UNIVERSITY OF CALIFORNIA,
IRVINE

In Silico Exploration and Experimental Validation: A Story of Protein Discovery, Aggregation and Degradation

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Chemistry

by

Megha Hemant Unhelkar

Dissertation Committee:
Professor Rachel W. Martin, Chair
Professor Carter T. Butts
Professor Douglas J. Tobias

2020

DEDICATION

To my parents, Aarti and Hemant Unhelkar,
for their constant encouragement, unwavering support and unconditional love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I must start by thanking my advisor, Prof. Rachel W. Martin, not only a fantastic scientist and advisor, who helped me achieve what I set out for, but also a good friend who always has my back. I also thank her for starting my lapel pins collection.

A big thank you to my collaborator and mentor, Prof. Carter T. Butts, who is my statistics guru, who supported me through the many papers we published together and taught me that the most important thing at any conference is networking and dinner.

I would also like to thank Prof. Douglas J. Tobias, another wonderful mentor, for always enthusiastically discussing science, encouraging innovative ideas and for patiently advising on science and life, as such.

It takes a village to finish a PhD and here is my tribe- Prof. Suvrajit Sengupta for teaching me all the NMR, Dr. Gianmarc Grazioli for teaching me all the MD, Dr. Jan Bierma for teaching me all the protein magic and Prof. Siddharth Joshi, for teaching me all the R and Matlab, and keeping me sane through the five years. This PhD would not have been possible without your friendship, or your brain. A shout out to all my collaborators, lab mates, case partners, friends and family for your encouragement and friendship.

Jerry Unhelkar and Jeeves Unhelkar, I don't love anyone as much as I love you- thank you for choosing me, rescuing me, saving me and loving me back.

This PhD is dedicated to my parents, Aarti and Hemant Unhelkar. Where would I be, had it not been for you *Aai Baba*? I can never thank you enough for your unconditional love, constant support and for having more faith in me than I ever did in myself. I can only try to make you proud of me, someday.

This work was supported by the following grants: in part, through access to the Genomic High Throughput Facility Shared Resource of the Cancer Center Support Grant (CA-62203) at the University of California, Irvine and NIH shared instrumentation grants 1S10RR025496-01 and 1S10OD010794-01; this research was also supported by NSF award DMS-1361425, DMS-1361425 and NIH shared instrumentation grants 1S10RR025496-01 and 1S10OD010794-01 and ARO award W911NF-14-1-0552.

CURRICULUM VITAE

Megha Hemant Unhelkar

EDUCATION

Doctor of Philosophy, Chemistry — **2020**
University of California, Irvine — *Irvine, California*

Master of Science, Pharmacology — **2013**
Northeastern University — *Boston, Massachusetts*

Bachelor of Pharmacy — **2011**
University of Mumbai — *Mumbai, India*

RESEARCH EXPERIENCE

Ph.D. Candidate — **2015–2020**
University of California, Irvine — *Irvine, California*

Associate Scientist — **2013–2015**
BASF — *San Diego, California*

Senior Research Associate — **2012–2013**
Celgene Avilomics Research — *Bedford, Massachusetts*

RELEVANT EXPERIENCE

Safety Minister — **2015- 2020**
Martin Lab, UCI — *Irvine, California*

Tech Transfer Fellow — **2018–2020**
Innovation Transfer Group — *Irvine, California*

Founder and CEO — **2018–2020**
The Science of Colors — *Irvine, California*

Consultant Participant — **2019**
1st Prize Winner, IXL Innovation Olympics — *Boston, Massachusetts*

PUBLICATIONS

- "Structure prediction and network analysis of chitinases from the Cape sundew, *Drosera capensis*" **Unhelkar, M. H.**; Duong, V. T.; Enendu, K. N.; Kelly, J. E.; Tahir, S.; Butts, C. T.; Martin, R. W. Biochimica et Biophysica Acta (2017): 636-643
- "Protein structure networks provide insight into active site flexibility in esterase/lipases from the carnivorous plant *Drosera capensis*" Duong, V. T.;**Unhelkar, M. H.**; Kelly, J. E..; Kim, S.; Butts, C. T.; Martin, R. W. Integrative Biology 10.12 (2018): 768-779
- "Network-based Classification and Modeling of Amyloid Fibrils" Grazioli, G.; Yu, Y.; **Unhelkar, M. H.**; Martin, R. W.; Butts, C. T. Biophysical Journal, 116(3), 559a
- "Sequence comparison, molecular modeling, and network analysis predict structural diversity in cysteine proteases from the Cape sundew, *Drosera capensis*" Butts C. T.; Zhang, X.;Kelly, J. E.; Roskamp, K.W.; **Unhelkar, M. H.**; Kelly J. E..; Kim S.; Freites, J. A.; Martin, R. W. Computational and structural biotechnology journal 14 (2016): 271-282
- "Network Hamiltonian Models Reveal Pathways to Amyloid Fibril Formation" Yu, Y.; Grazioli, G.; **Unhelkar, M. H.**; Martin, R. W.; Butts, C. T. Scientific Reports (2020): 1-11
- "Analysis of active site networks predicts variable substrate specificity in phospholipases from *Drosera capensis*" **Unhelkar, M. H.**; Zhuang, S.; Xu, M.; Kelly, J. E..; Butts, C. T.; Martin, R. W. *Manuscript in preparation, December 2020*
- "NMR structure determination and biophysical characterization of J2 crystallin- a unique eye lens protein from the box jellyfish" **Unhelkar, M. H.**; Khago, D.; Kelz, J; Bierma, J.; Sengupta, S.; Kingsley, C.;Valentic, T. R.; Dizon, P. A.; Martin, R. W. *Manuscript in preparation, December 2020*

INVITED TALKS

The fascinating story of *Drosera capensis* **November 2019**
Annual Gilliam Biomedical Research Conference, Irvine, California

Protein Discovery and Characterization using Molecular Modeling, Experimental Chemistry and Bioinformatics **March 2019**
Columbia University in the City of New York, NY

Structure Prediction and Network Analysis of Enzymes from Carnivorous Plant *Drosera capensis* **October 2017**
Women in Statistics and Data Science Conference, La Jolla, California

CONFERENCE PRESENTATIONS

Network-based modeling of amyloid fibril formation **March 2020**
American Chemical Society National Meeting, Philadelphia, Pennsylvania

Leveraging molecular modeling, experimental chemistry and bioinformatics to discover and characterize novel proteins **February 2020**
Biophysical Society Annual Meeting, San Diego, California

Biophysical characterization and solution-state structure determination using NMR of J2 crystallin: Novel eye lens protein from the box jellyfish **August 2019**
American Chemical Society National Meeting, San Diego, California

NMR structure determination and biophysical characterization of J2 crystallin- a unique eye lens protein from the box jellyfish **April 2019**
Experimental Nuclear Magnetic Resonance Conference, Asilomar, California

Protein Discovery and Characterization using Molecular Modeling, Experimental Chemistry and Bioinformatics **March 2019**
Biophysical Society Annual Meeting, Maryland, Baltimore

Design and Construction of a static ssNMR Probe for the Investigation of Oriented Membrane Samples and Single Crystals **March 2017**
Experimental Nuclear Magnetic Resonance Conference, Asilomar, California

Design and Construction of ssNMR Probes for the Investigation of Oriented Solids and Liquids **July 2016**
Rocky Mountain Conference on Magnetic Resonance Breckenridge, Colorada

ABSTRACT OF THE DISSERTATION

In Silico Exploration and Experimental Validation: A Story of Protein Discovery, Aggregation and Degradation

By

Megha Hemant Unhelkar

Doctor of Philosophy in Chemistry

University of California, Irvine, 2020

Professor Rachel W. Martin, Chair

Proteins are fundamental building blocks of life: understanding protein structure, function, aggregation, and degradation is, therefore, one of the central questions in biology. My work investigates protein aggregation and degradation through computational modeling, protein structure network analysis, and experimental verification.

One theme of my work is the discovery of new enzymes from the carnivorous plant, *Drosera capensis* (*D. capensis*). With the ever-expanding genomic data, it is imperative to swiftly move from raw genomic data to chemical results. Using the "target selection pipeline" that we invented, *in silico* protein structures can be predicted rapidly, to direct the subsequent experimental characterization of the promising candidates. Subsequent network analysis predicts interesting protein properties such as potential enzyme activity, enzyme specificity and the functional pH range, aiding the selection of functionally useful proteins for experimental characterization. This approach illustrates a generally applicable way to leverage the wealth of information provided by whole genome shotgun sequencing for proteomics. Computational techniques, despite their limitations, are now powerful enough to allow potentially useful proteins to be identified directly from the genome and filtered for strong indicators of biochemical function. So far, this work has resulted in three publications including proteases,

chitinases, and esterase/lipases.

The protease resistance of amyloid fibrils and their central role in more than 40 human diseases, including Alzheimer's, makes them an attractive target to test the activity of new proteases from *D.capensis*. To advance and streamline scientific discovery related to amyloid fibrils, it was crucial to have a standardized nomenclature. With collaborators, I introduced a systematic approach to the nomenclature of fibril topology using graph theoretic concepts to abstract the structure. The scheme encompasses all amyloid fibrils currently in the Protein Data Bank (PDB), and can be easily extended to accommodate newer discoveries. The work also showed that the vast majority of known fibril structures fall into just three topological categories, something that was previously unnoticed. My work has improved the discussion of fibril structures by condensing the descriptions of complicated structural features using a set of universal structural motifs.

The other theme of my work includes solving the protein structure of J2 crystallin, an aggregation resistant protein. J2-crystallin is a novel eye lens protein, highly expressed in *Tripedalia cystophora* (box jellyfish) and is an interesting target because of its very high stability and water-solubility. Unlike most non-cephalopod invertebrates, box jellyfish have camera-type eyes; therefore, their crystallins present an interesting system from an evolutionary biology perspective, making them an intriguing model system for vertebrates. Interestingly, Basic Local Alignment Search Tool (BLAST) search of J2 in the Protein Data Bank (PDB) found no proteins above 32 % similarity. Therefore, the structure determination of J2 is not only important from the evolutionary standpoint but also because of the hypothesis that J2 possesses a novel protein fold, due to lack of known homology. Here, I present the biophysical characterization, and solution-state NMR assignments of J2 crystallin, a previously uncharacterized eye lens protein, addressing the interplay of sequence, structure and function in the eye lens crystallins.

Chapter 1

Significance of rapid, *in silico* discovery of novel proteins from the carnivorous plant *Drosera capensis*

Proteins are fundamental building blocks of life: understanding protein structure, function, aggregation, and degradation is, therefore, one of the central questions in biology. It is not only important to discover new proteins as a source to study novel systems but also to expand the toolkit for chemical biology, biotechnology and proteomics. My work investigates protein aggregation and degradation through discovering new protein from the carnivorous plant *Drosera capensis*, using bioinformatics and genomics tools for analysis and testing the activity interesting of the newly discovered enzyme candidates experimentally. One may ask why *D. capensis*? The digestive enzymes of carnivorous plants have been a topic of biological interest at least since Darwin's 1875 monograph on insectivorous plants; in fact, in his book "Insectivorous plants" [17, 18] Darwin says he cares more about *Drosera* than the origin of all the species in the world!

Darwin observed that the mucilage secretions of plants in the genus *Drosera* contained a "ferment" that he conjectured to be similar to mammalian pepsin (now known to be an aspartic protease) [17, 18]. Only recently scientists are beginning to characterize the carnivorous plant digestive enzymes. Before *Drosera capensis*, only two carnivorous plant genomes, *Genlisea aurea* and *Utricularia gibba*, both members of the asterid order Lamiales, were sequenced [19, 20]. Both these plants feed on small, often microscopic, prey and perform their digestive functions in closed traps in a relatively thermostable environment (underground or under water), therefore they are less subject to the environmental constraints faced by carnivorous plants that perform their prey capture in exposed environments [3]. On the other hand, carnivorous plants require stable and highly active digestive enzymes that would allows the plant to digest its prey to its component amino acid over relatively long time spans and usually under milder chemical conditions than those of their animal counterparts [3]. As the digestive process must occur without any mechanical disruption of the prey tissue while competing with bacterial and fungal growth, carnivorous plants are attractive targets for enzyme discovery.

Carnivorous plant digestive enzymes are stable, are substrate specific, possess unique cleavage patterns, and have the ability to function over different pH ranges, presenting a rich resource for chemical biology and biotechnology laboratory applications [3, 21, 2]. We selected the Cape sundew *(Drosera capensis)*, native to the Cape region of South Africa and belonging to the order Caryophyllales [3, 21, 2]. *D. capensis* is an excellent model organism for the study of carnivory in plants as it can easily be cultivated, is capable of self-pollination, matures quickly, requires no period of dormancy, and is large and robust, facilitating tissue collection for multiple experiments from the same specimen [3, 21, 2].

The sequencing and assembly of a high-quality draft genome for *D. capensis* by the Martin lab and the Butts lab can be found in the paper [3] from which a plethora of enzymes were discovered. However, the journey of a protein sequence to an experimentally studied

protein takes many years, as observed in Uniprot which has more than 5 million protein sequences without a structure [22]. To study genomics effectively, it is imperative to quickly move from raw data to chemical results. Using the target selection pipeline, *in silico* protein structures can be predicted rapidly, to direct the subsequent experimental characterization of the promising candidates [3, 21, 2]. Subsequent network analysis predicts interesting protein properties such as potential enzyme activity, enzyme specificity and the functional pH range, aiding the selection of functionally useful proteins for experimental characterization. This approach illustrates a generally applicable way to leverage the wealth of information provided by whole genome shotgun sequencing for proteomics. Computational techniques, despite their limitations, are now powerful enough to allow potentially useful proteins to be identified directly from the genome and filtered for strong indicators of biochemical function [3, 21, 2]. So far, this work has resulted in three publications including proteases, chitinases, and esterase/lipases [18, 21, 2]. As the target selection pipeline forms a basis to my PhD, I will be discussing the target selection pipeline (Figure 1.1) in detail in this chapter.

1.0.1 Genomic DNA assembly and gene discovery

Genomic DNA was isolated using a protocol developed for recalcitrant plants by Prof. Rachel Martin [18] and details can be found in the paper.

1.0.2 Feature annotation

Sequence alignments are performed using ClustalOmega [24], with settings for gap open penalty = 10.0 and gap extension penalty = 0.05, hydrophilic residues = GPSNDQERK, and the BLOSUM weight matrix [23]. The presence and position of a signal sequence flagging the protein for secretion was predicted using the program SignalP 4.1 [24], while other localization sequences were identified using TargetP [25]. The alignment figures are annotated

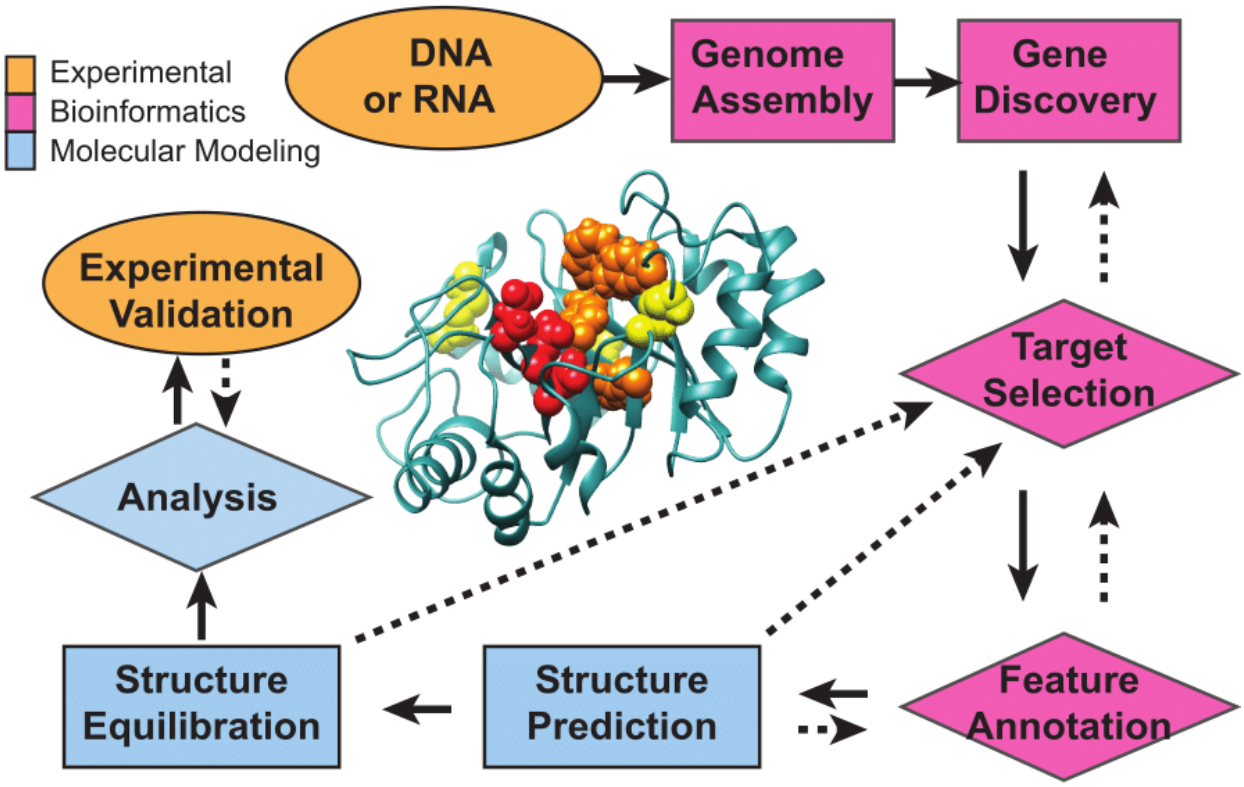

Figure 1.1: Flow chart illustrating the overall strategy for identifying enzymatic targets from genomic DNA. The workflow is indicated with solid arrows, while dotted arrows represent steps where information from a later stage of the pipeline enables refinement of earlier stages in an iterative manner. After genome sequencing, assembly, and gene discovery, target proteins are identified based on putative enzymatic activity. Functional sequence features are identified by analogy to annotation reference sequences found in the UniProt database. Structures are predicted using the Rosetta software, and equilibrated in explicit solvent after removal of sequence regions not present in the mature enzyme. Structures are compared using network analytic methods, enabling strategic selection of enzymes for experimental characterization in a future study. [1, 2, 3]

to highlight chemical properties of the amino acid residues as well as important sequence features. The amino acid attributes are color-coded as follows: cysteines are yellow, positively charged residues are blue, negatively charged residues are red, hydrophobic residues are green, and all others are black [1, 2, 3]. Highly conserved residues are indicated with a dot above the sequence position. The catalytic triad residues are marked with colored arrows. SignalP 4.1 [26] is used to predict the signal peptide cleavage site, which is specified by underlining the residues on either end of the cleavage point. The signal peptide itself is highlighted in light orange. Strikethrough text indicates sequence regions that are absent in the active enzyme, in this case the N-terminal signal peptide that is expressed but removed during maturation. Annotations were performed by homology to the annotations reference sequences from found in the UniProt database and identified by their UniProt IDs. An example of chitinase Family 18 is seen in Figure 2.2 [21]. More details on chitinases can be found in the following chapter.

