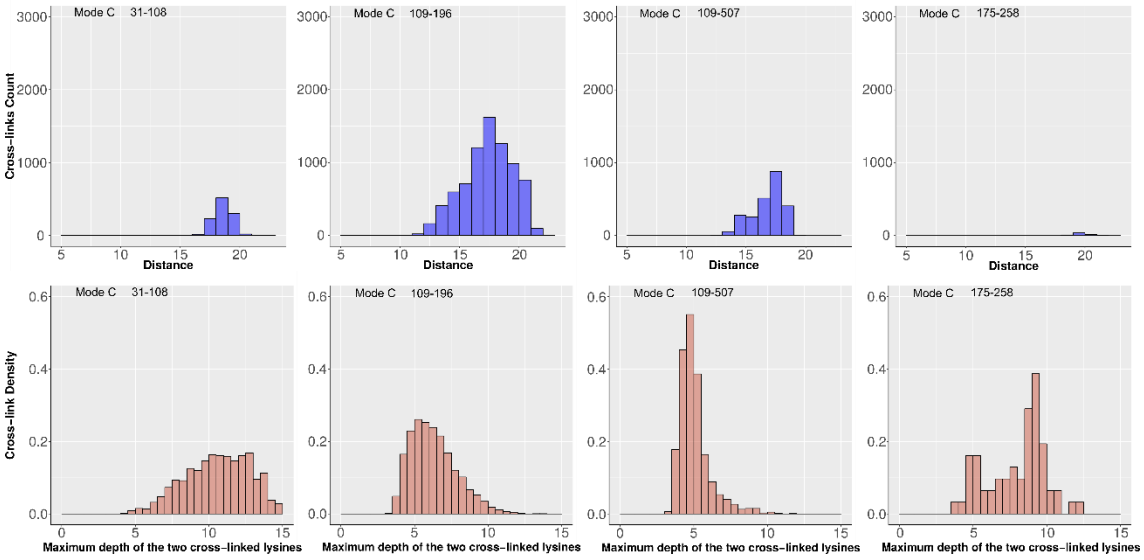


Figure 5-15 Distance and depth distribution of parallel mode cross-links

Examples of mode C or parallel mode cross-links showing distance (blue histograms) and depth distribution (red histograms) in the filament model trajectory.

5.4.7 K1-K10 crosslinks

To further validate the keratin K5-K14 filament model, the homolog K1-K10 intermediate filament cross-link data has been utilized [151]. The cross-linking data reports nineteen lysine pairs cross-linked with DST cross-linker. Since the K5-K14 and K1-K10 intermediate filaments are homologous with high sequence identity, their sequences were aligned and mapped to one another. The K1-K10 lysine amino acids crosslinks were mapped to the corresponding amino acids in the K5-K14 filament. The mapped amino acid could sometimes be amino acid other than lysine. Since only C^{α}-C^{α} and C^{β}-C^{β} distances were used to validate the cross-links, the residue identity was not significant. All nineteen cross-links (196-330, 124-495 and others listed in Figure 5-16) were reported with the lysine identities obtained according to the K5-K14 filament. The cross-links were sub-grouped according to K5-K14 filament model. Out of nineteen cross-links, seventeen cross-links (within a cutoff 22.0 Å as used for K5-K14) and eighteen cross-links (within a cutoff 22.3 Å) were satisfied with significant relative frequency (Figure 5-16). Among these eighteen cross-links, two crosslinks were common when compared to the K5-K14 filament. Hence, effectively the K1-K10 model satisfies the sixteen out of eighteen cross-links independently. The satisfaction of lysine-lysine cross-links in the homologous K1-K10 filament shows the robustness of the K5-K14 model and validates it at the residue level resolution.

Two cross-links 196-330 from mode A and 124-495 from mode B in K1-K10 filament could not satisfy the cross-link criteria. These cross-links were near the flexible loop region that were known to interact with head/tail domains. It was hypothesized that in K1-K10 filament, head/tail domains interactions with the flexible loop region could have locally changed the helix configuration. A slight change in the helix configuration would have been sufficient for lysine pairs to form cross-link. However, due to absence of head/tail domains in our models, a similar effect could not be observed.

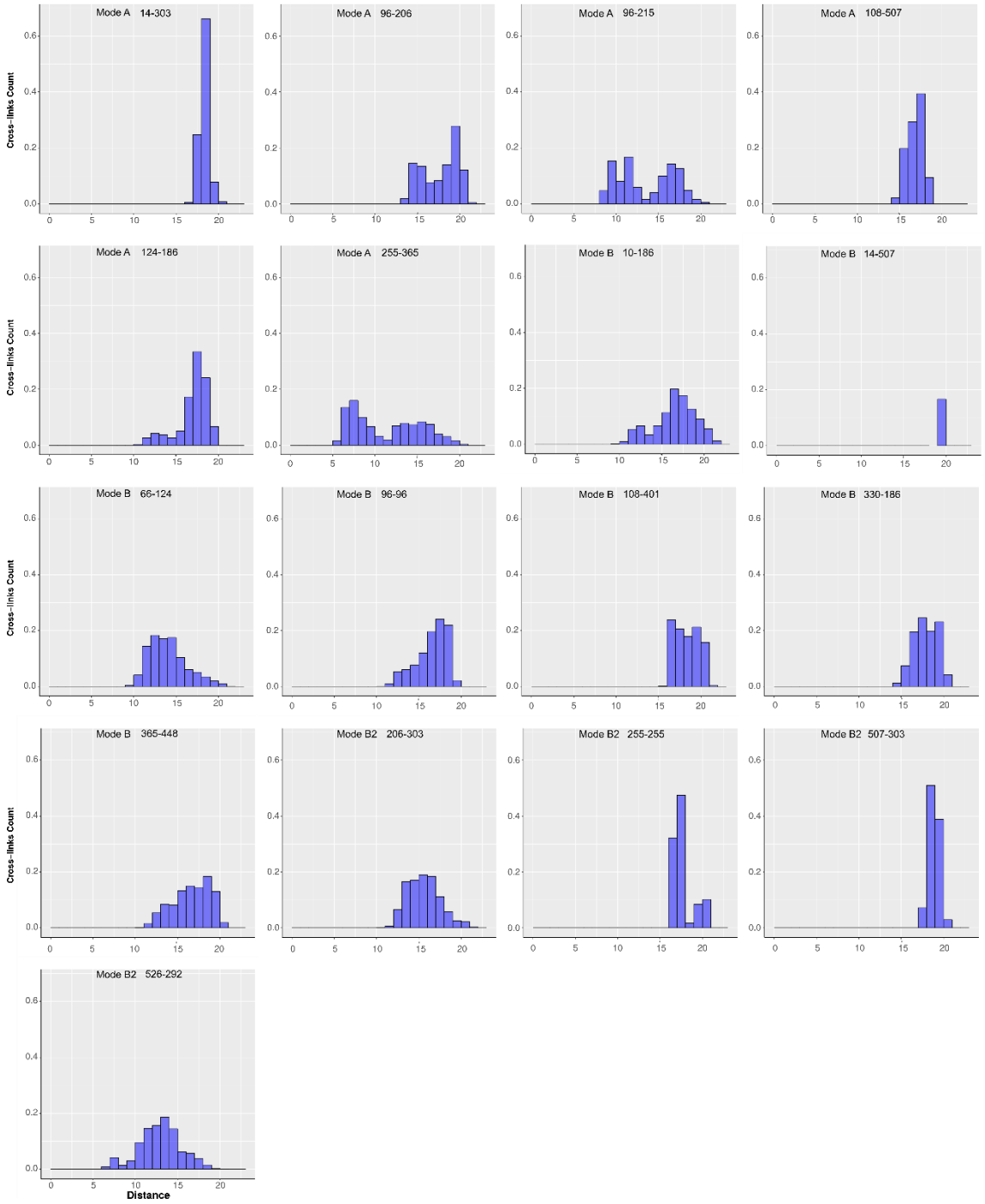


Figure 5-16 K1-K10 cross-links mapped on the K5-K14 filament

Each panel shows distribution of K1-K10 cross-links density (with y axis range 0 to 0.68) with respect to Cα- Cα distance of cross-linked amino acids (with x axis range 0 to 22 Å). Cross-links were sub-divided in mode A, mode B and mode B2 similar to K5-K14 cross-links. Two unsatisfied cross-links (196-330 from mode A, and 124-495 from mode B) was not shown in the plot.

5.5 Model properties using Voronoi packing volumes

5.5.1 Classification of protein into chemical groups

To understand the atomic level packing of the filament model, it was subjected to the volume measurements at the sub-residue level. A chemical group is a sub-residue grouping of atoms based on similar chemical properties (Figure 5-17). Every amino acid was sub divided into sixteen chemical groups (r1 to r16). This classification was essential to understand the packing of atoms having similar chemical properties. For instance, Lysine amino acids is composed of chemical group r1, r2, r2, r2, r5 (Figure 5-17). Chemical group r1 is a peptide bond, r2 is methylene carbon, and r5 is positively charged amino group. All three groups have different properties and packing volumes which could not be studied with the conventional classification.

5.5.2 Getting volume statistics of known protein structures

To generate the packing statistics of known protein structures, an old technique of partitioning the 3D space known as Voronoi diagrams have been utilized. This was coded in the software GARLIC (see section 9.2 for technical details). This software partitions the 3D space considering each chemical group as a single point. A database of 2579 single domain non redundant proteins was obtained from the PDB database [236]. To construct this database following parameters had been chosen: sequence identity <= 30%, resolution <= 1.81, R-factor <= 0.25 and sequence length 100 to 220. The missing residues in these structures were then modeled using MODELLER suite of programs with default parameters [26,211]. Voronoi volumes could be unbounded for chemical groups that were on the surface of the protein. To solve this issue, a single hydration layer was generated around the protein using the DEPTH [233–235] software. The hydrated structures were then subjected to Voronoi volume calculations using GARLIC. The Voronoi volumes of every chemical group in all the structures were clubbed and stored for comparison with the keratin filament model.

5.5.3 Getting volume statistics of K5-K14 filament model

Snapshots of the MD simulation trajectory of the filament model were extracted to generate the ensemble of models. Each structure was first solvated with one hydration layer and then subjected to GARLIC software (see section 9.2) for calculation of Voronoi volumes of the chemical groups. The comparison of the statistics from redundant protein structure database and K5-K14 filament model is discussed in the next section.

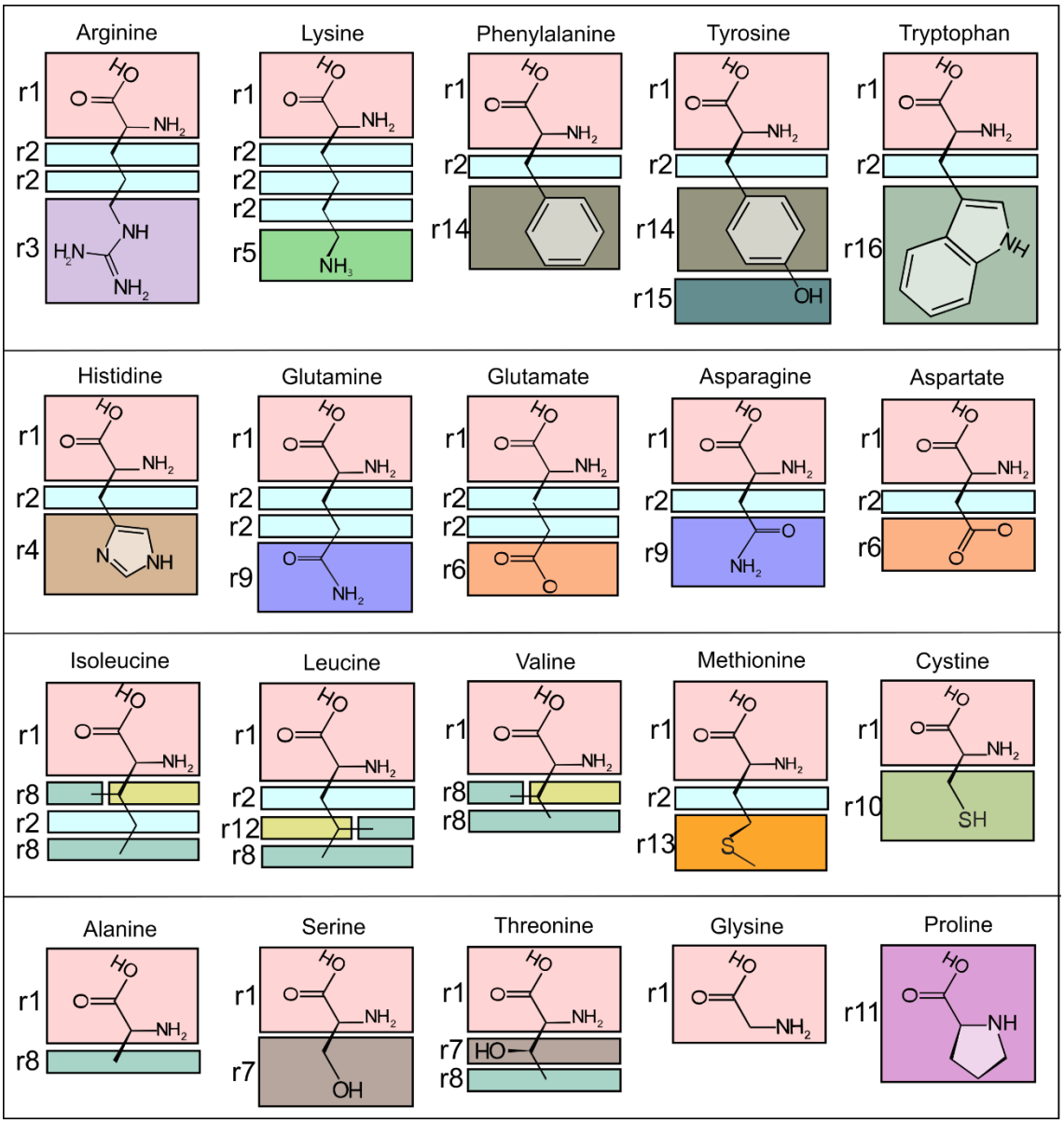


Figure 5-17 Sub-division of amino acids as chemical groups

Every amino acids (label at the top of boxes) was sub-divided into chemical groups (Boxes with 16 different colors). An amino acids structure could thus be represented in terms of various combinations of 16 chemical groups (labeled as r1 to r16 next to the colored box). Each of these different chemical groups have different chemical/biophysical properties.

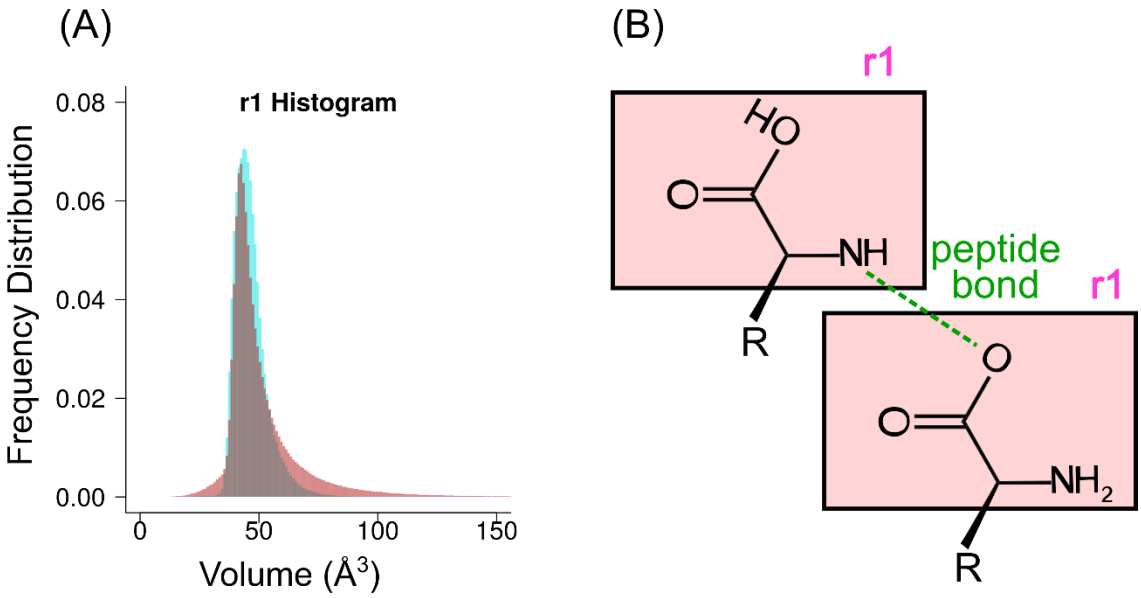


Figure 5-18 Assessing Voronoi volume distributions of K5-K14 filament

(A) Comparison of Voronoi volume distribution of chemical group r1 generated from a set of non-redundant protein database (cyan) with the ensemble of K5-K14 filament models (coral). (B) A schematic example of two covalently bonded amino acid with side chains R and chemical group r1.

5.5.4 Comparison of packing volume of the K5-K14 filament model

Voronoi volumes of all 16 chemical groups were extracted from non-redundant protein structure database (see section 5.5.2 for details). They have different means and standard deviations that was used to differentiate the chemical groups. Using the volume statistics, a quantitative score was developed to determine the quality of packing of chemical groups in any model. These distributions were used as to benchmark the Voronoi volume distribution of the K5-K14 filament model ensemble (see section 5.5.3).

Volume distribution of r1 chemical group (that form the peptide bond) generated from the ensemble of K5-K14 models shows characteristics similar to protein database (Figure 5-18). Packing artifacts due to thermal fluctuations (indicated by wide range of volume) would increase with an increase in the side chain length. Since chemical group r1 form the backbone, packing artifacts due to temperature fluctuations would be minimum for these groups. All these observations shows the packing of the filament backbone was in agreement with the protein database. The comparison of other chemical groups is currently under progress.

The protein database used in this study provides volume statistics for single domain proteins without considering any protein-protein interactions. However, the unit length

K5-K14 filament model is an assembly of 32 proteins dominated by charge-charge interactions. Thus comparing the two distributions governed by different interactions was likely to bias results. A possible solution to overcome this difficulty is by replacing single domain protein database (used in this study) with the protein-protein interface database. Along with this, the thermal fluctuation bias needs to be adjusted during MD simulations.

5.6 Summary

In the previous chapter, the ~500,000 atom Keratin K5-K14 filament model was constructed using Integrative modeling. However, being a computational model, it needs to be validated using the experimental data that have not been used to construct the model. In this chapter, the validation of the filament model was performed using five different types of available experimental data at various length scales. The validations included 1) Low resolution TEM micrographs of keratin filament bundles (length scale 100 nm to ~12 nm), 2) comparing lower and higher order structure dimensions (length scale ~13 nm to ~6 nm), 3) radial density of atoms using cryo-EM micrographs (length scale ~8.0 nm to ~0.5 nm), 4) identification of cross-linkable residues (length scale ~1.0 nm to ~0.5 nm) and 5) atomic packing volumes of chemical groups (length scale ~0.7 nm to ~0.3 nm). In the next chapter, validation of the keratin filament model through rationalization of disease causing mutations and prediction of new filament destabilizing substitutions would be discussed.

CHAPTER 6
IF Assembly and Stability Determinants

SALIENT FEATURES

- ***Exploring Micro-environments in the Filament***
  - Classification of chemical environments
  - Hydrogen bonding residues pairwise assignment
- ***Determining molecular determinants of filament stability***
  - Inter-dimer hydrogen bond depth and length distribution
  - Dimer-dimer interfaces
- ***Rationalization of disease causing mutations***
  - Analysis of depth and hydrogen bond profiles of mutants
- ***Prediction of filament disrupting mutations***
  - Classification of filament disrupting mutations
  - Prediction of new mutants by saturation mutagenesis
  - Prediction of temperature sensitive mutations
  - Experimental testing of new mutants
- ***Summary***

The previous two chapters (CHAPTER 4 and CHAPTER 5) provide the construction and validation details of the K5-K14 filament model. One of the main utilities of the 3D model is to help rationalize all known disease causing mutations (see Section 3.5). Previously (in CHAPTER 3), it was established that coiled-coil hydrophobic core packing could destabilize the filament assembly. The stability determinants were determined and validated by predicting substitutions in the hydrophobic core (at heptad positions **a/d**) (see Section 3.3.7 and Section 3.5.1). Predicted substitutions were tested by *in-vivo* tissue culture experiments (see Section 3.5.2). This also led to rationalization of most disease causing mutations observed at heptad positions **a/d**.

Studying the disease causing mutants at heptad position **a/d** requires only the intra-dimer interactions. However, dimer-dimer interactions were required to study disease-causing mutants at other heptad positions **b/c/e/f/g**. The inter-dimer interactions were obtained from the 3D structure of the Keratin K5-K14 model. This chapter discuss about the molecular determinants of stability at all non **a/d** heptad positions (Figure 6-1).

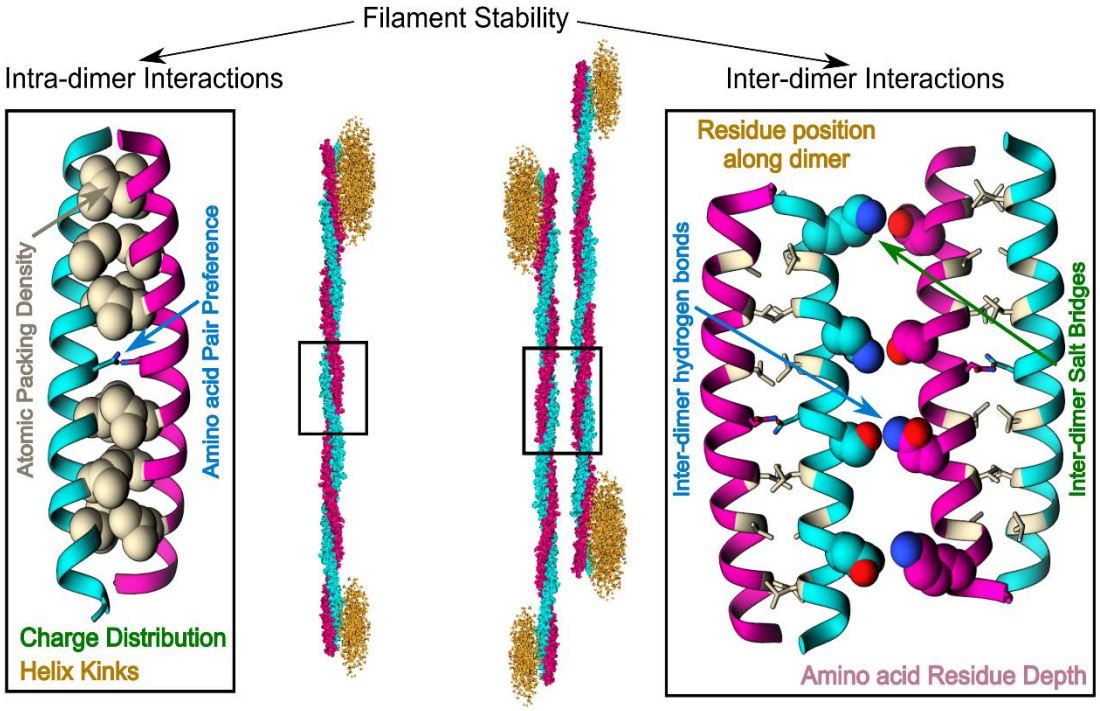


Figure 6-1 Molecular determinants of stability of filament.

A schematic of intra-dimer (left box) and inter-dimer (right box) interactions in the filament. Big boxes provide a zoomed picture of dimer and dimer-dimer interactions. Each box lists molecular determinants of stability (colored text) for both intra-dimer and inter-dimer interactions.

6.1 Exploring Micro-Environments in the Filament

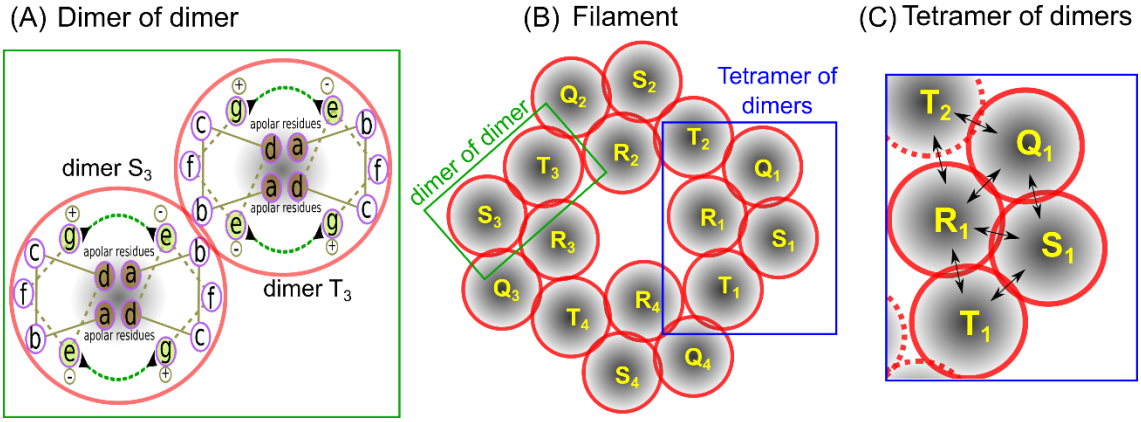


Figure 6-2 Schematic of inter-dimer interactions within the filament

(A) Two dimers (red circles) could possible interact only with amino acids at heptad positions ***b/c/e/f/g****. Amino acids at heptad positions* ***a/d*** *form hydrophobic core of coiled-coil. (B) Two dimensional filament model specifying four "tetramer of dimers" (octamer) with dimer id's* $Q_1/R_1/S_1/T_1$*. Possible interactions (black arrow) within "tetramer of dimer" represented with id's* $Q_1/R_1/S_1/T_1$*.*

6.1.1 Classification of chemical environments

In a coiled-coil dimer, amino acids at heptad positions **a/d** interact with one another to form hydrophobic core (see Figure 2-1). Aliphatic side chains of amino acids at heptad position **e/g** also corroborate this hydrophobic core (see Section 3.3.4). Any two dimers within a filament (Inter-dimer) interacts only with amino acids at heptad positions **b/c/f** (polar amino acids) along with charged amino acids at heptad position **e/g** (Figure 6-2 A). Thus, for extracting the inter-dimer molecular determinants of stability, the search space was restricted to polar interactions such as hydrogen bonds. Due to inherent symmetry in the model, only a "tetramer of dimers" (octamer) with id's "$Q_1/R_1/S_1/T_1$" was considered for predicting mutations (Figure 6-2 B, C). Since there were four identical K5-K14 dimers in every octamer, every amino acid could have as many as four different chemical environments depending upon the orientation of each dimer within the filament.

6.1.2 Hydrogen bonding pair assignment

To determine the inter-dimer hydrogen bonding pairs in the $Q_1/R_1/S_1/T_1$ octamer, all hydrogen bonding donor/acceptor pairs that were within the distance cutoff of 4Å and angle greater than 100˚ (donor-acceptor-acceptor antecedent angle). All donor/acceptor pairs that were part of the main chain would be contributing to the helix hydrogen bonds (keratin is an all alpha helix protein) and thus ignored. Only inter-dimer side chain

hydrogen bonds were considered for further analysis. Since hydrogen bond determination was dependent only on distance and angle cutoff, a large majority of donors/acceptors have multiple partners. However, at a given instant a donor/acceptor pair could form only one hydrogen bond. Thus, donors/acceptors need to be assigned appropriately to one another so that they form a unique hydrogen bond at a given instant.

To solve this unique assignment problem, the probability of every hydrogen bond forming donor-acceptor pairs were determined from 1000 snapshots of MD simulation trajectory of K5-K14 filament (see Section 5.4.1). Using this probability in the cost function, Hungarian Algorithm was used to obtain the unique assignment as explained in the following steps [237]. The cost of assigning a donor a to an acceptor d is given by $C_{ad} = -\log(P_{ad})$ where P_{ad} is the probability of hydrogen bond formed by acceptor-donor (ad) pair. Let A be a Boolean matrix such that $A[a, d] = 1$ if and only if acceptor a is assigned to donor d. Then the optimal assignment have the cost $min \sum_a \sum_d C_{ad} A_{ad}$ such that each donor is assigned to only one acceptor or vice versa. Solution to this equation provide an optimal set of hydrogen bonds with maximum probability of occurrence.

6.2 Determining molecular determinants of filament stability

6.2.1 Inter-dimer hydrogen bond depth distribution

The inter-dimer hydrogen bonds provide stability to the dimer-dimer interaction within the filament. Destabilizing these interactions would lead to reduced filament stability. Due to a disease causing mutation, the donor/acceptor pair that forms a hydrogen bond in the wild type would get destabilized with the presence of lone acceptor/donor atom. Higher residue depth of these lone donor/acceptor atoms (especially charged atoms) would contribute more to the instability. To ascertain the impact of substitution of buried amino acid on the stability of the filament, average residue depth of all hydrogen bonding residue pairs was calculated. The depth of all hydrogen bonds were compared to the depth of hydrogen bonds at disease causing positions (Figure 6-3). The disease causing mutations were more prominent in the terminal regions of the dimers (Residue positions K5:1-31, K5:298-313, K14:314-346, K14:609-624). These regions were used to rationalize/predict new substitutions that would destabilize the filament (Figure 6-3). The depth values ranging from 4 to 6.5 have the most disease causing mutations (Figure 6-3). However, the reported depth values were given for all the possible environments ($Q_1/R_1/S_1/T_1$) and doesn't necessary provide a correct profile of instability.

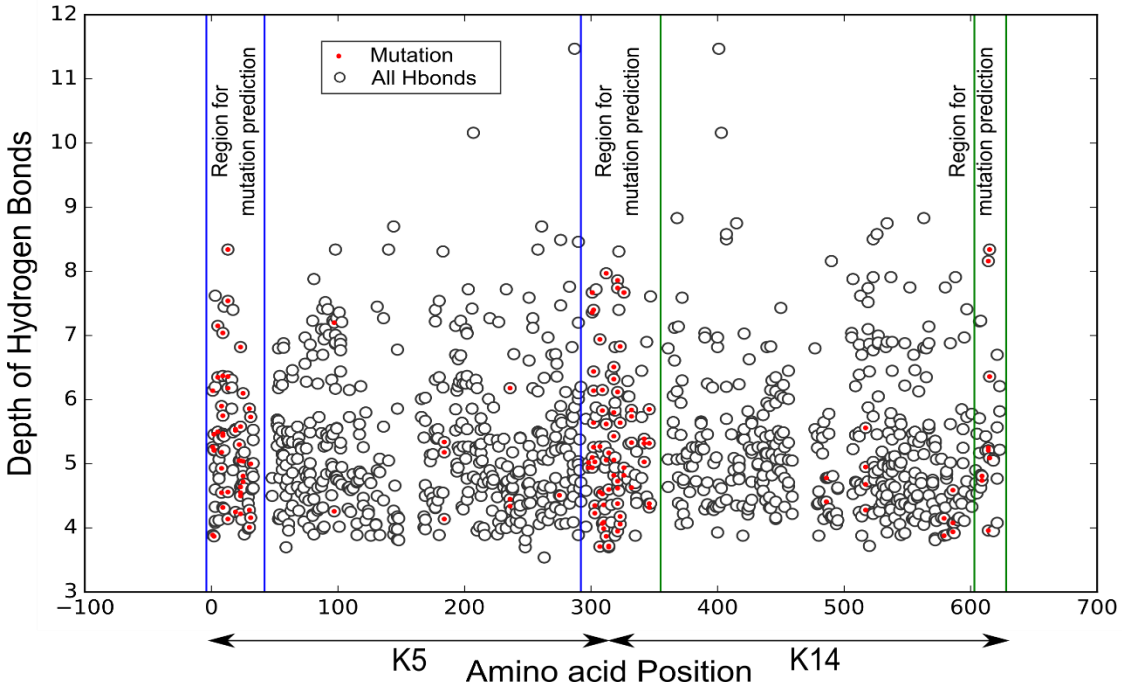


Figure 6-3 Depth distribution of hydrogen bonds across dimer length

The depth (y axis) of hydrogen bonding residue pairs (circles) from all four environments along the dimer length (x axis). Residue positions were labeled from 1-313 for K5 and 314-624 for K14. Residue pairs formed at disease mutant position (filled circles).

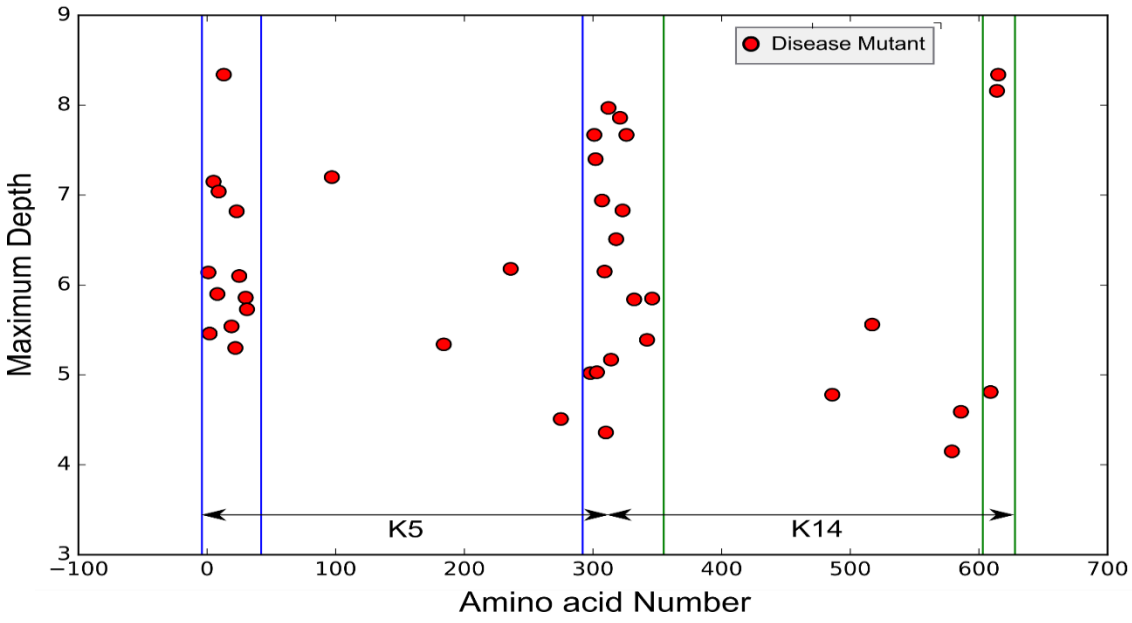


Figure 6-4 Maximum depth distribution for disease mutants

Maximum depth value (y axis) of hydrogen bonds across environments ($Q_1/R_1/S_1/T_1$) at positions (x axis) affected by disease mutants. Amino acid positions were labeled from 1-313 for keratin K5 and 314-624 for K14.

To verify the effect of residue depth, the maximum depth of hydrogen bonds across possible environments ($Q_1/R_1/S_1/T_1$) was calculated. Mutating a deeper residue should have higher impact on stability compared to surface residues. The minimum residue depth of disease causing mutations at which lone donor/acceptor atom was present without having a interacting partner was around 5.0 (Figure 6-4). Thus, any lone acceptor/donor having depth greater than 5.0 was considered unstable and was used for rationalizing/predicting intermediate filament disease causing mutations.

6.2.2 Inter-dimer hydrogen bond length

To further quantify the stability of the intermediate filament, the hydrogen bonds were characterized by measuring the "effective hydrogen bond lengths" (EHBL) (see 6.1.2 for hydrogen bond assignment). The EHBL for any hydrogen bond was calculated by adding the distances between C^{α}–donor and C^{α}–acceptor (Table 6-1) [238]. The change in EHBL due to mutation was calculated by subtracting EHBL of wild type and mutant hydrogen bond. A significant decrease in the EHBL would suggest the absence of hydrogen bond. The change in EHBL were estimated for every disease causing mutations (Figure 6-5). Most disease causing mutations have a positive change (greater than zero) in the EHBL suggesting disruption in the hydrogen bond. All negative changes were due to hydrophobic amino acids.

Table 6-1 Distance between C^{α} atom and donor/acceptor atom

Side chains of amino acids that do not have both acceptor and donor atoms were given value -1.0. Whereas value 0.0 was assigned if donor is present and acceptor is absent or vice versa. Distance values were obtained from [238].

Amino acid	C^{α} -Donor distance (Å)	C^{α} -Acceptor distance (Å)	Amino acid	C^{α} -Donor distance (Å)	C^{α} -Acceptor distance (Å)
CYS	2.805	0.0	GLN	4.918	4.313
HIS	3.702	0.0	TYR	6.438	6.438
TRP	4.520	0.0	GLY	-1.0	-1.0
LYS	6.334	0.0	PRO	-1.0	-1.0
ARG	6.634	0.0	MET	-1.0	-1.0
ASP	0.0	3.225	LEU	-1.0	-1.0
GLU	0.0	4.592	ALA	-1.0	-1.0
SER	2.427	2.427	VAL	-1.0	-1.0
THR	2.424	2.424	ILE	-1.0	-1.0
ASN	3.725	2.794	PHE	-1.0	-1.0

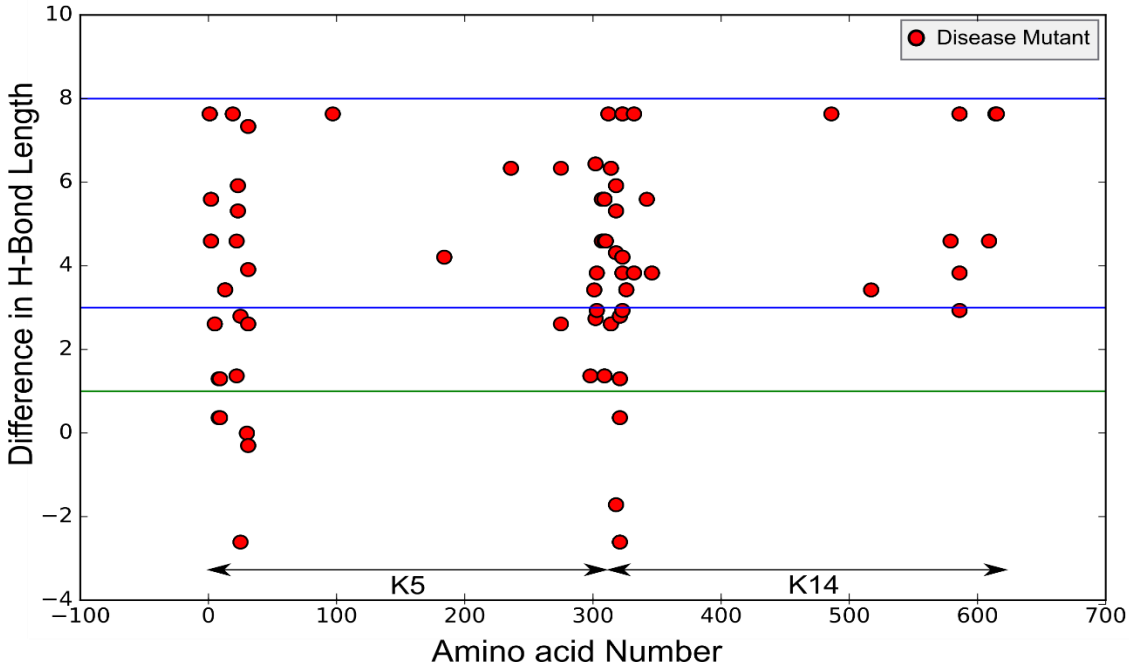


Figure 6-5 Change in effective H-bond length due to disease causing mutation.

Change in effective hydrogen bond length (EHBL) (y axis) calculated by subtracting the EHBL of wild type and disease causing mutation. Amino acid positions (x axis) were labeled from 1-313 for keratin K5 and 314-624 for K14.

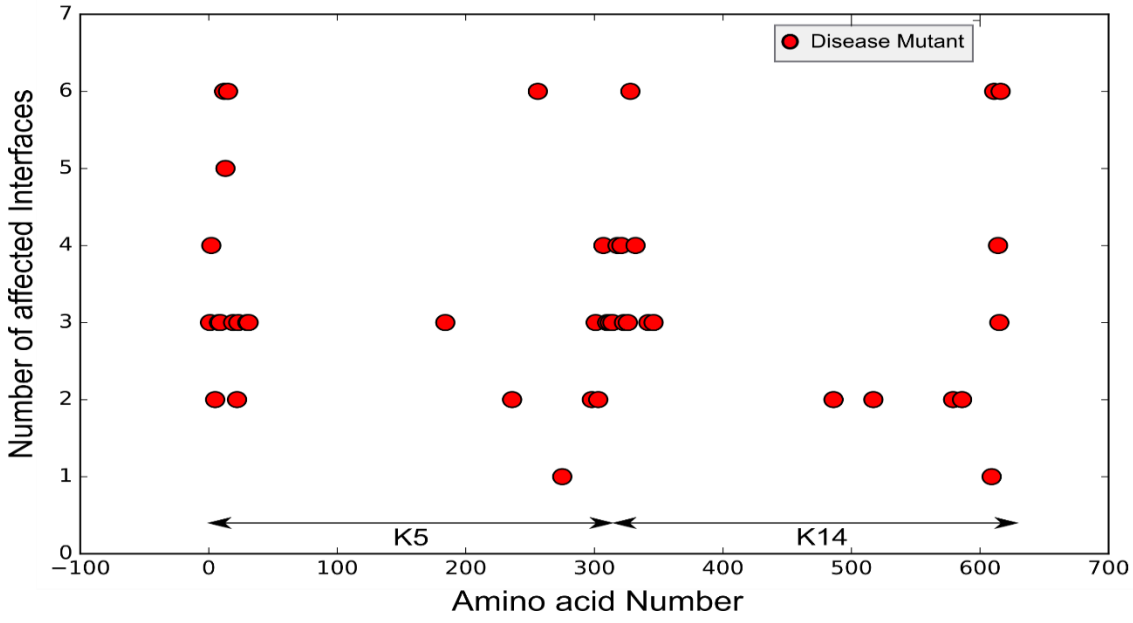


Figure 6-6 Number of affected interface due to disease causing mutation

Number of affected dimer-dimer interfaces (among Q_1R_1, Q_1S_1, Q_1T_2, R_1S_1, R_1T_1, R_1T_2, and S_1T_1*) in the filament (y axis) due to the disruption of hydrogen bonds caused by disease causing mutation.*

6.2.3 Number of dimer-dimer interacting interfaces

Since there were four copies of an amino acid in an octamer ($Q_1/R_1/S_1/T_1$), a single mutation could affect more than one interface (Figure 6-2 C). The number of affected dimer-dimer interfaces (namely Q_1R_1, Q_1S_1, Q_1T_2, R_1S_1, R_1T_1, R_1T_2, and S_1T_1) for a given mutant also contributes to the stability of the filament. For every disease causing mutation, the number of unique interfaces corresponding to destabilized hydrogen bonds were recorded (Figure 6-6). More than 78% (32/41) of disease causing mutations disrupt two to four dimer-dimer interfaces. Due to the absence of control mutations from non-disease positions, it was difficult to classify any amino acid substitutions as stable or unstable. However, a substitution that affects more than two interfaces was more likely to destabilize the filament.

6.3 Rationalization and Prediction of filament disrupting mutations

The molecular determinants governing the stability of the filaments was used to explain the mechanism of the disease causing mutations. Following molecular determinants were extracted from the model. 1) Number of hydrogen bonds, 2) Residue depth distribution of hydrogen bonds, 3) Position of mutation along the dimer, 4) Effective hydrogen bond length, and 5) Number of destabilized dimer-dimer interfaces. To predict all substitutions that would destabilize the dimer-dimer interfaces, an *in-silico* saturated mutagenesis was performed for all positions in a K5-K14 dimer. Each mutation was scored using the effective hydrogen bond length (EHBL) (see Section 6.2.2). The change in the EHBL suggest whether the hydrogen bond would form between the mutated K5-K14 dimers. An absence of hydrogen bond destabilizes the filament with different intensity level. Any substitution with the change in EHBL greater than 3.0 considered severely filament destabilizing (see Figure 6-5). Following presents a working example of severely destabilizing mutation.

For instance, in K5-K14 wild type filament Lys at position 5 was assigned to Glu at position 198 to form 5Lys-198Glu hydrogen bond (see assignment detail in Section 6.1.2). The EHBL for this hydrogen bond $EHBL_{WT} = 6.334 + 3.225 = 9.559$. With a mutation Lys5Ser, the new EHBL for this hydrogen bond $EHBL_{Mut} = 2.427 + 3.225 = 5.652$. The change in EHBL was $\Delta EHBL_{WT-Mut} = 9.559 - 5.652 = 3.907$. Since the change was greater than 3.0, the mutation Lys5Ser was considered severely destabilizing.

The strength of the destabilization depends on various other factors such as the affected number of hydrogen bonds, the number of affected dimer-dimer interfaces, residue depth profile and position along the dimer (see Section 6.2). Due to the absence of convincing proof in support of these factors, they were not considered for predicting filament destabilizing mutations. For every position in the dimer, all filament destabilizing mutations were compared with the known disease causing mutations.

In a K5-K14 dimer, heptad positions **b/c/e/f/g** have 71 disease causing mutations. Out of 71 mutations, wild type position in 15 mutations occupied by hydrophobic amino acids (Table 6-2) whereas 56 mutations have polar/charged amino acids (Table 6-3). Since majority of the heptad positions **b/c/e/f/g** occupied by polar/charge amino acids, the inter-dimer interactions were mostly hydrogen bonds and salt bridges (Figure 6-2 A). This chapter focus on the molecular determinants governed by polar/charged interactions. Hence, all 15 disease causing mutants having wild type position occupied by hydrophobic amino acids were eliminated. Thus, in this study only 56 disease-causing mutations were analyzed. The methodology used here could rationalize 54/56 disease causing mutations (Table 6-3). The two mutations that could not be explained were T30S and K31R. Since these mutations were near the linker domain L1, they might be interacting with head/tail domain and could have functional importance associated with them. All these 56 mutations were present in only 35 out of 624 possible positions in the dimer. The *in-silico* saturated mutagenesis scan lead to identification of 257 key interacting positions. Experimental validation of new filament disrupting mutations (selected from 257 key interacting positions) is currently being done through *in-vivo* setup (data not shown).

6.4 Prediction of temperature sensitive mutations

An amino acid substitution in a protein that features normal function/phenotype at low temperature but altered function/phenotype at higher temperature was generally referred as Temperature sensitive mutation. Keratin was known to have a E475G temperature sensitive mutation [239]. Keratin filament with E475G mutation in K5 remains intact at 37˚ Celsius but the filament reversibly disrupts at 43˚ Celsius. Such mutations would greatly help researchers in controlling the cytoskeleton assembly. This section provides the methodology to determine the temperature sensitive mutations for K5-K14 filament.

The previous section explained the *In-silico* saturated mutagenesis of the filament that provided filament-disrupting mutations. Many of these mutations would be disrupting the filament severely. However, there were substitutions with minimal change in donor-

acceptor distance (less than 2) as compared to wild type filaments (Figure 6-5). The hydrogen bonds in these cases could form by locally changing the pitch of the dimer, thus retaining the filament structure intact. For instance, consider E307D mutant in K5. The wild type Glu307 forms the hydrogen bond with neighboring Thr365 in K14 with a distance of 3.5 Å (Figure 6-7 A). However, the mutant E307D increases the distance between donor-acceptor pair to 4.8 Å (Figure 6-7 b). The hydrogen bond in the mutant structure could only form with the helix readjustment. Thus, taking the mutant filament to a high-energy state (due to readjustment) at room temperature. However, at higher temperatures such hydrogen bonds would be disrupted leading to filament fragmentation.

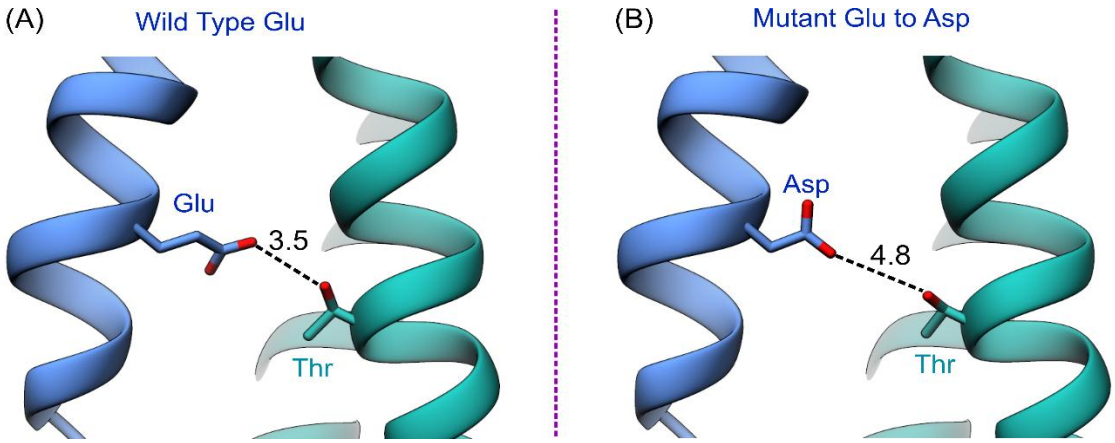


Figure 6-7 Schematic example of temperature sensitive of mutant

(A) Wild type protein with dimer-dimer (blue-turquoise) hydrogen bond (dotted line) between Glu-Thr amino acids. (B) Mutant protein (Glu to Asp) with equivalent hydrogen bond (dotted line) between Asp-Thr amino acids.

This idea forms the basis of predicting the temperature sensitive mutations. All predicted filament-destabilizing mutations were categorized into two parts. A) Severely filament disrupting mutations and, B) Partially filament disrupting mutations or temperature sensitive mutations. All mutations that have EBHL less than 3.0 (Figure 6-5) were chosen to be temperature sensitive mutations (see Table 6-3 for the list of mutations). Temperature sensitive mutations were more likely to be present near the terminal regions as a coiled-coil dimer are flexible for spatial adjustment only near the terminal domains. Currently, prediction of temperature sensitive mutations do not incorporate other parameters such as spatial position of the mutant along the dimer, position of head/tail domains etc. These parameters need to be included in the scoring scheme for predicting mutations with higher confidence.

Table 6-2 Disease causing mutants with wild type hydrophobic residue

Table lists the disease causing mutants at heptad position b/c/e/f/g having wild type hydrophobic residue. Columns: IF=Intermediate Filament, CC=Coiled-coil numbering, AA=Residue numbering, H=Heptad repeat, WT=Wild type residue.

IF	CC	AA	H	WT	Disease Mutations
K5	12	180	b	A	D, P
K5	15	183	e	I	F, V, M
K5	34	202	c	L	Q
K5	143	311	e	L	R
K5	256	424	g	A	P
K5	270	438	g	A	D
K5	308	476	c	G	D
K14	328	130	e	L	P
K14	611	413	b	A	T, P
K14	616	418	g	L	V, Q

Table 6-3 Predicted filament disrupting and temperature sensitive mutations

Table list the predicted filament disrupting mutations at sites previously known to have disease-causing mutations. All hydrophobic mutations were not listed in the table as these will completely disrupt the hydrogen bond interactions and were considered in severely filament disrupting category. Columns: IF=Intermediate Filament, CC=Coiled-coil numbering, AA=Residue numbering, H=Heptad repeat, WT=Wild type residue.

IF	CC	AA	H	WT	Disease Mutations	Predicted Severe Mutations	Predicted TS Mutations
K5	1	169	e	R	G, P	C, D, E, G, P, S, T	H, K, N, Q, W, Y
K5	2	170	f	E	G, K	C, G, H, K, P, R, W	D, N, Q, S, T
K5	5	173	b	K	N	C, D, E, G, P, S, T	H, N, Q, W
K5	8	176	e	N	S	D, E, G, H, K, P, R	C, S, T, W
K5	9	177	f	N	S	D, E, G, P	C, H, K,R,S,T, W
K5	13	181	c	S	P	D, E, G, H, K, P, R	C, T, W
K5	19	187	b	R	P	C, D, E, G, P, S, T	H, K, N, Q, W, Y
K5	22	190	e	E	D, K	C, G, H, K, P, R, W	D, N, Q, S, T
K5	23	191	f	Q	P	C, D, E, G,H,K,P, R, W	N, S, T
K5	30	198	f	T	S	D, E, G, H, K, P, R	C, W
K5	31	199	g	K	M, N, R, T	C, D, E, G, P, S, T	H, N, Q, W
K5	184	352	g	R	S	C, D, E, G, P, S, T	H, K, N, Q, W, Y
K5	236	404	e	K	E	C, D, E, G, P, S, T	H, N, Q, W
K5	275	443	e	K	E, N	C, D, E, G, P, S, T	H, N, Q, W
K5	298	466	g	E	D	C, G, H, K, P, R, W	D, N, Q, S, T
K5	301	469	c	T	P	D, E, G, P	C, H, K, R, W
K5	303	471	e	R	C, H	C, D, E, G, P, S, T	H, K, N, Q, W, Y
K5	307	475	b	E	G, K	C, G, H, K, P, R, W	D, N, Q, S, T
K5	310	478	e	E	K	C, G, H, K, P, R, W	D, N, Q, S, T

K5	312	480	g	R	G	C, D, E, G, P, S, T	H, K, N, Q, W, Y
K14	314	116	e	K	E, N	C, D, E, G, P, S, T	H, N, Q, W
K14	318	120	b	Q	P, R	C, D, E, G,H,K,P, R, W	N, S, T
K14	321	123	e	N	K, S	D, E, G, H, K, P, R	C, S, T, W
K14	323	125	g	R	C,G,H,L,P, S	C, D, E, G, P, S, T	H, K, N, Q, W, Y
K14	326	128	c	S	P	D, E, G, P	C, H, K, R, T, W
K14	332	134	b	R	C, P	C, D, E, G, P, S, T	H, K, N, Q, W, Y
K14	342	144	e	E	A	C, G, H, K, P, R, W	D, N, Q, S, T
K14	346	148	b	R	C	C, D, E, G, P, S, T	H, K, N, Q, W, Y
K14	486	288	e	R	L	C, D, E, G, P, S, T	H, K, N, Q, W, Y
K14	517	319	c	T	P	D, E, G, H, K, P, R	C, W
K14	579	381	e	E	K	C, G, H, K, P, R, W	D, N, Q, S, T
K14	586	388	e	R	C, G, H, P	C, D, E, G, P, S, T	H, K, N, Q, W, Y
K14	609	411	g	E	K	C, G, H, K, P, R, W	D, N, Q, S, T
K14	614	416	e	R	P	C, D, E, G, P, S, T	H, K, N, Q, W, Y
K14	615	417	f	R	P	C, D, E, G, P, S, T	H, K, N, Q, W, Y

6.5 Summary

The analysis of the 3D structure of keratin filaments provides the molecular basis of their stability and specific pairing. Molecular determinants of the Intermediate filament stability was classified into two types. 1) Intra-dimer stabilizing molecular determinants and 2) Inter-dimer stabilizing molecular determinants. Most intra-dimer interactions were comprised of hydrophobic interactions whereas inter-dimer interactions were polar/charged interactions. CHAPTER 3 discussed about the intra-dimer molecular determinants and their validation using *in-vivo* experiments. This chapter discussed the inter-dimer polar/charged interactions such as hydrogen bonding, residual depth etc. Starting from classifying the chemical environments in the filament that was coupled with the donor-acceptor assignment for hydrogen bonding residues. The hydrogen bond analysis provides five inter-dimer molecular determinants. 1) Number of hydrogen bonds, 2) Residue depth distribution of hydrogen bonds, 3) Position of mutation along the dimer, 4) Effective hydrogen bond length, and 5) Number of destabilized dimer-dimer interfaces. These molecular determinants provide 257 previously unknown critical interacting positions in the filament. The inter-dimer polar/charge interactions could explain 54/56 known disease causing mutations and predicted several temperature sensitive mutations that are currently under experimental testing. The study of inter-dimer hydrophobic interactions is required to rationalize the 15 mutations excluded in this study.

CHAPTER 7
Applications of IF Structural Properties

SALIENT FEATURES

- ***Designing Novel Coiled-coil proteins***
  - Analysis of dimer, trimer, and tetramer pair preferences
  - Selecting a novel coiled-coil topology (Anti-parallel octamer)
  - Optimization of coiled-coil rules to determine sequence composition
  - Modeling designed sequence and analyzing stability
- ***Designing inhibitor for Nipah Virus F protein***
  - Modeling F protein in two states (pre-fusion and post-fusion)
  - Modeling intermediate states
  - Applying coiled-coil rules in designing inhibitor
  - Assessment of the designed inhibitor
- ***Predicting Oligomeric state of coiled-coil assembly***
  - Obtaining data and conversion into sliding windows
  - Training and testing for oligomeric state prediction
  - Assessment of the predictions and drawbacks
- ***Summary***

The knowledge obtained during the construction and analysis of Intermediate filament structure was used to address several challenges related to coiled-coil proteins. The challenges includes, 1) Given a topology of the coiled-coil protein, determine the compatible amino acid sequence (Section 7.1), 2) Given an amino acid sequence, determine its coiled-coil structure and oligomerization state (Section 7.3), 3) Given a known coiled-coil structure/assembly, design an inhibitor peptide to disrupt the assembly (Section 7.2). Each of these challenges are discussed in the following sections.

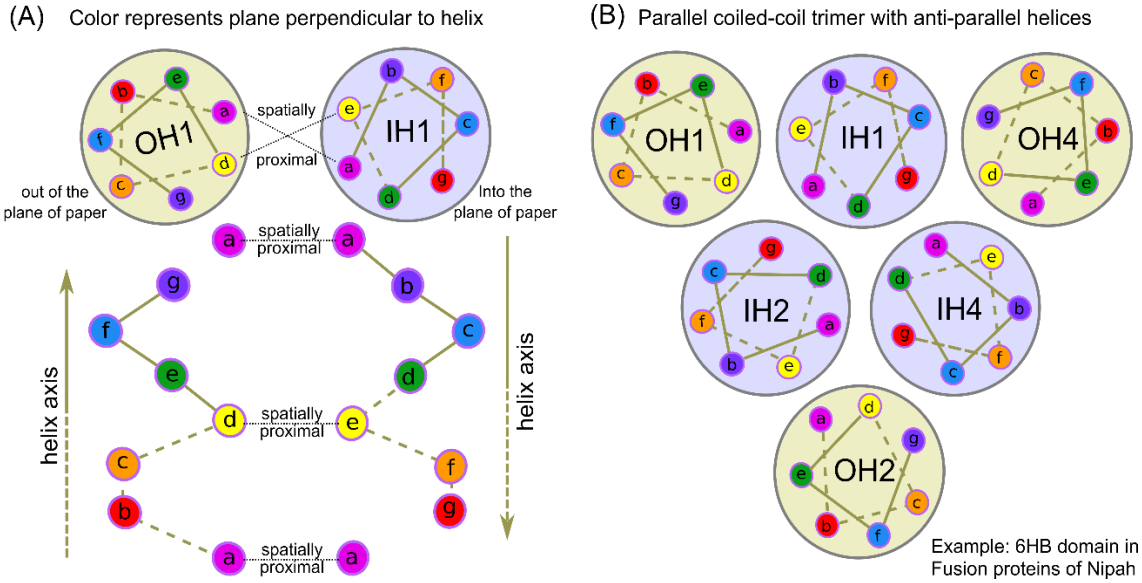


Figure 7-1 Schematic representation of coiled-coil in 3D

(A) A single big circle represents helix going into the plane of paper (light blue circle) and helix coming out of the plane of paper (light yellow circle). Every big circle has helical heptad repeats from ***a-b-c-d-e-f-g****. Rainbow colors in cyclic order (Violet, Indigo, Blue, Green, Yellow, Orange and Red) of the heptad repeats denote the plane perpendicular the helix axis. Cyclic separation between colors provides spatial proximity of residues along the helical axis (such as OH1(d) and IH1(e)). (B) An example of parallel coiled-coil trimer surrounded by anti-parallel helices.*

7.1 Designing novel coiled-coil proteins with given topology

The PDB database contain protein structures having different types of coiled-coil topologies (see Section 2.1.1). The coiled-coil structures were classified into various topologies using inter-helices interactions [91,92]. The most common coiled-coil topology was 2-helix homo/hetero dimers (see Figure 2-1). Other well-known coiled-coil topology is six-helix bundle (6HB) observed in fusion proteins of many viruses (see Section 7.2). A schematic example of 6HB domain shows three parallel helices forming

a coiled-coil trimer in the core along with additional anti-parallel helices on the outside (Figure 7-1 B). A similar 8HB coiled-coil topology with four parallel helices forming the tetramer core surrounded by four anti-parallel helices (anti-parallel octamer) was not observed in the PDB database [92]. Hence, this hypothetical anti-parallel octamer topology was chosen to design novel coiled-coil protein (see schematic representation in Figure 7-4). Since, there was no template available for anti-parallel octamer, amino acid interaction rules obtained from the analysis of coiled-coil structures would be utilized.

7.1.1 Determining the amino acids distribution across heptads

Pair preferences of amino acids at various heptad positions would determine the stability and specificity of any coiled-coil structure (see Section 3.3.1, Section 3.3.3, and Section 3.3.4 for details). A helix with a heptad repeat pattern could form dimer, trimer or tetramer depending upon the amino acid content. Amino acid preference for dimers (133)/trimers(69)/tetramers(13) at various heptad positions were obtained to determine the distribution (Figure 7-2). The frequency distribution was obtained from parallel coiled-coil homo-mers having two/three/four alpha helices. All coiled-coil sequences that were less than 50% redundant and had minimum 21 amino acids were considered. The raw amino acid frequency values were sum normalized across various heptad repeats.