Figure 1.2: Sequence alignment for Family 18 chitinases, annotated by homology to the reference sequence CHIT3_VITVI. The "DXDXE" motif, in which the acidic residues are marked with red arrows, is imperative for the enzyme activity. Orange arrows indicate residues implicated in substrate binding.

1.0.3 Structure prediction, equilibration and Protein Structure Network (PSN) Analysis

In collaboration with Prof. Carter Butts, we developed the *in silico* maturation method. Preliminary models of the newly discovered enzymes are produced using the online Robetta implementation [27] of Rosetta [28]. The Rosetta structures contain the full sequences, simulated in vacuum, and without any post-transnational modifications. *In silico* maturation matures the preliminary structures obtained from Rosetta, adds the post-translational modifications and chemical changes needed to make initial model match its native chemical environment, which includes, but not limited to pro-sequence removal, protonation state correction, and solvation, and adding disulfide bonds, introducing backbone cuts, oligomerization and metal coordination. During *in silico* maturation, the signal sequence is removed and the structure is equilibrated for 500 ps in explicit TIP3P solvent using NAMD[29], using the CHARMM22 forcefield with the CMAP correction and sodium or chlorine ions were added as necessary to neutralize the charge of the resulting structure [30]. Examples of chitinases, esterase/lipases, phospholipases and proteases are represented in the following chapters.

In addition to the presence or absence of specific features, identifying broader patterns of structural differentiation can be helpful when selecting putative proteins for expression and characterization: proteins within different structural subgroups may differ with respect to other biophysically important properties such as thermal stability, substrate affinity pattern, overall activity, or aggregation propensity, and choosing a structurally diverse sample thus has the potential to maximize the chance of identifying proteins with functionally significant variation [30]. PSNs are a useful tool for such exploration, as they directly represent patterns of potential interaction among chemical groups rather than e.g. side chain dihedral angles or other properties that may vary substantially without inducing significant changes in protein function [30]. PSNs are created for each protein structure by Prof. Carter Butts, using

custom scripts using VMD and statnet tools [18] and more details can be found in the paper [30].

1.0.4 Experimental validation

Once the target enzymes were selected using the target selection pipeline, the enzymes can be expressed and studied experimentally. Projects in the Martin lab include expression of chitinases [21] and proteases [30]. As seen in subsequent chapters, I wanted to test the activity of proteases and I talk in-depth in the following chapters.

Chapter 2

In silico structure prediction and network analysis of chitinases from *Drosera capensis*

2.1 Background

Chitin, a β-(1,4)-N acetylglucosamine (GlcNAc) biopolymer, is the second-most abundant biopolymer [31]. Chitinases (EC 3.2.1.14) are ubiquitous even among organisms that do not produce chitin, with the latter employing them for purposes of digestion and/or defense. These enzymes cleave chitin at the α-1,4 linkage of N-acetyl glucosamine units, although substantial variation in activity and substrate specificity exists. Some chitinases can also cleave peptidoglycans at β-1,4 linkages between N-acetylmuramic acid and N-acetyl-D-glucosamine, and chitodextrins between N-acetyl-D-glucosamine units. Plant chitinases sometimes have multiple functionalities; some display lysozyme activity [32], while others have a calcium storage function [33]. In humans, chitinases are produced in response to fungal infections, a

feature of the innate immune system that is suppressed in immunocompromised individuals, including AIDS patients, transplant recipients, and burn victims [34]. These enzymes and related chitin-binding proteins are expressed in human lung tissue, where they are dysregulated in cystic fibrosis and asthma [35].

In plants, these enzymes are expressed in response to environmental stress and pathogen or pest infestation [36], driving efforts to overexpress particularly effective examples in transgenic crop plants [37]. Carnivorous plants use chitinases as part of the prey capture response: active chitinases have been found in the pitcher fluid of *Nepenthes* [38, 39], and in the digestive fluids of the Venus flytrap [6]. However, the extent to which chitin is used as a nitrogen source remains controversial. *Drosera capensis* plants fed on chitin incorporate its nitrogen into their leaf tissue; however nutrient uptake is less efficient than for plants fed on protein [40]. Examination of insect carcasses after digestion reveals that 40-60% of the total nitrogen is unused [41, 42], consistent with the observation that the remains of insect exoskeletons appear mostly intact [43]. However, chitinase expression is upregulated in the presence of prey in the related species *Nepenthes alata*. In *Drosera rotundifolia*, an increase in both expression of chitinase mRNA and chitinase activity was induced by addition of crustacean chitin with mechanical stimulation of the traps [44]. The prey-induced induction of chitinase activity, despite the low efficiency of chitin use, may indicate that chitinases primarily function to inhibit fungal growth in the traps, just as cytotoxic peptides discourage microbial growth in the fluid of *Nepenthes* pitchers [45, 46].

In this work, I compare novel chitinases recently discovered from the genome of the Cape sundew (*D. capensis*) [47], to those from other carnivorous plants in order Caryophylalles. The conservation of the overall protein folds and active site architectures suggests that many of the *D. capensis* chitinase sequences form functional enzymes. Using the 'Target Selection Pipeline' described in Chapter 2 which involves sequence analysis, comparative modeling with all-atom refinement followed by *in silico* maturation [48], and investigation of protein

structure networks, structurally distinct subgroups of proteins for subsequent expression and biochemical characterization could be identified. Author contributions can be found in the paper [21]. It is important to understand the comparison of Family 19 and Family 18 and therefore, a portion of the results from the paper [21] are shown in this chapter.

2.2 Materials and Methods

2.2.1 Two Distinct Families of Carnivorous Plant Chitinases Are Found in *D. capensis*

Gene sequences annotated as coding for chitinases using the MAKER-P (v2.31.8) pipeline [49] and a BLAST search against SwissProt (downloaded 8/30/15) and InterProScan [50] were clustered by sequence similarity, along with chitinases previously identified from *Dionaea muscipula* [6] and various species of *Drosera* and *Nepenthes* [51]. Annotated sequence alignments of the Family 18 and Family 19 chitinases are shown in Figure 3.1. We have identified four fragments ranging from 41%-100% identity to the DcChit1_1 fragment previously found by Renner and Specht in *D. capensis* genomic DNA [51] (Figure2.4). Several well-characterized reference sequences (e.g chitinases from *Vitis vinifera*, *Brassica napus*, and *Hordeum vulgare*) are also included for comparison. Using the characterization scheme of the carbohydrate-active enzymes (CAZy) database [52, 53], the chitinases investigated here belong to Family 18 (orange) or Family 19 (green). Overall, the sequence identity among the Family 18 chitinases from Caryophylalles carnivorous plants is much higher than that of Family 19, as illustrated in Figure 3.1A and B. These two types of chitinases have different folds and are thought to have evolved independently, [54, 55], consistent with their separation into separate clusters (Figure 3.1C). Family 18 contains types III and V, while types I, II and IV belong to Family 19 [6]. My collaborator Vy Doung worked on Family 19. [1].

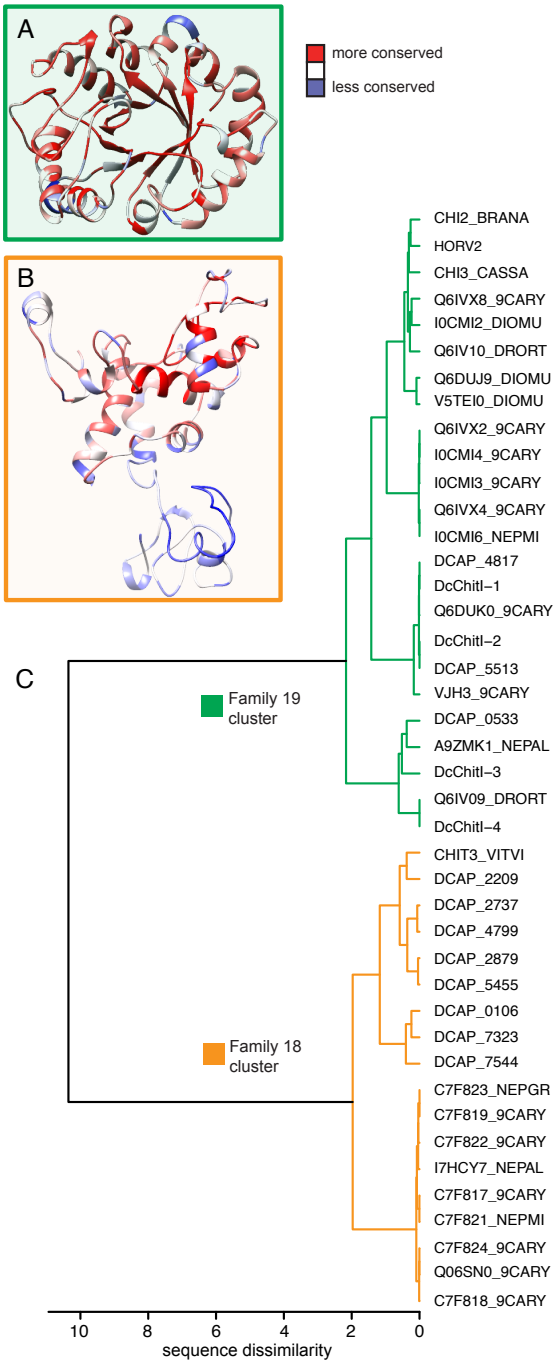

Figure 2.1: Clustering of chitinases identified from the *D. capensis* genome, compared with those from other Caryophylalles carnivorous plants and well-characterized reference sequences. All of the sequences examined belong to GH Families 18 or 19. The sequence dissimilarity used here is the e-distance metric of Székely and Rizzo [4] (with $\alpha = 1$). This parameter is a weighted function of within-cluster similarities and between-cluster differences with respect to a user-specified reference metric, defined here as the raw sequence dissimilarity (1 - (%identity)/100).

2.2.2 Sequence Alignment and Prediction of Putative Protein Structures

Sequences were aligned with ClustalOmega [56] (gap open penalty = 10.0, gap extension penalty = 0.05, hydrophilic residues = GPSNDQERK, weight matrix = BLOSUM). Secretion signal sequences were predicted using SignalP 4.1 [26]. Structure prediction was performed as in [47]. In the first stage, the Robetta [28] implementation of Rosetta [27] was used to produce an initial model for each protein. In the second stage, the model was subjected to "*in silico* maturation." Signal peptides were removed, and disulfide bonds identified by a combination of homology and distance constraints. Protonation states of active site residues were corrected to match literature values where necessary; for Family 18 chitinases, we approximate the sharing of a proton between active site residues D1 and D2 by protonation of D1 (which results in realistic side chain orientations and preserves the attractive interaction between D1 and D2). In the third and final stage, each matured enzyme model was equilibrated in explicit solvent (TIP3P water [57]) under periodic boundary conditions using NAMD [29]. Simulation was performed using the CHARMM36 forcefield [58], with each model being energy-minimized for 10,000 iterations and then simulated at 293K for 500ps; the final protein conformation was retained for subsequent analysis. For the one reference sequence for which a structure was available (HORV2, PDB ID 2BAA, [59]), this was used as the initial starting model (following removal of heteroatoms and protonation using REDUCE [60]). PDB files corresponding to the equilibrated structures for all the proteins discussed in this work available to download and discussed in Table 1 and Table 2.

A sequence alignment for Family 18 chitinases from Caryophylalles carnivorous plants is shown in Figure 2.2. The figure is annotated to highlight specific amino acid properties and important sequence features. The chemical properties of amino acids are color-coded as follows: cysteines are yellow, positively charged residues are blue, negatively charged residues are red, hydrophobic residues are green, and all others are black. Highly conserved

residues are indicated with a dot above the sequence position. Cysteine residues involved in structure-stabilizing disulfide bonds are indicated with yellow asterisks, while the active amino acid residues are marked with colored arrows. SignalP 4.1 is used to predict the signal peptide cleavage site, which is specified by underlining the residues on either of the cleavage point. The signal peptide itself is highlighted in light orange. Strikethrough text indicates sequence regions that are absent in the active enzyme, in this case the N-terminal signal peptide that is expressed but removed during maturation. Annotations were performed by homology to a well-characterized acidic endochitinase from *Vitis vinifera* (CHIT3_VITVI, Uniprot ID-P51614).

Figure 2.2: Sequence alignment for Family 18 chitinases, annotated by homology to the reference sequence CHIT3_VITVI. The "DXDXE" motif, in which the acidic residues are marked with red arrows, is imperative for the enzyme activity. Orange arrows indicate residues implicated in substrate binding.

Family 19 contains Class I, II, and IV chitinases, all of which are characterized by an anomeric inverting mechanism [61, 62]. Annotations for the Family 19 chitinases are shown in Figure 2.3. Family 19 contains Class I, II, and IV chitinases, all of which are characterized by an anomeric inverting mechanism [61, 62]. The N-terminal chitin-binding domain is present in Class I and absent in Class II, which are otherwise similar in sequence. Family 19 chitinases from plants have in common a catalytic domain with an active glutamic acid residue. The active site motif surrounding the active E is either HETT (type I and II) or HETG (type IV) [51], both of which are observed in this set of proteins. Annotations for the Family 19 chitinases are shown in Figure 2.3. Amino acid and sequence features are indicated as in Figure 2.2, with the following additions, when present: the C-rich domain is highlighted in light green, the P-rich hinge in light blue, and the C-terminal extension (CTE) in light gray. Both the C-rich domain and the P-rich hinge are highly variable in length and are absent in some sequences. Only three chitinases in this set contain the CTE, which targets those sequences to the vacuole. The reference sequences for this cluster are CHI3_CASSA (*Castanea sativa*), CHI2_BRANA (*Brassica napus*), and HORV2 (*Hordeum vulgare*).

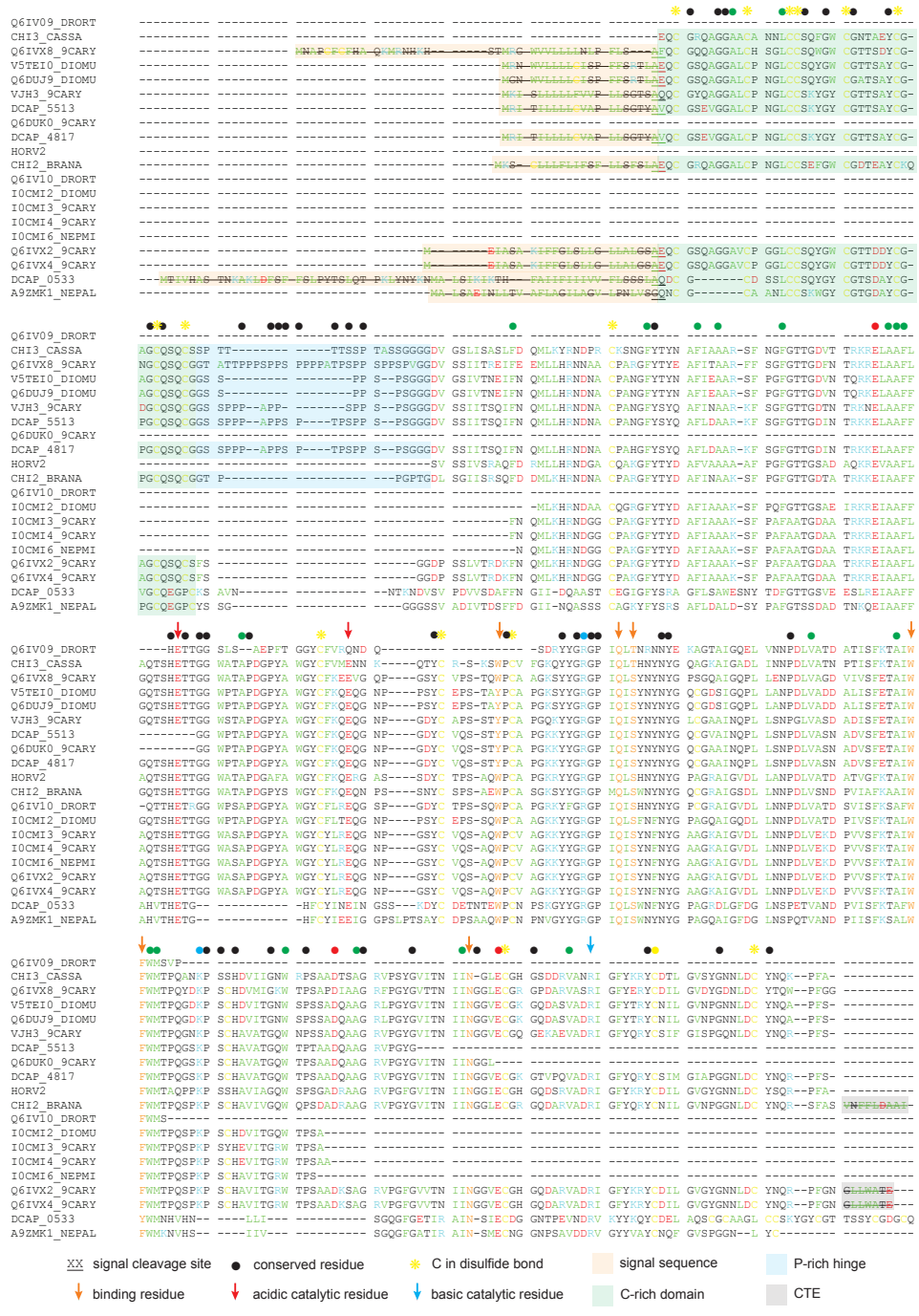

Figure 2.3: Sequence alignment and annotation for Family 19 chitinases. Many sequences in this cluster contain a chitin-binding C-rich domain (light green) that is connected to the active region by a P-rich hinge (light blue). Three sequences in this cluster contain a C-terminal extension (CTE) that causes the proteins to be targeted to the vacuole.

Four Family 19 chitinase fragments were identified from the *D. capensis* genome by performing a BLAST search for DcChit_1, a chitinase fragment previously identified from genomic DNA of the same organism [51]. Their sequences range from 41%-100% identity to DcChit1_1. These fragments contain part of the N-terminal region, including the C-rich domain and the P-rich hinge, neither of which was observed in the original fragment, along with part of the catalytic domain (Figure 2.4). However, these sequences are all truncated before the catalytic residues. Sequencing of the *D. capensis* transcriptome will clarify whether these are fragments of active genes containing one or more introns, or inactive pseudogenes, which are relatively common in gene families undergoing rapid evolution [63] (as is the case for many proteins associated with pathogen defense) [64].

```
DcChitI_2       ---------- ---------- ---------- -MRITILLLL CVAPLLSGTY AVQCGSEVGG ALCPNGLCCS KYGYCGTTSA YCGPGCQSQC GGSSPPPAPP
I0CMI1_DROCA    ---------- ---------- ---------- ---------- ---------- ---------- ---------- ---------- ----------
DcChitI_1       ---------- ---------- ---------- -MRITILLLL CVAPLLSGTY AVQCGSEVGG ALCPNGLCCS KYGYCGTTSA YCGPGCQSQC GGSSPPPAPP
DcChitI_3       MKTRSIPEIS STAPIISFTL DHTIQTRKIM SPPMKSIHMI CLVAAVIIFL TMPRHLAAQS CGCAAGLCCS KYGYCGTTSD YCGDGCQAGP CSSTPA----
DcChitI_4       ---------- ---------- ---------M SPPMRNYHMT CLVTAVIIFF TMPGHLAAQS CGCAAGLCCS KYGYCGTTSS YCGDGCQAGP CSSTPT----

DcChitI_2       SPTPSPPSPS GGGDVSSIIT SQIFNQMLLH RNDNACPANG FYSYQAFLDA ARKFSGFGTT GDINTRKKEL AAFFGQTSHE TTG------- ----------
I0CMI1_DROCA    ------PSPS GGGDVSSIIT SQIFNQMLLH RNDNACPAHG FYSYQAFLDA ARKFSGFGTT GDINTRKKEL AAFFGQTSHE TT-------- ----------
DcChitI_1       SPTPSPPSPS GGGDVSSIIT SQIFNQMLLH RNDNACPAHG FYSYQAFLDA ARKFSGFGTT GDINTRKKEL AAFFGQTSHE TT-------- ----------
DcChitI_3       ---------G SGVSVPAVVT VAFF-NGIIN KAGSGCPGTG FYSRSAFLSA IGSYPSFGTT GTSDAAKREI AAFFAHVTHE TGCKHIHFFL SKFYAVLYRV
DcChitI_4       ---------S SGVSVPAVVT DAFF-NGIIN QAGSGCPGKG FYSRSAFLSA IGSYPSFGTT GTTDASKQEI AAFFAHVTHE T--------- ----------

DcChitI_2       ---------- ---
I0CMI1_DROCA    ---------- ---
DcChitI_1       ---------- ---
DcChitI_3       IILYAWIKDE AID
DcChitI_4       ---------- ---
```

Figure 2.4: Chitinase 1 fragments discovered using a BLAST search of the *D. capensis* genome against the DcChitI_1 fragment previously identified by Renner and Specht from *D. capensis* genomic DNA.

2.2.3 Preliminary Structural Models and *In silico* Maturation

Preliminary models for both Family 18 and Family 19 chitinases were produced using Rosetta [28], implemented in the online Robetta server [27]. The Rosetta structures contain the full sequences, including the N-terminal signal peptides, and in some cases, C-terminal targeting peptides that are also cleaved during maturation. These Rosetta models then underwent the *in silico* maturation process [48], and the process is illustrated in Figure 2.5 for a represen-

tative family 18 chitinase, DCAP_2209. The initial Rosetta sequence, including the signal peptide and lacking post-translational modifications, is shown in Figure 2.5. In order to generate the equilibrated structure Figure 2.5b, which more closely approximates the active form of the enzyme in solution, the signal sequence is removed, disulfide bonds are added using homology to a reference sequence (in this case CHIT3_VITVI), and the structure is equilibrated in explicit solvent. Many Family 18 chitinases from plants contain three disulfide bonds [65, 66], although examples without any disulfide bonds also exist [67]. Three are found in all the Family 18 chitinases in this set, as in CHIT3_VITVI [36], and hevamine from *Hevea brasiliensis* (PDB ID: 2HVM) [68]. The functionally important cis peptide bonds are captured by the molecular models for all the Family 18 chitinases examined here except for DCAP_7323, which unlikely to be active in any case because it is truncated at the N-terminal end.

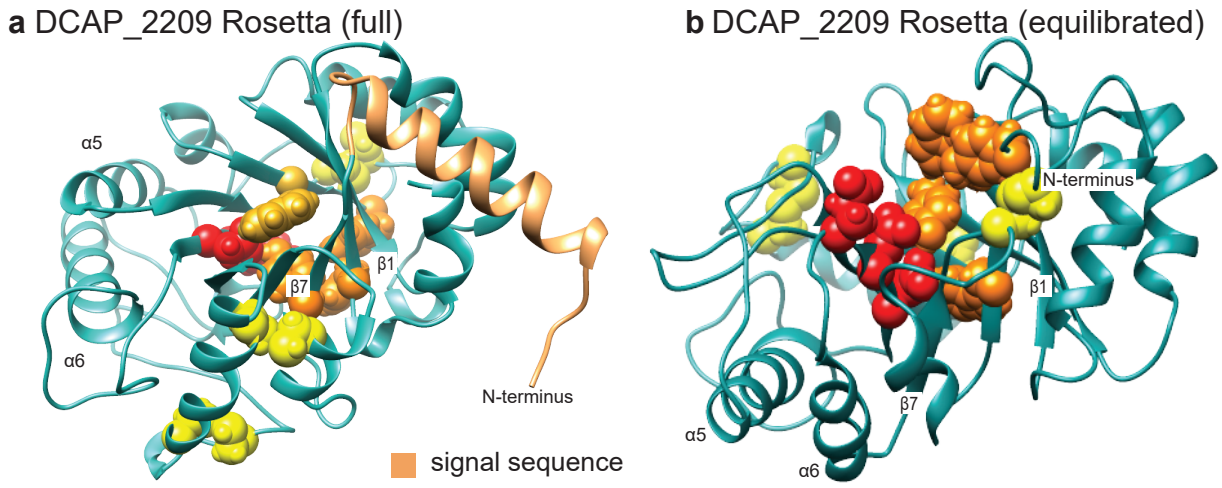

Figure 2.5: DCAP_2209 (a) before and (b) after *in silico* maturation. The light orange helix in part a is the N-terminal signal sequence. Important residues are color-coded as follows: Red: catalytically active residues of the "DXDXE" motif. Orange: aromatic substrate-binding residues. Yellow: Cysteines in disulfide bonds.

Figure 2.6 shows full-length structures for Q6IVX8_9CARY and Q6IVX2_9CARY (Family 19) from *Drosera spatulata*. The N-terminal and C-terminal targeting sequences are exposed on the surface of the protein, as expected. The P-rich hinge in these proteins is variable

in length, and highly flexible, as illustrated by the different relative conformations of of the catalytic and C-rich chitin binding domains observed here.

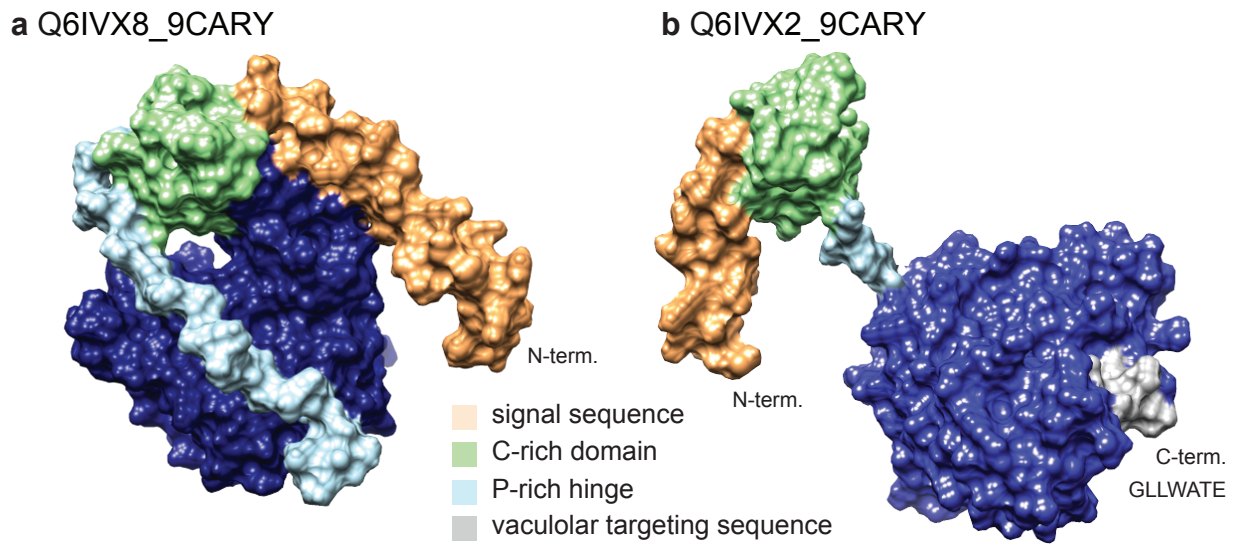

Figure 2.6: Initial Rosetta structures for two class I chitinases from *Drosera spatulata*, Q6IVX8_9CARY and Q6IVX2_9CARY, illustrating positioning of the N-terminal and C-terminal targeting sequences and the variability in length and conformation for the P-rich hinge.

All initial and equilibrated structures are available for download as PDB files from our paper online [1]. The available structures for Families 18 and 19 are tabulated in Tables 1 and 2, respectively.

Protein	Organism	Sequence Elements included	File Name
CHIT3_VITVI	*Vitis vinifera*	signal, active region	CHIT3_VITVI_m1.pdb
CHIT3_VITVI	*Vitis vinifera*	active region	CHIT3_VITVI_mature_m1.pdb
DCAP_7323	*D. capensis*	active region	DCAP_7323_m1.pdb
DCAP_7323	*D. capensis*	active region	DCAP_7323_mature_m1.pdb
DCAP_0106	*D. capensis*	signal, active region	DCAP_0106_m1.pdb
DCAP_0106	*D. capensis*	active region	DCAP_0106_mature_m1.pdb
DCAP_7544	*D. capensis*	signal, active region	DCAP_7544_m1.pdb
DCAP_7544	*D. capensis*	active region	DCAP_7544_mature_m1.pdb
DCAP_2209	*D. capensis*	signal, active region	DCAP_2209_m1.pdb
DCAP_2209	*D. capensis*	active region	DCAP_2209_mature_m1.pdb
C7F821_NEPMI	*N. mirabilis*	signal, active region	C7F821_NEPMI_m1.pdb
C7F821_NEPMI	*N. mirabilis*	active region	C7F821_NEPMI_mature_m1.pdb
C7F817_9CARY	*D. spatulata*	signal, active region	C7F817_9CARY_m1.pdb
C7F817_9CARY	*D. spatulata*	active region	C7F817_9CARY_mature_m1.pdb
I7HCY7_NEPAL	*N. alata*	signal, active region	I7HCY7_NEPAL_m1.pdb
I7HCY7_NEPAL	*N. alata*	active region	I7HCY7_NEPAL_mature_m1.pdb
C7F818_9CARY	*D. spatulata*	signal, active region	C7F818_9CARY_m1.pdb
C7F818_9CARY	*D. spatulata*	active region	C7F818_9CARY_mature_m1.pdb
Q06SN0_9CARY	*D. spatulata*	signal, active region	Q06SN0_9CARY_m1.pdb
Q06SN0_9CARY	*D. spatulata*	active region	Q06SN0_9CARY_mature_m1.pdb
C7F824_9CARY	*D. spatulata*	signal, active region	C7F824_9CARY_m1.pdb
C7F824_9CARY	*D. spatulata*	active region	C7F824_9CARY_mature_m1.pdb
C7F822_9CARY	*D. spatulata*	signal, active region	C7F822_9CARY_m1.pdb
C7F822_9CARY	*D. spatulata*	active region	C7F822_9CARY_mature_m1.pdb
C7F819_9CARY	*D. spatulata*	signal, active region	C7F819_9CARY_m1.pdb
C7F819_9CARY	*D. spatulata*	active region	C7F819_9CARY_mature_m1.pdb
C7F823_NEPGR	*N. gracilis*	signal, active region	C7F823_NEPGR_m1.pdb
C7F823_NEPGR	*N. gracilis*	active region	C7F823_NEPGR_mature_m1.pdb
DCAP_5455	*D. capensis*	signal, active region	DCAP_5455_m1.pdb
DCAP_5455	*D. capensis*	active region	DCAP_5455_mature_m1.pdb
DCAP_2879	*D. capensis*	signal, active region	DCAP_2879_m1.pdb
DCAP_2879	*D. capensis*	active region	DCAP_2879_mature_m1.pdb
DCAP_4799	*D. capensis*	signal, active region	DCAP_4799_m1.pdb
DCAP_4799	*D. capensis*	active region	DCAP_4799_mature_m1.pdb
DCAP_2737	*D. capensis*	signal, active region	DCAP_2737_m1.pdb
DCAP_2737	*D. capensis*	active region	DCAP_2737_mature_m1.pdb

Protein	Organism	Sequence Elements included	File Name
HORV2	*H. vulgare*	active region	HORV2 PDBID: 2BAA
HORV2	*H. vulgare*	active region	HORV2_crystal_struc_mature_m1.pdb
Q6IV09_DRORT	*D. rotundifolia*	active region	Q6IV09_DRORT_m1.pdb
Q6IV09_DRORT	*D. rotundifolia*	active region	Q6IV09_DRORT_mature_m1.pdb
CHI3_CASSA	*Castanea sativa*	C-rich domain, P-rich hinge, active region	CHI3_CASSA_m1.pdb
CHI3_CASSA	*Castanea sativa*	C-rich domain, P-rich hinge, active region	CHI3_CASSA_mature_m1.pdb
Q6IVX8_9CARY	*D. spatulata*	signal, C-rich domain, P-rich hinge, active region	Q6IVX8_9CARY_m1.pdb
Q6IVX8_9CARY	*D. spatulata*	C-rich domain, P-rich hinge, active region	Q6IVX8_9CARY_mature_m1.pdb
V5TEI0_DIOMU	*D. muscipula*	signal, C-rich domain, P-rich hinge, active region	V5TEI0_DIOMU_m1.pdb
V5TEI0_DIOMU	*D. muscipula*	C-rich domain, P-rich hinge, active region	V5TEI0_DIOMU_mature_m1.pdb
Q6DUJ9_DIOMU	*D. muscipula*	signal, C-rich domain, P-rich hinge, active region	Q6DUJ9_DIOMU_m1.pdb
Q6DUJ9_DIOMU	*D. muscipula*	C-rich domain, P-rich hinge, active region	6DUJ9_DIOMU_mature_m1.pdb
VJH3_9CARY	*D. spatulata*	signal, C-rich domain, P-rich hinge, active region	VJH3_9CARY_m1.pdb
VJH3_9CARY	*D. spatulata*	C-rich domain, P-rich hinge, active region	VJH3_9CARY_mature_m1.pdb
DCAP_5513	*D. capensis*	signal, C-rich domain, P-rich hinge, active region	DCAP_5513_m1.pdb
DCAP_5513	*D. capensis*	C-rich domain, P-rich hinge, active region	DCAP_5513_mature_m1.pdb
Q6DUKO_9CARY	*D. spatulata*	active region	Q6DUKO_9CARY_m1.pdb
Q6DUKO_9CARY	*D. spatulata*	active region	Q6DUKO_9CARY_mature_m1.pdb
DCAP_4817	*D. capensis*	signal, C-rich domain, P-rich hinge, active region	DCAP_4817_m1.pdb
DCAP_4817	*D. capensis*	C-rich domain, P-rich hinge, active region	DCAP_4817_mature_m1.pdb
CHI2_BRANA	*B. napus*	signal, C-rich domain, P-rich hinge, active region, CTE	CHI2_BRANA_m1.pdb
CHI2_BRANA	*B. napus*	C-rich domain, P-rich hinge, active region	CHI2_BRANA_mature_m1.pdb
Q6IV10_DRORT	*D. rotundifolia*	active region	Q6IV10_DRORT_m1.pdb
Q6IV10_DRORT	*D. rotundifolia*	active region	Q6IV10_DRORT_mature_m1.pdb
I0CMI2_DIOMU	*D. muscipula*	active region	I0CMI2_DIOMU_m1.pdb
I0CMI2_DIOMU	*D. muscipula*	active region	I0CMI2_DIOMU_mature_m1.pdb
I0CMI3_9CARY	*D. spatulata*	active region	I0CMI3_9CARY_m1.pdb
I0CMI3_9CARY	*D. spatulata*	active region	I0CMI3_9CARY_mature_m1.pdb
I0CMI4_9CARY	*D. spatulata*	active region	I0CMI4_9CARY_m1.pdb
I0CMI4_9CARY	*D. spatulata*	active region	I0CMI4_9CARY_mature_m1.pdb
I0CMI6_NEPMI	*N. mirabilis*	active region	I0CMI6_NEPMI_m1.pdb
I0CMI6_NEPMI	*N. mirabilis*	active region	I0CMI6_NEPMI_mature_m1.pdb
Q6IVX2_9CARY	*D. spatulata*	signal, C-rich domain, P-rich hinge, active region, CTE	Q6IVX2_9CARY_m1.pdb
Q6IVX2_9CARY	*D. spatulata*	C-rich domain, P-rich hinge, active region	Q6IVX2_9CARY_mature_m1.pdb
Q6IVX4_9CARY	*D. spatulata*	signal, C-rich domain, P-rich hinge, active region, CTE	Q6IVX4_9CARY_m1.pdb
Q6IVX4_9CARY	*D. spatulata*	C-rich domain, P-rich hinge, active region	Q6IVX4_9CARY_mature_m1.pdb
DCAP_0533	*D. capensis*	signal, C-rich domain, P-rich hinge, active region, C-terminal domain	DCAP_0533_m1.pdb
DCAP_0533	*D. capensis*	C-rich domain, P-rich hinge, active region, C-terminal domain	DCAP_0533_mature_m1.pdb
A9ZMK1_NEPAL	*N. alata*	signal, C-rich domain, P-rich hinge, active region	A9ZMK1_NEPAL_m1.pdb
A9ZMK1_NEPAL	*N. alata*	C-rich domain, P-rich hinge, active region	A9ZMK1_NEPAL_mature_m1.pdb