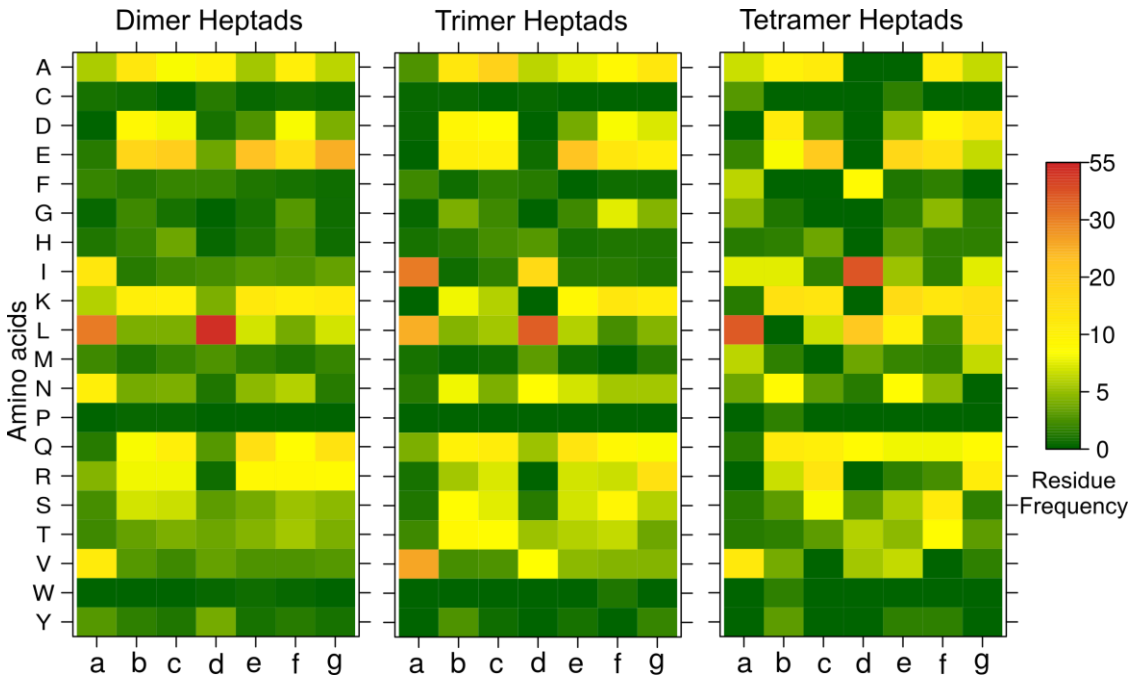


Figure 7-2 Amino acid distribution in parallel dimer, trimer and tetramers

Amino acid (y axis) distribution across heptad positions for dimer, trimer and tetramer (x axis). For each oligomer the frequency values were sum normalized across heptad repeat and multiplied by hundred. The residue frequency range from 0 (green) to 55 (red).

The final values were converted to percentage by multiplying with hundred (Figure 7-2). The comparison of the three distribution (dimer/trimer/tetramer) shows the striking difference only at heptad positions **a/d** and of amino acids I/L/V. The dimer distribution shows abundance of Leu at heptad **a/d** whereas the trimer distribution shows abundance of Leu in heptad **d** but Ile in heptad **a**. However, tetramer distribution shows a contrasting trend with Ile abundant at heptad **d** and Leu abundant at heptad **a**. These results corroborated the previous conclusions about hydrophobic core density formed by heptad positions **a/d** (see Section 3.3.3 for more details). The striking difference between the amino acid distributions of various oligomers ensures confident selection of amino acids, and thus driving the assembly towards specific oligomerization state. These frequency plots were also able to distinguish parallel dimer/trimer/tetramer coiled-coil forming sequences from one another with high accuracy (see Section 7.3).

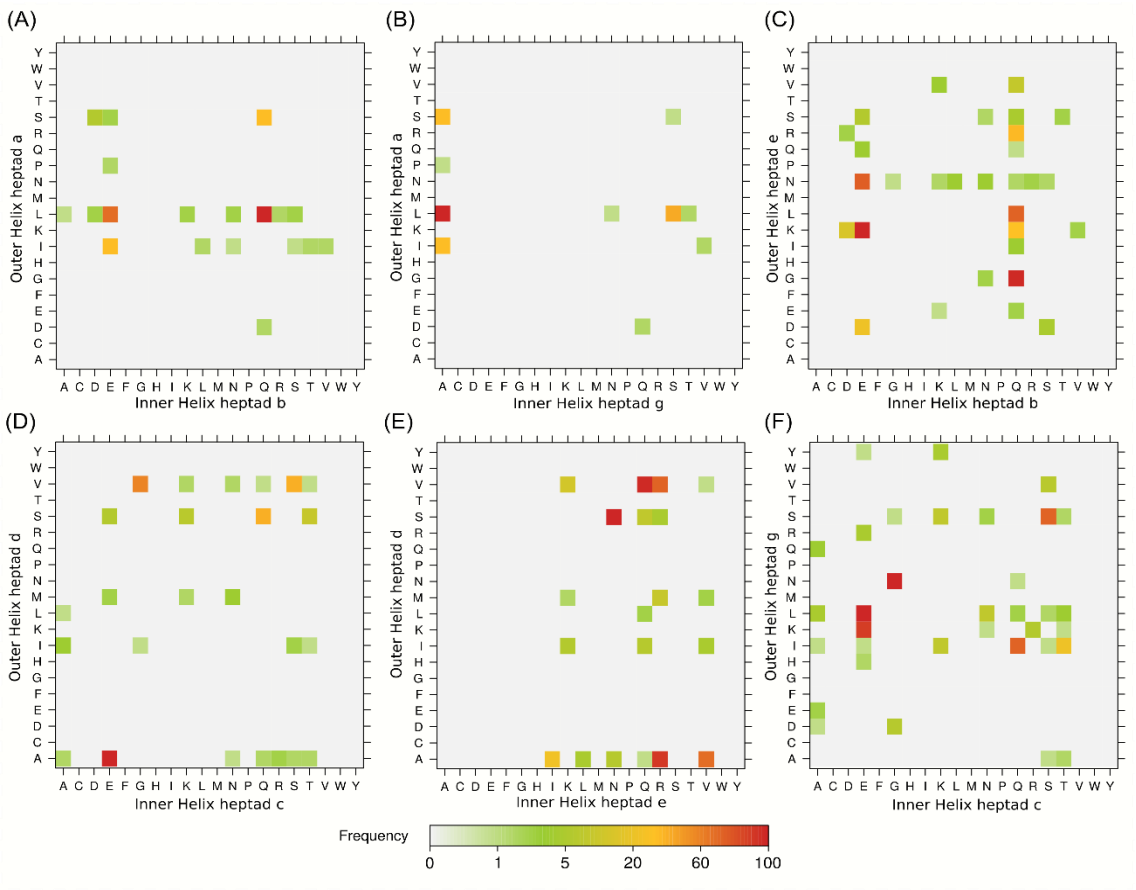


Figure 7-3 Frequency of interacting amino acids in anti-parallel trimers.

Frequency (max-normalized) of amino acid pairs in anti-parallel trimers between two parallel Inner helices (IH1 & IH2) (x axis) and one anti-parallel Outer helix (OH) (y axis). (A), (B), (D), and (E) Hydrophobic interactions at heptad pairs ***a-b***, ***a-g***, ***d-e*** *and* ***d-c*** *respectively. (C), (F) Polar interactions at heptad pairs* ***e-b*** *and* ***g-c****.*

The anti-parallel octamer topology consists of two parts; A) a parallel tetramer core and B) anti-parallel helices bound on the parallel tetramer (Figure 7-4). The amino acid sequence for the part A (parallel tetramer core) would be based on the frequency distribution obtained from the parallel dimer/trimer/tetramers. The sequence should follow tetramer distribution but negate the dimer/trimer distribution. Whereas the amino acids sequence for the part B (anti-parallel helices) were based on the frequency distribution of anti-parallel trimers. Since the three helices (IH1, IH2, OH1) in the 6HB domain of fusion proteins contains anti-parallel trimers (Figure 7-1 B). Thus, 250 homologs of 6HB domain was used to obtain the anti-parallel trimer distribution (Figure 7-3). The homologs (Refer file: APT_UP_ID.txt in Supplemental Material) were obtained by using the 6HB domain of Nipah virus fusion protein (UniProtKB id: Q9IH63) as a query against the UniProtKB database [240].

7.1.2 Designing Anti-parallel octamer

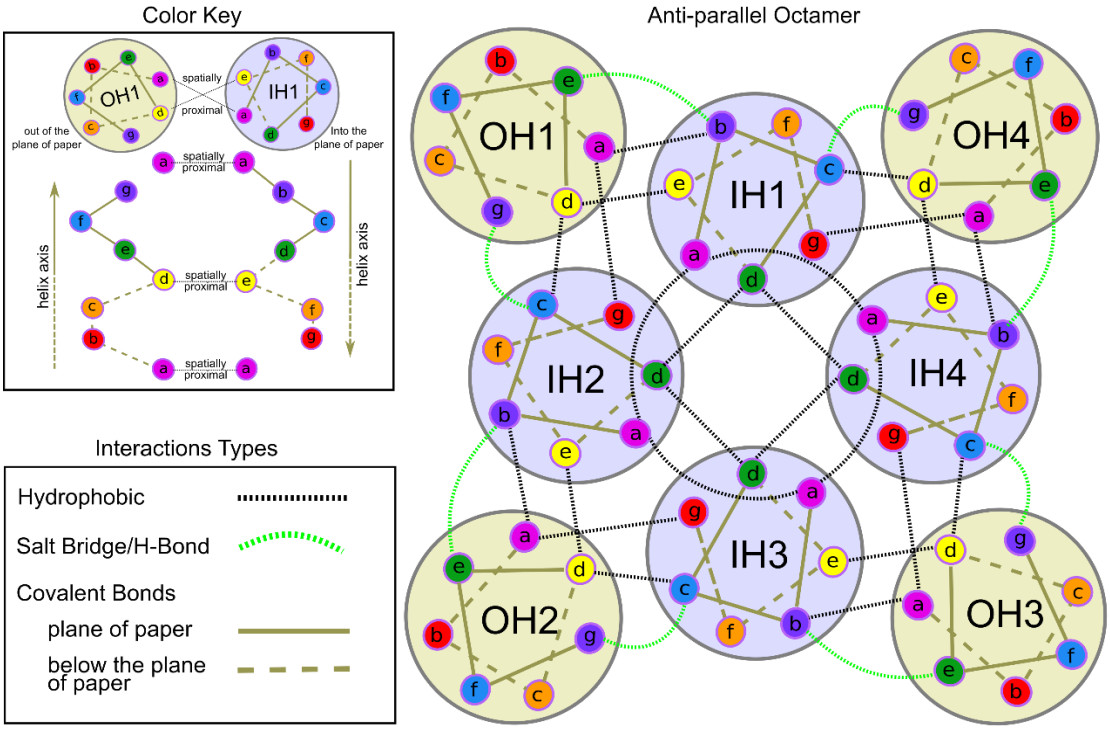


Figure 7-4 Schematic representation of Anti-parallel octamer and interactions

A schematic representation of a hypothetical anti-parallel octamer with four Inner Helix (IH) (light blue circles) and four Outer Helix (OH) (light yellow circles). Colored lines and dots represent different interaction types between helices at each heptad position. The color key gives spatially proximal interacting heptad positions (see color key details in Figure 7-1 A). For instance, OH(d) and IH(e) both have yellow color, denoting they are spatially proximal.

A schematic representation of hypothetical anti-parallel octamer with possible inter-helical interactions (Figure 7-4). Four blue right-handed helices going into the plane of paper, hereby referred as Inner Helices (IH). Four gray colored right-handed helices coming up from the plane of paper forming an anti-parallel interaction with parallel tetramer, hereby referred as Outer Helices (OH). Four IH forms the hydrophobic core through heptad positions **a-a** and **d-d**. Every OH interacts with two IH with hydrophobic interactions at heptad pairs **a-b**, **a-g**, **d-e** and **d-c** and polar interactions at heptad pairs **e-b** and **g-c**. Selecting an appropriate sequence for OH and IH satisfying these interactions would result in a stable structure.

Anti-parallel octamer similar to the schematic shown in Figure 7-4 was constructed in two steps. Step 1 determines the optimal amino acid sequence corresponding to the anti-parallel octamer. Step 2 models the structure of anti-parallel octamer corresponding to the sequence determined in step 1. The following paragraphs discuss the two steps in details.

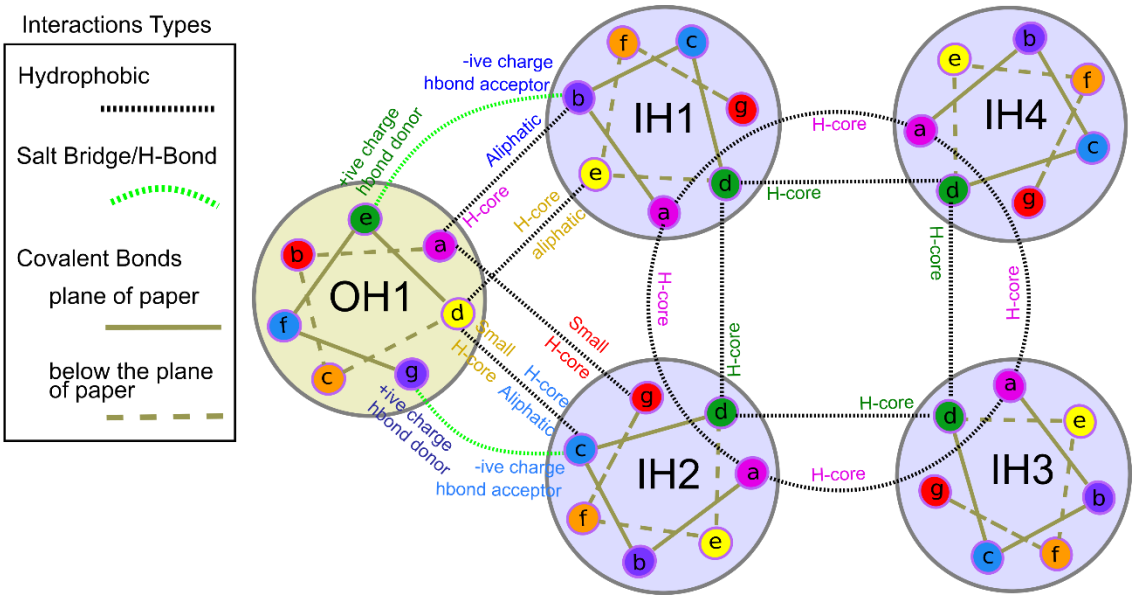


Figure 7-5 Different types of interactions in Anti-parallel octamer.

*Different types of interactions observed between two Inner helices (IH) and one outer helix (OH) in the anti-parallel octamer. Other OHs interactions with IHs will be identical to the shown interaction owing to symmetry. Every OH interacts with two IH through hydrophobic amino acids at heptad pairs (**a-g**, **d-e**) or aliphatic side chains at heptad pairs (**a-b**, **d-c**). The hydrophobic core formed by these positions are surrounded by polar/charged interactions (**e-b** and **g-c**).*

1. Sequence Determination:

A heptad repeat template for all the eight helices in the proposed Anti-parallel octamer topology was constructed (Figure 7-4). This topology was chosen as it was similar to fusion protein in many viruses. Hence, a part of the fusion protein was utilized in modeling template (see part 2 of this section). The heptad repeat template should specify the orientation of helices with respect to other helices (Figure 7-4).

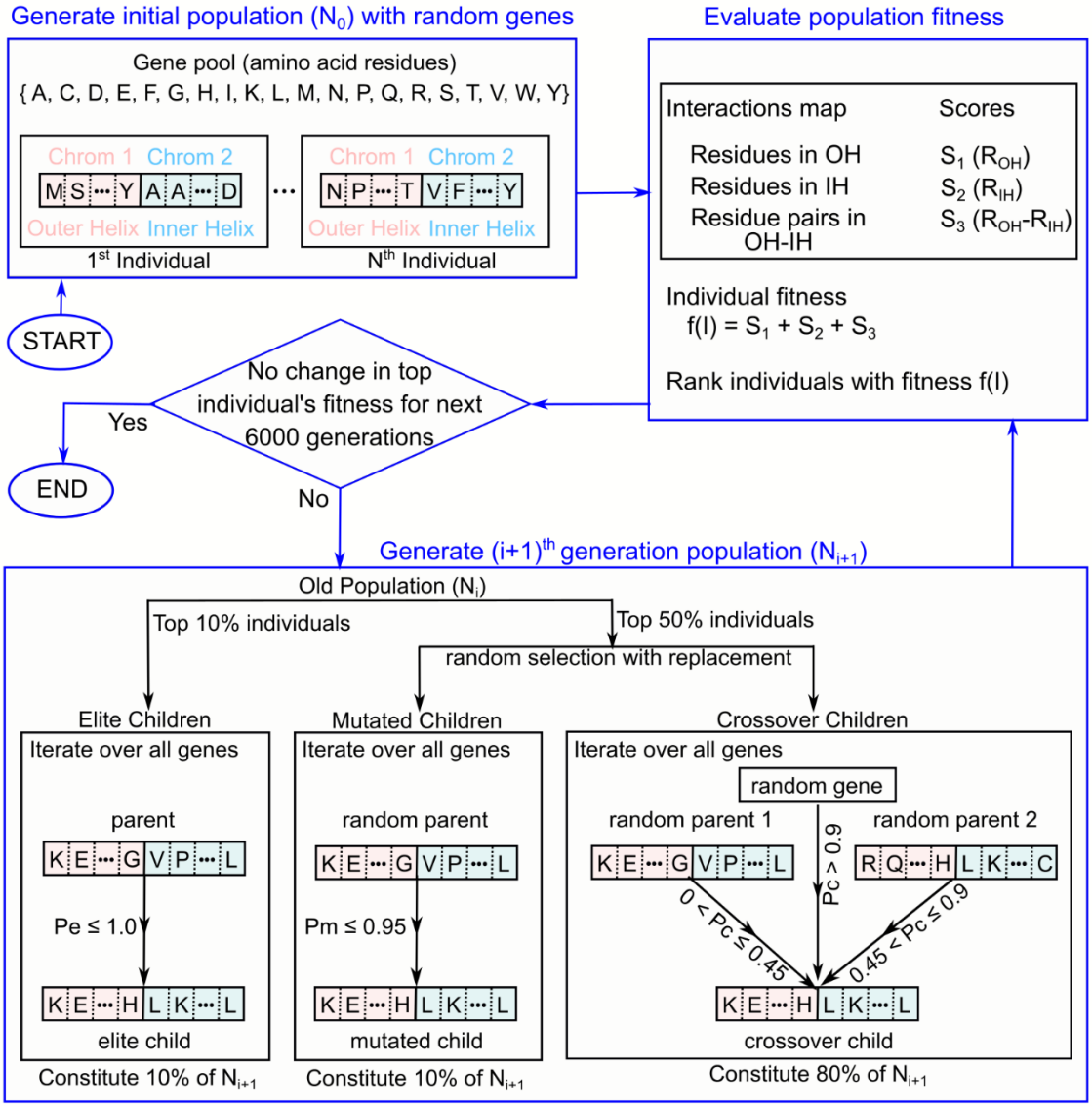


Figure 7-6 Genetic algorithm to determine OH and IH sequence.

Algorithm has four major parts (blue labels). 1) Generating Initial population, 2) Evaluation of population and ranking of individuals, 3) Termination criteria, and 4) Constructing next generation. Following representation was used in the figure. Genes: Amino acid sequence; Chromosomes (coral and turquoise): OH/IH sequence; Individual: OH-IH sequence pair; Population: set of OH-IH sequence.

The heptad repeat template guides the possible interactions between the various heptad positions. For instance, heptad positions **a/d** of all Inner helices IHs (**a/d**) forms the hydrophobic core of the tetramer. Similarly, heptad positions **b/c/e/g** of IH interacts with heptad positions **a/d/e/g** of the neighboring anti-parallel helix OH (Figure 7-5).

Starting with the random sequences of Inner helix and outer helix, a search for the optimal sequence was performed using genetic algorithm (Figure 7-6). An individual in the algorithm has two chromosomes (coral and turquoise) containing one inner helix (IH) and one outer helix (OH) amino acid sequence. Each individual has genes that could be any of the twenty amino acids. An initial population of 100 individuals (N_0) with randomly assigned genes were initialized (Figure 7-6 Initialization box).

Evaluation of an individual form the population has three components. IH residue scores, OH residue scores and IH-OH interacting residue scores. Residues in IH and OH were obtained directly from the individual's sequence. OH-IH interacting residue pairs were obtained from the anti-parallel octamer topology. Following were the scoring functions for evaluating the three components (Figure 7-6 Evaluation box).

$$"IH\ residue\ scores"\ \equiv S_2(R_{IH}) = \begin{cases} 1.0/FD_{tet}(R_{IH}), & heptad \in a,d \\ 0.1, & HR, heptad \in e,g \\ 1, & non\ HR, heptad \in e,g \\ 0.1, & NPR, heptad \in b,c \\ 1, & non\ NPR, heptad \in b,c \\ 0, & Serine, heptad \in f \\ 1, & non\ Serine, heptad \in f \end{cases}$$

Where HR means Hydrophobic residues, NPR stands for negative/polar residues and FD_{tet} stands for frequency distribution of parallel tetramers (Figure 7-2).

$$"OH\ residue\ scores"\ \equiv S_1(R_{OH}) = \begin{cases} 0.1, & HR, heptad \in a,d \\ 1, & non\ HR, heptad \in a,d \\ 0.1, & PPR, heptad \in e,g \\ 1, & non\ PPR, heptad \in e,g \\ 0, & Serine, heptad \in b,c,f \\ 1, & non\ Serine, heptad \in b,c,f \end{cases}$$

Where HR means Hydrophobic residues, PPR stands for positive/polar residues.

$$"OH-IH\ residue\ scores" \equiv S_3(R_{OH-IH}) = \begin{cases} 1.0/FD_{trimer}(R_{IH}), & \forall\ IP \\ 0.1, & OC, heptad\ pairs \in eb, gc \\ 1, non\ OC, heptad\ pairs \in eb, gc \end{cases}$$

Where OC stands for Opposite Charge, IP stands Interacting pairs and FD_{trimer} stands for frequency distribution of anti-parallel trimer (Figure 7-3).

The algorithm stops when the score of the fittest individual in the population do not change over 6000 generations. If the optimal solution was not obtained during the optimization, then the next generation individuals (children) were constructed in three ways (Figure 7-6 Generate new population box). 1) Elite children (10%), 2) Mutation children (10%) and 3) Crossover children (80%). Top 10% of the individuals were 10 elite children and directly go to the next generation without any change to the sequence (genes). The top 50% of the old population was used to generate 10 mutated children and 80 crossover children. The mutated children were generated by randomly selecting a parent from the old population and copy its genes to the child with a mutation frequency of 0.05. The crossover children were generated by mating of two randomly selected parents from the population. During mating, each parent provide 45% of its randomly selected genes while rest 10% of genes were obtained by random sampling from the gene pool. The size of the next population remains identical to the old population.

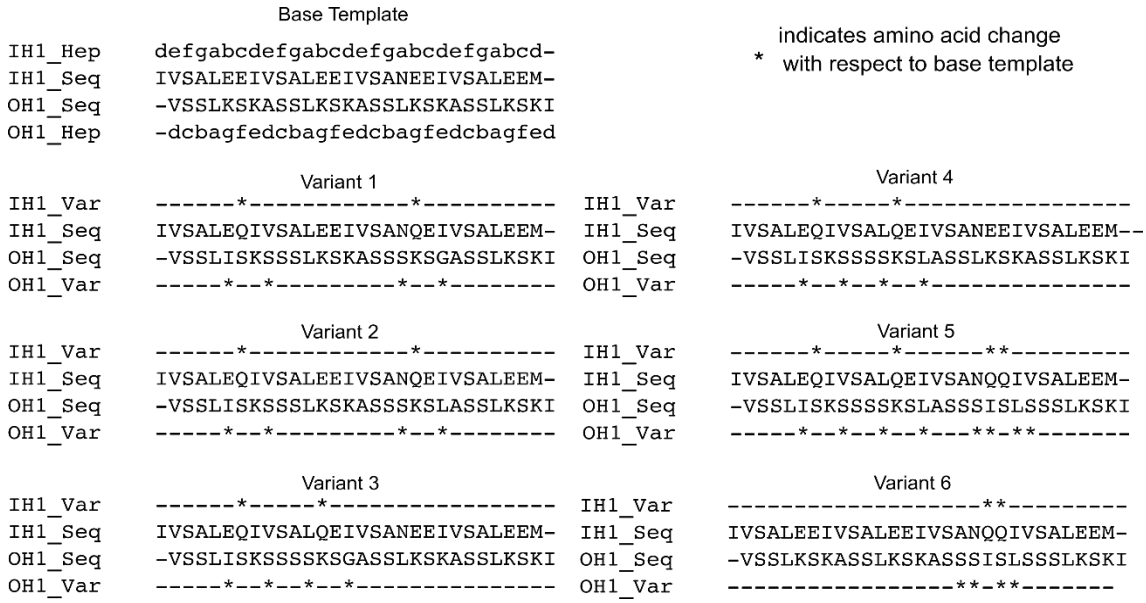


Figure 7-7 Anti-parallel octamer Inner and Outer helix sequences.

Anti-parallel octamer base template sequences (OH1 and IH1) were designed to construct variant sequences (OH1 and IH1). Sequence in all variant (1 to 6) were marked with (*) indicating change in amino acid with respect to Base template sequence. Sequences of other IHs and OHs was identical to OH1 and IH1.

The fittest individual corresponds to the lowest score for the given heptad repeat template of anti-parallel octamer (Figure 7-7, Base template). Since outer helix (OH) sequence could also independently form dimer/trimer/tetramer, the sequence was modified to prevent independent oligomerization without Inner helix. Serine amino acid was not preferred at heptad position **a/d** in homo-dimers/trimer/tetramers but preferred in anti-

parallel trimers (Figure 7-2 and Figure 7-3). Thus, serine amino acid was added in inner helix at **a/d** heptad positions and corresponding changes were made in the other interacting positions to generate six variants of Anti-parallel octamers (Figure 7-7).

2. Structure Modeling

The sequence determined in the previous section was optimized for the Anti-parallel octamer topology. The 8-helix bundle topology was chosen, as it was similar to fusion protein in many viruses. An Anti-parallel octamer template was constructed by structurally superimposing "a parallel homo-tetramer similar to four Inner Helices (IHs)" and "an anti-parallel trimer similar to one Outer Helix (OH) and two Inner Helices (IHs)". Parallel homo-tetramer coordinates were obtained from Nipah virus phosphoprotein tetramerization domain (PDB id: 4N5B) and anti-parallel trimer coordinates were obtained from Nipah virus fusion core (PDB id: 1WP7). Two parallel helices from parallel homo-tetramer (IH1-IH2) were superimposed with two parallel helices of anti-parallel trimer using CLICK [187,190]. The procedure was repeated for IH2-IH3, IH3-IH4, IH4-IH1 parallel helix pairs thereby generating Anti-parallel octamer structure. This structure was used as a template to model the Anti-parallel octamer with the sequence determined in the previous section.

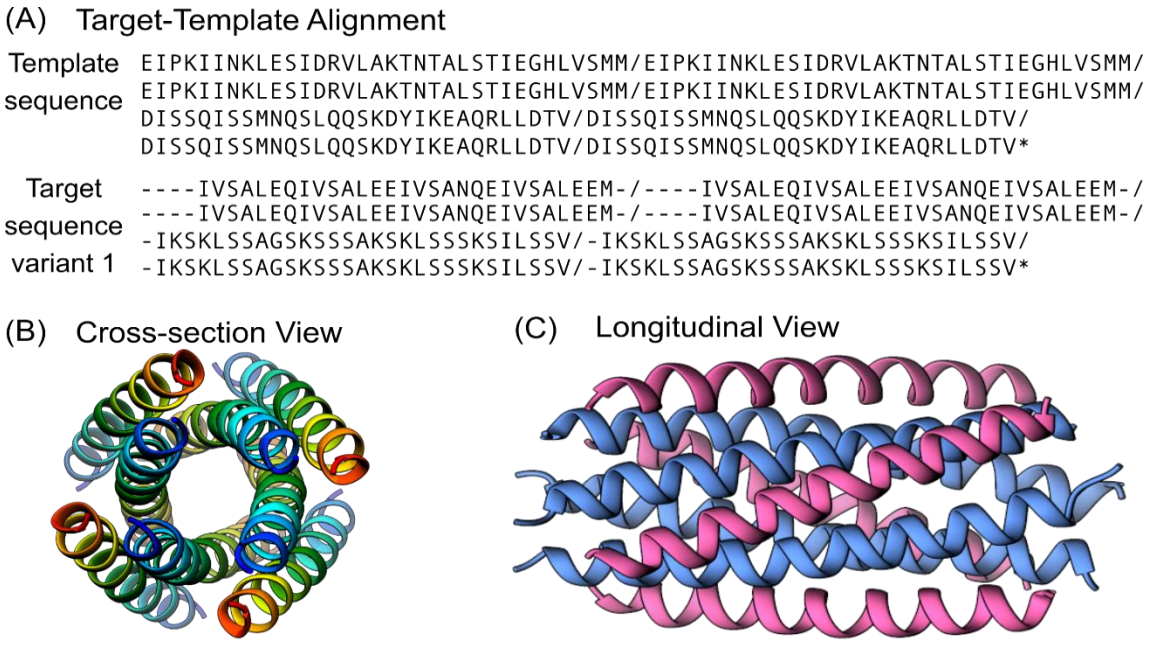


Figure 7-8 Ribbon representation of Anti-parallel octamer model.

(A) Target sequence and variant one sequence alignment. '/' and '' indicates chain and sequence break respectively. Other variants also align identically (B) Cross-sectional view of Anti-parallel octamer model with N-terminus (blue) to C-terminus (red). (C) Longitudinal view with Inner helices (blue) and Outer helices (pink).*

The template was aligned with all target sequences (variants 1 to 6 in Figure 7-7 of "Anti-parallel octamer" using heptad repeats (Figure 7-8 A). Targets were then modeled using Modeller set of routines with default parameters [26,211]. Total five models of each variants were generated and the model with the least DOPE score was chosen as the best model (Figure 7-8). Every variant model were energy minimized for 200ps followed by 200ps equilibration through NVT and NPT simulations (see technical details of MD simulations in Section 0and section 3.6.2). The molecular dynamics simulations of all variants were performed for 100ns for stability analysis.

7.1.3 Structural stability analysis of all Anti-parallel octamer variants

Structures of all six variants of anti-parallel octamer (Figure 7-7) were subjected to molecular dynamics simulation. The 100ns trajectories were analyzed for Root Mean Square Deviation (RMSD) (Figure 7-9) and Root Mean Square Fluctuations (RMSF) (Figure 7-10).

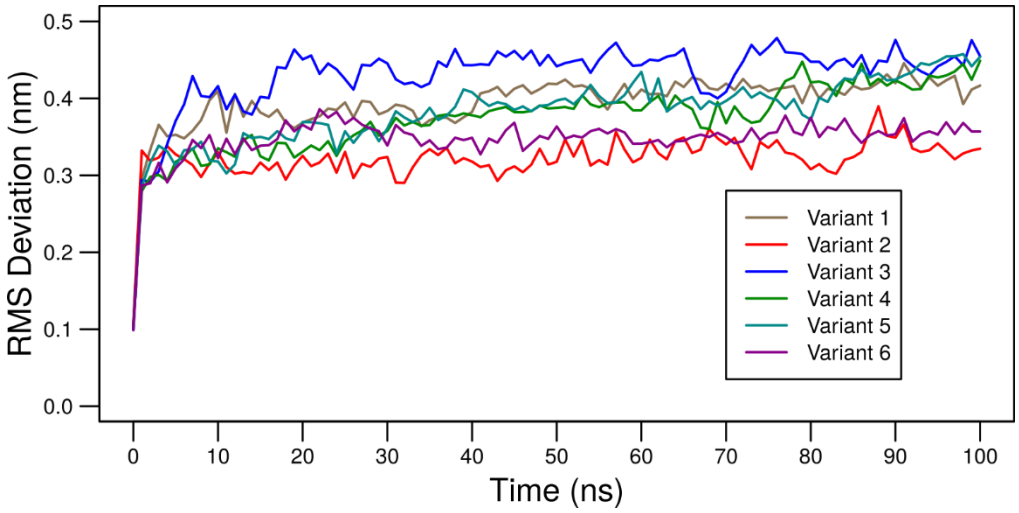


Figure 7-9 RMSD of Anti-parallel octamer variants.