2.2.4 Network Modeling and Analysis

In collaboration with Prof. Carter T. Butts, we mapped each equilibrated protein structure to a protein structure network (PSN) as defined by the representation of [69] using software tools from [47]; these in turn make use of VMD [70] and the `statnet` toolkit [71, 72] within the R statistical computing system [73]. To compare PSNs, we use the structural distance approach of [74], which defines a metric on graph pairs that is in our case equal to the number of edges in one graph that would need to be altered in order to make it isomorphic to the other. (Isolate addition was performed when comparing graphs with differing numbers of vertices.) To remove size effects, the raw distance between each pair of PSNs was normalized by the number of vertices, yielding a metric corresponding to edge changes per vertex. These normalized structural distances were analyzed using hierarchical clustering using R. Additional network analysis and visualization was performed using the `network` and `sna` libraries within `statnet` [72, 75].

2.3 Results

This work reports molecular models and functional predictions from protein structure networks for eleven new chitinases from *D. capensis*, including a novel class IV chitinase with two active domains. This architecture has previously been observed in microorganisms but not in plants. This work used a combination of comparative and de novo structure prediction followed by molecular dynamics simulation to produce models of the mature forms of these proteins in aqueous solution. Protein structure network analysis of these and other plant chitinases reveal characteristic features of the two major chitinase families. Vy Duong worked mostly on Family 19 and my work was mostly concentrated on Family 18. It is important to understand the comparison of Family 19 and Family 18 and therefore, a portion of the results from the paper [21] are shown in this chapter.

2.3.1 *D. capensis* Chitinases are Predicted to Adopt Folds Consistent with Active Enzymes

The catalytic action of family 18 chitinases, which retains the β-anomeric carbon stereochemistry from the substrate to the product, is based on substrate-assisted hydrolysis of the glycosidic bond [76, 77, 78]. Catalysis is initiated by distorting the -1 sugar ring subsite adjacent to the glycosidic bond. Next, Asp 123 rotates to form hydrogen bonds with both Glu 127 and the N-acetyl group of the +1 sugar. This step protonates Glu 127. Then, the anomeric carbon is subjected to a nucleophilic attack by the oxygen from the N-acetyl group, forming an oxazolinium ion as an intermediate, followed by cleavage of the glycosidic bond by hydrolysis to generate smaller fragments. The DXDXE motif is essential for activity, hence fragments that were lacking this sequence due to truncation were excluded from the protein set.

Family 18 chitinases, which retain the β-anomeric carbon stereochemistry from the substrate to the product, adopt the $(\alpha\text{-}\beta)_8$ triosephosphateisomerase (TIM)-barrel fold [66, 76], shown for DCAP_0106 in Figure 2.7A. The active site (Figure 2.7B), consists of a characteristic DXXDXDXE motif [66, 76]. The "tunnel" containing the active site is shaped by an unusual structural feature, two non-proline cis peptide bonds that are highly conserved, although the particular residues involved are somewhat variable [68, 33]. The cis peptide bonds (shown in black in Figure 2.7C), are captured by the molecular models for all full-length Family 18 chitinases examined here. The shape of the tunnel and the surface formed by the aromatic rings opposite the catalytic D and E residues acts to guide the chitin polymer chains into the active site, leading to processive activity [79]. The ability of Family 18 chitinases to keep the strand that is currently being degraded from re-encountering solid substrate is thought to be a key determinant of their ability to hydrolyze crystalline polysaccharides [80].

The Family 19 chitinases, all of which are characterized by an anomeric inverting mechanism

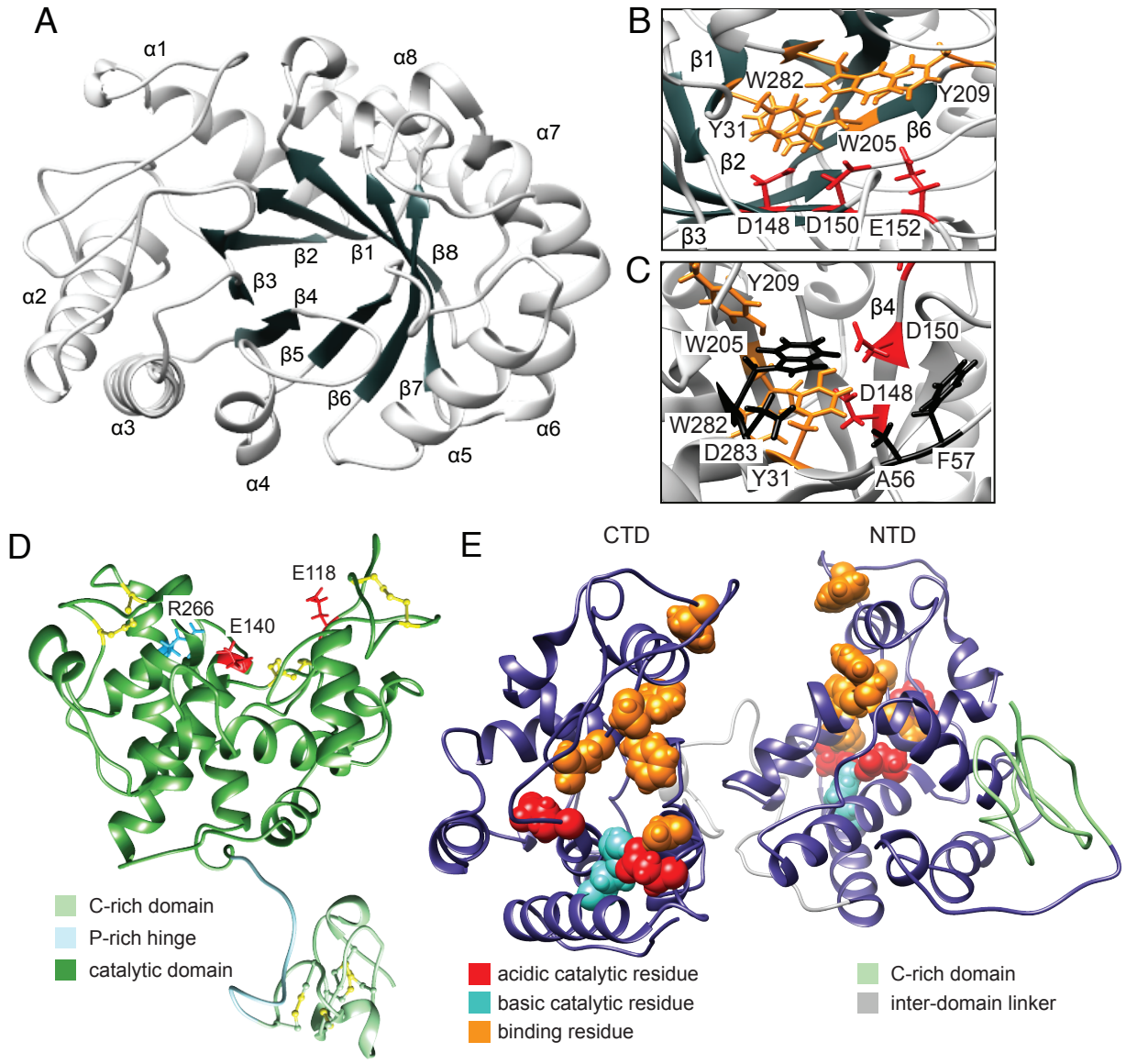

Figure 2.7: Equilibrated structures of the mature sequences of chitinases from carnivorous plants. A. DCAP_0106, a representative Family 18 chitinase, after *in silico* maturation. Numbering of secondary structure elements follows the convention of Si et al. [5]. B. Notably, the tunnel containing the active site has two surfaces with different chemical properties; the aromatic rings (orange) hold the more hydrophobic face of the chitin polymer in place, while the acidic residues (red) perform hydrolysis of the glycosidic linkages. C. Two conserved non-proline cis peptide bonds (black) are critical to shaping the active site tunnel in Family 18 chitinases. D. Chitinase VF-1 from *Dionaea muscipula* V5TEI0_DIOMU [6], with important sequence features and active site residues labeled (red: acidic active residue. blue: basic active residue. yellow: disulfide bond). E. The two-domain chitinase DCAP_0533. Color coding is as in D, with the addition of substrate-binding residues in orange.[1]

[61], have diverse structural features. Much of the structural and functional diversity results from two highly variable regions, the C-rich chitin-binding domain and the P-rich hinge [81, 82], each of which may vary in length or be absent altogether. We have identified two class I chitinases (DCAP_4817 and DCAP_5513) and one class IV chitinase (DCAP_0533) from the *D. capensis* genome. Most of the sequences in this set contain N-terminal secretion signals, however two *D. spatulata* sequences (Q6IVX2_9CARY and Q6IVX4_9CARY) and the reference sequence CHI2_BRANA contain short C-terminal extensions indicating targeting to the vacuole, consistent with their playing a purely defensive role. One sequence each from *D. capensis* (DCAP_5513), *D. rotundifolia* (Q6IV09_DRORT), and *D. spatulata* (Q6DUK0_9CARY) is missing one or more critical active site residues; in other organisms, enzymatically non-functional chitinase homologs are often present and can serve as chitin-binding proteins [83]. The predicted structure after *in silico* maturation for a representative chitinase, VF-1 from *D. muscipula* (Figure 2.7 is in good agreement overall with the homology model of Paszota et al. [6], with the active site residues positioned in a shallow cleft on the surface of the active domain. The two models do differ in the relative orientations of the domains; however examination of the other models in this set suggests that the P-rich hinge is highly flexible as seen in Figure 2.6.

2.3.2 The Novel Class IV Chitinase DCAP_0533 Has Two Functional Domains

This work identified a new class IV chitinase from *D. capensis*, DCAP_0533. A class IV chitinase has previously been described as one of the most abundant proteins in the pitcher fluid *N. alata* [45], where it preferentially hydrolyzes small GlcNAc oligomers over larger polymeric substrates [84]. Unlike other known plant chitinases, DCAP_0533 contains two class IV catalytic domains. The N-terminal domain appears to be fully active, while the C-terminal domain lacks one of the active residues but contains a full complement of substrate-

binding residues (Figure 2.7E, Figures 2.9, Figure 2.10). Multidomain chitinases containing dedicated substrate-binding domains have previously been observed in microbes [85]. For example, ChiA from the thermophilic archeon *Pyrococcus kodakaraensis*, has two chitinase domains and three catalytically inactive substrate binding domains, allowing separate optimizaton of substrate binding and catalytic function [86]. AFM data suggests the binding is mostly determined by interaction of the aromatic residues in the binding site (orange in Figure 2.7E) with the pyranose rings of the substrate [87]. This type of functionality has not been previously observed in plants; we hypothesize that it is an adaptation associated with carnivory, perhaps related to more effective breakdown of small oligosaccharides to components that can be used as a nitrogen source.

Structurally, each domain consists of two lobes with eight helices each, separated by a large active site cleft Figure 2.10(a). In Figure 2.10(b), the two domains of this protein are shown overlaid with the crystal structures of class IV chitinases from *Zea mays* (PDBID: 4MCK, 60% identity with the NTD) and *Picea abies* (PDBID: 3HBE, 64% identity with the CTD). The NTD Figure 2.10(c) has an N-terminal signal peptide, a conserved C-rich binding domain, and a catalytic domain that appears to be functional. In its homolog CHIA_MAIZE, Chaudet et. al. characterized four catalytic residues (E62, E71, E165, and R171) [88], all of which have counterparts in the NTD of DCAP_0533 (E173, E182, E278, R290) Figures 2.9 and 2.10. Previous modeling studies of well-characterized class I chitinases from barley, mustard, and chestnut seed homologs (barley: E67, mustard: E212, chestnut: E124) suggest the necessity of E62 in CHIA_MAIZE and E173 in the NTD of DCAP_0533 as a proton donor [89, 90, 91]. Overall, mutagenesis studies highlight the significance of E62 as an essential residue of the catalytic triad (E62, E165, R171 in CHIA_MAIZE) which we use to infer an equivalent catalytic triad in the NTD of DCAP_0533 (E173, E278, and R290). It has also been hypothesized that purpose of the triad is to alter the surrounding environment to induce activation of the glutamic acid in the HETG/I (class IV) or HETT (class I/II) motif by changing its pKa [91].

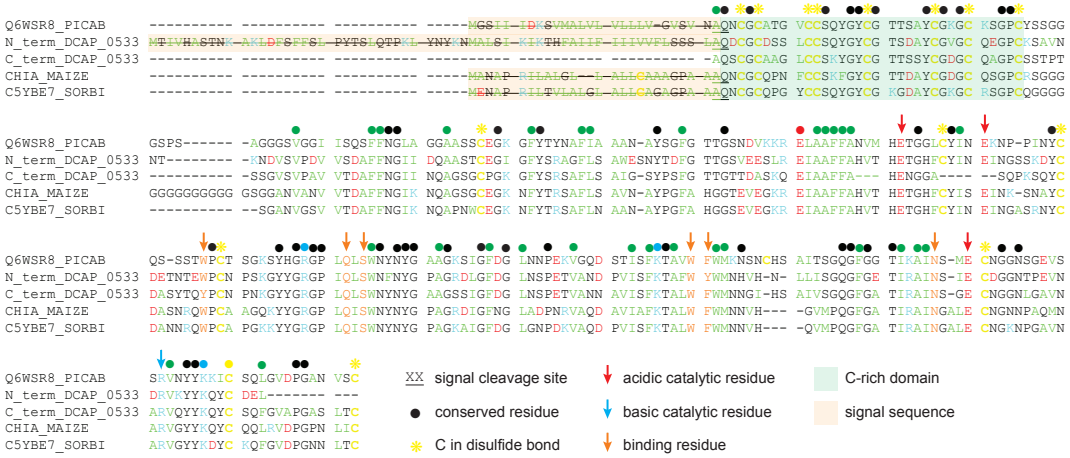

Figure 2.8: Sequence alignment and annotation of Q6WSR8_PICAB, CHIA_MAIZE, and the N-terminal domain (NTD) and C-terminal domain (CTD) of DCAP_0533. For the purpose of comparison, the sequence is manually separated above. We observe high sequence conservation regarding: the signal cleavage site, C-rich domain length and location, cysteines composing disulfide bonds, other binding site residues surrounding the main binding site residues (orange arrows), and catalytic residues except Glu407 of the CTD which is unaligned with Glu113 of Q6WSR8_PICAB

Class IV chitinases exhibit an amino acid substitution in the first active site region relative to Class I chitinases, resulting in a HETG/I motif instead of the HETT motif [51]. A deletion of four amino acids in the Cys-rich binding domain is also observed in class IV chitinases, as shown for a class IV chitinase from *Nepenthes alata* (A9ZMK1_NEPAL) [84] and DCAP_0533 in 2.3. Figure 2.9 shows a sequence alignment of the N- and C-terminal domains of the Class IV chitinase DCAP_0533 with single domain class IV chitinases from *Picea abies* (Q6WSR8_PICAB), *Zea mays* (CHIA_MAIZE), and *Sorghum bicolor* (C5YBE7_SORBI). The two domains of DCAP_0533 were aligned with the most closely related annotated class IV chitinases, those from *Picea abies* (EC: 3.2.1.14, Uniprot: Q6WSR8_PICAB), *Zea mays* (EC: 3.2.1.14, Uniprot: CHIA_MAIZE), and *Sorghum bicolor* (Uniprot: C5YBE7_SORBI) [92, 51, 88] (Figure 2.10).

Linked to the NTD by a cysteine and glycine-rich linker sequence, the CTD of DCAP_0533

(Figure 2.10(d) potentially houses a second catalytic domain or binding domain whose closest structural homolog is Q6WSR8_PICAB from Norway spruce (*Picea abies*) (Figure 2.9). Binding site residues and cysteines involved in disulfide bond formation are conserved in both chitinases. Comparing this sequence with the catalytic triad of Q6WSR8_PICAB (E113, R230, E218), we observe a potentially equivalent triad in the CTD (E407, E507, R519) (Figure 2.10). Ubhayasekera et. al. describe the flexibility of E113 and demonstrate two conformations that it can adopt during catalysis [92]. Although E407 is not located in the equivalent sequence position to E113, the flexibility of this residue in Q6WSR8_PICAB suggests that Glu407 may be at an appropriate distance to function as part of the CTD triad. Alternatively, the CTD may lack catalytic activity and act as a binding domain as in multidomain chitinases from archaea and bacteria.

2.3.3 Description of a Novel Two-Domain Class IV Chitinase

Class IV chitinases exhibit an amino acid substitution in the first active site region relative to Class I chitinases, resulting in a HETG/I motif instead of the HETT motif [51]. A deletion of four amino acids in the Cys-rich binding domain is also observed in class IV chitinases, as shown for a class IV chitinase from *Nepenthes alata* (A9ZMK1_NEPAL) [84] and DCAP_0533 in 2.3. Figure 2.9 shows a sequence alignment of the N- and C-terminal domains of the Class IV chitinase DCAP_0533 with single domain class IV chitinases from *Picea abies* (Q6WSR8_PICAB), *Zea mays* (CHIA_MAIZE), and *Sorghum bicolor* (C5YBE7_SORBI). The two domains of DCAP_0533 were aligned with the most closely related annotated class IV chitinases, those from *Picea abies* (EC: 3.2.1.14, Uniprot: Q6WSR8_PICAB), *Zea mays* (EC: 3.2.1.14, Uniprot: CHIA_MAIZE), and *Sorghum bicolor* (Uniprot: C5YBE7_SORBI) [92, 51, 88] (Figure 2.10).

Structurally, each domain consists of two lobes with eight helices each, separated by a large active site cleft (Figure 2.10(a). In Figure 2.10(b), the two domains of this protein are shown overlaid with the crystal structures of class IV chitinases from *Zea mays* (PDBID: 4MCK, 60% identity with the NTD) and *Picea abies* (PDBID: 3HBE, 64% identity with the CTD). The NTD Figure 2.10(c) has an N-terminal signal peptide, a conserved C-rich binding domain, and a catalytic domain that appears to be functional. In its homolog CHIA_MAIZE, Chaudet et. al. characterized four catalytic residues (E62, E71, E165, and R171) [88], all of which have counterparts in the NTD of DCAP_0533 (E173, E182, E278, R290) (Figures 2.9, 2.10. Previous modeling studies of well-characterized class I chitinases from barley, mustard, and chestnut seed homologs (barley: E67, mustard: E212, chestnut: E124) suggest the necessity of E62 in CHIA_MAIZE and E173 in the NTD of DCAP_0533 as a proton donor [89, 90, 91]. Overall, mutagenesis studies highlight the significance of E62 as an essential residue of the catalytic triad (E62, E165, R171 in CHIA_MAIZE) which we use to infer an equivalent catalytic triad in the NTD of DCAP_0533 (E173, E278, and R290). It has also been hypothesized that purpose of the triad is to alter the surrounding environment to induce activation of the glutamic acid in the HETG/I (class IV) or HETT (class I/II) motif by changing its pKa [91].

Linked to the NTD by a cysteine and glycine-rich linker sequence, the CTD of DCAP_0533 (Figure 2.10(d) potentially houses a second catalytic domain or binding domain whose closest structural homolog is Q6WSR8_PICAB from Norway spruce (*Picea abies*) (Figure 2.9). Binding site residues and cysteines involved in disulfide bond formation are conserved in both chitinases. Comparing this sequence with the catalytic triad of Q6WSR8_PICAB (E113, R230, E218), we observe a potentially equivalent triad in the CTD (E407, E507, R519) (Figure 2.10). Ubhayasekera et. al. describe the flexibility of E113 and demonstrate two conformations that it can adopt during catalysis [92]. Although E407 is not located in the equivalent sequence position to E113, the flexibility of this residue in Q6WSR8_PICAB

suggests that Glu407 may be at an appropriate distance to function as part of the CTD triad. Alternatively, the CTD may lack catalytic activity and act as a binding domain as in multidomain chitinases from archaea and bacteria.

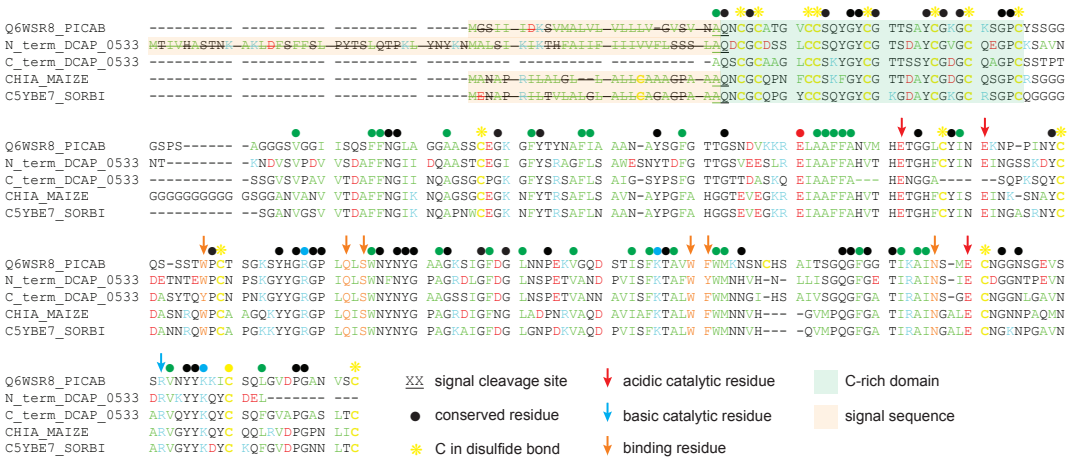

Figure 2.9: Sequence alignment and annotation of Q6WSR8_PICAB, CHIA_MAIZE, and the N-terminal domain (NTD) and C-terminal domain (CTD) of DCAP_0533. For the purpose of comparison, the sequence is manually separated above. We observe high sequence conservation regarding: the signal cleavage site, C-rich domain length and location, cysteines composing disulfide bonds, other binding site residues surrounding the main binding site residues (orange arrows), and catalytic residues except Glu407 of the CTD which is unaligned with Glu113 of Q6WSR8_PICAB

2.3.4 Network Analysis Shows Substantial Topological Differences by Family and within Proteins

When selecting potential targets for biophysical characterization, it is useful to consider general patterns of structural similarity or difference within and between families that may correlate with functional differences. Protein structure networks are useful for this purpose, as they directly encode the potential for direct physical interaction between functional groups (rather than representing detailed structure through properties such as side chain dihedral angles that can often vary substantially and dynamically without impacting protein function). In collaboration with Prof. Carter T. Butts, we employ the PSN representation of

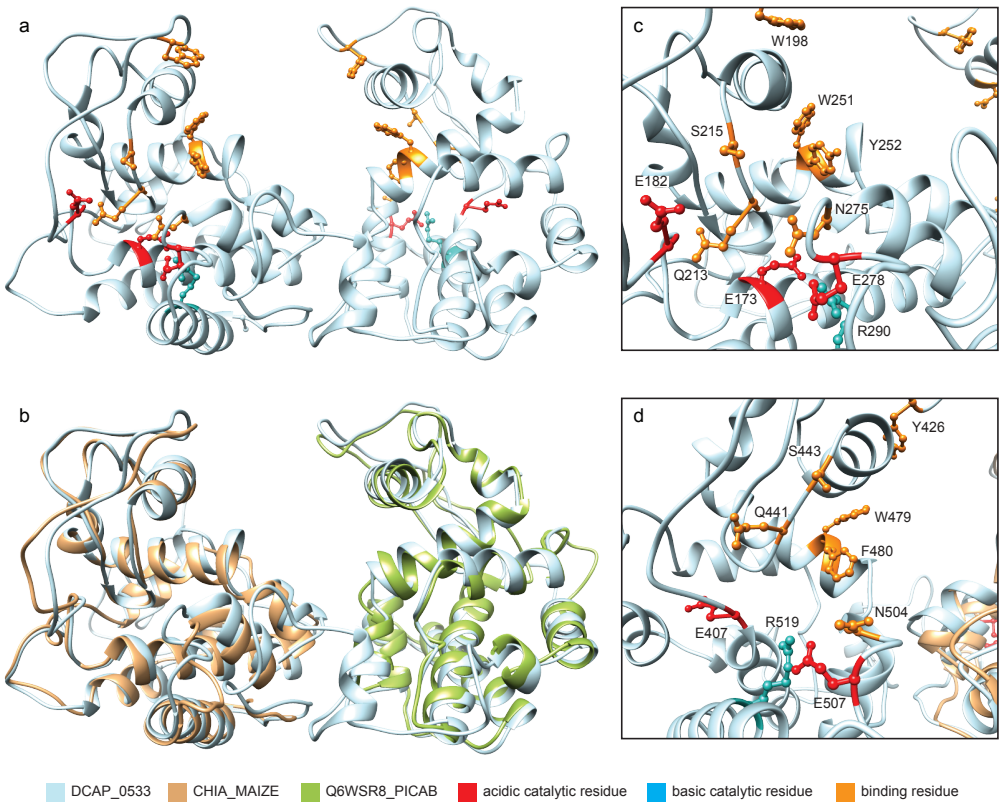

Figure 2.10: DCAP_0533 comparison with CHIA_MAIZE (4MCK) and Q6WSR8_PICAB (3HBE) and close up of catalytic residues and binding residues: (a) Robetta generated predicted structure with highlighted catalytic residues and binding residues. (b) Superimposition of CHIA_MAIZE and Q6WSR8_PICAB against DCAP_0533. (c) Catalytic site of NTD with 1-letter residue code and specifier. Catalytic triad consists of E173, E278, R290. (d) Catalytic site of CTD with 1-letter residue code and specifier. Catalytic triad consists of E407, E507, R519.

[69], where vertices represent small moieties and edges represent the potential for direct interaction (as determined by moiety-specific proximity constraints). Given two or more such PSNs, we may compare their topology by the structural distance method of [74], identifying the smallest number of edge changes (i.e. altered inter-moiety interactions) needed to make one PSN isomorphic to the other. Figure 2.11 depicts respective hierarchical clusterings of the Family 18 (panel A) and Family 19 (panel B) chitinases based on this notion of structural similarity, with distances normalized by the number of vertices to yield a metric with units of average changed interactions per moiety. For Family 18, the pattern of topological similarity is strikingly close to the pattern of sequence similarity, although somewhat more diversity

can be seen among structures than among sequences (compare with Figure 4.9). By contrast, topological clustering of Family 19 chitinases shows substantial differences from the equivalent sequence-based clustering. For instance, while DCAP_0533, A9ZMK1_NEPAL, and Q6IV09_DRORT belong to an outlying but internally cohesive cluster with respect to sequence similarity, the three show markedly different topologies (and, indeed, are split between the two large structural clusters characterizing the family). More broadly, we find that the Family 19 chitinases divide structurally into two primary clusters (rather than the four obtained from sequence similarity), both of which are internally heterogeneous and neither of which maps cleanly onto the clusters found by sequence similarity. The relationship between sequence and structure is thus much more tightly coupled for Family 18 than Family 19.

Further insight into the structural differences between the two families can be obtained by considering variation in the properties of their respective PSNs. Here, we examine four basic graph-level indices (GLIs) related to protein network organization. *Transitivity* [93] is defined as the fraction of (i, j, k) two-paths for which there exists an (i, k) edge, and is a standard measure of triadic closure; in the PSN context, higher levels of transitivity are associated with structures that are closely and uniformly packed, with few cavities or extended regions. *Degree* is defined as the number of edges incident on a given vertex; for a PSN, this corresponds to the number of other moieties with which a given chemical group is in contact. The standard deviation of the degree distribution within a PSN then provides a measure of the level of heterogeneity in local packing around chemical groups, and we employ it here as a second GLI. At a somewhat less local level, the (degree) *core number* of a given vertex [94] provides a measure of the extent to which that vertex is embedded in a region of high cohesion within the graph. More precisely, the k-th core (or k-core) of a graph is defined as the maximum set of vertices having at least k neighbors within the set. The core number of a vertex is then the number of the highest-order k-core to which it belongs. Although each k-core is not necessarily cohesive as a whole, cores with $k \geq 2$ are composed of *unions* of cohesive subgraphs, such that all vertices with high core numbers necessarily belong to highly

cohesive subgroups. In a PSN context, cohesive subgroups of moieties are joined by multiple, redundant paths and cannot be pulled apart without severing large numbers of edges. At the level of the entire PSN, then, the standard deviation of the core number serves as an indicator of the degree of heterogeneity in structural cohesion, and distinguishes between highly organized structures and structures that combine rigidly and loosely bound regions. Finally, we consider an indicator of the global path structure within the PSN, which we call *M-eccentricity*. The *eccentricity* of a vertex is the maximum geodesic distance from that vertex to any other vertex in the graph [95]; we here refer to the corresponding mean geodesic distance as the M-eccentricity. Vertices with high M-eccentricity are on average peripheral to the graph structure, while those with low M-eccentricity are relatively centrally located. At the level of the PSN as a whole, the standard deviation of the M-eccentricity distinguishes between uniformly globular structures and structures with deformations or other elongations, and we employ it as our fourth GLI.

Panel C of Figure 2.11 shows the distribution of the above GLI values for both chitinase families. All GLIs were calculated using the `sna` library [75]; to facilitate visualization, each GLI was standardized across the combined set of PSNs by subtracting the mean and dividing by the standard deviation prior to analysis. As is clear from Figure 2.11, the two families differ markedly on these four characteristics. On average, the Family 18 structures are substantially more homogeneous with respect to extended structure, local packing, and cohesion, while also being less transitive ($p < 0.001$ for all measures, two-tailed t-test). With respect to variation within family, the Family 18 structures show significantly less variability in eccentricity heterogeneity and transitivity (permutation test of logged IQR ratios, respective p values $< 1e-5$ and 0.007), but comparable variability with respect to heterogeneity in local packing and cohesion (respectively $p = 0.146$ and $p = 0.064$).

To provide an intuition for how these patterns play out in specific cases, Figure 2.12 shows vertex-level core numbers and M-eccentricity scores for the structures of CF821_NEPMI

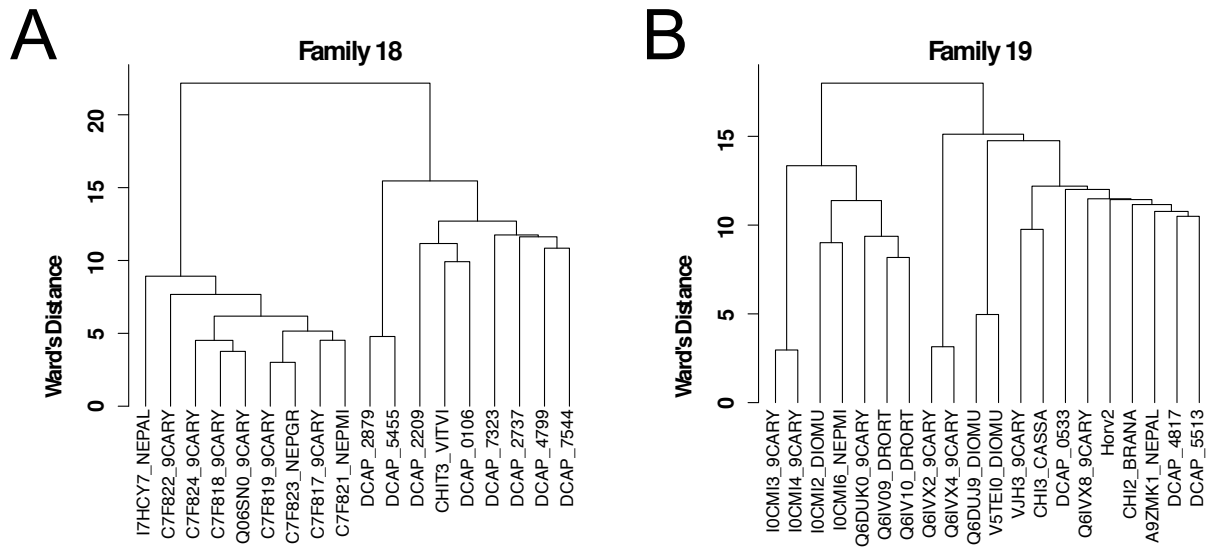

Figure 2.11: (a)-(b) Within-family clustering of chitinases by normalized structural distances. Ward's method (in the generalization of [4]) was employed to construct a hierarchical clustering of Family 18 (a) and Family 19 (b) chitinases based on topological dissimilarity. Sequence similarity is broadly recapitulated by the structural distances in Family 18, while Family 19 shows distinct patterns of variation.

(Family 18) and DCAP_5513 (Family 19). These structures were chosen by finding the PSN in each family with the smallest median distance to each other structure in the family, and are hence broadly representative of the classes in question. The core number visualizations of panels (a) and (b) clearly show that CF821_NEPMI is dominated be a large and uniformly cohesive core region, with few vertices in the outer region (i.e., lower cores). By contrast, the highly irregular structure of DCAP_5513 has numerous areas of low cohesion (including much of the C-rich domain) as well as the highly cohesive region associated with the central helices (compare with Figure 2.7). Differences in global structure are brought into sharp relief by the M-eccentricity visualizations of panels (c) and (d). The uniform and tightly connected topology of CF821_NEPMI results in a large number of vertices with short path distances to nearly all other chemical groups in the protein, and relatively little overall variation. Moieties in DCAP_5513, on the other hand, may be at an average distance of more than 9 steps from the rest of the protein, with large differences between the relatively central vertices in the

helical region and those in the outer portions of the C-rich domain or the P-rich hinge.