Root Mean Square Deviation (RMSD) profile (y axis) of all six variants of Anti-parallel octamer calculated across 100ns. RMSD values saturates between 0.3nm to 0.5nm.

RMSD values of all six anti-parallel octamer variants shows an increase in RMSD values to 3Å in the first 3ns. After 3ns, the RMSD values saturates between 3Å to 4.5Å. The initial rise in the RMSD values was due to inappropriate packing of outer helices (OHs) with respect to inner helices (IHs) obtained from the artificially made template structure. However, after 3-4ns the outer helices repacks against inner helices to obtain the low energy state. All anti-parallel octamer variants do not show significant structure change over the course of simulation and remain stable.

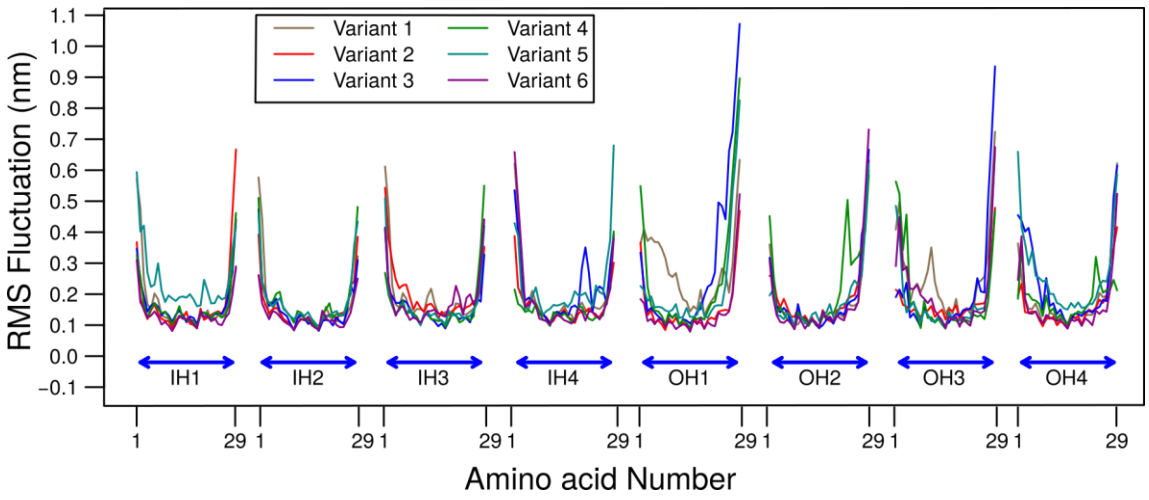


Figure 7-10 RMSF profile of Anti-parallel octamer variants.

Root mean Square Fluctuation (RMSF) profile (y axis) of all six variants of Anti-parallel octamer calculated over 100ns. IH1 to IH6 stands for Inner Helices, and OH1 to OH6 stands for Outer Helices with their residue numbers (1 to 29) indicated on x axis.

RMSF values of all six anti-parallel octamer variants show fluctuations between 1Å to 2Å for all the amino acids (except terminal residues). Two to three terminal amino acids of all eight helices show high RMSF values between 3Å to 7Å due to unfolding of helices during the simulation. Since there were no anchoring residues at the terminal domains to dampen the thermal fluctuations, such simulation dynamics were expected from any coiled-coil helices including simulation of high-resolution crystal structures (see an example in Figure 3-14). A more robust way to test is to do free energy-based simulations that would provide more confidence in the design and stability. Apart from the terminal regions, the 8-helix bundle assembly of anti-parallel octamer remains stable over the simulation. The experimental validation of this novel coiled-coil protein is currently being done in the lab of Dr. Raghavan Varadarajan, IISc Bangalore.

7.2 Designing inhibitor for Nipah Virus F protein

The infection caused by Nipah Virus (NiV) leads to the death of many individuals [241–245]. Currently there are no drugs to treat or prevent NiV infections [241]. The NiV genome encodes six structural proteins (G, F, M, N, L, P) and three non-structural proteins (C, W, V) [246,247]. Each of these proteins are involved in mediating the viral infection. In our recent study, the peptide/drug inhibitors corresponding to each of these proteins was designed using various computational approaches [248]. This section describes the construction of the peptide inhibitor corresponding to the coiled-coil domain of F protein.

F protein mediate the fusion of viral and host membrane through its coiled-coil HRA (blue) and HRB domain (red) (Figure 7-11). The fusion mechanism involves the transition of F protein in pre-fusion conformation to intermediate state followed by post-fusion conformation (Figure 7-11) [247,249,250]. During the fusion event, pre-fusion conformation bound to the viral membrane (via HRB domain coiled-coil trimer) changes its conformation to an intermediate state that forms a coiled-coil trimer of HRA domain. HRA domain punctures the host membrane while the HRB domain still binds to the viral membrane. HRA domain along with HRB domains subsequently helps lipid membranes to fuse while having change in conformation from intermediate state to post-fusion state. This results in the formation of a stable 6-helix bundle (6HB) coiled-coil assembly. The aim of this study is to determine the helical peptide that would bind to HRA domain stronger than HRB domain, thereby inhibiting the host-viral membrane fusion.

7.2.1 Modeling F protein Post-Fusion conformation

NiV F protein is known to exist as homo-trimer in two stable states. Pre-fusion/post-fusion states indicates its conformation before/after the fusion of viral and host membrane [251]. The structure of the pre-fusion complex was solved experimentally (PDB id: 5EVM) (Figure 7-11 A). However, the structure of the post-fusion complex is unknown that could provide vital information in designing an inhibitor for NiV F protein. Since the NiV F protein shares 40.5% similarity and 26.4% identity with Human Parainfluenza Virus (HPIV) F protein, the structure of the post-fusion state was modeled using homology modeling considering HPIV F protein (PDB ID: 1ZTM) as a template. The alignment between target and template sequences was done using CLUSTAL O (v1.2.4) [252]. A trimer of the post-fusion state was modeled using Modeller suite of programs with default parameters (Figure 7-11 C).

7.2.2 Modeling F protein intermediate state conformation

The pre-fusion and post-fusion complex structures were used to model the intermediate state conformation using linear morphing of the two structures (Figure 7-11 B). The morphing was done using AIMorpher tool (see section 9.4 for details and algorithm). The morphing tool generates a range of intermediate structures using linear/non-linear interpolation. The tool could generate morph structures between multiple meta-stable intermediate structures. Considering meta-stable states during morphing provides more accurate morphed structures compared to conventional morphing. In case of F protein, it was known that at least one meta-stable conformation was present [253] (Figure 7-11 B).

Thus, a two-step morphing considering one intermediate state and one final state was done to obtain 50 intermediate states between pre-fusion and post-fusion conformations. The intermediate structures were generated to identify the flexible residues/domains during large-scale movements. The interpolated structures may not corresponds to the actual transition trajectory. A more accurate way to determine the intermediates are using tools that generate interpolated structures along with minimization [254]. The interpolated structures were not used in any form to design the inhibitor.

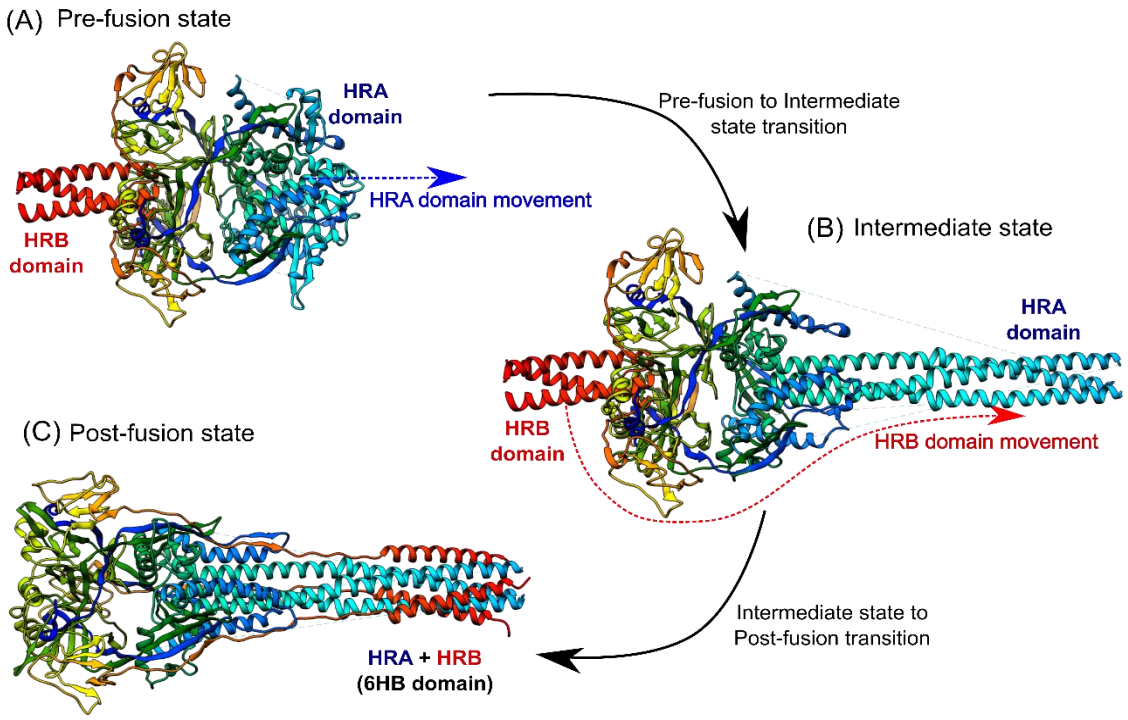


Figure 7-11 Pre-fusion state to post-fusion state transition of F protein

Schematic representation of F protein transition from pre-fusion state (A) to intermediate state (B) followed by post-fusion state (C). The blue arrow indicates the movement of HRA domain (3 cyan helices) from pre-fusion to intermediate state. The red arrow indicates the movement of HRB domain (3 cyan helices) from intermediate state to post-fusion state to form 6HB domain. In 6HB domain, HRA domain formed by three helices (cyan) were sandwiched by three helices of HRB domain (red).

7.2.3 Modeling Inhibitor of F protein transition event

F protein contains two helical domains namely HRA and HRB. Initially, HRA domain forms a three-helix coiled-coil bundle which then associates with three helices of HRB domain to form a six-helix coiled-coil bundle (6HB) [250] (Figure 7-11). Formation of six-helix bundle domain is essential for fusion of viral and host membrane. Both HRA and HRB that fold into a coiled-coil structure in the post fusion state following a heptad

repeat pattern (see Section 2.1 for details on coiled-coils) [255]. A coiled-coil and ribbon representation of both the HRA (blue) and HRB (yellow) domains shows the position of each heptad relative to the six-helix bundle (Figure 7-12, A, C). An alignment of HRA and HRB domain sequence provides the amino acids that forms the core of the six-helix bundle (Figure 7-12, E).

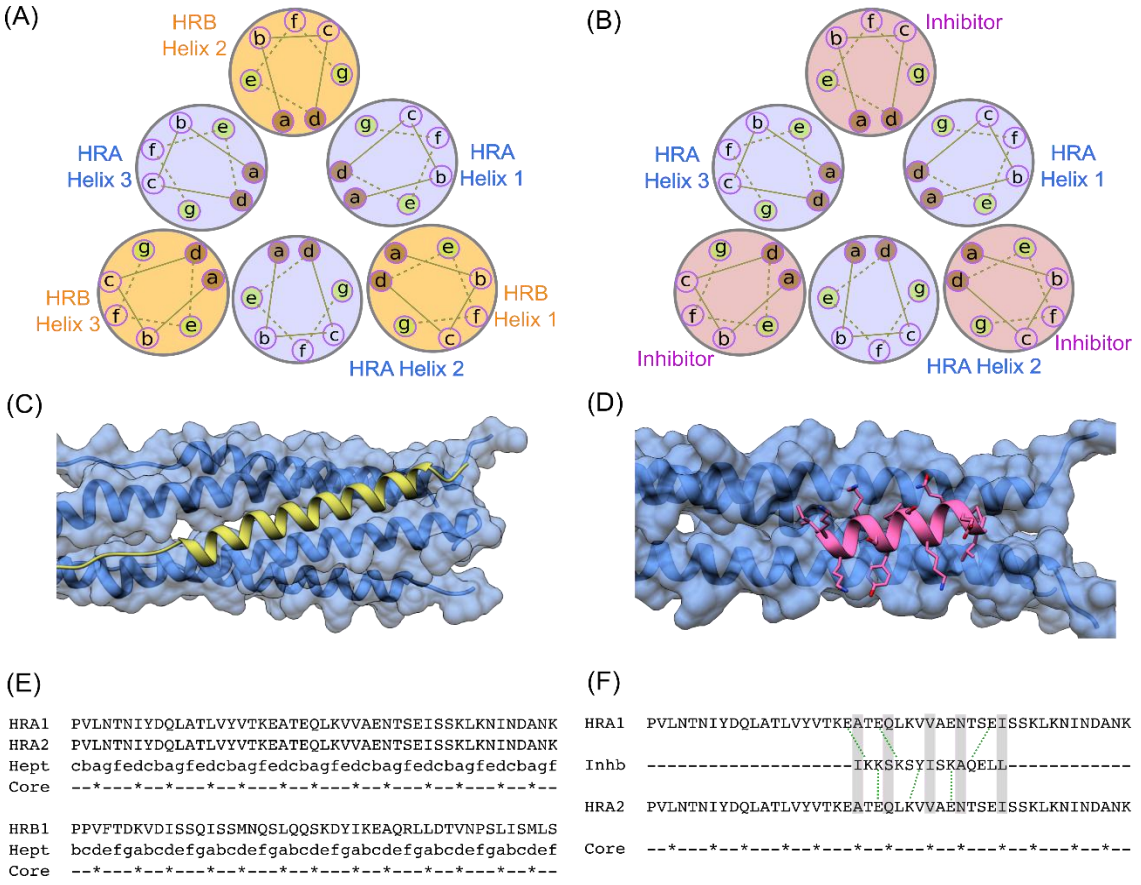


Figure 7-12 Six-Helix bundle domain inhibitor design and interactions.

*A) Heptad repeat representation of the 6-HB domain formed by HRA (blue) and HRB (yellow) domain helices. Every helix (big blue/yellow circle) have amino acid (small circles) at heptad positions **a-g**. B) Heptad repeat representation of HRA domain (blue) and bound inhibitor (pink) replacing HRB domain. C) Crystal structure of 6HB bundle with HRA domain (blue) in surface representation and HRB domain (yellow) in ribbon representation. D) Full atom model of inhibitor (pink) in ribbon representation bound to HRA domain (blue) in surface representation. E) Heptad repeat (Hept) assignment of HRA domain helices (HRA1 and HRA2) with HRB domain helix (HRB1) along with hydrophobic core **a/d** positions (Asterisk) F) Interactions of an inhibitor bound to two helices of the HRA domain (HRA1 and HRA2). The inhibitor interacts through the salt bridges (green dotted lines) and the hydrophobic core (gray bar).*

An amino acid sequence mimicking the heptad repeat of the HRB domain was designed such that all hydrophobic amino acids occupy **a/d** heptad positions (Figure 7-12 B, D). The amino acids were selected such that inhibitor forms the same number of salt bridges (green dotted lines) to each helix of HRA domain (Figure 7-12 F). Polar amino acids occupy the heptad positions that were exposed to the solvent and were not making any interaction with the HRA domain. Once the sequence was determined for the inhibitor (Figure 7-12 F, Inhb), its heptad repeat guided alignment was performed against six-helix bundle domain (Figure 7-13). The alignment was used to model the inhibitor using Modeller with default parameters [26,211]. Total 5 models were built and the model with the least DOPE [57] score was selected as final model (Figure 7-12, Panel D).

Target-Template Alignment

Template sequence	HRA1	KNADNINKLKSSIESTNEAVVKLQETAEKTVYVLTALQDYINTNLVPTIDKISCK/
	HRB1	PPVFTDKVDISSQISSMNQSLQQSKDYIKEAQRLLDTVN/
	HRA2	KNADNINKLKSSIESTNEAVVKLQETAEKTVYVLTALQDYINTNLVPTIDKISCK*
Target sequence	HRA1	KNADNINKLKSSIESTNEAVVKLQETAEKTVYVLTALQDYINTNLVPTIDKISCK/
	Inhibitor	--------------------IKKSKSYISKAQELL----/
	HRA2	KNADNINKLKSSIESTNEAVVKLQETAEKTVYVLTALQDYINTNLVPTIDKISCK*

Figure 7-13 Alignment of Inhibitor against six-helix bundle domain

Target sequence comprise of HRA1 helix and HRA2 helix and neighboring HRB1 helix. Target sequence has both HRA1 and HRA2 helices along with Inhibitor sequence mimicking HRB1 helix. '/' and '' indicates chain and sequence breaks.*

7.2.4 Assessment of six-helix bundle-inhibitor complex

MD simulations were carried out in triplicates for the predicted six-helix bundle inhibitor complex. The simulations were carried out using GROMACS [193,194] with the Amber99SB-ILDN force field [256]. A cubic water box whose sides were at a minimum distance of 12 Å from any protein atom (total 3417 protein atoms) was used for solvating each of the systems (total 37176 solvent atoms). In all nine chloride ions were added to achieve charge neutrality. Electrostatic interactions were treated using the particle mesh Ewald sum method [231] and LINCS[232] was used to constrain hydrogen bond lengths. A time step of 2 fs was used for the integration. The whole system was minimized for 5000 steps or until the maximum force was less than 1000 kJ/mol/nm. Using a Berendsen thermostat, the system was then heated to 300K in an NVT ensemble simulation for 100 ps. The pressure was stabilized in an NPT ensemble simulation for 100 ps using a

Parrinello-Rahman barostat. The system was simulated for a maximum of 100 ns and structures were stored after every 10 ps. The temperature, potential energy and kinetic energy were monitored during the simulation to check for anomalies.

7.2.5 Stability of Six-helix bundle domain inhibitor complex

The inhibitor against F protein was designed with the amino acid sequence IKKSKSYISKAQELL following the heptad repeat pattern, to prevent the formation of six-helix bundle (6HB) (see Section 7.2.3 for construction details).

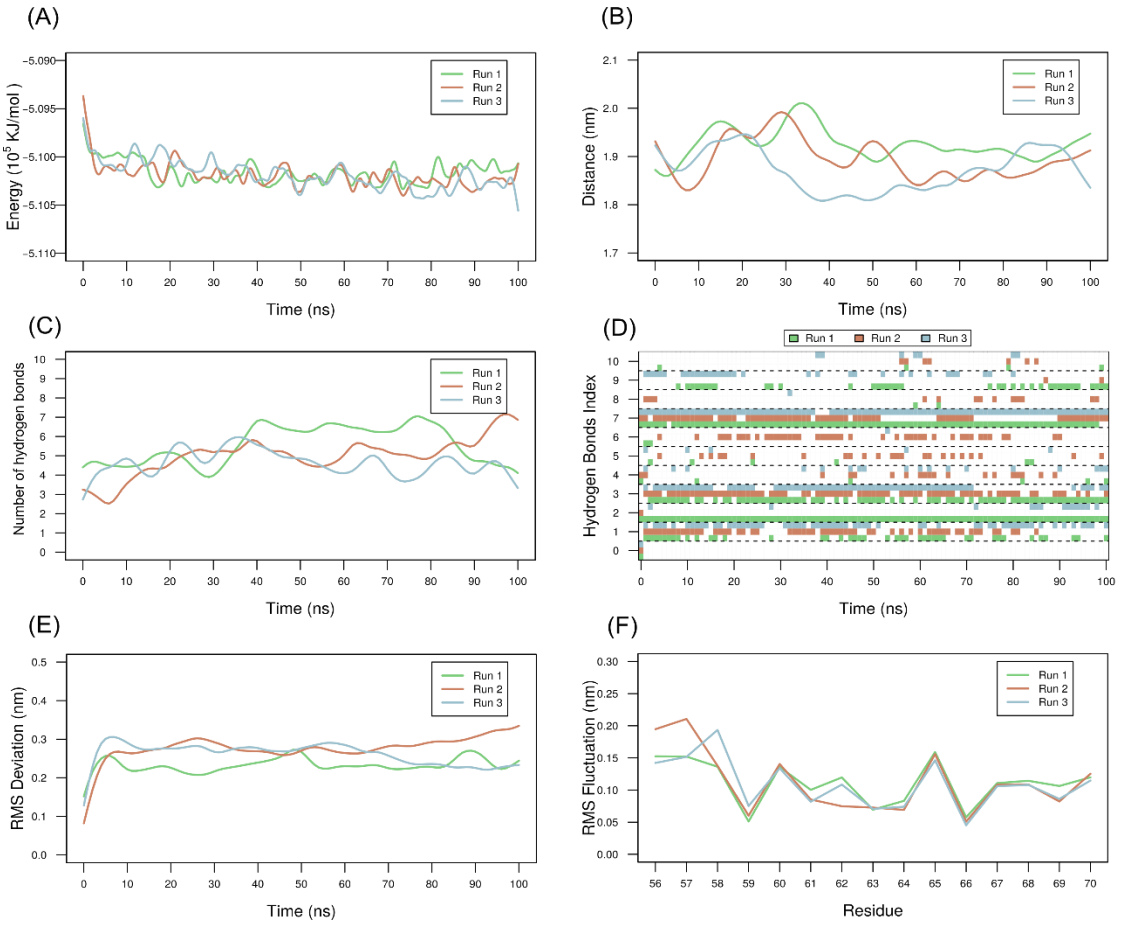


Figure 7-14 MD simulation stability analysis of F protein inhibitor

Triplicate MD simulations (red, green, blue) over 100ns (x axis). A) Energy (y axis) of the F protein-inhibitor complex. B) Distance (y axis) of the centers of the inhibitor and F protein. C), D) Number (y axis) and existence matrix of hydrogen bonds (rectangular color box) between the F protein-inhibitor complex. E), F) Root mean square deviation (y axis) and fluctuation (y axis) of the inhibitor residues (x axis) during the simulations.

The formation of 6HB is essential for the fusion of the virus in the endocytic vesicle membrane. The designed peptide would hence bind competitively preventing viral entry

into the cell. Three independent MD simulations for 100 ns were performed to assert the stability of the complex. For each of the trajectories the total potential energy, distance between protein-peptide, total number of protein-peptide hydrogen bonds, existence matrix of hydrogen bond, RMSD and RMSF of the peptide were analyzed. The total potential energy of the system remains at an average value of -510178 kJ/mol (Figure 7-14 A). The peptide remain bound to the protein with an average center-to-center distance of 1.9 nm (Figure 7-14 B). The F protein peptide inhibitors forms an average of five hydrogen bonds (Figure 7-14 C). The existence matrix of the hydrogen bond provided 11 uniquely indexed hydrogen bonds (see Table 7-2 for details of indexes) across 100 ns simulation (Figure 7-14 D). Each rectangular color box represent the presence of hydrogen bond for a particular run. Hydrogen bonds with index 2, 3, 4, and 8 provide consistent main chain – side chain interactions and thus crucial in stabilizing the protein-peptide inhibitor complex. The strength of binding were also corroborated by an average peptide RMSD of 0.25 nm (Figure 7-14 E) and average peptide RMSF (Figure 7-14 F) of 0.11 nm. The mean and standard deviation values for all structural properties shows the consistency across independent runs (Table 7-1). The F protein inhibitor appears stable as estimated from the hydrogen bonding analysis. This inhibitor would putatively prevent the pre to post fusion transition of F protein required for viral entry and therefore could be used as a lead compound for further drug development.

Table 7-1 Stability parameter values of F protein inhibitor

Mean and standard deviation of the energy, distance of the center of the inhibitor with the center of the protein, number of hydrogen bonds between the inhibitor and the protein, RMSD and RMSF of the peptide residues obtained from the three 100ns MD simulations of F protein-inhibitor complex.

Run	**Energy (kJ/mol)**		**Protein-peptide distance (nm)**		**Number of Hydrogen Bonds**		**RMSD (nm)**		**RMSF (nm)**	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
1	-510162	1093	1.93	0.05	5.54	1.48	0.23	0.04	0.11	0.03
2	-510194	1084	1.89	0.05	4.99	1.74	0.28	0.04	0.11	0.05
3	-510179	1090	1.87	0.06	4.62	1.32	0.26	0.04	0.11	0.04
Mean	-510178		1.90		5.05		0.26		0.11	

Table 7-2 Hydrogen bonds existence percentage across different independent runs

F protein HRA domain has chain ids A, C, E and Inhibitor has chain ids chain B, D, F. Each row corresponds to the number of times a hydrogen bond. Inhibitor residue numbers in chain D (56.D-70.D) and neighboring protein chains C and E.

Hydrogen Bond Index	Hydrogen Bond	Run 1 (in percentage)	Run 2 (in percentage)	Run 3 (in percentage)
1	56.D-26.E	1	1	1
2	56.D-28.C	41.6	55.4	59.4
3	59.D-19.E	100	1	12.9
4	59.D-24.C	77.2	75.2	68.3
5	60.D-24.C	5.9	18.8	17.8
6	65.D-15.E	5.9	22.8	5
7	65.D-18.E	2	42.6	1
8	66.D-12.E	97	64.4	97
9	67.D-14.C	2	12.9	1
10	67.D-17.C	36.6	2	31.7
11	69.D-8.E	4	6.9	6.9

7.3 Predicting Oligomeric state of coiled-coil assembly

Various software exist for determining the oligomeric states of coiled-coil proteins [82,257–260]. However, except LOGICOIL, other software could only predict the oligomerization state of parallel dimers or trimers. LOGICOIL works by using bayesian variable selection approach and outperforms most software with AUC value 0.9 and accuracy 0.95 [257]. LOGICOIL addressed the oligomerization state prediction problem for a comparatively small set of coiled-coil dimers, trimers, and tetramers. However, coiled-coil assemblies are much more complex and are not limited to dimer/trimer/tetramers [92]. This section describes a method “RFMCOIL” (on a much bigger dataset) that predicts the oligomerization state of any coiled-coil assembly irrespective of the number of monomers. The coiled-coil structural rules described previously in this report were used for determining the oligomerization state.

7.3.1 Obtaining the coiled-coil data

All canonical parallel homo-oligomeric 2/3 helices coiled-coil structures, greater than 14 amino acid and less than 50% redundancy were obtained from the CCPLUS database [92]. A total of 196 dimers, 88 trimers and 22 tetramers were obtained from the database.

7.3.2 Conversion of dataset into sliding windows

For each coiled-coil structure from the above dataset, its amino acid sequence and heptad repeat pattern were extracted. The mapped amino acid sequence and the corresponding heptad repeat pattern were translated into sliding window dataset. The amino acid sliding windows were obtained by considering only continuous **a/d** heptad positions, i.e. "**adad**" and "**dada**". The known states (dimer/trimer/tetramer) of every sliding window in the complete database were recorded. Each amino acid in the sliding window was converted to a numerical score with the Atomic density packing feature established earlier (see section 3.3.3 for details).

7.3.3 Training and testing of Random Forest classifier model

A 10-fold cross-validation method was used to train and test the random forest ensemble model with 500 decision trees. Random 90% from the total number of dimers/trimmers/tetramers were used for training the model whereas rest 10% dimers/trimmers/tetramers were used to test the predictions. After selection of the random training and testing set, the data were converted to sliding windows as explained in the next section. The random forest model was then trained on the training sliding windows set. After training the model, it was used to predict the state of each sliding windows in the testing set. Predicted state of a dimer/trimer/tetramer in the training set was determined as the state having the most number of identical predictions. After each prediction, Matthews Correlation Coefficient (MCC) and confusion matrix were computed. The process was repeated for 10 times in accordance with the standard 10-fold cross-validation methodology.

7.3.4 Results and Drawbacks of the RFMCOIL method

Across 10-fold cross-validation, average MCC value is 0.97±0.04 indicating a near perfect classification of dimer, trimers and tetramers. Table 7-3 highlights various prediction parameters of a single run (having MCC 0.98) of the random forest model.

Table 7-3 Classification report of random forest model

	Accuracy	**Precision**	**Recall**	**F1 score**
Dimer	0.99	0.99	1.00	0.99
Trimer	0.99	1.00	0.99	0.99
Tetramer	0.99	1.00	0.95	0.98
Average	0.99	0.99	0.98	0.98

Table 7-4 Comparison with existing oligomerization prediction software

First column list out the software name, second column gives area under curve value for predictions; third and fourth column provides recall and accuracy respectively. Table adapted from [257]

Algorithm	**AUC**	**Recall (dimer/trimer/tetramer)**	**Accuracy**
LOGICOIL	0.90	1/0.88	0.95
SCORER 2.0	0.85	0.86/0.24	0.71
PrOcoil	0.89	0.97/0.60	0.88
Multicoil	0.591	0.09/0.00	0.06
Multicoil2	0.66	0.35/0.20	0.32
RFMCOIL (this work)	0.99	1/0.90/0.95	0.99

The results obtained from RFMCOIL were compared with the known software (Table 7-4). However, the comparison with other software was not appropriate due to very different datasets. The method described here out performs all other software. However, the main utility of this work is to be able to predict the oligomerization state of any coiled-coil assembly. The number of structures for higher order coiled-coils were very less (one or two in some cases during testing). The sliding window approach makes use of previous knowledge of the oligomerization state (see Section 7.3.2). Thus, the method needs to be redesigned such that it works without explicitly using the structural information and utilize only the coiled-coil rules (see Section 3.3.7, Section 3.5 and Section 6.2 for coiled-coil structural rules).

7.4 Summary

Coiled-coil structural rules obtained from the Keratin models and other coiled-coil structures were employed to three applications. 1) Designing novel coiled-coil structures, 2) Designing inhibitors for coiled-oil proteins and 3) Predicting the oligomeric state of a coiled-coil assembly.

In the first application, a novel coiled-coil assembly of 8-helix bundle was constructed without using any known template (Not seen in the PDB database until date). The sequence and structure was determined purely from the structural rules developed in this work. Genetic algorithm was used to determine the best possible sequence for the specified geometry. The scoring scheme for the optimization was developed using the structural rules. The model for the best scoring sequence was analyzed for structural stability.

In the second application, the coiled-coil structural rules were applied to design inhibitor for Nipah Virus Fusion protein (F protein). Nipah F protein has a coiled-coil 6 helix bundle (6HB) domain that exists in two different states. A small helix (inhibitor) was designed that would competitively bind to this domain. The inhibitor would be interfering with the transition process of the protein, thereby affecting the virus life cycle.

In the third application, the structural rules were used to predict the oligomeric state of a coiled-coil assembly with near perfect accuracy. The method outperforms the existing methods until date and is currently in development to incorporate all kinds of coiled-coil assemblies.

CHAPTER 8
Discussions and Outlook

SALIENT FEATURES

- ***Summaries and Conclusions of the study on Intermediate Filaments***

- ***Following main topics and possible future prospects were covered***
  - Intermediate filament dimerization principles
  - Integrative modeling of Keratin filament
  - Assessment of Keratin filament models
  - Applications of knowledge gained from Intermediate filaments

This chapter summarizes the work done on the Intermediate filaments and discusses the key findings and future prospects of the study (Figure 8-1). The study was divided into four main topics. 1) Principles of Intermediate filament dimerization, 2) Integrative modeling of Keratin filament, 3) Assessment of the Keratin filament models 4) Applications of Intermediate filaments structural rules. These topics would be discussed in the individual sections along with their possible future directions.

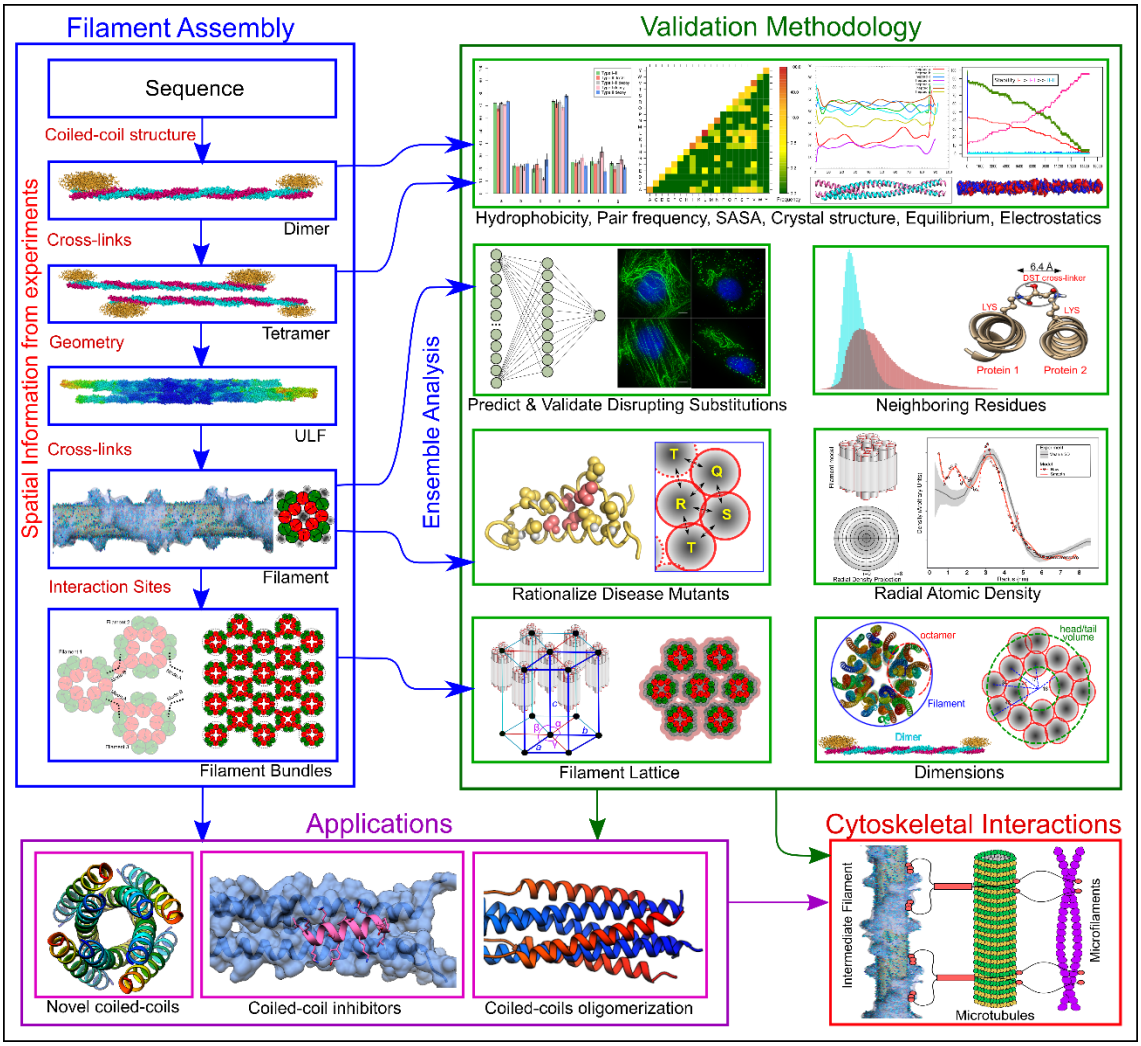


Figure 8-1 IF assembly, validation, applications and significance

Schematic diagram represents Keratin intermediate filament assembly pathway (blue box), the validation methodology performed at each step of filament construction (green box), the possible applications of the knowledge obtained during filament modeling (purple box) and the future outlook in cytoskeletal interactions (red box). The arrows indicate the knowledge transfer from different structural representation. The study at various level of details eventually leads to the knowledge of interactions between various cytoskeletal elements.

8.1 Intermediate Filament Dimerization Principles

8.1.1 Overview of Dimerization Principles

The study started with understanding the molecular basis of preferential dimerization of Intermediate filaments (IFs) and their associated diseases. A database of 61 Intermediate filaments was constructed, and their coiled-coil domains were aligned against each other. Using this database, the molecular determinants influencing the stability and specificity of IF dimers were derived. The molecular determinant includes atomic density fluctuation, pair preferences, volume/charge distributions, hydrophobicity and kinks in the helices. The numerical form of these molecular determinants was learned using an artificial neural network schema model. The trained neural network models were used for classification and prediction of dimer destabilizing residue pairs at heptad positions **a/d**. The neural network model rationalized 35% (153/438) of disease mutants, accounting for all mutations in the heptad positions **a/d**. The models were utilized to score new filament destabilizing mutations obtained through *in-silico* saturated mutagenesis of all IF family members. K5-K14 IF family mutations were chosen to validate these findings. 16 mutations were validated through *in-vivo* tissue culture experiments. Molecular determinants for artificial type II IF homodimers suggest unstable residue pairings. This leads us to rationalize preferential pairing of IFs. i.e., why type III to type VI intermediate filaments form homodimers whereas type I and type II intermediate filaments form heterodimers using kinetic Monte Carlo simulations

8.1.2 Future prospects

The molecular determinants derived from the Intermediate filament database classify the dimer/filament destabilizing residue pairs. The stability of residue pairs near terminal regions is less when compared to residues pairs at the center of a coiled-coil dimer [112,115,117]. This was also evident from the distribution of intermediate filament disease causing mutations (Figure 3-1). A residue pair could be stable at the center of the dimer but may become unstable when present near the terminus. The position dependency on the stability of residue pairs were not incorporated in the neural network schema. This addition would further improve the classification accuracy. The stability of residue pairs were used to rationalize the preferential dimerization of K5/K14 monomers using kinetic Monte Carlo approach. However, the major limitation of this approach was the absence of rate constant values. It is impossible to scan all possible rate constant

values for any system. However, an analytical solution of this system is currently in progress. It would provide a robust solution to the preferential dimerization problem and would be applicable to other systems. Preferential dimerization problem was also addressed through quantitative electrostatic calculations using APBS software [215]. This package seems to work fine for globular proteins but fails for unusually elongated proteins in one dimension (such as IF dimer). Electrostatic potentials were largely inconsistent when compared with coulomb potential. PBEQ online server [216] was giving much better results with long dimers but it was unable to support large system size with complex calculations. The probable origin of the erratic performance is due to the uneven density of grid points in different dimensions. The length of the Intermediate filament dimer is about ~50 times more compared to its width. The APBS works using equidistant grid points that forms a 3D mesh around the protein. Thus, for a given number of mesh points in the smaller dimension, the number of mesh points in larger dimensions are separated very far leading to inaccurate results. Other factor could be the implementation of decay of permittivity as pointed out by the reviewer. Implementing decay of permittivity specifically for fibrous protein could be another solution to this problem. Thus, there is a need for a software that could calculate electrostatic potential accurately for such systems.

8.2 Integrative Modeling of Keratin Filament

8.2.1 Overview of Modeling Protocol

This part of the work discusses about the Integrative modeling applied for determining the structure of Intermediate filament. Integrative modeling requires spatial information obtained by inferring experimental data. Initially, a hetero-meric K5-K14 coiled-coil dimer was constructed using homology modeling. Known coiled-coil structures and heptad repeat guided sequence alignment guide the homology modeling protocol. The K5-K14 dimer model was validated using various structural properties and compared with the available segments of the X-ray crystal structure. The number of possible dimers in the K5-K14 filament cross-section was determined using mass and volume measurements. The K5-K14 dimer model was sampled exhaustively against its duplicate to obtain valid dimer-dimer configurations. The configurations that parsimoniously satisfy cross-links observed in the experiment were recorded. Each K5-K14 dimer was converted to a 2D disc by taking a longitudinal projection on a plane. Disctransformation

software was then used to sample and score all possible configurations exhaustively (10^{16} possibilities). 699 models were obtained that were clustered in six topologically different configurations. The cluster with the optimal packing of the dimers was chosen as the final model of keratin filament. The 2D model was then mapped back to obtain the full atom 3D model.

8.2.2 Possible Future Directions

In this study, the keratin filament model at near atomic resolution was successfully constructed. Robustness of these models could be tested by excluding a few cross-links during the modeling procedure. These excluded cross-links should be satisfied in the final model. The relative frequency of cross-links could not be ascertained due to the limitation of the HPLC based cross-link data. Due to the presence of sixteen identical dimers in the assembly, the number of possibilities between the dimers satisfying the cross-links became extremely large. To solve this issue, the lateral distance between the dimers had to be fixed depending upon the inter-atomic clashes. A more robust experiment such as mass spectrometry instead of HPLC would have provided the exact relative frequency of each cross-link. This would decrease the number of configurations to search and would increase the confidence in the model.

During modeling, symmetry was not explicitly incorporated in the models. All possible sampling was done to avoid missing the valid models. The head and tail domains were not accounted during the filament modeling. Head/tails domains were added later to the final model. They were represented by the density of virtual atoms (equivalent to the head/tail domains) near N- and C- terminus of each dimer. These models could be used to superimpose on the low-resolution electron density map of keratin filaments. A good match of the electronic density would validate the models. These models could also help in explaining the discontinuity in the regularity of molecular packing. The protruding head and tail domains on the filament surface could produce a seam along the axis of the keratin filament observed in experiments [261]. Initial filament models have a similar seam along the filaments produced by the head and tail domains. Analysis with the exact measurement of the spiral pitch needs to be done as a model validation procedure. Prediction of the exact location of the head/tail domains is among the future aims of this work. Such prediction will greatly help in understanding the mechanism of filament assembly and growth. Electrostatics of keratin full-atom model could help explain the

interaction region on the filament surface. This could help explain why intermediate filaments cannot keep growing larger in diameter.

8.3 Validation of Keratin Filament Models

8.3.1 Overview of Validation Methodologies

In the previous section, the ~500,000 atom Keratin K5-K14 filament model was constructed using Integrative modeling. However, being a computational model, it needs to be validated using the experimental data that was not used to construct the model. This section talks about the validation of the filament model performed using six different types of available experimental data at various length scales. The experimental validations includes 1) comparing low-resolution TEM micrographs of keratin filament bundles, 2) comparing lower and higher order structure dimensions, 3) comparing radial density of atoms from cryo-EM micrographs, 4) identification of cross-linkable residues, 5) atomic packing volumes of chemical groups, and 6) determining key residues important for filament assembly.

The analysis of the 3D structure of keratin filament provides the molecular basis of their stability and specific pairing. Molecular determinants of the Intermediate filament stability were classified into two types. 1) Dimer stabilizing molecular determinants and 2) Dimer-dimer stabilizing molecular determinants. Dimer level interactions comprised of hydrophobic interactions whereas dimer-dimer level interactions were polar/charged interactions. Dimer stabilizing molecular determinants were extracted from dimer models and validated using *in-vivo* experiments. However, dimer-dimer stabilizing polar/charged interactions such as hydrogen bonding, residue depth, etc. were extracted from the filament models. The classification of the chemical environments in the filament was coupled with the donor-acceptor assignment for hydrogen bonding residues. The analysis of various hydrogen bonding residues provides 257 previously unknown critical interacting positions in the K5-K14 filament. The dimer-dimer interactions rationalized K5-K14 disease-causing mutations (54/56 disease mutants). The interactions were also used to predict several temperature-sensitive mutations that are currently under experimental testing.

8.3.2 Future Prospects

The keratin K5-K14 intermediate filament model was constructed and validated using experimental data available in the literature. Some of the experimental data used for

validating epithelial K5-K14 keratin filament model was obtained from both hard (trichocyte keratins) and soft (epithelial) keratins. Though the two types of keratins have ~55% identical coiled-coil domains, their head/tail domains share ~17% sequence identity. This shows that the overall structure of the two types of keratins has identical topology given by the coiled-coil domains. However, their head/tail domains could pack very differently thereby providing large variability across different keratin types. In hard keratins, head/tails could pack tighter compared to soft keratins. Thus, comparing the experimental data across different types of keratins are justified if it involves coiled-coil domains. The best solution to this problem is to generate experimental data targeted for a particular system and use it to build models along with head/tail domains.

Head/tail domains are known to interact with various cytoskeletal elements (Example plectin). A keratin filament model with accurate placement of head/tail domains would be of extreme importance in the context of cytoskeletal interactions. The position of head/tail domains could be predicted accurately by iteratively optimizing their position with respect to the coiled-coil domains. The configuration that fits across a range of experimental data such as EM, SAXS, DXMS, Cryo-EM, etc. would be the best representative of the filament model with head/tail domains.

Relative frequencies of intra-dimer and inter-dimer cross-links could be used to validate the models. However, it is a difficult and tricky task. The relative frequency of the models should be similar to the relative frequency of the cross-links in the experiment. If the model shows inconsistencies in the relative frequency of cross-links, then this data itself could be used to refine the models. However, relative frequency values obtained from HPLC experiment were not accurate and could not be used for validation.

The keratin K5-K14 model built in this study could be used as a template for building other intermediate filament models using homology modeling. A range of mutations for all intermediate filament family members could be predicted that would greatly help researchers to study the effects of filament assembly. Along with these, temperature-sensitive mutations for all intermediate filaments would provide an alternative route to study cytoskeletal reorganization at restrictive temperatures.

8.4 Applications of Structural Rules from IF modeling

8.4.1 Summary and Future prospects

Coiled-coil structural rules obtained from the Keratin models and other coiled-coil structures were employed to three applications. 1) Designing novel coiled-coil structures,

2) Designing peptide inhibitors for coiled-oil proteins and 3) Predicting the oligomeric state of a coiled-coil assembly.

In the first application, a novel coiled-coil assembly of 8-helix bundle was constructed without using any known template (Not seen in the PDB database until date). The sequence and structure was determined purely from the structural rules developed in this work. Genetic algorithm was used to determine the best scoring sequence for the specified geometry. The structural model of the best scoring sequence was analyzed for stability. The completely automated version of this software is currently in development that has inbuilt propensities inferred from the known coiled-coil interactions. Such software could be used to design any coiled-coil assembly with maximum stability. This would be of tremendous importance in developing targeted drug delivery systems.

In the second application, the coiled-coil structural rules were applied to design peptide inhibitor for Nipah Virus Fusion protein (F protein). Many viruses have the class 1 fusion protein that have domains similar to NiV fusion protein 6HB domain. In this study, a small helix (inhibitor) was designed that would competitively bind to 6HB domain and interfere with the virus life cycle. Rigorous experimental testing of the inhibitor could potentially be useful as a template for the development of inhibitors for other viruses as well.

In the third application, the structural rules were used to predict the oligomeric state of a coiled-coil assembly. The method outperforms the existing methods until date and currently in development to incorporate all kinds of coiled-coil assemblies. The method would be useful for modeling of various protein assemblies that follow heptad repeat pattern but do not give a clue about the number of helices in the assembly. Once a geometry and oligomerization state is predicted, it could be clubbed with the coiled-coil generator software to build and validate the model.

8.5 Concluding Remarks

Using integrative modeling, 3 unit lengths of the whole Keratin (K5-K14) filament at near atomic resolution was assembled along with low-resolution head/tail domains. This ~ 500,000-atom model of Keratin was validated by inferring experimental data that were not used in structure construction. The validations included identification of cross-linkable residues, comparing the TEM profile with the experiments, the lower/higher order structure dimensions, comparing radial density distribution from Cryo-EM micrographs, amino acid packing and identification of substitutions that could lead to

filament disruption. The model could further be validated by fitting into low-resolution EM maps and SAXS profile. This is among the first 3D structure of an Intermediate filament and could be used as a template to model other Intermediate filaments. This structure could also serve as an attractive starting point for the rational design of therapeutic agents.

Apart from determining the filament structure, the study also tries to decode the assembly mechanism of intermediate filaments. The analysis of the 3D structures of the keratin filaments and coiled-coil proteins provide the molecular basis of their stability and specific pairing. The structural properties include coiled-coil core packing density, charge distribution, amino acids pair preferences, hydrogen bonding, and salt bridges. These structural properties were tested thoroughly on various molecular dynamic simulations and subsequently used for the prediction of mutants. The structural rules derived from the IF family dimers and Keratin IF structure explains 43% (215/502) of disease mutants. These rules were then used to score new filament destabilizing mutations obtained through *in-silico* saturated mutagenesis of all IF family. To validate these findings, K5-K14 IF system mutations was tested through in-vivo cell culture experiments. 16 different substitutions were validated, that also agree with the computational outcome. Temperature-sensitive mutations for the K5-K14 filament were also predicted, that would greatly help researchers to study the effects of filament assembly at restrictive temperatures.

The methods developed in this study were shown to be used for a variety of applications including designing novel coiled-coil proteins, oligomerization state prediction, and designing inhibitors of coiled-coil proteins. Most importantly, the method can be used to predict complementary substitutions that will stabilize the unstable disease-causing IF variants.

Intermediate filaments are integral elements of the cytoskeleton. They interact with other cytoskeletal elements with proteins known as Intermediate Filament Associated Proteins (IFAP). Some of the examples of IFAPs are plectin, desmosomes, hemidesmosomes, etc. [262–264]. Plectin cross-links microtubules and various IFs (Vimentin, Lamins, etc.). Desmosomes directly connect to IFs of two neighboring cells at cell-cell junction whereas hemidesmosomes directly connect IFs in one cell to the extracellular matrix. Until date, very few IFAPs have been discovered due to difficulty in controlling the IF organization and assembly. Three-dimensional (3D) structural information of these filaments could give insights into the interaction mechanism. A library of possible IFAPs could be

developed by predicting the interactions of IFs with other proteins. Interactions could be predicted by scanning through all the protein sequences along with their heptad repeats (if present) combined with the filament sequence/structure in the genetic algorithm based software. The 3D structure of keratin intermediate filament determined in this study opens a wide variety of research fields that were previously not possible. The structural information about intermediate filaments provide crucial information in the context of cytoskeletal interactions. This study would make a huge impact on the current and future research on the Intermediate filaments and coiled-coil proteins in general.

CHAPTER 9
APPENDICES

SALIENT FEATURES

- ***Detailed description of Software developed during this study***

- ***Following software/algorithms discussed***

1. Disctransformation
2. GARLIC
3. Cell_List
4. AIMorpher
5. MonteCarlo Sampler
6. ATP Binding Predictor
7. SUMO Server

In the previous chapters, software and algorithms discussed includes Coiled-coil sequence visualizer (section 3.3.2), Coiled-coil stability predictor (section 3.4), Disctransformation (section 4.8), Coiled-coil assembler (section 7.1), and Coiled-coil oligomerization detector (section 7.3). This chapter provide brief descriptions of software/algorithms that was not discussed in earlier chapters.

9.1 Disctransformation

Disctransformation is a software package which packs the 2D objects in a given geometry. The 2D objects can be approximations of molecular objects (entire molecules, domains of molecules, etc.) whose size can be altered by the user. The objects are packed in accordance with a user-specified arrangement that follows a particular interaction topology. Sampling algorithms such as Monte Carlo, Divide and conquer can be used in conjunction with Disctransformation to get the optimal packing of the objects. Disctransformation can be used for the development of new model representations and combining various scoring schemes to solve the packing optimization problems. The result can be visually inspected and evaluated using the output written by the code. Disctransformation could be used with front-end graphical user interface (GUI) or directly from the command line (Figure 9-1).

9.1.1 Relevance to the molecular architectures

Disctransformation is useful in instances where molecular interactions can be envisaged as quasi-2D problems, or when the focus of the problem is on a cross-section of the structure. For instance, the packing of helices in a bundle. The modes of interactions between helices, each of which could be approximated by a disc, are obtained from experimental sources (e.g., FRET or chemical cross-linking). DiscTransformation has the capacity to distinguish between different vertical orientations of the helices (N above and C below or vice versa could be denoted as parallel and anti-parallel orientations respectively in the input file). The program is hence a quick visual guide to the possible solutions to the molecular arrangement problem. For further analysis, other (external) scoring schemes could be used to distinguish between the different solutions.

Code stats: ~4000 lines of modular code which can be integrated with external scripts easily. This software is under LGPL license and available for download from the following link. https://github.com/neeleshsoni21/disctransformation

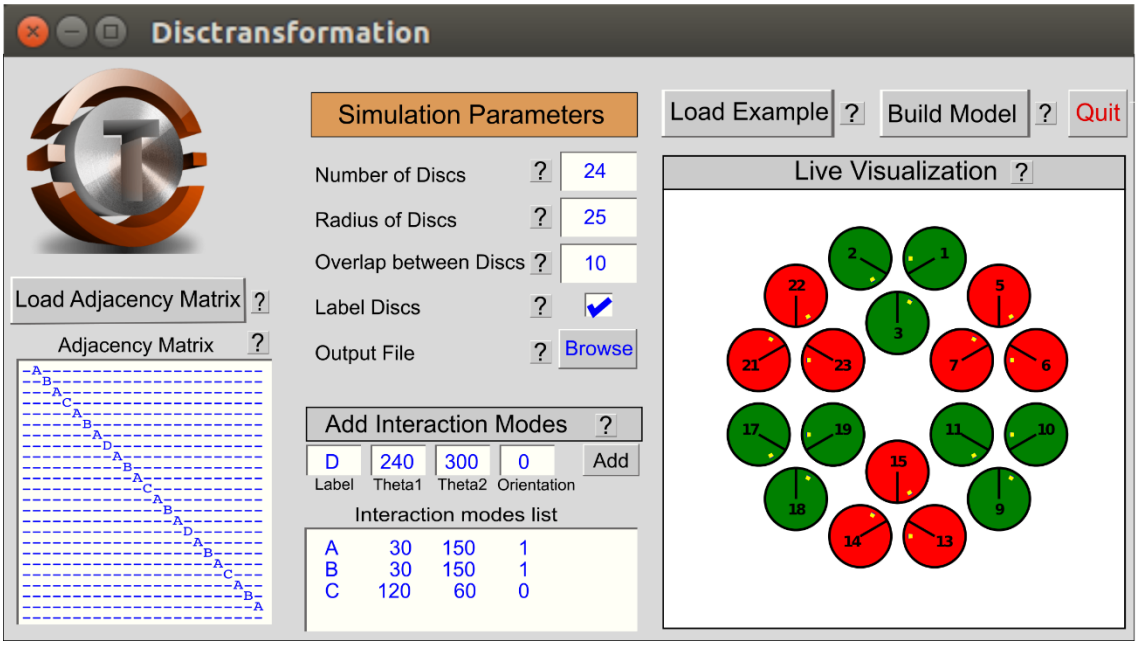


Figure 9-1 Disctransformation GUI interface

Disctransformation could be used with the frontend GUI with user specified example. The user needs to provide simulation parameters, interacting modes and adjacency matrix of various interacting discs (or molecular objects). The GUI would then provide the visualization of the model upon clicking on the "Build Model" button. All simulation parameters could also be provided from the input file in a specified format. A user can use external scoring schemes provided in the code to score a model and optimize it with sampling algorithms.

9.2 GARLIC

GARLIC stands for Geometric Assessment of Residues using Locally Interacting Chemical groups. It is a software package that can efficiently predict the packing defects in a protein. The software uses chemical groups (can be defined in the input file) as the basic building blocks of a protein. Chemical groups were obtained by partitioning amino acids according to their chemical properties (Figure 5-17). This software is particularly useful in determining the improperly packed regions (or side chains) in a protein model. The scoring scheme provided by this software could then be used in any optimizer to rebuild the model with minimum packing defects. A schematic algorithm of the GARLIC was provided with training and testing methodology (Figure 9-2).

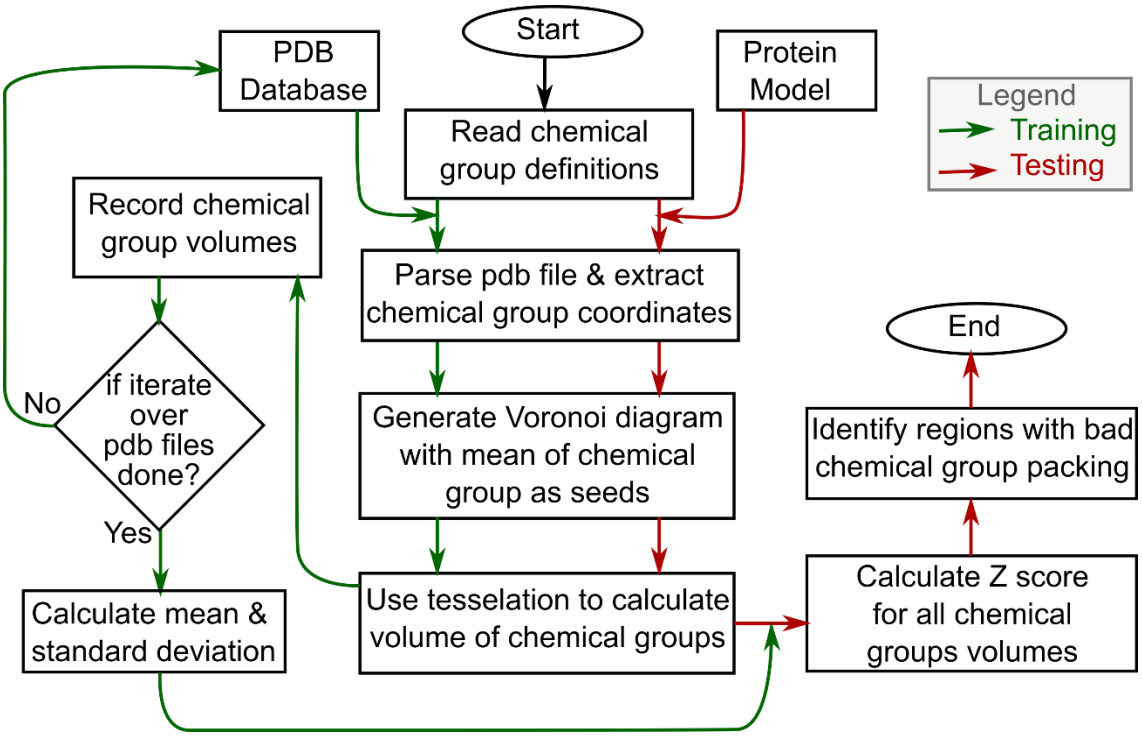


Figure 9-2 Flowchart of GARLIC showing training and testing methodology

The software is divided into two parts namely training or collecting statistics (green arrows) and testing any protein model (red arrows). During training protein coordinates obtained from PDB database were used to determine the packing of chemical groups. The partitioning and volume calculations of every chemical group were done using Voronoi and Delaunay tessellations respectively. After iterating through all protein coordinates the mean and standard deviations were determined for every chemical group for calculation of Z score. During testing protein model would be accessed through previously calculated Z score to identify improperly packed regions.

The software is based on an old technique of partitioning the 3D space known as Voronoi diagrams [213]. Voronoi diagrams divide the 3D space by using perpendicularly bisecting planes between any two seed points. In this case, centroid of every chemical group were considered as seed points. Once the Voronoi diagram was generated, every chemical group would have a 3D Voronoi cell having specific volumes. A properly packed chemical group expected to have narrow distribution of volumes. This distribution provides the basis for differentiating between high resolution protein coordinates (PDB file) and a protein model. High-resolution PDB files was assumed to have an effective packing density of the chemical groups. The software generates statistics on the database of 2579 single domain high-resolution non-redundant proteins.

The software was tested on two different datasets. 1) Incorrectly threaded model, in which all beta sheet structure threaded to all alpha helix structure (Figure 9-3), and 2) Melo's dataset, that contain both good and bad models (Table 9-1) [265].

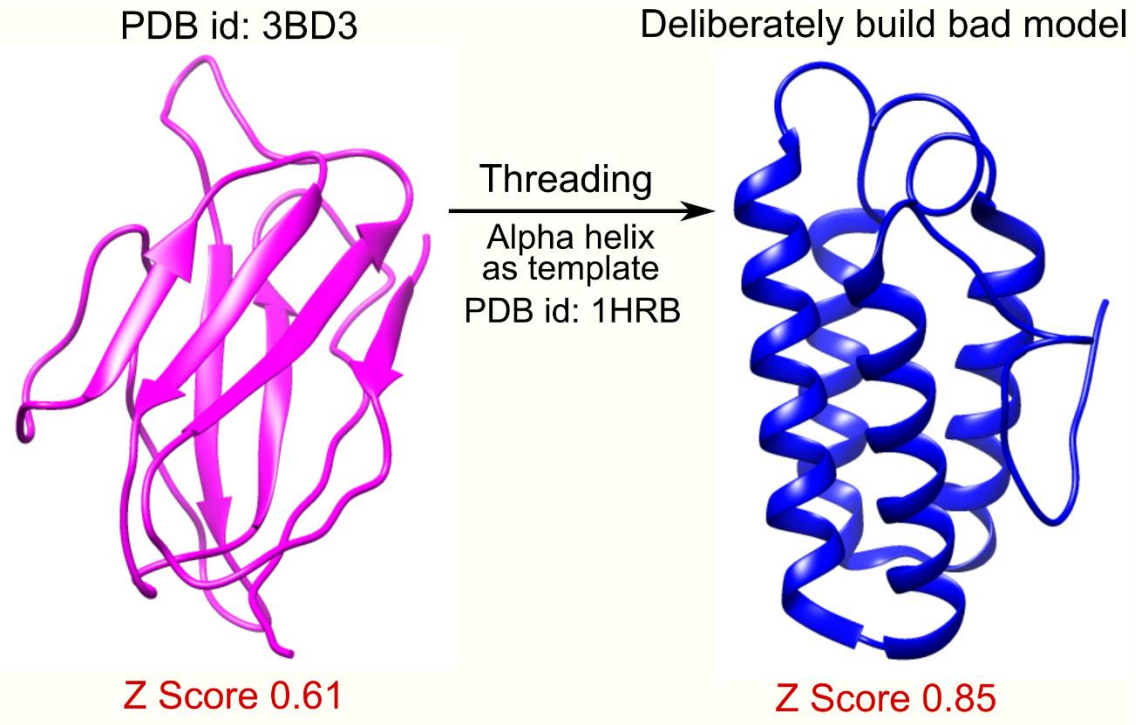


Figure 9-3 GARLIC testing on deliberately built bad model

An all beta sheet protein structure (PDB id: 3BD3) (pink) was remodeled with template (PDB id: 1HRB) having all alpha helices (blue). This deliberately built bad model should have higher Z score (red text) value compared to native structure.

Table 9-1 GARLIC testing on Melo's dataset.

Melo's dataset contains five different sets represented by their corresponding X-Ray structures. Each of these sets has deliberately built good and bad models. GARLIC provided Z Scores for every model in the dataset.