Taken together with the findings above and Prof. Carter T. Butt's PSN analysis of chitinases from our paper [1], findings suggest substantial structural differences in the basic organization of the Family 18 and Family 19 chitinases, with the former having more internally homogeneous structures, and with structural differences being more closely related to differences in sequence. Family 19 is on the whole more diverse, and contains members that are on average less internally homogeneous. The presence of a higher volume of low-cohesion regions in the Family 19 chitinases suggests that these enzymes may be more prone to thermal denaturation than those in Family 18 (since low-cohesion regions require fewer disrupted edges to pull apart), but may also have functional significance (e.g., by allowing enhanced flexibility). Such structural insights from PSN topology complement those gained by studying specific features, and are more easily extended to analyzing large numbers of sequences.

2.4 Conclusion

Modeling and analysis of Family 18 and 19 chitinases from *D. capensis* and several related species reveal a number of novel enzymes that present promising targets for subsequent expression and biophysical characterization. These include what is to our knowledge the first plant chitinase found with multiple active domains, as well as several proteins that differ in more conventional ways from others in their class. Comparative network analysis of these structures reveals within- and between-family differences in structural properties, with Family 18 chitinases tending to be substantially more homogeneous in internal structure and Family 19 chitinases showing variation in cohesion and packing with possible implications for both function and thermal stability. These results also demonstrate the potential of *in silico* pipelines to move rapidly from genomic DNA to predictions of tertiary structure and

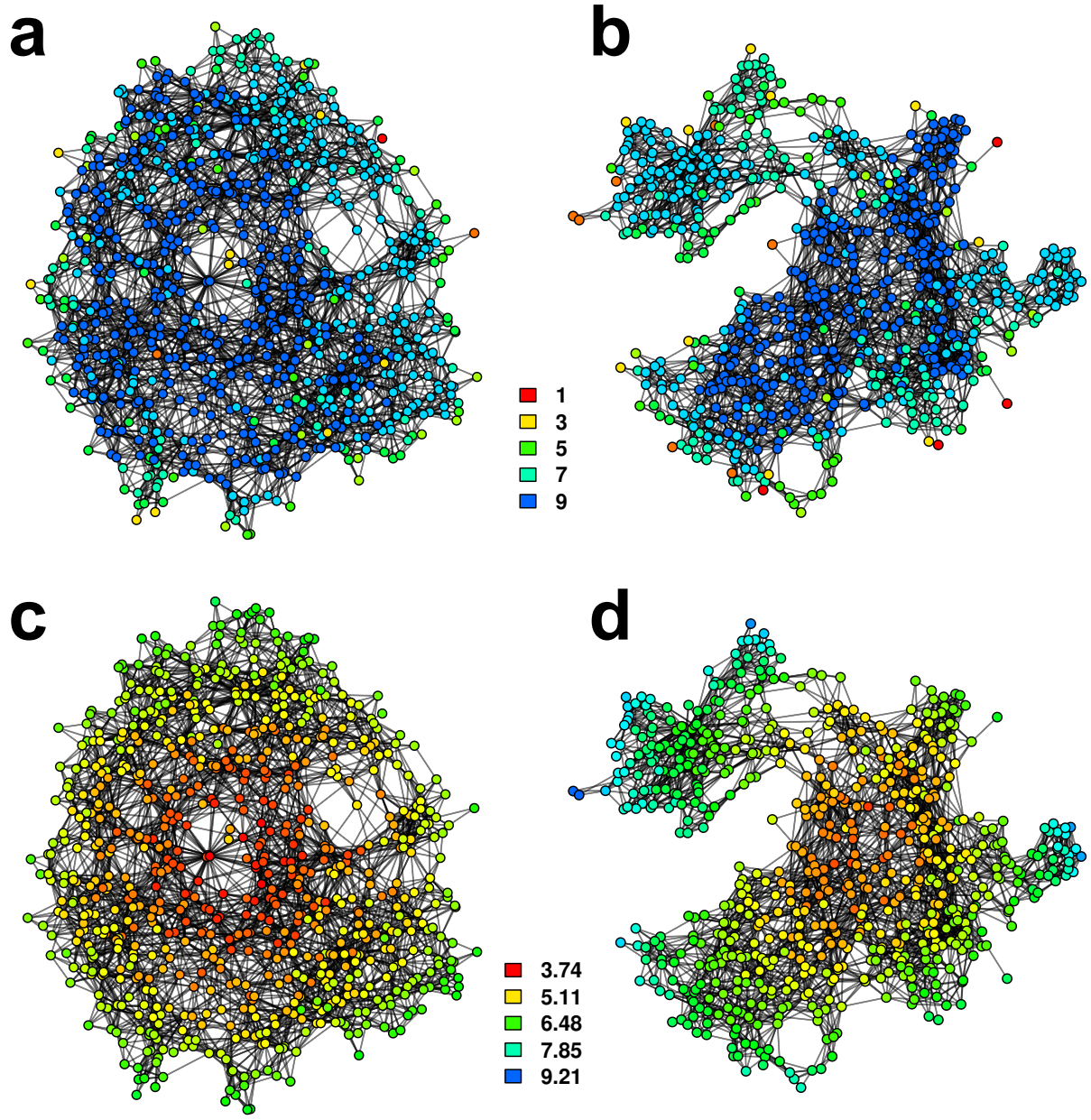

Figure 2.12: PSN Visualizations for family-representative structures C7F821_NEPMI (Family 18, (a) and (c)) and DCAP_5513 (Family 19, (b) and (d)). In panels (a) and (b), vertices are colored by k-core number; vertices with higher core numbers are embedded in more strongly cohesive local structures. Panels (c) and (d) show vertices by M-eccentricity (with higher values indicating a higher mean distance to other vertices in the network). The much higher level of internal heterogeneity in DCAP_5513 versus C7F821_NEPMI is immediately evident, with the former containing complex and irregular structure that subjects some vertices to higher levels of both cohesion and proximity than others.

comparative analysis thereof. As the "genomic revolution" makes such data available at an ever-increasing rate, such pipelines will become critical to our ability to exploit this scientific resource.

Chapter 3

Insights into esterase/lipases, phospholipases and nucleases from the carnivorous plant *Drosera capensis*

Drosera capensis represents a so far underexploited reservoir of novel enzymes with potentially useful activities. In this chapter I present the results from estarase/lipases, phospholipases and nucleases found in *Drosera capensis*.

3.1 Protein structure networks provide insight into active site flexibility in esterase/lipases from the carnivorous plant *Drosera capensis*

Esterase/lipases play important roles in plant defense, stress responses, and drought tolerance. Plant genomes and transcriptomes have provided a wealth of data about expression

patterns and the circumstances under which these enzymes are upregulated, e.g. pathogen defense and response to drought; however predicting the function of these enzymes from genomic or transcriptome data is challenging due to weak sequence conservation among the diverse members of this group. Functional blocks mediating enzyme activity have been identified; however progress to date has been hampered by the paucity of information on the structural relationships among these sequence regions and their relationships with substrate specificity and enzymatic activity. We have developed a methodology for efficient target selection based on molecular modeling and analysis of protein structure networks. Here this approach is demonstrated for 26 previously uncharacterized esterase/lipases from the genome of the carnivorous plant *Drosera capensis*. Analysis of the network relationships among functional blocks and among the chemical moieties making up the catalytic triad reveals potentially functionally significant differences that are not apparent from sequence analysis alone.

I worked on choosing the protein set, determining the functional regions of interest, generating the predicted structures, analyzing sequence and structure data, performing sequence annotation and comparisons and wrote the manuscript with my co-authors on this paper. A portion of the paper [2] is reproduced in this chapter to understand the importance of the results found in this study. For more details and in-depth analysis, please refer the paper [2].

3.1.1 Background

In land plants, tissues that are exposed to air are protected by the cuticle, a composite biomaterial comprising a cross-linked polyester scaffold interpenetrated by wax components [96]. The cuticle provides a barrier that minimizes water loss and protects the plant from pathogen infection. The relative quantities of hydrophilic and hydrophobic components

must be appropriately balanced and spatially located to adhere to the underlying cell walls while presenting a hydrophobic surface to the air interface [97]. Numerous enzymes are involved in producing the polymer components of this material, including esterases, lipases, and GDSL esterase/lipases. Esterase/lipases belong to the large α/β hydrolase enzyme superfamily, in which the catalytic triad consists of a nucleophile, an acid, and a stabilizing histidine (in this case Ser-Asp-His). This study focusses on the GDSL esterase/lipases, characterized by the proximity of the active serine residue to the N-terminus, as well as by its surrounding residues (canonically GDSL). Numerous plant GDSL esterase/lipases have been discovered from genome and transcriptome data [98, 99]; however their specific functions and substrate preferences remain relatively unexplored despite their potential commercial and technological importance. Much of what is known about the specific enzymatic activities of proteins in this family comes from studies of crop plants that produce large fruits. [100]. Esterase/lipases present attractive targets for biotechnology applications because of their potential for producing robust yet ultimately biodegradable polyester materials and hydrophobic surface coatings.

This study presents molecular modeling and functional analyses of 26 esterase/lipases recently discovered from the genome of the Cape sundew (*Drosera capensis*) [47]. The conservation of active site residues, key functional sequence blocks, and overall protein folds suggests that many of the *D. capensis* esterase/lipase sequences form functional enzymes; however the diversity of sequence and structural features indicates a range of potential molecular targets and enzymatic activities. The study uses sequence analysis, comparative modeling with all-atom refinement followed by *in silico* maturation and comparison of protein structure networks to identify distinct subgroups of proteins as a first step toward target selection for subsequent expression and biochemical characterization. To enable analysis of structural features with potential functional relevance, the study defines two novel types of *functionally-targeted protein structure networks* (FT-PSNs) generated using functional information specific to this protein class. Clustering of FT-PSNs based on connectivity among

functional blocks reveals several classes with distinct structural characteristics, which we hypothesize are related to enzyme flexibility and activity. Network connectivity among functional sequence blocks acts as a useful descriptor of protein structure and a predictor of global flexibility, while FT-PSNs based on connectivity among chemical moieties in the neighborhood of the active site are used to construct measures hypothesized to correlate with active site flexibility and hence enzyme promiscuity. Comparison with well-characterized reference sequences suggests that most of the *D. capensis* esterase/lipases have relatively rigid active sites, consistent with the specific functionality of the tomato GDSL1 enzyme, the best-characterized member of this class so far.

3.1.2 Methods

Sequence Alignments

Sequence alignments for the esterase/lipases from *D. capensis* are shown along with annotation reference sequences from other plants. Cluster 1 (Figure 3.1) contains enzymes with the traditional GDSL motif, including GDL1_CARPA from *Carica papaya*. Cluster 2 (Figure 3.2) contains only sequences from *D. capensis*, while Cluster 3 (Figure 3.3) contains two reference sequences from *Arabidopsis thaliana*. Cluster 4 is split into two figures for legibility (Figures 3.4 and 3.5). The alignment figures are annotated the same as chitinases (Chapter 2). Additionally, strikethrough text indicates sequence regions that are absent in the active enzyme, in this case the N-terminal signal peptide that is expressed but removed during maturation. Functional blocks I-IV are highlighted with colored boxes. Annotations were performed by homology to the annotations reference sequences from *C. papaya* and *A. thaliana* found in the UniProt database and identified by their UniProt IDs.

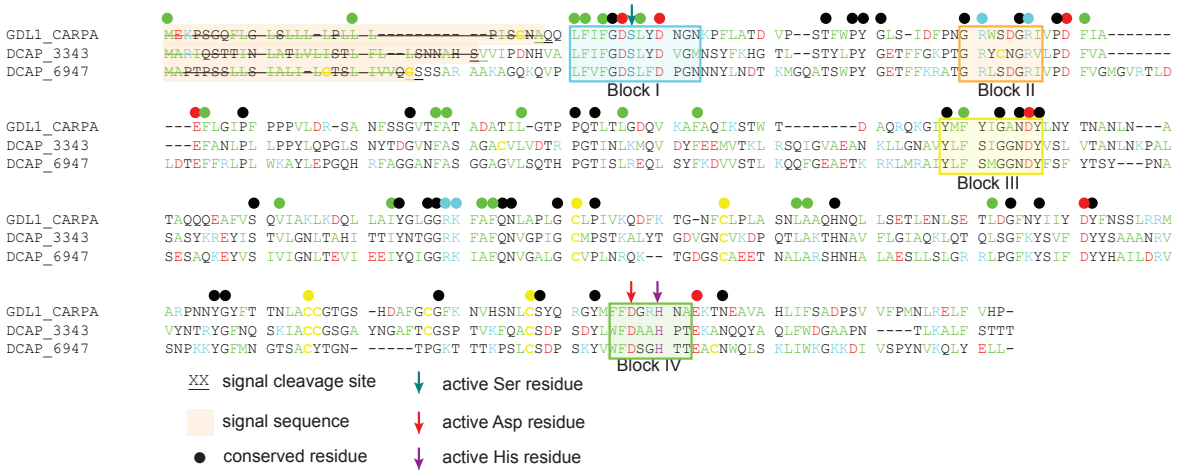

Figure 3.1: Sequence alignment for Cluster 1 esterase/lipases, annotated by homology to the reference sequence GDL1_CARPA. The four functional blocks that are critical for enzyme function are highlighted using outlined colored boxes. The N-terminal signal peptide is highlighted in light orange. Colored arrows indicate the catalytic triad residues. Conserved residues are marked using colored dots: acidic (red), basic (blue), hydrophobic (green), and hydrophilic (black) residues.

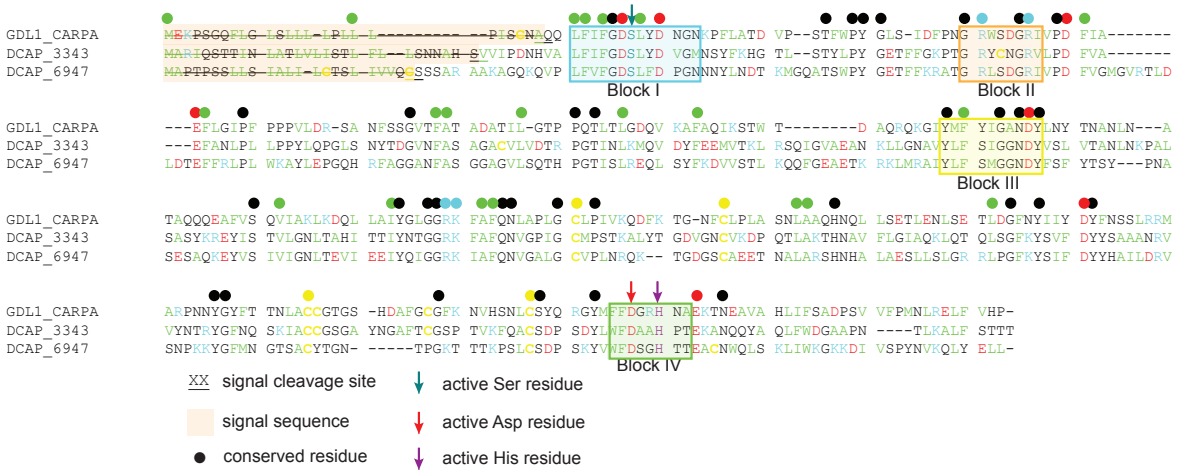

Figure 3.2: Sequence alignment and annotation for Cluster 2. The four block regions are determined by sequence conservation and outlined with colored boxes. Three *D. capensis* esterase/lipases contain the N-terminal signal sequence (highlighted in light orange) and three lack it. The catalytic triad is indicated using colored arrows. Colored dots denote conserved residues.

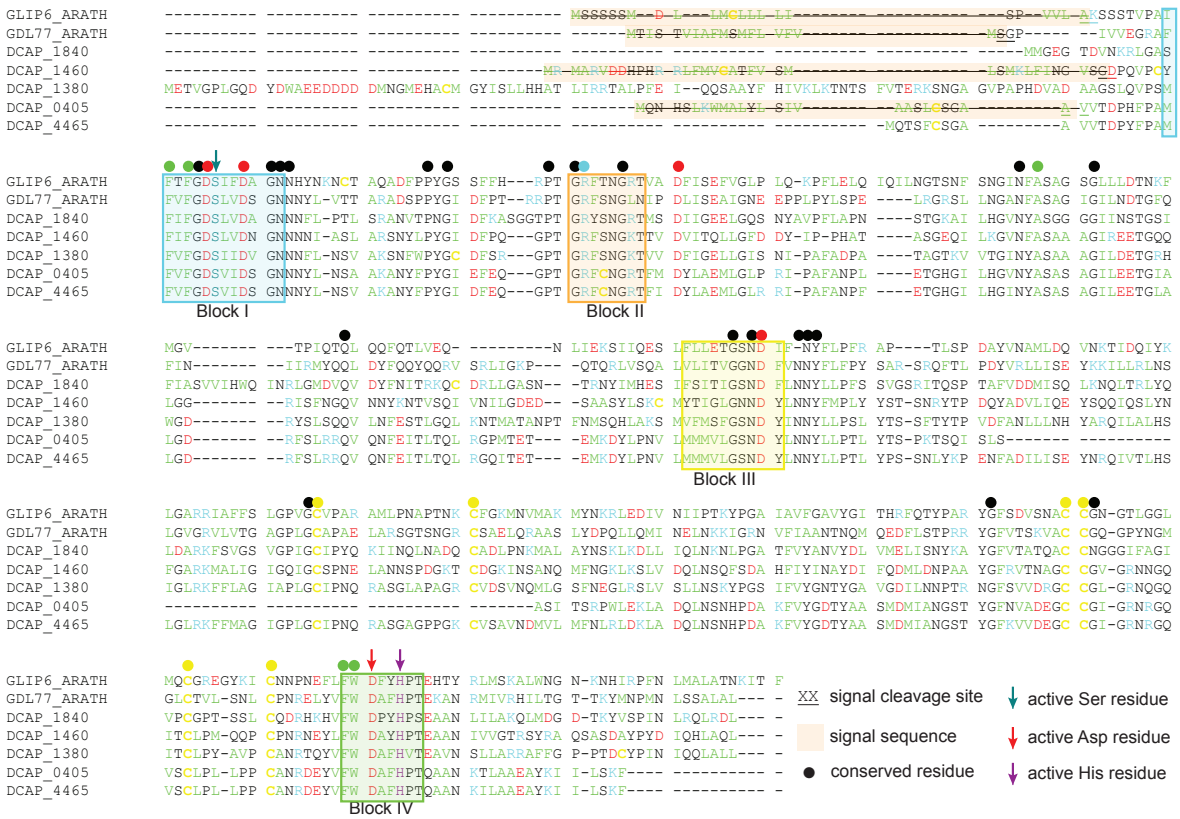

Figure 3.3: Sequence alignment and annotation for Cluster 3. Reference sequences are GLIP6_ARATH and GDL77_ARATH. All but three Cluster 3 esterase/lipases contain a N-terminal signal peptide (highlighted in light orange). Functional block regions are outlined using colored boxes. Colored dots indicate conserved residues.