Dataset	Structure Type	Number of structures	Average Z Score Range
Melo's Set 1	X-Ray	1	0.63
	Good Models	9	0.83 to 1.70
	Bad Models	3	3.1 to 9.04
Melo's Set 2	X-Ray	1	1.1
	Good Models	2	0.82 to 1.47
	Bad Models	4	1.52 to 2.76
Melo's Set 3	X-Ray	1	0.71
	Good Models	3	0.71 to 0.82
	Bad Models	8	1.12 to 22.4
Melo's Set 4	X-Ray	1	1.13
	Good Models	4	0.88 to 6.77
	Bad Models	9	0.89 to 2.63
Melo's Set 5	X-Ray	1	0.96
	Good Models	66	0.6 to 2.9
	Bad Models	7	0.88 to 8.4

In both the test datasets, the software seems to identify the incorrectly packed regions identified by the high Z scores compare to the Z scores of native (Table 9-1). However, the Z score vary with the protein structure. Thus, an absolute Z score that do not vary with protein structure was required to identify the incorrectly packed regions in an unknown protein model. Currently this software is under development under LGPL license. **Code stats:** ~5000 lines of code.

9.3 CellList

CellList is a tool to determine pairwise distances between atoms within some cutoff distance efficiently. A brute force approach for finding pairwise distances has complexity $\Theta(N^2)$ whereas improved cell list algorithm used in this tool does it in linear time with complexity $\Theta(N)$. Using a brute force approach for large protein system takes enormous amount of time. CellList is especially helpful for large proteins where one has to find the pairwise distances repeatedly. For instance, for determining the clashes between atoms or for calculating the electrostatic energy for a set of atoms. CellList tool was built considering the flexibility needed for the users working with protein system. It includes various features such as finding distances between specific pair or atoms/chemical groups/residues/chains without the need for modifying the input pdb file. CellList could be used directly as python library after installation. It is currently under LGPL license.

Code stats: ~500 lines of code.

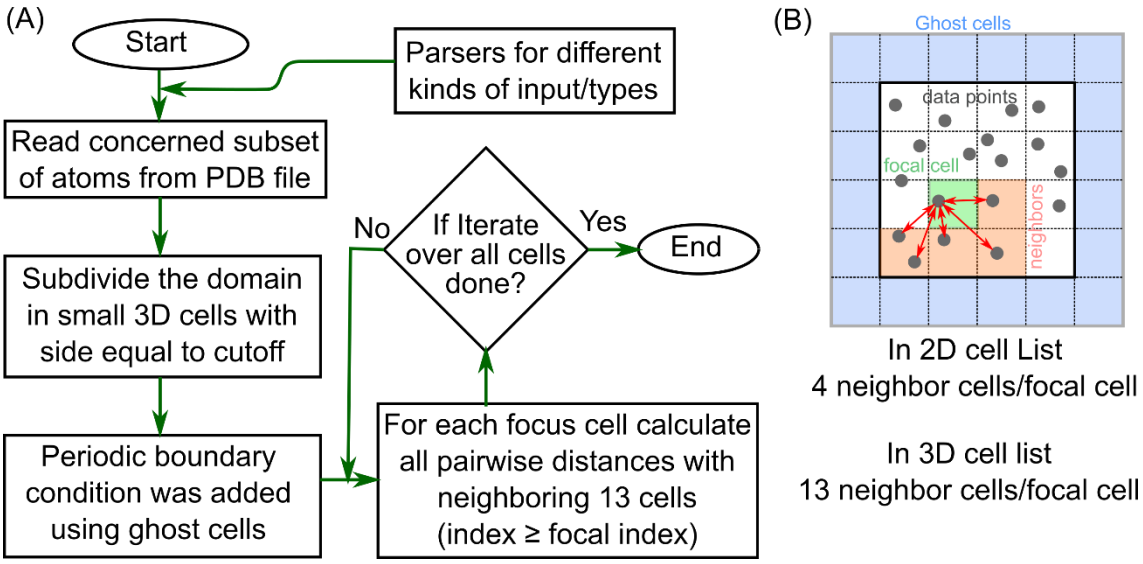


Figure 9-4 CellList tool flowchart and schematic of distance calculation

(A) Flowchart of CellList tool specifying the methodology used for algorithm implementation. (B) Schematic showing focal cell (green), neighboring cells (red), and ghost cells (blue) along with data points (gray circles) for pairwise distance (red arrows) calculation.

9.4 AIMorpher

AIMorpher is an ultrafast linear morphing tool specifically designed for macro-molecular assemblies. As an input, AIMorpher requires initial and final state of any proteins along with the number of intermediate states. Unlike other morphing tools, AIMorpher do not require two input states to have identical sequence. Any two un-related proteins could be used to generate linearly morphed conformations. The mapping between amino acids required for morphing was done by sequence alignment between the two states. The software uses its own substitution matrix for global alignment. However, standard substitution matrices such as BLOSUM, PAM, etc. could also be provided as an input. The utility of AIMorpher was shown by morphing Nipah virus fusion protein (Figure 9-5). The AIMorpher was used to generate fifty intermediate states between pre-fusion and post-fusion complex of fusion protein. The software is currently under development for improvisation such as adding flexibility for macromolecular assemblies, sampling along with linear morphing, and a scoring scheme for the selection of best conformation during sampling.

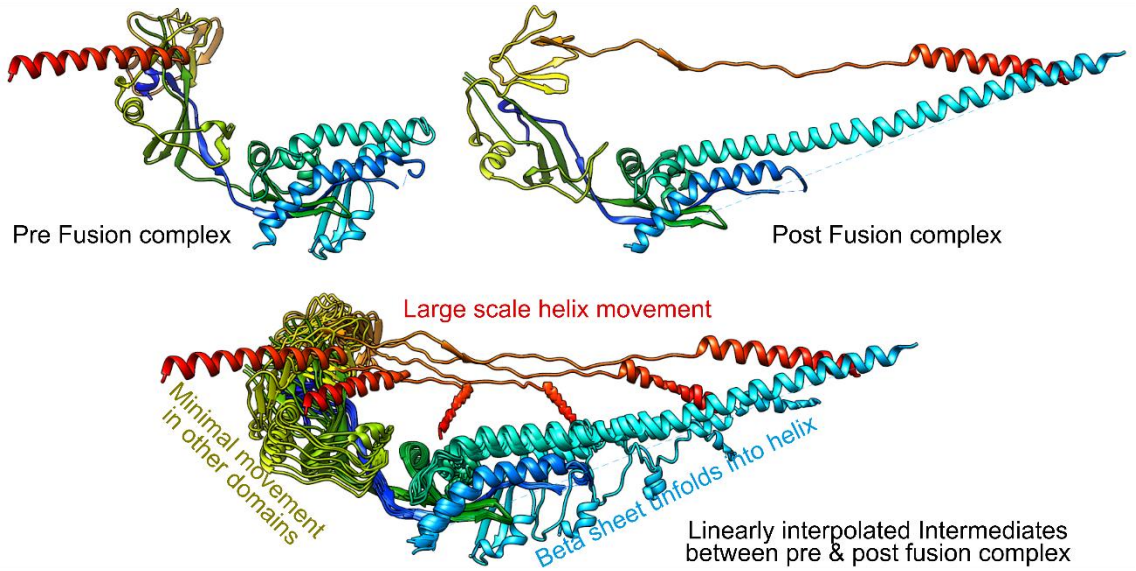


Figure 9-5 Example of morphing by AIMorpher

A pre-fusion and post-fusion complex of Nipah fusion protein shown in ribbon representation with amino acids colored from blue (N-terminus) to red (C-terminus). Large-scale movement of domains were shown by the trajectory of intermediate states. Only five intermediate states are shown for clarity.

9.5 Monte Carlo Sampler

Monte Carlo Sampler is a sampling and scoring software designed to randomly sample conformations along with a scoring scheme. The software is currently used in Disctransformation and AIMorpher as an external sampling and scoring methodology. It was used to generate possible head/tail domain configurations of Keratin filament. The utility of this software is shown with the help of PPII peptide-protein system. An approximate location for modeling PPII peptide in the query protein was done with the help of CLICK and various structural rules (unpublished work done by Dr. Binh Thanh Nguyen). Monte Carlo Sampler was then used to build the PPII in the predicted location. The following section describes the sampling and scoring steps used by the Monte Carlo Sampler.

9.5.1 Sampling and Scoring

In PPII-protein system, the rigid body sampling was done on PPII coordinates.

- Translation of random amplitude (default range -5 Å and +5 Å) along a random direction vector passing through the mean coordinates of the PPII.
- Rotation of random angle (default range -20° to 20°) around a randomly chosen direction vector passing through the mean coordinates of the PPII.

The default scoring scheme in the Monte Carlo Sampler comprised of Hydrogen bond (number and distance), Atomic clashes and Pseudo Van der Waals interactions as described below.

1. Number of Hydrogen Bonds Score ($NHBS$)

$NHBS = hb^2$ where hb is the total number of hydrogen bonds.

Square of the number of hydrogen bonds was used to harmonically weight the score. A hydrogen bond is assumed to form if the distance between acceptor and donor distance is less than 3.5 Å and the angle between donor, acceptor and acceptor antecedent is greater than 100°. NHBS score maximizes the total number of hydrogen bonds at any given point of the simulation.

2. Restrained Hydrogen Bonds Score ($RHBS$)

RHBS is the sum of distance score ($D(x)$) and angle score ($A(x)$) calculated as follows.

$$D(x) = \begin{matrix} 6400 * (x - 2.5) * (x - 4.2) & \forall\, x < 2.5 \; or \; x > 4.2 \\ 0 & \forall\, 2.5 \leq x \leq 4.2 \end{matrix}$$

where x is the distance between donor and acceptor. A factor of 6400 was multiplied to give more weight to distance score.

$$A(x) = \begin{matrix} 1 * (x - 80) * (x - 180) & \forall\ x < 80\ or\ x > 180 \\ 0 & \forall\ 80\ \leq x \leq 180 \end{matrix}$$

where x is the angle between donor, acceptor and acceptor antecedent.

$$RHBS = D(x) + A(x)$$

$RHBS$ ensures the formation of hydrogen bond between Tryptophan amino acid and Donor atoms of PPII.

3. Clash Score (CS)

Any two atoms were considered to be clashing if the distance between them was less than 2.8 Å. Depending upon the clashing atoms type, these clashes were categorized as Main chain-Main chain clash ($MCMC$), Main chain-Side chain clash ($MCSC$) and side chain-side chain clash ($SCSC$). Clash score was calculated as follows:

$$CS = 100 * MCMC^2 + 5 * MCSC^2 + 5 * SCSC$$

Square terms were used for $MCMC$ and $MCSC$ to harmonically increase the score compared to the $SCSC$ clashes that could be tolerated up to some extent. The scaling factors for each term were chosen approximately such that their scores follow the following inequalities.

$$MCMCscore \gg MCSCscore \gg SCSCscore$$

4. Pseudo Van der Waals Score (PVWS)

PVWS was the total number of atom pairs whose distances were in the interval of 2.8 to 5.0 Å. This score is used to maximize the Van der Waals energy term by increasing the density of protein atoms around PPII.

Total energy (E) of the system at any given point was calculated as follows:

$$E = -NHBS + RHBS + CS - PVWS$$

Temperature of the simulation was kept constant at 400°C (T) and Gas constant (R) as 2*10-3 kcal*mol-1*degree-1

9.5.2 Selection/Rejection criteria.

Metropolis criteria were used with a probability (P),

$P = \exp(-beta * \Delta E)$ where $beta = \frac{1}{R*T}$

40 independent runs (initialized by random seed) of Monte Carlo Simulations was performed with 4000 Monte Carlo steps in every simulation. Every independent run leads to a unique model.

All 40 models were clustered into two sets using scipy hierarchical clustering "fcluster" module with "maxclust" criterion. The linkage matrix was created with "average" method. The model that had minimum RMSD with other models in the largest cluster was chosen as the representative model. A pictorial representation of sampling is shown with PPII-Protein system.

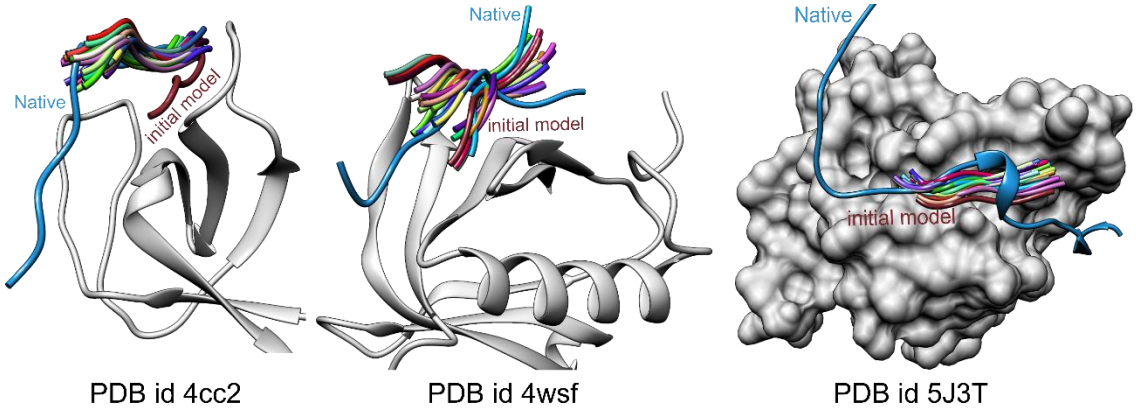


Figure 9-6 Model PPII-Protein system using Monte Carlo Sampler.

The initial predicted conformation of PPII (brown) on three different proteins (PDB ids 4CC2, 4WSF, 5J3T) was optimized using Monte Carlo Sampler. The cluster with the maximum number of PPII peptides was chosen as the representative model. The native position of PPII peptides (blue) overlap with the predictions.

9.6 Predicting ATP/ADP Ligand Binding Site

In this work, the method for predicting the ATP/ADP binding site was developed. A library of descriptor set from the ATP (and its homologs) binding protein was created which was used to determine the ATP binding site in 14-3-3 family of proteins. Every representative descriptor from the library defines the possible donor/acceptor in an ATP bound state. Using CLICK every descriptor was searched to get possible alignments on the surface/cavities of the query protein. All possible alignments were scored using number of hydrogen bonds, LJ potential and charge complementarity. DEPTH software and other possible interactions was used to filter out most of the binding sites predicted by CLICK. In case of 14-3-3 family of proteins, matches was scored using the presence of nucleophile around ATP as 14-3-3 family is known to hydrolyze ATP. ATP coordinates as given by CLICK were inserted in the query protein. The final model was energy minimized for 10 ns. All hydrogen bond interactions of ATP with the query protein was recorded. These interactions were listed as possible sites for experimental testing. Preliminary experimental results in collaboration with Dr. V. Prasanna validates the predicted ATP binding site (Figure 9-7).

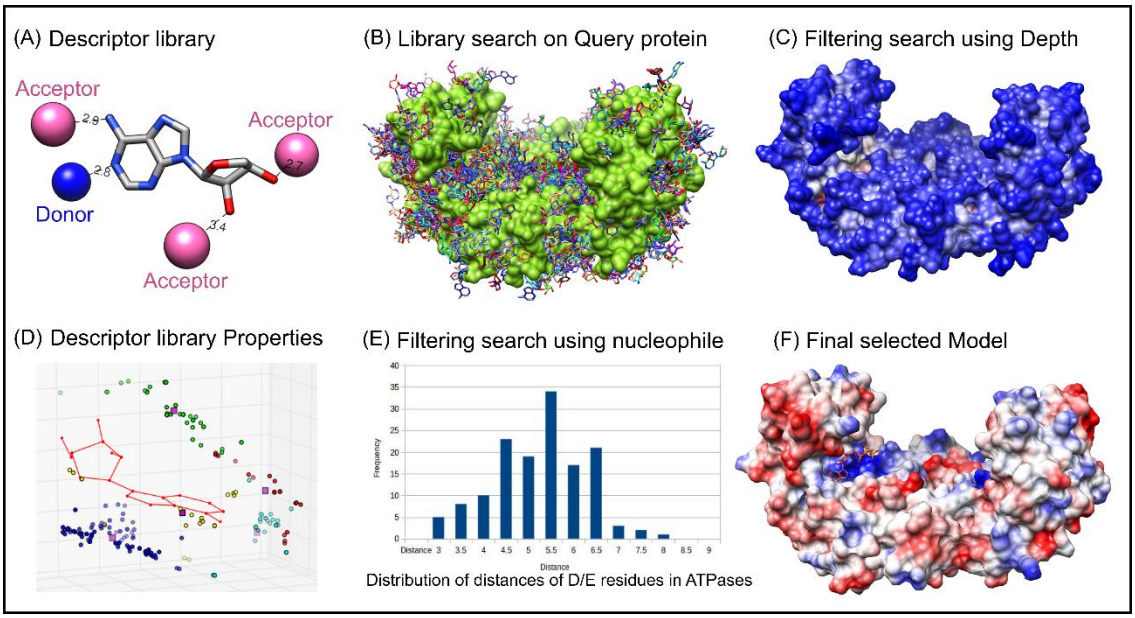
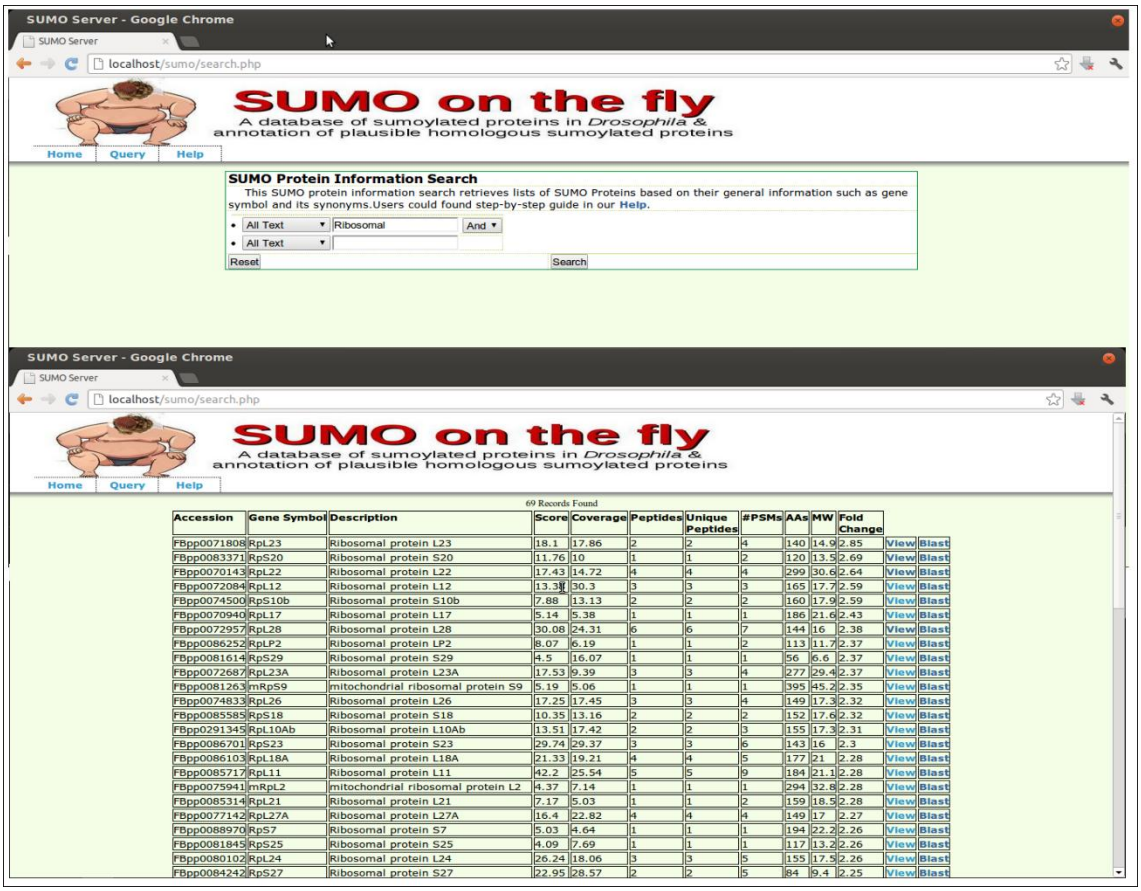


Figure 9-7 Schematic of methodology adopted for predicting ATP binding site

(A) The construction of descriptor library of hydrogen bonds donor and acceptor. (B) Every descriptor was searched on the query protein using CLICK. (C) Filtering of the possible models using DEPTH predicted cavities. (D, E) Models scored according to the type of interactions descriptor has with query protein. F() Top scoring model was selected as representative of ATP/ADP binding site.

9.7 SUMO Server

SUMO server is a database of SUMOylated proteins in Drosophila and annotations of plausible homologous sumoylated proteins (Figure 9-8). The database was previously published and was obtained in collaboration with Dr. Girish Ratnaparakhi [266]. Along with the database annotation, the aim of this work is to predict the sumoylation sites on various proteins. Many sumolyation sites have the consensus motif ΨKXE [267–269]. There are many proteins that have ΨKXE motif but still do not get sumoylated. While, there are others that do not have the ΨKXE motif and still get sumoylated. The idea is that these proteins have a 3D motif i.e. the ΨKXE residues come close together from different parts of the protein sequence in specific geometry. Prediction of sumoylation sites is currently under development. Currently the database server is available at the following link (local IISER network): http://192.168.8.70/sumo-server.com/public_html/



Accession	Gene Symbol	Description	Score	Coverage	Peptides	Unique Peptides	#PSMs	AAs	MW	Fold Change		
FBpp0071808	RpL23	Ribosomal protein L23	18.1	17.86	2	2	4	140	14.9	2.85	View	Blast
FBpp0083371	RpS20	Ribosomal protein S20	11.76	10	1	1	2	120	13.5	2.69	View	Blast
FBpp0070143	RpL22	Ribosomal protein L22	17.43	14.72	4	4	4	299	30.6	2.64	View	Blast
FBpp0072084	RpL12	Ribosomal protein L12	13.38	30.3	3	3	3	165	17.7	2.59	View	Blast
FBpp0074500	RpS10b	Ribosomal protein S10b	7.88	13.13	2	2	2	160	17.9	2.59	View	Blast
FBpp0070940	RpL17	Ribosomal protein L17	5.14	5.38	1	1	1	186	21.6	2.43	View	Blast
FBpp0072957	RpL28	Ribosomal protein L28	30.08	24.31	6	6	7	144	16	2.38	View	Blast
FBpp0086252	RpLP2	Ribosomal protein LP2	8.07	6.19	1	1	2	113	11.7	2.37	View	Blast
FBpp0081614	RpS29	Ribosomal protein S29	4.5	16.07	1	1	1	56	6.6	2.37	View	Blast
FBpp0072687	RpL23A	Ribosomal protein L23A	17.53	9.39	3	3	4	277	29.4	2.37	View	Blast
FBpp0081263	mRpS9	mitochondrial ribosomal protein S9	5.19	5.06	1	1	1	395	45.2	2.35	View	Blast
FBpp0074833	RpL26	Ribosomal protein L26	17.25	17.45	3	3	4	149	17.3	2.32	View	Blast
FBpp0085585	RpS18	Ribosomal protein S18	10.35	13.16	2	2	2	152	17.6	2.32	View	Blast
FBpp0291345	RpL10Ab	Ribosomal protein L10Ab	13.51	17.42	2	2	3	155	17.3	2.31	View	Blast
FBpp0086701	RpS23	Ribosomal protein S23	29.74	29.37	3	3	6	143	16	2.3	View	Blast
FBpp0086103	RpL18A	Ribosomal protein L18A	21.33	19.21	4	4	5	177	21	2.28	View	Blast
FBpp0085717	RpL11	Ribosomal protein L11	42.2	25.54	5	5	9	184	21.1	2.28	View	Blast
FBpp0075941	mRpL2	mitochondrial ribosomal protein L2	4.37	7.14	1	1	1	294	32.8	2.28	View	Blast
FBpp0085314	RpL21	Ribosomal protein L21	7.17	5.03	1	1	2	159	18.5	2.28	View	Blast
FBpp0077142	RpL27A	Ribosomal protein L27A	16.4	22.82	4	4	4	149	17	2.27	View	Blast
FBpp0088970	RpS7	Ribosomal protein S7	5.03	4.64	1	1	1	194	22.2	2.26	View	Blast
FBpp0081845	RpS25	Ribosomal protein S25	4.09	7.69	1	1	1	117	13.2	2.26	View	Blast
FBpp0080102	RpL24	Ribosomal protein L24	26.24	18.06	3	3	5	155	17.5	2.26	View	Blast
FBpp0084242	RpS27	Ribosomal protein S27	22.95	28.57	2	2	5	84	9.4	2.25	View	Blast

Figure 9-8 SUMO server example screen image

(Top panel) Home page of the SUMO server provide input search boxes and classifications. (Bottom panel) Output of an example search for "Ribosomal" that result in various proteins with "ribosomal" in their descriptions.

9.8 Example structures solved through Integrative Modeling

Table 9-2 List of studies solved through Integrative modeling.

List includes only those studies that incorporate experimental methods along with computational technique.

Model System	Structural Data Used	Reference
Phosphoinositide 3-Kinase p85α Homodimer	SAXS, Cross-linking	[270]
Yeast Mediator complex	Cross-linking, Mass spectrometry, X-Ray crystallography, homology modeling, Cryo-EM	[271]
Chromatin	Mass spectrometry, Cryo-EM	[70,71,272,273]
Human and Yeast TFIIH	Cross-linking, Mass spectrometry, biochemical analysis, Electron microscopy	[274]
Bilitranslocase Transmembrane Domain	Bioinformatics, NMR, FRET	[275]
Yeast eIF3:eIF5 complex	Cross-linking, Mass spectrometry	[276]
DNA Repair Complex of Human Rad9-Hus1-Rad1/FEN1/DN	X-Ray crystallography, Single particle EM	[277,278]
Ubiquitin-modified PCNA	SAXS	[277,279]
Photoreceptor phosphodiesterase PDE6 catalytic dimer (αβ)	Cross-linking and Mass spectrometry	[280]
Actin-scruin complex	X-Ray crystallography, Cross-linking	[69]
SEA complex	Affinity purification, Cross-linking	[281]

40S·eIF1·eIF3 Translation Initiation Complex	Electron microscopy, Cross-linking, Mass spectrometry	[282]
26S proteosome	X-Ray crystallography, Cross-linking, Mass spectrometry, Cryo-EM	[50]
Nucleo-pore Complex, Nup84 complex	Ultracentrifugation, Quantitative immunoblotting, Affinity purification, overlay assay, Electron microscopy, Immuno-EM, Bio-informatics, Membrane fractionation, Cross-linking, Mass spectrometry, X-Ray crystallography, homology modeling	[2,283,284]
Proteasomal lid (back calculate)	Mass Spectrometry, Proteomics, ion mobility-MS, Cross-linking	[285]
Proteasome interacting proteins (PIPs)(back calculate)	Mass Spectrometry, Proteomics, ion mobility-MS, Cross-linking	[285]
Chaperone involved in formation proteasomal base complex (back calculate)	Mass Spectrometry, Proteomics, ion mobility-MS, Cross-linking	[285]
Collicin-immunity protein and dnase complex (back calculate)	X-ray crystallography, Mass spectrometry	[9]
Eukaryotic Ribosome	Cryo-EM, Homology Modeling	[72]
Mammalian Ribosome	Cryo-EM, Homology Modeling	[67]
Exosome on polyribonucleotide phosphorylase (PNPase)	Homology modeling, Electron microscopy	[286]

RNA polymerase II with RPB4/7	X-Ray crystallography, Homology modeling, Electron microscopy	[286]
Chaperonin CCT and phophoducin 2(PLP2) and G protein homolog VID27	X-Ray crystallography, Electron microscopy	[286]
Ski complex	Homology modeling, Electron microscopy	[286]
RNase P (POP) complex	Electron microscopy	[286]
Bacterial type II pilus	Cryo-EM, EM, NMR, Cross-linking	[68,287]
Ryanodine receptor channel	Cryo-EM, Homology Modeling	[51]
TriC/CCT chaperonin	Cryo-EM, Homology Modeling	[66]
γ-tubulin small complex	Cryo-EM, Cross-linking, Homology Modeling	[288]
Triprptidyl peptidase II	Cryo-EM, Homology Modeling, MD fitting, Single particle analysis	[289]
Dam1 kinetochore complex	Cross-linking	[290]

Bibliography

1. Soni N, Madhusudhan MS: **Computational modeling of protein assemblies**. *Curr. Opin. Struct. Biol.* 2017, **44**:179–189.
2. Karni-Schmidt O, Alber F, Zhang W, Sali A, Chait BT, Williams R, Kipper J, Devos D, Suprapto A, Veenhoff LM, et al.: **Determining the architectures of macromolecular assemblies**. *Nature* 2007, **450**:683–694.
3. Rout MP, Sali A: **Principles for Integrative Structural Biology Studies**. *Cell* 2019, **177**:1384–1403.
4. Braitbard M, Schneidman-Duhovny D, Kalisman N: **Integrative Structure Modeling: Overview and Assessment**. *Annu. Rev. Biochem.* 2019, **88**:113–135.
5. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, et al.: **The molecular architecture of the nuclear pore complex**. *Nature* 2007, **450**:695–701.
6. Tamura K, Hara-Nishimura I: **The molecular architecture of the plant nuclear pore complex**. *J. Exp. Bot.* 2013, **64**:823–832.
7. Topf M, Sali A: **Combining electron microscopy and comparative protein structure modeling**. *Curr. Opin. Struct. Biol.* 2005, **15**:578–585.
8. Joseph AP, Polles G, Alber F, Topf M: **Integrative modelling of cellular assemblies**. *Curr. Opin. Struct. Biol.* 2017, **46**:102–109.
9. Karaca E, Bonvin AMJJ: **Advances in integrative modeling of biomolecular complexes**. *Methods* 2013, **59**:372–381.
10. Trnka MJ, Pellarin R, Robinson PJ: **Role of integrative structural biology in understanding transcriptional initiation**. *Methods* 2019, **159–160**:4–22.
11. Fuchs E, Weber K: **Intermediate filaments: structure, dynamics, function, and disease**. *Annu Rev Biochem* 1994, **63**:345–382.
12. Katsumoto T, Mitsushima A, Kurimura T: **The role of the vimentin intermediate filaments in rat 3Y1 cells elucidated by immunoelectron microscopy and computer-graphic reconstruction**. *Biol. Cell* 1990, **68**:139–146.
13. Peng WR, Lin WP, Chien HC, Chi S: **A novel scheme for better performance in optical code-division multiple access**. In *Pacific Rim Conference on Lasers and Electro-Optics, CLEO - Technical Digest*. . 2005:1408–1409.
14. Vaidya MM, Kanojia D: **Keratins: Markers of cell differentiation or regulators of cell differentiation?** *J. Biosci.* 2007, **32**:629–634.

15. Oshima RG: **Intermediate filaments: A historical perspective**. *Exp. Cell Res.* 2007, **313**:1981–1994.
16. Herrmann H, Aebi U: **Intermediate Filaments: Molecular Structure, Assembly Mechanism, and Integration Into Functionally Distinct Intracellular Scaffolds**. *Annu. Rev. Biochem.* 2004, **73**:749–789.
17. Steinert PM, Marekov LN, Parry DAD: **Diversity of intermediate filament structure**. *J. Biol. Chem.* 1993, **268**:24916–24925.
18. Parry DAD, North ACT: **Hard α-keratin intermediate filament chains: substructure of the N-and C-terminal domains and the predicted structure and function of the C-terminal domains of type I and type II chains**. *J. Struct. Biol.* 1998, **122**:67–75.
19. Bray DJ, Walsh TR, Noro MG, Notman R: **Complete structure of an epithelial keratin dimer: Implications for intermediate filament assembly**. *PLoS One* 2015, **10**:e0132706.
20. Crick FHC: **The packing of α-helices: simple coiled-coils**. *Acta Crystallogr.* 1953, **6**:689–697.
21. Szeverenyi I, Cassidy AJ, Cheuk WC, Lee BTK, Common JEA, Ogg SC, Chen H, Shu YS, Goh WLP, Kee WN, et al.: **The human intermediate filament database: Comprehensive information on a gene family involved in many human diseases**. *Hum. Mutat.* 2008, **29**:351–360.
22. Lane EB, McLean WHI: **Keratins and skin disorders**. *J. Pathol.* 2004, **204**:355–366.
23. Petoukhov M V., Svergun DI: **Applications of small-angle X-ray scattering to biomacromolecular solutions**. *Int. J. Biochem. Cell Biol.* 2013, **45**:429–437.
24. Rohl CA, Strauss CEM, Misura KMS, Baker D: **Protein Structure Prediction Using Rosetta**. *Methods Enzymol.* 2004, **383**:66–93.
25. Roy A, Kucukural A, Zhang Y: **I-TASSER: A unified platform for automated protein structure and function prediction**. *Nat. Protoc.* 2010, **5**:725–738.
26. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints.** *J. Mol. Biol.* 1993, **234**:779–815.
27. Renom MMA, Sali A, Fiser A, Sanchez R, Melo F, Stuart AC: **Comparative protein structure modeling of genes and genomes.** *Annu Rev Biophys Biomol Struct* 2000, **29**:291–325.
28. Webb B, Sali A: **Comparative protein structure modeling using MODELLER**.

Curr. Protoc. Bioinforma. 2016, **2016**:5.6.1-5.6.37.

29. Neogi U, Rao SD, Bontell I, Verheyen J, Rao VR, Gore SC, Soni N, Shet A, Schülter E, Ekstrand ML, et al.: **Novel tetra-peptide insertion in Gag-p6 ALIXbinding motif in HIV-1 subtype C associated with protease inhibitor failure in Indian patients**. *Aids* 2014, **28**:2319–2322.
30. Short JM, Liu Y, Chen S, Soni N, Madhusudhan MS, Shivji MKK, Venkitaraman AR: **High-resolution structure of the presynaptic RAD51 filament on single-stranded DNA by electron cryo-microscopy**. *Nucleic Acids Res.* 2016, **44**:9017–9030.
31. Manjunath GP, Soni N, Vaddavalli PL, Shewale DJ, Madhusudhan MS, Muniyappa K: **Molecular Mechanism Underlying ATP-Induced Conformational Changes in the Nucleoprotein Filament of Mycobacterium smegmatis RecA**. *Biochemistry* 2016, **55**:1850–1862.
32. Narcis FF, Andras F: **Hybrid Methods for Protein Structure Prediction**. *Introd. to Protein Struct. Predict. Methods Algorithms* 2010, doi:10.1002/9780470882207.ch13.
33. Zheng W, Doniach S: **Fold recognition aided by constraints from small angle X-ray scattering data**. *Protein Eng. Des. Sel.* 2005, **18**:209–219.
34. Zheng W, Doniach S: **Protein structure prediction constrained by solution X-ray scattering data and structural homology identification**. *J. Mol. Biol.* 2002, **316**:173–187.
35. Lees JG, Janes RW: **Combining sequence-based prediction methods and circular dichroism and infrared spectroscopic data to improve protein secondary structure determinations**. *BMC Bioinformatics* 2008, **9**:24.
36. Schröder GFF, Grubmüller H: **FRETsg: Biomolecular structure model building from multiple FRET experiments**. *Comput. Phys. Commun.* 2004, **158**:150–157.
37. Alexander N, Al-Mestarihi A, Bortolus M, Mchaourab H, Meiler J: **De Novo High-Resolution Protein Structure Determination from Sparse Spin-Labeling EPR Data**. *Structure* 2008, **16**:181–195.
38. Schmitz C, Vernon R, Otting G, Baker D, Huber T: **Protein Structure Determination from Pseudocontact Shifts Using ROSETTA**. *J. Mol. Biol.* 2012, **416**:668–677.
39. Hamuro Y, Burns LL, Canaves JM, Hoffman RC, Taylor SS, Woods VL: **Domain organization of D-AKAP2 revealed by enhanced deuterium exchange-mass**

spectrometry (DXMS). *J. Mol. Biol.* 2002, **321**:703–714.

40. Kamal JKA, Chance MR: **Modeling of protein binary complexes using structural mass spectrometry data**. *Protein Sci.* 2007, **17**:79–94.
41. Mouradov D, Craven A, Forwood JK, Flanagan JU, García-Castellanos R, Gomis-Rüth FX, Hume DA, Martin JL, Kobe B, Huber T: **Modelling the structure of latexin–carboxypeptidase A complex based on chemical cross-linking and molecular docking**. *Protein Eng. Des. Sel.* 2006, **19**:9–16.
42. Mouradov D, King G, Ross IL, Forwood JK, Hume DA, Sinz A, Martin JL, Kobe B, Huber T: **Protein Structure Determination Using a Combination of Cross-linking, Mass Spectrometry, and Molecular Modeling**. 2008:459–474.
43. Forwood JK, Thakur AS, Guncar G, Marfori M, Mouradov D, Meng W, Robinson J, Huber T, Kellie S, Martin JL, et al.: **Structural basis for recruitment of tandem hotdog domains in acyl-CoA thioesterase 7 and its role in inflammation**. *Proc. Natl. Acad. Sci.* 2007, **104**:10382–10387.
44. Koukos PI, Bonvin AMJJ: **Integrative modelling of biomolecular complexes**. *J. Mol. Biol.* 2019, doi:10.1016/j.jmb.2019.11.009.
45. Vajdos FF, Adams CW, Breece TN, Presta LG, De Vos AM, Sidhu SS: **Comprehensive functional maps of the antigen-binding site of an anti-ErbB2 antibody obtained with shotgun scanning mutagenesis**. *J. Mol. Biol.* 2002, **320**:415–428.
46. Yu EW, Aires JR, McDermott G, Nikaido H: **A periplasmic drug-binding site of the AcrB multidrug efflux pump: A crystallographic and site-directed mutagenesis study**. *J. Bacteriol.* 2005, **187**:6804–6815.
47. Sutto L, Marsili S, Valencia A, Gervasio FL: **From residue coevolution to protein conformational ensembles and functional dynamics**. *Proc. Natl. Acad. Sci. U. S. A.* 2015, **112**:13567–13572.
48. Szurmant H, Weigt M: **Inter-residue, inter-protein and inter-family coevolution: bridging the scales**. *Curr. Opin. Struct. Biol.* 2018, **50**:26–32.
49. Sánchez R, Sali A: **Large-scale protein structure modeling of the Saccharomyces cerevisiae genome**. *Proc. Natl. Acad. Sci.* 1998, **95**:13597–13602.
50. Lasker K, Forster F, Bohn S, Walzthoeni T, Villa E, Unverdorben P, Beck F, Aebersold R, Sali A, Baumeister W: **Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach**. *Proc. Natl.*

Acad. Sci. 2012, **109**:1380–1387.

51. Serysheva II, Ludtke SJ, Baker ML, Cong Y, Topf M, Eramian D, Sali A, Hamilton SL, Chiu W: **Subnanometer-resolution electron cryomicroscopy-based domain models for the cytoplasmic region of skeletal muscle RyR channel**. *Proc. Natl. Acad. Sci.* 2008, **105**:9610–9615.
52. Spahn CMT, Beckmann R, Eswar N, Penczek PA, Sali A, Blobel G, Frank J: **Structure of the 80S ribosome from Saccharomyces cerevisiae - tRNA-ribosome and subunit-subunit interactions**. *Cell* 2001, **107**:373–386.
53. Lim RY, Fahrenkrog B: **The nuclear pore complex up close**. *Curr. Opin. Cell Biol.* 2006, **18**:342–347.
54. Rout MP, Aitchison JD, Suprapto A, Hjertaas K, Zhao Y, Chait BT: **The yeast nuclear pore complex: Composition, architecture, transport mechanism**. *J. Cell Biol.* 2000, **148**:635–651.
55. Strambio-de-Castillia C, Blobel G, Rout MP: **Isolation and characterization of nuclear envelopes from the yeast Saccharomyces.** *J. Cell Biol.* 1995, **131**:19–31.
56. Alber F, Kim MF, Sali A: **Structural characterization of assemblies from overall shape and subcomplex compositions**. *Structure* 2005, **13**:435–445.
57. Shen M, Sali A: **Statistical potential for assessment and prediction of protein structures**. *Protein Sci.* 2006, **15**:2507–2524.
58. Beck M, Förster F, Ecke M, Plitzko JM, Melchior F, Gerisch G, Baumeister W, Medalia O: **Nuclear pore complex structure and dynamics revealed by cryoelectron tomography**. *Science (80-.).* 2004, **306**:1387–1390.
59. Kiseleva E, Allen TD, Rutherford S, Bucci M, Wente SR, Goldberg MW: **Yeast nuclear pore complexes have a cytoplasmic ring and internal filaments**. *J. Struct. Biol.* 2004, **145**:272–288.
60. Yang Q, Rout MP, Akey CW: **Three-Dimensional Architecture of the Isolated Yeast Nuclear Pore Complex: Functional and Evolutionary Implications**. *Mol. Cell* 1998, **1**:223–234.
61. Nelly Pante MK: **Nuclear Pore Complex Is Able to Transport Macromolecules with Diameters of j39 nm**. *Mol. Biol. Cell* 2002, **13**:425–434.
62. Gamini R, Han W, Stone JE, Schulten K: **Assembly of Nsp1 Nucleoporins Provides Insight into Nuclear Pore Complex Gating**. *PLoS Comput. Biol.* 2014, **10**:e1003488.

63. Labokha AA, Fassati AF: **Viruses challenge selectivity barrier of nuclear pores**. *Viruses* 2013, **5**:2410–2423.

64. Makio T, Stanton LH, Lin CC, Goldfarb DS, Weis K, Wozniak RW: **The nucleoporins Nup170p and Nup157p are essential for nuclear pore complex assembly**. *J. Cell Biol.* 2009, **185**:459–473.

65. Beck M, Lǔí V, Förster F, Baumeister W, Medalia O: **Snapshots of nuclear pore complexes in action captured by cryo-electron tomography**. *Nature* 2007, **449**:611–615.

66. Booth CR, Meyer AS, Cong Y, Topf M, Sali A, Ludtke SJ, Chiu W, Frydman J: **Mechanism of lid closure in the eukaryotic chaperonin TRiC/CCT**. *Nat. Struct. Mol. Biol.* 2008, **15**:746–753.

67. Jiang L, Ruan Q, Chen W: **The complete mitochondrial genome sequence of the Xizang Plateau frog, Nanorana parkeri (Anura: Dicroglossidae)**. *Mitochondrial DNA* 2016, **27**:3184–3185.

68. Campos M, Francetic O, Nilges M: **Modeling pilus structures from sparse data**. *J. Struct. Biol.* 2011, **173**:436–444.

69. Cong Y, Topf M, Sali A, Matsudaira P, Dougherty M, Chiu W, Schmid MF: **Crystallographic Conformers of Actin in a Biologically Active Bundle of Filaments**. *J. Mol. Biol.* 2008, **375**:331–336.

70. Baú D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA: **The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules**. *Nat. Struct. Mol. Biol.* 2011, **18**:107–115.

71. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L: **Genome architectures revealed by tethered chromosome conformation capture and population-based modeling**. *Nat. Biotechnol.* 2012, **30**:90–98.

72. Taylor DJ, Devkota B, Huang AD, Topf M, Narayanan E, Sali A, Harvey SC, Frank J: **Comprehensive Molecular Structure of the Eukaryotic Ribosome**. *Structure* 2009, **17**:1591–1604.

73. Chandramouli P, Topf M, Ménétret JF, Eswar N, Cannone JJ, Gutell RR, Sali A, Akey CW: **Structure of the Mammalian 80S Ribosome at 8.7 Å Resolution**. *Structure* 2008, **16**:535–548.

74. Russel D, Lasker K, Webb B, Velázquez-Muriel J, Tjioe E, Schneidman-Duhovny D, Peterson B, Sali A: **Putting the pieces together: Integrative modeling**

platform software for structure determination of macromolecular assemblies. *PLoS Biol.* 2012, **10**:1–5.

75. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E: **Equation of state calculations by fast computing machines**. *J. Chem. Phys.* 1953, **21**:1087–1092.
76. Seiler MC, Seiler FA: **Numerical Recipes in C: The Art of Scientific Computing**. *Risk Anal.* 1989, **9**:415–416.
77. Galassi M, Davies J, Gough B, Jungman G, Alken P: *GNU Scientific Library: Reference Manual 1.15*. Network Theory; 2011.
78. Schlick T: *Molecular Modeling and Simulation: An Interdisciplinary Guide*. Springer Science+Business Media, LLC; 2010.
79. Rieping W, Nilges M, Habeck M: **ISD: A software package for Bayesian NMR structure calculation**. *Bioinformatics* 2008, **24**:1104–1105.
80. Sugita Y, Okamoto Y, Sugita Y: **Replica-exchange molecular dynamics method for protein folding**. *Chem. Phys. Lett.* 1999, **314**:141–151.
81. Lasker K, Topf M, Sali A, Wolfson HJ: **Inferential Optimization for Simultaneous Fitting of Multiple Components into a CryoEM Map of Their Assembly**. *J. Mol. Biol.* 2009, **388**:180–194.
82. Wolf E, Kim PS, Berger B: **MultiCoil: A program for predicting two- and three-stranded coiled coils**. *Protein Sci.* 1997, doi:10.1002/pro.5560060606.
83. Mason JM, Arndt KM: **Coiled coil domains: Stability, specificity, and biological implications**. *ChemBioChem* 2004, **5**:170–176.
84. Liu J, Zheng Q, Deng Y, Cheng C-S, Kallenbach NR, Lu M: **A seven-helix coiled coil**. *Proc. Natl. Acad. Sci.* 2006, **103**:15457–15462.
85. PAULING L, COREY RB: **Compound Helical Configurations of Polypeptide Chains: Structure of Proteins of the α-Keratin Type**. *Nature* 1953, **171**:59–61.
86. Ludwiczak J, Winski A, Szczepaniak K, Alva V, Dunin-Horkawicz S: **DeepCoil – a fast and accurate prediction of coiled-coil domains in protein sequences**. *Bioinformatics* 2019, doi:10.1093/bioinformatics/bty1062.
87. Lupas a, Van Dyke M, Stock J: **Predicting coiled coils from protein sequences.** *Science* 1991, **252**:1162–4.
88. Delorenzi M, Speed T: **An HMM model for coiled-coil domains and a comparison with PSSM-based predictions**. *Bioinformatics* 2002, doi:10.1093/bioinformatics/18.4.617.

89. Gruber M, Söding J, Lupas AN: **Comparative analysis of coiled-coil prediction methods**. *J. Struct. Biol.* 2006, doi:10.1016/j.jsb.2006.03.009.

90. McDonnell A V., Jiang T, Keating AE, Berger B: **Paircoil2: Improved prediction of coiled coils from sequence**. *Bioinformatics* 2006, doi:10.1093/bioinformatics/bti797.

91. Walshaw J, Woolfson DN: **Socket: a program for identifying and analysing coiled-coil motifs within protein structures.** *J. Mol. Biol.* 2001, **307**:1427–1450.

92. Testa OD, Moutevelis E, Woolfson DN: **CC+: A relational database of coiled-coil structures**. *Nucleic Acids Res.* 2009, **37**:D315-22.

93. Crick FHC: **The Fourier transform of a coiled-coil**. *Acta Crystallogr.* 1953, **6**:685–689.

94. Chen Z, Boyken SE, Jia M, Busch F, Flores-Solis D, Bick MJ, Lu P, VanAernum ZL, Sahasrabuddhe A, Langan RA, et al.: **Programmable design of orthogonal protein heterodimers**. *Nature* 2018, doi:10.1038/s41586-018-0802-y.

95. Huang PS, Oberdorfer G, Xu C, Pei XY, Nannenga BL, Rogers JM, DiMaio F, Gonen T, Luisi B, Baker D: **High thermodynamic stability of parametrically designed helical bundles**. *Science (80-.).* 2014, doi:10.1126/science.1257481.

96. Anderson GP, Shriver-Lake LC, Liu JL, Goldman ER: **Orthogonal Synthetic Zippers as Protein Scaffolds**. *ACS Omega* 2018, doi:10.1021/acsomega.8b00156.

97. Acharya A, Rishi V, Vinson C: **Stability of 100 homo and heterotypic coiled-coil a-a′ pairs for ten amino acids (A, L, I, V, N, K, S, T, E, and R)**. *Biochemistry* 2006, **45**:11324–11332.

98. De Crescenzo G, Litowski JR, Hodges RS, O'Connor-McCourt MD: **Real-time monitoring of the interactions of two-stranded de novo designed coiled-coils: Effect of chain length on the kinetic and thermodynamic constants of binding**. *Biochemistry* 2003, **42**:1754–1763.

99. Zhu H, Celinski S a, Scholtz JM, Hu JC: **The contribution of buried polar groups to the conformational stability of the GCN4 coiled coil.** *J. Mol. Biol.* 2000, **300**:1377–87.

100. Tripet B, Wagschal K, Lavigne P, Mant CT, Hodges RS: **Effects of side-chain characteristics on stability and oligomerization state of a de Novo-designed model coiled-coil: 20 amino acid substitutions in position "d."** *J. Mol. Biol.* 2000, **300**:377–402.

101. Moitra J, Szilák L, Krylov D, Vinson C: **Leucine is the most stabilizing aliphatic**

amino acid in the d position of a dimeric leucine zipper coiled coil. *Biochemistry* 1997, **36**:12567–12573.

102. Oakley MG, Kim PS: **A buried polar interaction can direct the relative orientation of helices in a coiled coil**. *Biochemistry* 1998, **37**:12603–12610.
103. Apgar JR, Gutwin KN, Keating AE: **Predicting helix orientation for coiled-coil dimers**. *Proteins Struct. Funct. Genet.* 2008, **72**:1048–1065.
104. Akey DL, Malashkevich VN, Kim PS: **Buried polar residues in coiled-coil interfaces**. *Biochemistry* 2001, **40**:6352–6360.
105. Kim BW, Jung YO, Kim MK, Kwon DH, Park SH, Kim JH, Kuk YB, Oh SJ, Kim L, Kim BH, et al.: **ACCORD: An assessment tool to determine the orientation of homodimeric coiled-coils**. *Sci. Rep.* 2017, **7**:43318.
106. Thomas F, Niitsu A, Oregioni A, Bartlett GJ, Woolfson DN: **Conformational Dynamics of Asparagine at Coiled-Coil Interfaces**. *Biochemistry* 2017, **56**:6544–6554.
107. Keating AE, Malashkevich VN, Tidor B, Kim PS: **Side-chain repacking calculations for predicting structures and stabilities of heterodimeric coiled coils**. *Proc. Natl. Acad. Sci.* 2001, **98**:14825–14830.
108. Woolfson DN, Alber T: **Predicting oligomerization states of coiled coils**. *Protein Sci.* 1995, **4**:1596–1607.
109. Szczepaniak K, Ludwiczak J, Winski A, Dunin-Horkawicz S: **Variability of the core geometry in parallel coiled-coil bundles**. *J. Struct. Biol.* 2018, **204**:117–124.
110. Arndt KM, Pelletier JN, Müller KM, Plückthun A, Alber T: **Comparison of in vivo selection and rational design of heterodimeric coiled coils**. *Structure* 2002, **10**:1235–1248.
111. Ciani B, Bjelic S, Honnappa S, Jawhari H, Jaussi R, Payapilly A, Jowitt T, Steinmetz MO, Kammerer RA: **Molecular basis of coiled-coil oligomerization-state specificity**. *Proc. Natl. Acad. Sci.* 2010, **107**:19850–19855.
112. Jeffrey YS, Hodges RS, Kay CM: **Effect of Chain Length on the Formation and Stability of Synthetic α-Helical Coiled Coils**. *Biochemistry* 1994, doi:10.1021/bi00255a032.
113. Lumb KJ, Kim PS: **A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil.** *Biochemistry* 1995, **34**:8642–8648.
114. Harbury PB, Zhang T, Kim PS, Alber T: **A switch between two-, three-, and**

four-stranded coiled coils in GCN4 leucine zipper mutants. *Science (80-.).* 1993, **262**:1401–1407.

115. Kwok SC, Hodges RS: **Effect of chain length on coiled-coil stability: Decreasing stability with increasing chain length**. *Biopolym. - Pept. Sci. Sect.* 2004, doi:10.1002/bip.20141.

116. Tropsha a, Bowen JP, Brown FK, Kizer JS: **Do interhelical side chain-backbone hydrogen bonds participate in formation of leucine zipper coiled coils?** *Proc. Natl. Acad. Sci. U. S. A.* 1991, **88**:9488–9492.

117. Litowski JR, Hodges RS: **Designing heterodimeric two-stranded β-helical coiled-coils: The effect of chain length on protein folding, stability and specificity**. *J. Pept. Res.* 2001, doi:10.1034/j.1399-3011.2001.10972.x.

118. Steinkruger JD, Woolfson DN, Gellman SH: **Side-chain pairing preferences in the parallel coiled-coil dimer motif: Insight on ion pairing between core and flanking sites**. *J. Am. Chem. Soc.* 2010, **132**:7586–7588.

119. Krylov D, Mikhailenko I, Vinson C: **A thermodynamic scale for leucine zipper stability and dimerization specificity: e and g interhelical interactions.** *EMBO J.* 1994, **13**:2849–61.

120. Parry DAD: **Coiled-coils in alpha-helix-containing proteins: analysis of the residue types within the heptad repeat and the use of these data in the prediction of coiled-coils in other proteins**. *Biosci. Rep.* 1982, **2**:1017–1024.

121. O'Shea EK, Lumb KJ, Kim PS: **Peptide "Velcro": Design of a heterodimeric coiled coil**. *Curr. Biol.* 1993, **3**:658–667.

122. Faix J, Steinmetz M, Boves H, Kammerer RA, Lottspeich F, Mintert U, Murphy J, Stock A, Aebi U, Gerisch G: **Cortexillins, major determinants of cell shape and size, are actin- bundling proteins with a parallel coiled-coil tail**. *Cell* 1996, **86**:631–642.

123. Kohler JJ, Schepartz A: **Kinetic studies of Fos??Jun??DNA complex formation: DNA binding prior to dimerization**. *Biochemistry* 2001, **40**:130–142.

124. Porotto M, Rockx B, Yokoyama CC, Talekar A, DeVito I, Palermo LM, Liu J, Cortese R, Lu M, Feldmann H, et al.: **Inhibition of Nipah Virus Infection In Vivo: Targeting an early stage of paramyxovirus fusion activation during viral entry**. *PLoS Pathog.* 2010, doi:10.1371/journal.ppat.1001168.

125. Lupas AN, Bassler J: **Coiled Coils – A Model System for the 21st Century**. *Trends Biochem. Sci.* 2017, **42**:130–140.

126. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J: **Intermediate Filaments**. In *Molecular Cell Biology. 4th edition.* . W. H. Freeman; 2000.

127. Ishikawa H, Bischoff R, Holtzer H: **Mitosis and intermediate-sized filaments in developing skeletal muscle.** *J. Cell Biol.* 1968, **38**:538–555.

128. Greenwold MJ, Bao W, Jarvis ED, Hu H, Li C, Gilbert MTP, Zhang G, Sawyer RH: **Dynamic evolution of the alpha (α) and beta (β) keratins has accompanied integument diversification and the adaptation of birds into novel lifestyles**. *BMC Evol. Biol.* 2014, **14**:249.

129. Marchuk D, McCrohon S, Fuchs E: **Remarkable conservation of structure among intermediate filament genes.** *Cell* 1984, **39**:491–8.

130. van de KLUNDERT FAJM, RAATS JMH, BLOEMENDAL H: **Intermediate filaments: regulation of gene expression and assembly**. *Eur. J. Biochem.* 1993, **214**:351–366.

131. Herrmann H, Bär H, Kreplak L, Strelkov S V., Aebi U: **Intermediate filaments: From cell architecture to nanomechanics**. *Nat. Rev. Mol. Cell Biol.* 2007, **8**:562–573.

132. Chang L, Goldman RD: **Intermediate filaments mediate cytoskeletal crosstalk.** *Nat. Rev. Mol. Cell Biol.* 2004, **5**:601–13.

133. Herrmann H, Aebi U: **Intermediate filaments and their associates: Multi-talented structural elements specifying cytoarchitecture and cytodynamics**. *Curr. Opin. Cell Biol.* 2000, **12**:79–90.

134. Kim S, Wong P, Coulombe PA: **A keratin cytoskeletal protein regulates protein synthesis and epithelial cell growth**. *Nature* 2006, **441**:362–365.

135. Herrmann H, Aebi U: **Intermediate Filaments: Structure and Assembly.** *Cold Spring Harb. Perspect. Biol.* 2019, **8**:a018242.

136. Moll R, Divo M, Langbein L: **The human keratins: Biology and pathology**. *Histochem. Cell Biol.* 2008, **129**:705–733.

137. Herrmann H, Strelkov S V., Burkhard P, Aebi U: **Intermediate filaments: Primary determinants of cell architecture and plasticity**. *J. Clin. Invest.* 2009, **119**:1772–1783.

138. Coulombe PA, Omary MB: **"Hard" and "soft" principles defining the structure, function and regulation of keratin intermediate filaments.** *Curr. Opin. Cell Biol.* 2002, **14**:110–22.

139. Geisler F, Leube R: **Epithelial Intermediate Filaments: Guardians against Microbial Infection?** *Cells* 2016, **5**:29.
140. Block J, Schroeder V, Pawelzyk P, Willenbacher N, Köster S: **Physical properties of cytoplasmic intermediate filaments**. *Biochim. Biophys. Acta - Mol. Cell Res.* 2015, **1853**:3053–3064.
141. Kreplak L, Herrmann H, Aebi U: **Tensile properties of single desmin intermediate filaments**. *Biophys. J.* 2008, **94**:2790–2799.
142. Sokolova A V., Kreplak L, Wedig T, Mucke N, Svergun DI, Herrmann H, Aebi U, Strelkov S V.: **Monitoring intermediate filament assembly by small-angle x-ray scattering reveals the molecular architecture of assembly intermediates**. *Proc. Natl. Acad. Sci.* 2006, **103**:16206–16211.
143. Quinlan RA, Franke WW: **Heteropolymer filaments of vimentin and desmin in vascular smooth muscle tissue and cultured baby hamster kidney cells demonstrated by chemical crosslinking.** *Proc. Natl. Acad. Sci. U. S. A.* 1982, **79**:3452–6.
144. Hanukoglu I, Fuchs E: **The cDNA sequence of a Type II cytoskeletal keratin reveals constant and variable structural domains among keratins.** *Cell* 1983, **33**:915–24.
145. Aebi U, Häner M, Troncoso J, Eichner R, Engel A: **Unifying principles in intermediate filament (IF) structure and assembly**. *Protoplasma* 1988, **145**:73–81.
146. Parry DAD, Steinert PM: **Intermediate filaments: molecular architecture, assembly, dynamics and polymorphism.** *Q. Rev. Biophys.* 1999, **32**:99–187.
147. Strelkov S V., Herrmann H, Geisler N, Wedig T, Zimbelmann R, Aebi U, Burkhard P: **Conserved segments 1A and 2B of the intermediate filament dimer: Their atomic structures and role in filament assembly**. *EMBO J.* 2002, **21**:1255–1266.
148. Strnad P, Usachov V, Debes C, Gräter F, Parry DAD, Omary MB: **Unique amino acid signatures that are evolutionarily conserved distinguish simple-type, epidermal and hair keratins**. *J. Cell Sci.* 2011, **124**:4221–4232.
149. Herrmann H, Häner M, Brettel M, Müller SA, Goldie KN, Fedtke B, Lustig A, Franke WW, Aebi U: **Structure and Assembly Properties of the Intermediate Filament Protein Vimentin: The Role of its Head, Rod and Tail Domains**. *J. Mol. Biol.* 1996, **264**:933–953.
150. Rogers KR, Herrmann H, Franke WW: **Characterization of disulfide crosslink**

formation of human vimentin at the dimer, tetramer, and intermediate filament levels. *J. Struct. Biol.* 1996, **117**:55–69.

151. Steinert PM, Marekov LN, Fraser RDBD, Parry DAD: **Keratin intermediate filament structure. Crosslinking studies yield quantitative information on molecular dimensions and mechanism of assembly.** *J. Mol. Biol.* 1993, **230**:436–52.
152. Steinert PM, North AC, Parry DAD: **Structural features of keratin intermediate filaments.** *J. Invest. Dermatol.* 1994, **103**:19S-24S.
153. Parry DAD: **Fifty years of fibrous protein research: A personal retrospective**. *J. Struct. Biol.* 2014, **186**:320–334.
154. Jacob JT, Coulombe PA, Kwan R, Omary MB: **Types I and II keratin intermediate filaments**. *Cold Spring Harb. Perspect. Biol.* 2018, **10**:a018275.
155. Fraser RDB, Parry DAD: **Structural hierarchy of trichocyte keratin intermediate filaments**. In *Advances in Experimental Medicine and Biology*. . 2018:57–70.
156. Tarbet HJ, Dolat L, Smith TJ, Condon BM, O'Brien ET, Valdivia RH, Boyce M: **Site-specific glycosylation regulates the form and function of the intermediate filament cytoskeleton**. *Elife* 2018, **7**.
157. Kirmse R, Qin Z, Weinert CM, Hoenger A, Buehler MJ, Kreplak L: **Plasticity of intermediate filament subunits**. *PLoS One* 2010, **5**:e12115.
158. Sharma P, Alsharif S, Fallatah A, Chung BM: **Intermediate Filaments as Effectors of Cancer Development and Metastasis: A Focus on Keratins, Vimentin, and Nestin**. *Cells* 2019, **8**:497.
159. Chernyatina AA, Nicolet S, Aebi U, Herrmann H, Strelkov S V.: **Atomic structure of the vimentin central -helical domain and its implications for intermediate filament assembly**. *Proc. Natl. Acad. Sci.* 2012, **109**:13620–13625.
160. Aziz A, Hess JF, Budamagunta MS, Voss JC, Kuzin AP, Huang YJ, Xiao R, Montelione GT, FitzGerald PG, Hunt JF: **The structure of vimentin linker 1 and rod 1B domains characterized by site-directed spin-labeling electron paramagnetic resonance (SDSL-EPR) and X-ray crystallography**. *J. Biol. Chem.* 2012, **287**:28349–28361.
161. Meier M, Padilla GP, Herrmann H, Wedig T, Hergt M, Patel TR, Stetefeld J, Aebi U, Burkhard P: **Vimentin Coil 1A-A Molecular Switch Involved in the Initiation of Filament Elongation**. *J. Mol. Biol.* 2009, **390**:245–261.

162. Chernyatina AA, Strelkov S V: **Stabilization of vimentin coil2 fragment via an engineered disulfide**. *J. Struct. Biol.* 2012, **177**:46–53.

163. Bunick CG, Milstone LM: **The X-Ray Crystal Structure of the Keratin 1-Keratin 10 Helix 2B Heterodimer Reveals Molecular Surface Properties and Biochemical Insights into Human Skin Disease**. *J. Invest. Dermatol.* 2017, **137**:142–150.