Preliminary Structural Models and *In Silico* Maturation

The study performs *in silico* maturation, which I have previously described for chitinases [21], for each protein. Figure 3.6A shows the workflow of the overall enzyme discovery process as described in Chapter 1. Panels (B) and (C) show an example of a Cluster 2 esterase/lipase, DCAP_8086, before (B) and after (C) the *in silico* maturation process. Further comparison of a Cluster 3 esterase/lipase (DCAP_1460) to Cluster 4 enzymes and a cutin synthase from *Solanum lycopersicum* (tomato), G1DEX3_SOLLC, is shown in Figure 3.7. Functional sequence blocks DCAP_1460 and G1DEX3_SOLLC are highlighted by color (Figure S3.7). DCAP_4076, has an additional C-terminal domain. A PSI-BLAST search for the sequence of this domain indicated that it is related to the negative regulator of systemic acquired

Figure 3.4: Sequence alignment and annotation of Cluster 4a (first set), annotated by homology to EXL3_ARATH. Cluster 4 is separated into two parts (4a and 4b) for clarity. Block regions I-IV are shown in colored boxes with active site residues marked by colored arrows. Colored dots indicate conserved residues. When present, the N-terminal signal peptide is highlighted in light orange.

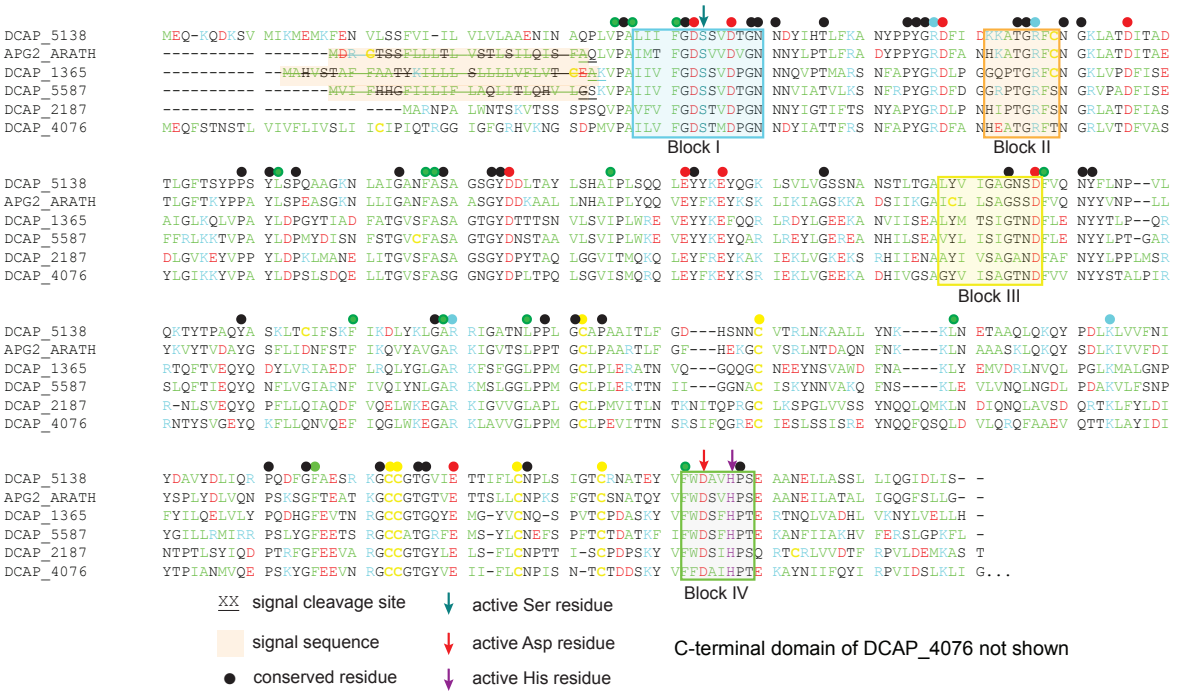

Figure 3.5: Sequence alignment and annotation of Cluster 4b (second set), annotated by homology to APG2_ARATH. Cluster 4 is separated into two parts (4a and 4b) for clarity. Block regions I-IV are shown in colored boxes with active site residues marked by colored arrows. Colored dots indicate conserved residues. When present, the N-terminal signal peptide is highlighted in light orange. DCAP_4076 has an additional C-terminal domain (shown in Figure S3.8).

resistance proteins previously discovered in other plants [2], with approximately 36% sequence identity to the SNI1 proteins from *Arabidopsis thaliana* (Uniprot ID: SNI1_ARATH) and *Glycine max* (Uniprot ID: Q0ZFU8_SOYBN). The *Arabidopsis* protein negatively regulates DNA recombination and gene expression during short-term stress responses. It has been suggested that SNI1_ARATH provides a scaffold for other proteins involved in regulation of transcription to bind; it is possible that this domain is playing a similar role here. DCAP_4076 lacks the N-terminal secretion signal common to many of the esterase/lipases, suggesting an intracellular function (Figure 3.8). The above results were mostly collected by Vy, Prof. Martin and Prof. Butts and are reproduced here to illustrate the importance of estarase/lipases.

The template structures used by Rosetta to calculate the predicted structures for a representative esterase / lipase, DCAP_0434, are tabulated in Table 1.

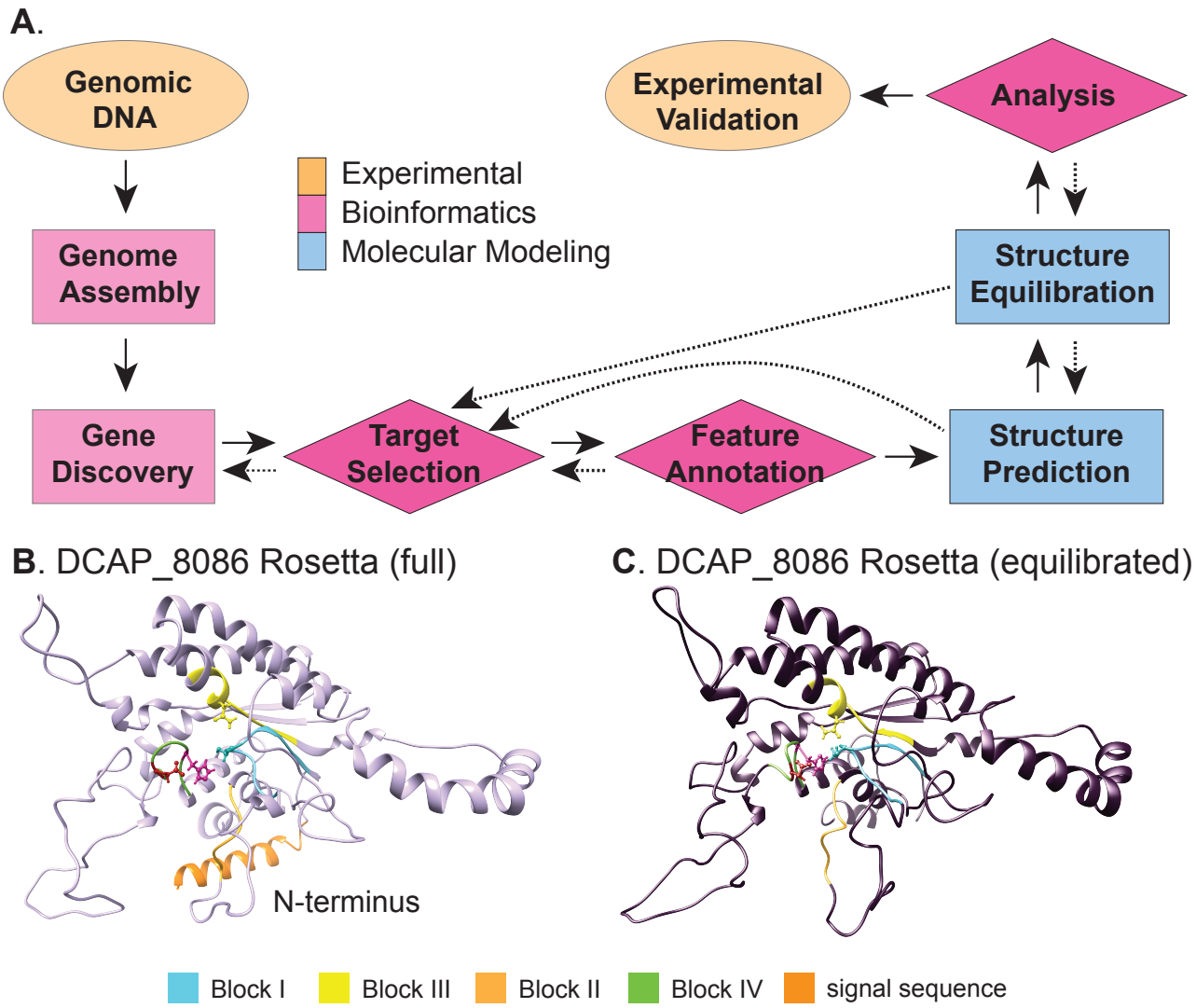

Figure 3.6: (A) Flow chart, made by me, illustrating the overall strategy for identifying enzymatic targets from genomic DNA. The workflow is indicated with solid arrows, while dotted arrows represent steps where information from a later stage of the pipeline enables refinement of earlier stages in an iterative manner. After genome sequencing, assembly, and gene discovery, target proteins are identified based on putative enzymatic activity. Functional sequence features are identified by analogy to annotation reference sequences found in the UniProt database. Structures are predicted using the Rosetta software, and equilibrated in explicit solvent after removal of sequence regions not present in the mature enzyme. Structures are compared using network analytic methods, enabling strategic selection of enzymes for experimental characterization in a future study. (B) DCAP_8086 before and (C) after *in silico* maturation. The light orange helix in part A is the N-terminal signal sequence, which is cleaved upon maturation. Important residues are color-coded as follows: dark cyan (catalytically active serine), red (active site aspartic acid), purple (active site histidine).

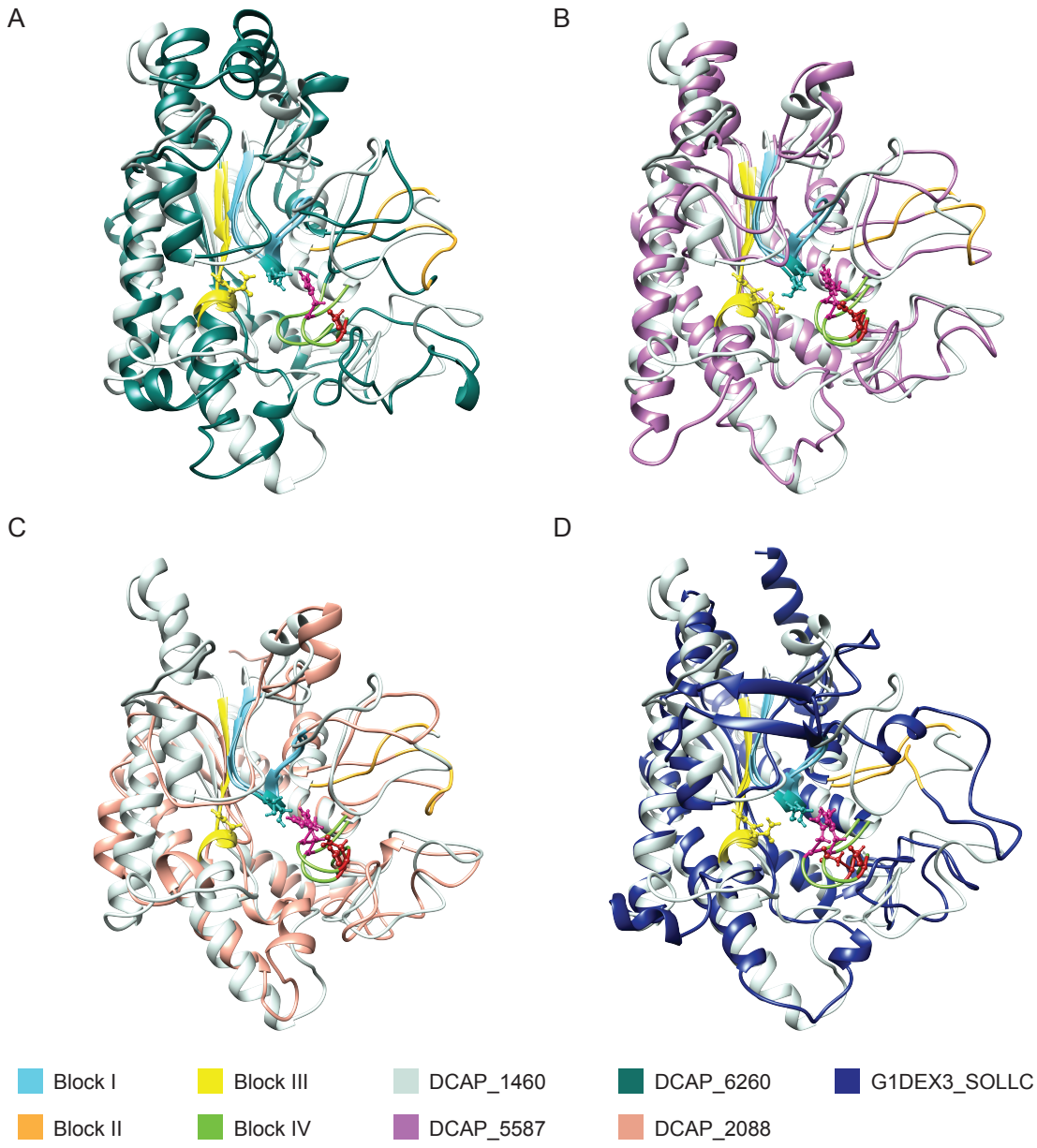

Figure 3.7: Comparison of DCAP_1460 (Cluster 3) to *D. capensis* esterase/lipases from each of the other clusters. These pairwise alignments of structural models provide an indication of the type and magnitude of structural differences between clusters: in general, the overall fold and secondary structural elements is conserved, although considerable variation can be observed in their relative positions and the conformations of loops and termini. Alignment was performed using the matchmaker feature of Chimera with default settings. Functional block regions I-IV are colored accordingly while the catalytic triad (Ser-His-Asp) residues are colored dark cyan, red, and purple. Active site residues are located in block I and IV, binding residues in block II-III. A. Comparison of DCAP_1460 to esterase/lipase DCAP_6260 (Cluster 4a). B. Comparison of DCAP_1460 to DCAP_5587 (Cluster 4b). C. Comparison of DCAP_1460 to DCAP_2088 (Cluster 4a). D. Comparison of DCAP_1460 to model esterase/lipase, G1DEX3_SOLLC, from *Solanum lycopersicum* (tomato).