164. Nicolet S, Herrmann H, Aebi U, Strelkov S V.: **Atomic structure of vimentin coil 2**. *J. Struct. Biol.* 2010, **170**:369–376.

165. Chernyatina AA, Guzenko D, Strelkov S V.: **Intermediate filament structure: The bottom-up approach**. *Curr. Opin. Cell Biol.* 2015, **32**:65–72.

166. Rafik ME, Doucet J, Briki F: **The intermediate filament architecture as determined by X-ray diffraction modeling of hard α-keratin**. *Biophys. J.* 2004, **86**:3893–3904.

167. Premchandar A, Kupniewska A, Tarnowski K, Mücke N, Mauermann M, Kaus-Drobek M, Edelman A, Herrmann H, Dadlez M: **Analysis of distinct molecular assembly complexes of keratin K8 and K18 by hydrogen-deuterium exchange**. *J. Struct. Biol.* 2015, **192**:426–440.

168. Fraser RDB, Parry DAD: **Intermediate filament structure in fully differentiated (oxidised) trichocyte keratin**. *J. Struct. Biol.* 2017, **200**:45–53.

169. Kolsch A, Windoffer R, Wurflinger T, Aach T, Leube RE: **The keratin-filament cycle of assembly and disassembly**. *J. Cell Sci.* 2010, **123**:2266–2272.

170. Omary MB, Ku NO, Strnad P, Hanada S: **Toward unraveling the complexity of simple epithelial keratins in human disease**. *J. Clin. Invest.* 2009, **119**:1794–1805.

171. Gu LH, Coulombe PA: **Keratin function in skin epithelia: a broadening palette with surprising shades**. *Curr. Opin. Cell Biol.* 2007, **19**:13–23.

172. Fraser RDB, Parry DAD: **Structural changes in the trichocyte intermediate filaments accompanying the transition from the reduced to the oxidized form**. *J. Struct. Biol.* 2007, **159**:36–45.

173. Fraser RDB, MacRae TP, Sparrow LG, Parry DAD: **Disulphide bonding in α-keratin**. *Int. J. Biol. Macromol.* 1988, **10**:106–112.

174. Russell D, Andrews PD, James J, Lane EB: **Mechanical stress induces profound remodelling of keratin filaments and cell junctions in epidermolysis bullosa simplex keratinocytes**. *J. Cell Sci.* 2004, **117**:5233–5243.

175. Lee CH, Kim MS, Chung BM, Leahy DJ, Coulombe PA: **Structural basis for heteromeric assembly and perinuclear organization of keratin filaments**. *Nat. Struct. Mol. Biol.* 2012, **19**:707–715.

176. Peters B, Kirfel J, Büssow H, Vidal M, Magin TM: **Complete cytolysis and neonatal lethality in keratin 5 knockout mice reveal its fundamental role in skin integrity and in epidermolysis bullosa simplex.** *Mol. Biol. Cell* 2001, **12**:1775–89.

177. Wagschal K, Tripet B, Lavigne P, Mant C, Hodges RS: **The role of position a in determining the stability and oligomerization state of alpha-helical coiled coils: 20 amino acid stability coefficients in the hydrophobic core of proteins.** *Protein Sci.* 1999, **8**:2312–2329.

178. Bjelić S, Wieser M, Frey D, Stirnimann CU, Chance MR, Jaussi R, Steinmetz MO, Kammerer RA: **Structural Basis for the Oligomerization-State Switch from a Dimer to a Trimer of an Engineered Cortexillin-1 Coiled-Coil Variant**. *PLoS One* 2013, **8**:e63370.

179. Kenne M, Dejean A: **Diet and foraging activity in Myrmicaria opaciventris (Hymenoptera: Formicidae: Myrmicinae)**. *Sociobiology* 1999, **33**:171–184.

180. Lamb D, Schüttelkopf AW, van Aalten DMF, Brighty DW: **Charge-surrounded pockets and electrostatic interactions with small ions modulate the activity of retroviral fusion proteins**. *PLoS Pathog.* 2011, **7**.

181. Oshaben KM, Salari R, McCaslin DR, Chong LT, Horne WS: **The native GCN4 leucine-zipper domain does not uniquely specify a dimeric oligomerization state**. *Biochemistry* 2012, **51**:9581–9591.

182. Zamyatnin AA: **Protein volume in solution**. *Prog. Biophys. Mol. Biol.* 1972, **24**:107–123.

183. Chang DK, Cheng SF, Trivedi VD, Lin KL: **Proline affects oligomerization of a coiled coil by inducing a kink in a long helix**. *J. Struct. Biol.* 1999, **128**:270–279.

184. Martínková N, Nová P, Sablina O V., Graphodatsky AS, Zima J: **Karyotypic relationships of the Tatra vole (Microtus tatricus)**. *Folia Zool.* 2004, **53**:279–284.

185. Walen KH, Brown SW: **Chromosomes in a marsupial (Potorous tridactylis) tissue culture**. *Nature* 1962, **194**:406.

186. Lee DL, Ivaninskii S, Burkhard P, Hodges RS: **Unique stabilizing interactions identified in the two-stranded α-helical coiled-coil: Crystal structure of a**

cortexillin I/GCN4 hybrid coiled-coil peptide. *Protein Sci.* 2003, **12**:1395–1405.

187. Nguyen MN, Tan KP, Madhusudhan MS: **CLICK - Topology-independent comparison of biomolecular 3D structures**. *Nucleic Acids Res.* 2011, **39**:W24-8.
188. Tan J, Ng I, Lim J, Wong HB, Yeo TT: **Patients who talk and deteriorate: A new look at an old problem**. *Ann. Acad. Med. Singapore* 2004, **33**:489–493.
189. Zhang L, Hermans J: **Calculation of the pitch of the alpha-helical coiled coil: An addendum**. *Proteins Struct. Funct. Genet.* 1993, **17**:217–218.
190. Nguyen MN, Madhusudhan MS: **Biological insights from topology independent comparison of protein 3D structures**. *Nucleic Acids Res.* 2011, **39**:e94–e94.
191. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE: **UCSF Chimera - A visualization system for exploratory research and analysis**. *J. Comput. Chem.* 2004, **25**:1605–1612.
192. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC: **GROMACS: Fast, flexible, and free**. *J. Comput. Chem.* 2005, **26**:1701–1718.
193. Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, Shirts MR, Smith JC, Kasson PM, van der Spoel D, et al.: **GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit**. *Bioinformatics* 2013, **29**:845–854.
194. Berendsen HJC, van der Spoel D, van Drunen R: **GROMACS: A message-passing parallel molecular dynamics implementation**. *Comput. Phys. Commun.* 1995, **91**:43–56.
195. Gillespie D: **Exact stochastic simulation of coupled chemical reactions**. *J. Phys. Chem.* 1977, **81**:2340–2361.
196. O'Shea EK, Rutkowski R, Stafford WF, Kim PS: **Preferential heterodimer formation by isolated leucine zippers from fos and jun.** *Science* 1989, **245**:646–648.
197. Hatzfeld M, Weber K: **The coiled coil of in vitro assembled keratin filaments is a heterodimer of type I and II keratins: Use of site-specific mutagenesis and recombinant protein expression**. *J. Cell Biol.* 1990, **110**:1199–1210.
198. Schnabel J, Weber K, Hatzfeld M: **Protein-protein interactions between keratin polypeptides expressed in the yeast two-hybrid system**. *Biochim. Biophys. Acta - Mol. Cell Res.* 1998, **1403**:158–168.
199. Trost A, Costa I, Jakab M, Ritter M, Haim M, Hintner H, Bauer JW, Önder K: **K16 is a further new candidate for homotypic intermediate filament protein**

interactions. *Exp. Dermatol.* 2010, **19**:e241-50.

200. Renner W, Franke WW, Schmid E, Geisler N, Weber K, Mandelkow E: **Reconstitution of intermediate-sized filaments from denatured monomeric vimentin**. *J. Mol. Biol.* 1981, **149**:285–306.
201. Steinert PM, Zimmerman SB, Starger JM, Goldman RD: **Ten-nanometer filaments of hamster BHK-21 cells and epidermal keratin filaments have similar structures.** *Proc. Natl. Acad. Sci. U. S. A.* 1978, **75**:6098–101.
202. Steinert PM, Idler WW, Goldman RD: **Intermediate filaments of baby hamster kidney (BHK-21) cells and bovine epidermal keratinocytes have similar ultrastructures and subunit domain structures.** *Proc. Natl. Acad. Sci. U. S. A.* 1980, **77**:4534–4538.
203. Parry DAD, Marekov LN, Steinert PM: **Subfilamentous Protofibril Structures in Fibrous Proteins: Cross-linking evidence for protofibrils in intermediate filaments**. *J. Biol. Chem.* 2001, **276**:39253–39258.
204. Steinert PM, Marekov LN, Parry DAD: **Conservation of the structure of keratin intermediate filaments: Molecular mechanism by which different keratin molecules integrate into preexisting keratin intermediate filaments during differentiation**. *Biochemistry* 1993, **32**:10046–10056.
205. Steinert PM, Marekov LN, Fraser RDB, Parry DAD: **Keratin Intermediate Filament Structure**. *J. Mol. Biol.* 1993, **230**:436–452.
206. Bragg PD, Hou C: **A cross-linking study of the Ca2+, Mg2+-activated adenosine triphosphatase of Escherichia coli.** *Eur. J. Biochem.* 1980, **106**:495–503.
207. Norlén L, Al-Amoudi A: **Stratum corneum keratin structure, function, and formation: The cubic rod-packing and membrane templating model**. *J. Invest. Dermatol.* 2004, **123**:715–732.
208. Aebi U, Fowler WE, Rew P, Sun TT: **The fibrillar substructure of keratin filaments unraveled**. *J. Cell Biol.* 1983, **97**:1131–1143.
209. Liang J, Dill KA: **Are proteins well-packed?** *Biophys. J.* 2001, **81**:751–766.
210. Herrmann H, Häner M, Brettel M, Ku NO, Aebi U: **Characterization of distinct early assembly units of different intermediate filament proteins.** *J. Mol. Biol.* 1999, **286**:1403–1420.
211. Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, Pieper U, Stuart AC, Marti-Renom MA, Madhusudhan MS, Yerkovich B, et al.: **Tools for comparative**

protein structure modeling and analysis. *Nucleic Acids Res.* 2003, **31**:3375–3380.

212. Gaines JC, Clark AH, Regan L, O'Hern CS: **Packing in protein cores**. *J. Phys. Condens. Matter* 2017, **29**:293001.
213. Richards FM: **The interpretation of protein structures: Total volume, group volume distributions and packing density**. *J. Mol. Biol.* 1974, **82**:1–14.
214. Harpaz Y, Gerstein M, Chothia C: **Volume changes on protein folding**. *Structure* 1994, **2**:641–649.
215. Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA: **Electrostatics of nanosystems: Application to microtubules and the ribosome**. *Proc. Natl. Acad. Sci.* 2001, **98**:10037–10041.
216. Jo S, Vargyas M, Vasko-Szedlar J, Roux B, Im W: **PBEQ-Solver for online visualization of electrostatic potential of biomolecules.** *Nucleic Acids Res.* 2008, **36**:W270-5.
217. Geisler N, Schunemann J, Weber K: **Chemical cross-linking indicates a staggered and antiparallel protofilament of desmin intermediate filaments and characterizes one higher-level complex between protofilaments**. *Eur. J. Biochem.* 1992, **206**:841–852.
218. Steinert PM: **Analysis of the mechanism of assembly of mouse keratin 1/keratin 10 intermediate filaments in vitro suggests that intermediate filaments are built from multiple oligomeric units rather than a unique tetrameric building block**. *J. Struct. Biol.* 1991, **107**:175–188.
219. Fraser RDB, Parry DAD: **The three-dimensional structure of trichocyte (hard α-) keratin intermediate filaments: Features of the molecular packing deduced from the sites of induced crosslinks**. *J. Struct. Biol.* 2005, **151**:171–181.
220. Watts NR, Jones LN, Cheng N, Wall JS, Parry DAD, Steven AC: **Cryo-electron microscopy of trichocyte (hard α-keratin) intermediate filaments reveals a low-density core**. *J. Struct. Biol.* 2002, **137**:109–118.
221. Nolting JF, Möbius W, Köster S: **Mechanics of individual keratin bundles in living cells**. *Biophys. J.* 2015, **107**:2693–2699.
222. Matoltsy AG: **Keratinization**. *J. Invest. Dermatol.* 1976, **67**:20–25.
223. Gillard BK, Thurmon LT, Marcus DM: **Association of glycosphingolipids with intermediate filaments of mesenchymal, epithelial, glial, and muscle cells**. *Cell Motil. Cytoskeleton* 1992, **21**:255–271.

224. Traub P, Perides G, Schimmel H, Scherbarth A: **Interaction in vitro of nonepithelial intermediate filament proteins with total cellular lipids, individual phospholipids, and a phospholipid mixture.** *J. Biol. Chem.* 1986, **261**:10558–68.

225. Norlén L, Emilson A, Forslind B: **Stratum corneum swelling. Biophysical and computer assisted quantitative assessments**. *Arch. Dermatol. Res.* 1997, **289**:506–13.

226. Mücke N, Wedig T, Bürer A, Marekov LN, Steinert PM, Langowski J, Aebi U, Herrmann H: **Molecular and biophysical characterization of assembly-starter units of human vimentin**. *J. Mol. Biol.* 2004, **340**:97–114.

227. Steinert PM, Parry DAD: **The conserved H1 domain of the type II keratin 1 chain plays an essential role in the alignment of nearest neighbor molecules in mouse and human keratin 1/keratin 10 intermediate filaments at the two- to four-molecule level of structure**. *J. Biol. Chem.* 1993, **268**:2878–2887.

228. Steven AC: **Intermediate Filament Structure**. In *Cellular and Molecular Biology of Intermediate Filaments*. . Springer US; 1990:233–263.

229. Schmid N, Eichenberger AP, Choutko A, Riniker S, Winger M, Mark AE, Van Gunsteren WF: **Definition and testing of the GROMOS force-field versions 54A7 and 54B7**. *Eur. Biophys. J.* 2011, **40**:843–856.

230. Huang W, Lin Z, van Gunsteren WF: **Validation of the GROMOS 54A7 Force Field with Respect to β-Peptide Folding**. *J. Chem. Theory Comput.* 2011, **7**:1237–1243.

231. Darden T, York D, Pedersen L: **Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems**. *J. Chem. Phys.* 1993, **98**:10089–10092.

232. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM: **LINCS: A Linear Constraint Solver for molecular simulations**. *J. Comput. Chem.* 1997, **18**:1463–1472.

233. Tan KP, Varadarajan R, Madhusudhan MS: **DEPTH: A web server to compute depth and predict small-molecule binding cavities in proteins**. *Nucleic Acids Res.* 2011, **39**:W242–W248.

234. Chakravarty S, Varadarajan R: **Residue depth: A novel parameter for the analysis of protein structure and stability**. *Structure* 1999, **7**:723–732.

235. Tan KP, Nguyen TB, Patel S, Varadarajan R, Madhusudhan MS: **Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins.** *Nucleic*

Acids Res. 2013, **41**:W314-21.

236. Berman HM, Westbrook JD, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucl. Acids Res.* 2000, **28**:235–242.
237. Kuhn HW: **The Hungarian method for the assignment problem**. In *50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art*. . 2010.
238. Rainey JK, Goh MC: **Statistically based reduced representation of amino acid side chains**. *J. Chem. Inf. Comput. Sci.* 2004, **44**:817–830.
239. Morley SM, Dundas SR, James JL, Gupta T, Brown RA, Sexton CJ, Navsaria HA, Leigh IM, Lane EB: **Temperature sensitivity of the keratin cytoskeleton and delayed spreading of keratinocyte lines derived from EBS patients.** *J. Cell Sci.* 1995, **108 (Pt 1**:3463–71.
240. Bateman A: **UniProt: A worldwide hub of protein knowledge**. *Nucleic Acids Res.* 2019, **47**:D506–D515.
241. who: **WHO | Nipah virus infection [Internet]**. *WHO* 2018, [no volume].
242. Looi L-M, Chua K-B: **Lessons from the Nipah virus outbreak in Malaysia.** *Malays. J. Pathol.* 2007, **29**:63–7.
243. Chew MHL, Arguin PM, Shay DK, Goh K, Rollin PE, Shieh W, Zaki SR, Rota PA, Ling A, Ksiazek TG, et al.: **Risk Factors for Nipah Virus Infection among Abattoir Workers in Singapore**. *J. Infect. Dis.* 2000, **181**:1760–1763.
244. **WHO | Nipah R&D**. *Who* 2018, [no volume].
245. Rajbari M, Rajbari N, Faridpur F: **Morbidity and mortality due to Nipah or Nipah-like virus encephalitis in WHO South-East Asia Region Country: India**. 2018, [no volume].
246. Yoneda M, Guillaume V, Sato H, Fujita K, Georges-Courbot M-C, Ikeda F, Omi M, Muto-Terao Y, Wild TF, Kai C: **The Nonstructural Proteins of Nipah Virus Play a Key Role in Pathogenicity in Experimentally Infected Animals**. *PLoS One* 2010, **5**:e12709.
247. Rockx B, Winegar R, Freiberg AN: **Recent progress in henipavirus research: Molecular biology, genetic diversity, animal models**. *Antiviral Res.* 2012, **95**:135–149.
248. Sen N, Kanitkar TR, Roy AA, Soni N, Amritkar K, Supekar S, Nair S, Singh G, Madhusudhan MS: **Predicting and designing therapeutics against the Nipah virus**. *PLoS Negl. Trop. Dis.* 2019, **13**:e0007419.

249. Welch BD, Yuan P, Bose S, Kors CA, Lamb RA, Jardetzky TS: **Structure of the Parainfluenza Virus 5 (PIV5) Hemagglutinin-Neuraminidase (HN) Ectodomain**. *PLoS Pathog.* 2013, **9**:e1003534.

250. Steffen DL, Xu K, Nikolov DB, Broder CC: **Henipavirus Mediated Membrane Fusion, Virus Entry and Targeted Therapeutics**. *Viruses* 2012, **4**:280–308.

251. Xu K, Chan YP, Bradel-Tretheway B, Akyol-Ataman Z, Zhu Y, Dutta S, Yan L, Feng YR, Wang LF, Skiniotis G, et al.: **Crystal Structure of the Pre-fusion Nipah Virus Fusion Glycoprotein Reveals a Novel Hexamer-of-Trimers Assembly**. *PLoS Pathog.* 2015, **11**:e1005322.

252. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al.: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.** *Mol. Syst. Biol.* 2011, **7**:539.

253. Aguilar HC, Aspericueta V, Robinson LR, Aanensen KE, Lee B: **A Quantitative and Kinetic Fusion Protein-Triggering Assay Can Discern Distinct Steps in the Nipah Virus Membrane Fusion Cascade**. *J. Virol.* 2010, **84**:8033–8041.

254. Weiss DR, Levitt M: **Can Morphing Methods Predict Intermediate Structures?** *J. Mol. Biol.* 2009, **385**:665–674.

255. Chambers P, Pringle CR, Easton AJ: **Heptad Repeat Sequences are Located Adjacent to Hydrophobic Regions in Several Types of Virus Fusion Glycoproteins**. *J. Gen. Virol.* 1990, **71**:3075–3080.

256. Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE: **Improved side-chain torsion potentials for the Amber ff99SB protein force field**. *Proteins Struct. Funct. Bioinforma.* 2010, **78**:1950–1958.

257. Vincent TL, Green PJ, Woolfson DN: **LOGICOIL - Multi-state prediction of coiled-coil oligomeric state**. *Bioinformatics* 2013, **29**:69–76.

258. Mahrenholz* C, Abfalter* I, Volkmer R, Bodenhofer U, Hochreiter S: **PrOCoil – Advances in predicting two- and three-stranded coiled coils**. *Nat. Preced.* 2010, doi:10.1038/npre.2010.4677.

259. Trigg J, Gutwin K, Keating AE, Berger B: **Multicoil2: Predicting coiled coils and their oligomerization states from sequence in the twilight zone**. *PLoS One* 2011, **6**.

260. Armstrong CT, Vincent TL, Green PJ, Woolfson DN: **SCORER 2.0: An algorithm for distinguishing parallel dimeric and trimeric coiled-coil**

sequences. *Bioinformatics* 2011, **27**:1908–1914.

261. Fraser RDB, MacRae TP: **The structure of the α-keratin microfibril**. *Biosci. Rep.* 1983, doi:10.1007/BF01120695.
262. Svitkina TM, Verkhovsky AB, Borisy GG: **Plectin sidearms mediate interaction of intermediate filaments with microtubules and other components of the cytoskeleton**. *J. Cell Biol.* 1996, **135**:991–1007.
263. Kouklis PD, Hutton E, Fuchs E: **Making a connection: Direct binding between keratin intermediate filaments and desmosomal proteins**. *J. Cell Biol.* 1994, **127**:1049–1060.
264. Fuchs E, Cleveland DW: **A Structural Scaffolding of Intermediate in Health and Disease**. *Science (80-.).* 1998, **279**:514–519.
265. Melo F, Sánchez R, Sali A: **Statistical potentials for fold assessment.** *Protein Sci.* 2002, **11**:430–48.
266. Handu M, Kaduskar B, Ravindranathan R, Soory A, Giri R, Elango VB, Gowda H, Ratnaparkhi GS: **SUMO-Enriched Proteome for Drosophila Innate Immune Response.** *G3 (Bethesda).* 2015, **5**:2137–54.
267. Isogai S, Shirakawa M: **Protein modification by SUMO**. *Seikagaku* 2007, **79**:1120–1130.
268. Kurepa J, Walker JM, Smalle J, Gosink MM, Davis SJ, Durham TL, Sung DY, Vierstra RD: **The small ubiquitin-like modifier (SUMO) protein modification system in Arabidopsis. Accumulation of sumo1 and -2 conjugates is increased by stress**. *J. Biol. Chem.* 2003, **278**:6862–6872.
269. Hay RT: **SUMO: A history of modification**. *Mol. Cell* 2005, **18**:1–12.
270. LoPiccolo J, Kim SJ, Shi Y, Wu B, Wu H, Chait BT, Singer RH, Sali A, Brenowitz M, Bresnick AR, et al.: **Assembly and molecular architecture of the phosphoinositide 3-kinase p85α homodimer**. *J. Biol. Chem.* 2015, **290**:30390–30405.
271. Robinson PJ, Trnka MJ, Pellarin R, Greenberg CH, Bushnell DA, Davis R, Burlingame AL, Sali A, Kornberg RD: **Molecular architecture of the yeast Mediator complex**. *Elife* 2015, **4**.
272. Thalassinos K, Pandurangan AP, Xu M, Alber F, Topf M: **Conformational states of macromolecular assemblies explored by integrative structure calculation**. *Structure* 2013, **21**:1500–1508.
273. Zhou XJ, Alber F: **Zooming in on genome organization**. *Nat. Methods* 2012,

9:961–963.

274. Luo J, Cimermancic P, Viswanath S, Ebmeier CC, Kim B, Dehecq M, Raman V, Greenberg CH, Pellarin R, Sali A, et al.: **Architecture of the Human and Yeast General Transcription and DNA Repair Factor TFIIH**. *Mol. Cell* 2015, **59**:794–806.
275. Choudhury AR, Sikorska E, Van Den Boom J, Bayer P, Popenda Ł, Szutkowski K, Jurga S, Bonomi M, Sali A, Zhukov I, et al.: **Structural model of the bilitranslocase transmembrane domain supported by NMR and FRET data**. *PLoS One* 2015, **10**.
276. Politis A, Schmidt C, Tjioe E, Sandercock AM, Lasker K, Gordiyenko Y, Russel D, Sali A, Robinson C V.: **Topological models of heteromeric protein assemblies from mass spectrometry: Application to the yeast eIF3:eIF5 complex**. *Chem. Biol.* 2015, **22**:117–128.
277. Xu X, Yan C, Wohlhueter R, Ivanov I: **Integrative Modeling of Macromolecular Assemblies from Low to Near-Atomic Resolution**. *Comput. Struct. Biotechnol. J.* 2015, **13**:492–503.
278. Querol-Audi J, Yan C, Xu X, Tsutakawa SE, Tsai M-S, Tainer JA, Cooper PK, Nogales E, Ivanov I: **Repair complexes of FEN1 endonuclease, DNA, and Rad9-Hus1-Rad1 are distinguished from their PCNA counterparts by functionally important stability**. *Proc. Natl. Acad. Sci.* 2012, **109**:8528–8533.
279. Tsutakawa SE, Van Wynsberghe AW, Freudenthal BD, Weinacht CP, Gakhar L, Washington MT, Zhuang Z, Tainer JA, Ivanov I: **Solution X-ray scattering combined with computational modeling reveals multiple conformations of covalently bound ubiquitin on PCNA**. *Proc. Natl. Acad. Sci.* 2011, **108**:17672–17677.
280. Zeng-Elmore X, Gao XZ, Pellarin R, Schneidman-Duhovny D, Zhang XJ, Kozacka KA, Tang Y, Sali A, Chalkley RJ, Cote RH, et al.: **Molecular architecture of photoreceptor phosphodiesterase elucidated by chemical cross-linking and integrative modeling**. *J. Mol. Biol.* 2014, **426**:3713–3728.
281. Algret R, Fernandez-Martinez J, Shi Y, Kim SJ, Pellarin R, Cimermancic P, Cochet E, Sali A, Chait BT, Rout MP, et al.: **Molecular Architecture and Function of the SEA Complex, a Modulator of the TORC1 Pathway**. *Mol. Cell. Proteomics* 2014, **13**:2855–2870.
282. Erzberger JP, Stengel F, Pellarin R, Zhang S, Schaefer T, Aylett CHS,

Cimermančič P, Boehringer D, Sali A, Aebersold R, et al.: **Molecular architecture of the 40S · eIF1 · eIF3 translation initiation complex**. *Cell* 2014, **158**:1123–1135.

283. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, et al.: **The molecular architecture of the nuclear pore complex**. *Nature* 2007, **450**:695–701.
284. Shi Y, Fernandez-Martinez J, Tjioe E, Pellarin R, Kim SJ, Williams R, Schneidman-Duhovny D, Sali A, Rout MP, Chait BT: **Structural Characterization by Cross-linking Reveals the Detailed Architecture of a Coatomer-related Heptameric Module from the Nuclear Pore Complex**. *Mol. Cell. Proteomics* 2014, **13**:2927–2943.
285. Politis A, Stengel F, Hall Z, Hernández H, Leitner A, Walzthoeni T, Robinson C V, Aebersold R: **A mass spectrometry-based hybrid method for structural modeling of protein complexes.** *Nat. Methods* 2014, **11**:403–406.
286. Aloy P, Böttcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin A-C, Bork P, Superti-Furga G, Serrano L, et al.: **Structure-based assembly of protein complexes in yeast.** *Science* 2004, **303**:2026–2029.
287. Campos M, Nilges M, Cisneros DA, Francetic O: **Detailed structural and assembly model of the type II secretion pilus from sparse data**. *Proc. Natl. Acad. Sci.* 2010, **107**:13081–13086.
288. Greenberg CH, Kollman J, Zelter A, Johnson R, MacCoss MJ, Davis TN, Agard DA, Sali A: **Structure of γ-tubulin small complex based on a cryo-EM map, chemical cross-links, and a remotely related structure**. *J. Struct. Biol.* 2016, **194**:303–310.
289. Schönegge AM, Villa E, Förster F, Hegerl R, Peters J, Baumeister W, Rockel B: **The structure of human tripeptidyl peptidase II as determined by a hybrid approach**. *Structure* 2012, **20**:593–603.
290. Zelter A, Bonomi M, Kim JO, Umbreit NT, Hoopmann MR, Johnson R, Riffle M, Jaschob D, MacCoss MJ, Moritz RL, et al.: **The molecular architecture of the Dam1 kinetochore complex is defined by cross-linking based structural modelling**. *Nat. Commun.* 2015, **6**:8673.

Publications

1. Neogi U, Rao SD, Bontell I, Verheyen J, Rao VR, Gore SC, **Soni N**, et al. Research letter: Novel tetra-peptide insertion in Gag-p6 ALIX binding motif in HIV-1 subtype C associated with protease inhibitor failure in Indian patients. Aids. 2014;28: 2319–2322.

2. Hotta K, Ranganathan S, Liu R, Wu F, Machiyama H, Gao R, **Soni N**, et al. Biophysical Properties of Intrinsically Disordered p130Cas Substrate Domain - Implication in Mechanosensing. PLoS Comput Biol. 2014;10: e1003532.

3. Manjunath GP, **Soni N**, Vaddavalli PL, Shewale DJ, Madhusudhan MS, Muniyappa K. Molecular Mechanism Underlying ATP-Induced Conformational Changes in the Nucleoprotein Filament of Mycobacterium smegmatis RecA. Biochemistry. American Chemical Society; 2016;55: 1850–1862.

4. Short JM, Liu Y, Chen S, **Soni N**, Madhusudhan MS, Shivji MKK, et al. High-resolution structure of the presynaptic RAD51 filament on single-stranded DNA by electron cryo-microscopy. Nucleic Acids Res. Oxford University Press; 2016;44: 9017–9030.

5. **Soni N**, Madhusudhan MS. Computational modeling of protein assemblies. Curr Opin Struct Biol. Elsevier Ltd; 2017;44: 179–189.

6. Roy A, Nair S, Sen N, **Soni N**, Madhusudhan MS. In silico methods for design of biological therapeutics. Methods. Elsevier Ltd; 2017;131: 33–65.

7. Sen N, Kanitkar T, Roy A, **Soni N**, Amritkar K, Supekar S, Nair S, Singh G, Madhusudhan M.S. Predicting and Designing therapeutics against the Nipah virus. PLoS Negl. Trop. Dis. 2019, 13:e0007419.

8. *Nguyen M.N., Sharma P, **Soni N**, Verma C, Madhusudhan MS. CLICK: a web server to mine structural databases for similar sub-structures of biological macromolecules and their complexes.*

9. ***Soni N**, Madhusudhan MS. Molecular basis of IF related diseases and their dimerization principles; Predictions and validation of novel filament disruptive mutations.*

10. ***Soni N***, *Madhusudhan M.S. Three-dimensional structure of Keratin K5-K14 filament at near atomic level resolution.*

11. ***Soni N***. *Disctransformation: A software to efficiently pack and score 2D objects in a given geometry.*

12. ***Soni N***, *Magin T, Madhusudhan M.S. Predicting temperature sensitive mutations for Intermediate filament family.*

13. ***Soni N***, *Madhusudhan M.S. GARLIC: A package for Geometric Assessment of Residues using Locally Interacting Chemical groups.*

14. ***Soni N***, *Madhusudhan M.S. AIMorpher: A fast and intelligent morphing tool for macro-molecular structures.*

15. ***Soni N***, *Ramtirtha Y., Prasanna V, Madhusudhan M.S. ATP binding site prediction and validation for 14-3-3 family of proteins.*

16. ***Soni N***, *Ramtirtha Y., Ratnaparkhi G, Madhusudhan M.S. SUMO Server. A database of Sumoylated proteins in drosophila and annotation of plausible homologous sumoylated proteins.*

Supplemental Material

Details of the additional research data can be obtained from the following link.

http://cospi.iiserpune.ac.in/cospi/IF/IFresearchdata.zip

File Password: K#R@T!NK5K14