Table 1: Rosetta structures for esterase / lipases (PDB files available for download)

Protein	Organism	Sequence Elements included	File Name
GDL1_CARPA	*Carica papaya*	signal, active region	GDL1_CARPA_m1.pdb
DCAP_3343	*D. capensis*	signal, active region	DCAP_3343_m1.pdb
DCAP_6947	*D. capensis*	signal, active region	DCAP_6947_m1.pdb
DCAP_0448	*D. capensis*	signal, active region	DCAP_0448_m1.pdb
DCAP_8086	*D. capensis*	signal, active region	DCAP_8086_m1.pdb
DCAP_0434	*D. capensis*	active region	DCAP_0434_m1.pdb
DCAP_4098	*D. capensis*	active region	DCAP_4098_m1.pdb
DCAP_5529	*D. capensis*	signal, active region	DCAP_5529_m1.pdb
DCAP_5165	*D. capensis*	active region	DCAP_5165_m1.pdb
GLIP6_ARATH	*A. thaliana*	signal, active region	GLIP6_ARATH_m1.pdb
GDL77_ARATH	*A. thaliana*	signal, active region	GDL77_ARATH_m1.pdb
DCAP_1840	*D. capensis*	active region	DCAP_1840_m1.pdb
DCAP_1460	*D. capensis*	signal, active region	DCAP_1460_m1.pdb
DCAP_1380	*D. capensis*	active region	DCAP_1380_m1.pdb
DCAP_0405	*D. capensis*	signal, active region	DCAP_0405_m1.pdb
DCAP_4465	*D. capensis*	active region	DCAP_4465_m1.pdb
DCAP_6218	*D. capensis*	active region	DCAP_6218_m1.pdb
DCAP_6260	*D. capensis*	active region	DCAP_6260_m1.pdb
EXL3_ARATH	*A. thaliana*	signal, active region	EXL3_ARATH _m1.pdb
DCAP_1761	*D. capensis*	active region	DCAP_1761_m1.pdb
DCAP_6217	*D. capensis*	signal, active region	DCAP_6217_m1.pdb
DCAP_5461	*D. capensis*	signal, active region	DCAP_5461_m1.pdb
DCAP_0158	*D. capensis*	signal, active region	DCAP_0158_m1.pdb
DCAP_2088	*D. capensis*	active region	DCAP_2088_m1.pdb
DCAP_2089	*D. capensis*	active region	DCAP_2089_m1.pdb
DCAP_5138	*D. capensis*	active region	DCAP_5138_m1.pdb
APG2_ARATH	*A. thaliana*	signal, active region	APG2_ARATH _m1.pdb
DCAP_1365	*D. capensis*	signal, active region	DCAP_1365_m1.pdb
DCAP_5587	*D. capensis*	signal, active region	DCAP_5587_m1.pdb
DCAP_2187	*D. capensis*	active region	DCAP_2187_m1.pdb
DCAP_4076	*D. capensis*	active region	DCAP_4076_m1.pdb

All initial and equilibrated structures available for download as PDB files are tabulated in Tables 1 and 2, respectively.

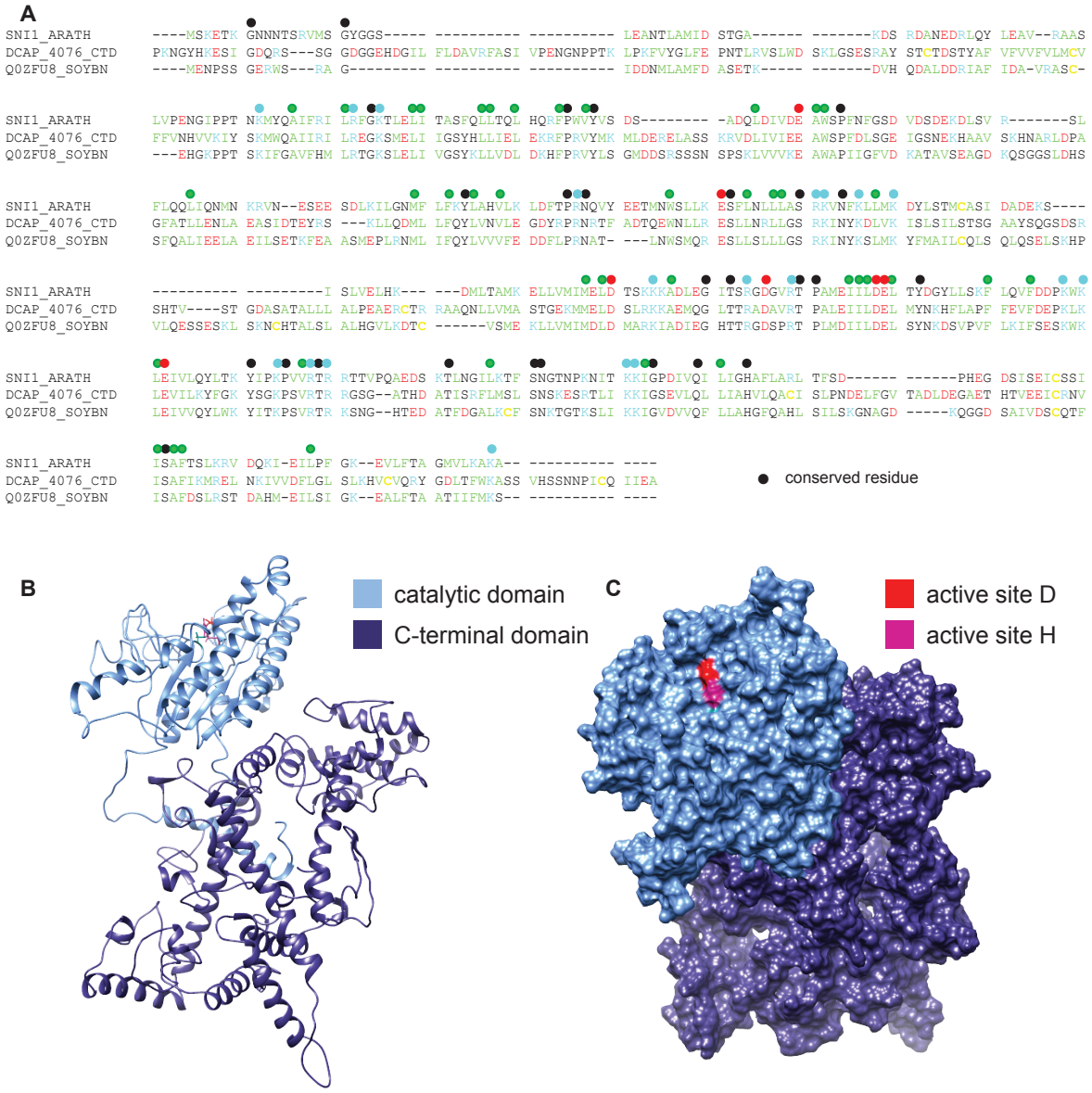

Figure 3.8: A. Sequence alignment of the C-terminal domain of DCAP_4076 with the SNI1 proteins from *Arabidopsis thaliana* (Uniprot ID: SNI1_ARATH) and *Glycine max* (Uniprot ID: Q0ZFU8_SOYBN). B. Ribbon structure of DCAP_4076, with the catalytic domain in light blue and the C-terminal domain in dark blue. C. Structural model of DCAP_4076 showing the surface representation. The active site D (red) and H (magenta) residues are visible at the top of the model.

Table 2: Mature structures for esterase / lipases (PDB files available for download)

Protein	Organism	Sequence Elements included	File Name
GDL1_CARPA	*Carica papaya*	active region	GDL1_CARPA_mature_m1.pdb
DCAP_3343	*D. capensis*	active region	DCAP_3343_mature_m1.pdb
DCAP_6947	*D. capensis*	active region	DCAP_6947_mature_m1.pdb
DCAP_0448	*D. capensis*	active region	DCAP_0448_mature_m1.pdb
DCAP_8086	*D. capensis*	active region	DCAP_8086_mature_m1.pdb
DCAP_0434	*D. capensis*	active region	DCAP_0434_mature_m1.pdb
DCAP_4098	*D. capensis*	active region	DCAP_4098_mature_m1.pdb
DCAP_5529	*D. capensis*	active region	DCAP_5529_mature_m1.pdb
DCAP_5165	*D. capensis*	active region	DCAP_5165_mature_m1.pdb
GLIP6_ARATH	*A. thaliana*	active region	GLIP6_ARATH_mature_m1.pdb
GDL77_ARATH	*A. thaliana*	active region	GDL77_ARATH_mature_m1.pdb
DCAP_1840	*D. capensis*	active region	DCAP_1840_mature_m1.pdb
DCAP_1460	*D. capensis*	active region	DCAP_1460_mature_m1.pdb
DCAP_1380	*D. capensis*	active region	DCAP_1380_mature_m1.pdb
DCAP_0405	*D. capensis*	active region	DCAP_0405_mature_m1.pdb
DCAP_4465	*D. capensis*	active region	DCAP_4465_mature_m1.pdb
DCAP_6218	*D. capensis*	active region	DCAP_6218_mature_m1.pdb
DCAP_6260	*D. capensis*	active region	DCAP_6260_mature_m1.pdb
EXL3_ARATH	*A. thaliana*	active region	EXL3_ARATH _mature_m1.pdb
DCAP_1761	*D. capensis*	active region	DCAP_1761_mature_m1.pdb
DCAP_6217	*D. capensis*	active region	DCAP_6217_mature_m1.pdb
DCAP_5461	*D. capensis*	active region	DCAP_5461_mature_m1.pdb
DCAP_0158	*D. capensis*	active region	DCAP_0158_mature_m1.pdb
DCAP_2088	*D. capensis*	active region	DCAP_2088_mature_m1.pdb
DCAP_2089	*D. capensis*	active region	DCAP_2089_mature_m1.pdb
DCAP_5138	*D. capensis*	active region	DCAP_5138_mature_m1.pdb
APG2_ARATH	*A. thaliana*	active region	APG2_ARATH _mature_m1.pdb
DCAP_1365	*D. capensis*	active region	DCAP_1365_mature_m1.pdb
DCAP_5587	*D. capensis*	active region	DCAP_5587_mature_m1.pdb
DCAP_2187	*D. capensis*	active region	DCAP_2187_mature_m1.pdb
DCAP_4076	*D. capensis*	active region	DCAP_4076_mature_m1.pdb

Network Modeling and Analysis

A protein structure network for each protein was calculated as described in chaitinases (Chapter 2) by Dr. Carter Butts. These structures were then secondarily processed to construct functionally targeted PSNs (FT-PSNs) using the `sna` library [75] within `statnet`. For further details, please look at the paper [2].

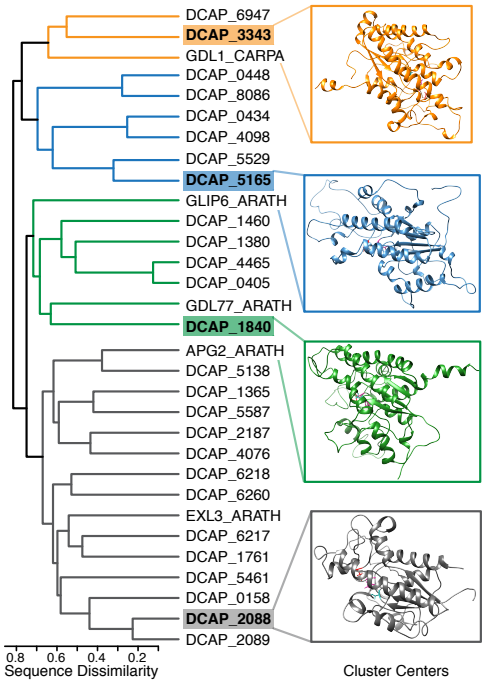

Figure 3.9: Clustering of esterase/lipase sequences identified from the *D. capensis* genome along with reference sequences from other plants.

3.1.3 Results and Discussion

D. capensis Esterase/Lipases Cluster Into Distinct Subfamilies Based on Sequence Features

Cluster 1 (Figure 4.9) contains sequences that have the canonical GDSL motif, as found in the reference sequence GDL1_CARPA, which was isolated from papaya latex [101] and has been proposed as a "naturally immobilized" biocatalyst for performing regioselective esterification and transesterification reactions [102]. The enzymes in cluster 2 instead have GDSN in the first functional block. Clusters 3 and 4 contain the motif GDSX, where X is usually a hydrophobic residue, but is Ser or Thr in some cases. Overall, the presence of the three active site residues in 24 of the 25 *D. capensis* esterase/lipases suggests they are functionally active enzymes.

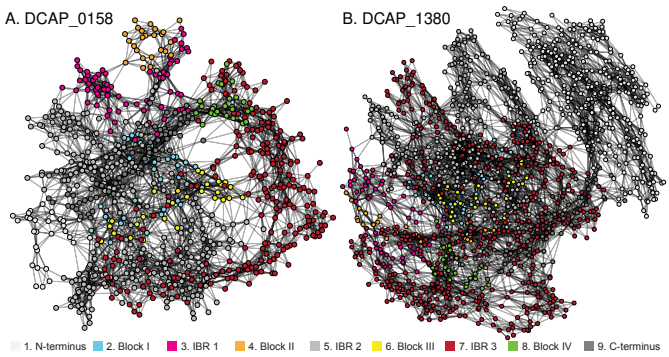

Figure 3.10: A. The sequences of the four conserved blocks. The sizes of the residue labels correlate with the fraction of sequences in the cluster having that residue in the indicated position. Amino acid properties are color coded as follows: hydrophobic-green, positive-blue, negative-red, cysteine-yellow, other-black. B. A representative molecular model of a *D. capensis* esterase/lipase (DCAP_0434) with the four functional blocks highlighted. C. The active site catalytic triad for a typical esterase/lipase (DCAP_0434).

Conserved Active Site Residues Suggest Functional Enzymes

In general, esterase/lipases are characterized by four moderately conserved sequence blocks of length 8-13 residues that contain the cataytic triad, the oxyanion hole proton donors, and other functionally important residues [103]. Functional sequence blocks I-IV are highlighted in the sequence alignments in the methods section. In Figure 3.10A, these functional blocks are represented as sequence logos, where the size of each residue label correlates with the number of instances at that sequence position within each cluster. The Ser-Asp-His catalytic triad is located within two block regions: block I (Ser) and block IV (Asp-His). The remaining two blocks contain conserved oxyanion hole residues, Gly in block II and Asn in block III [100]. Most of the proteins in this set contain the expected functional residues, as exemplified by the reference sequences GDL1_CARPA, GLIP6_ARATH, and GDL7_ARATH, as well as the well-characterized tomato GDSL esterase/lipase G1DEX3_SOLLC. Other results, including but not limited to PSN's and molecular modelling can be found in [2]; I have only included the results that I was directly involved in.

3.1.4 Conclusion

In summary, molecular modeling and protein structure network analysis of 26 esterase/lipases identified from the genomic DNA of *Drosera capensis* suggest that—with the exception of one protein, DCAP_3343—the active site regions of these enzymes are less flexible than those of related microbial proteins. The four clusters produced by the initial sequence analysis and clustering provides a stark contrast to the more limited esterase/lipase inventory typical of microbes (as evidenced by Uniprot searches). Subsequent principal component analysis of active site moieties generated from PSNs further categorized the *D. capensis* and reference sequences from decreasing to increasing active site rigidity. Together, these findings from comparative sequence and structural analyses demonstrate the diverse, *Drosera capensis* esterase/lipase landscape employed in carnivory, defense, and a plethora of functionalities with potentially significant, biotechnological applications.

3.2 The Phospholipases Found in *D. capensis* Form Four Clusters with Homology to Known Sequences

Phospholipases are a diverse set of enzymes that hydrolyze phospholipids. In plants, phospholipase D, phospholipase C, phospholipase A1 (PLA1), and phospholipase A2 (PLA2) have been characterized, that hydrolyze glycerophospholipids at different ester bonds as seen in Figure 3.12 [104, 105]. These enzymes are involved in a broad range of functions in cellular regulation and development, lipid metabolism, abiotic and biotic stress responses and membrane remodeling [104, 105, 106]. Within each type of phospholipase family, there are different families or subfamilies of enzymes that can differ in substrate specificity, cofactor requirement, and/or reaction conditions. These differences provide insights into determining the cellular function of specific phospholipases in plants, and they can be explored for

different industrial applications[104, 105, 106].

With Prof. Rachel Martin, I choose the protein set, generated the predicted structures, and analyzed the sequence and structure data as seen in Figure 3.11. The other team members include Shanon Zhuang, Michelle Xu, and Dr. John E. Kelly with Prof. Carter T. Butts. I performed sequence annotation and comparisons with the team as seen in Figure 3.14 and Figure 3.15. We are currently in the process of analyzing the structures and writing the manuscript.

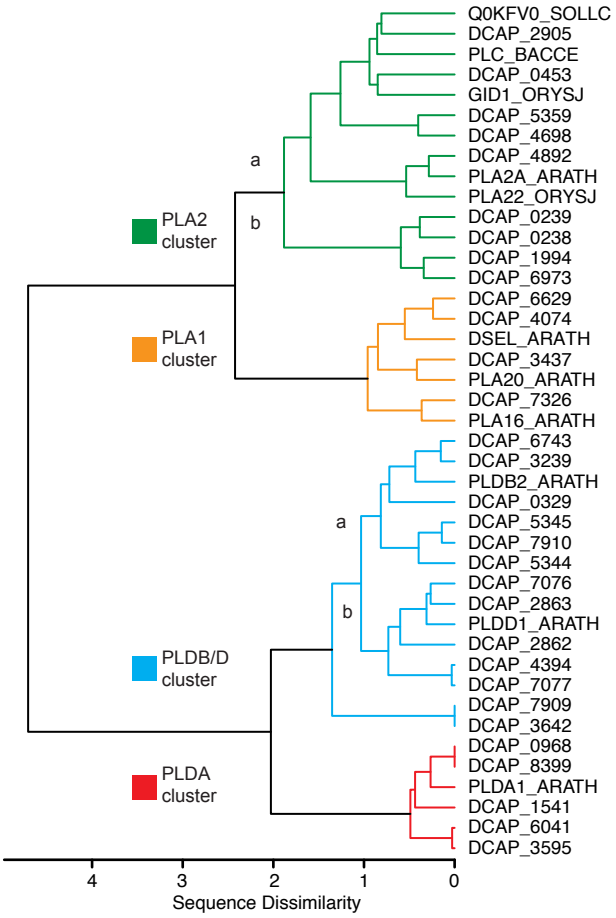

Figure 3.11: The current chosen set for phospholipases is seen in the Figure 3.11 with four different families, PLA2 (shown in green), PLA1 (shown in orange), PLDB/D (shown in blue) and PLDA (shown in red) found in *D. capensis*

The current chosen set for phospholipases is seen in the Figure 3.11 with four different families, PLA2 (shown in green), PLA1 (shown in orange), PLDB/D (shown in blue) and PLDA

(shown in red) in *D. capensis*. An example of the phospholipases found in *D. capensis* is seen in Figure 3.12 where the active site residues are highlighted and lablelled, and the propeptide, C2 and a PLD domain are highlighted in salmon, green and aqua colors respectively. The figure also shows the cut sites of different enzymes on a phospholipase.

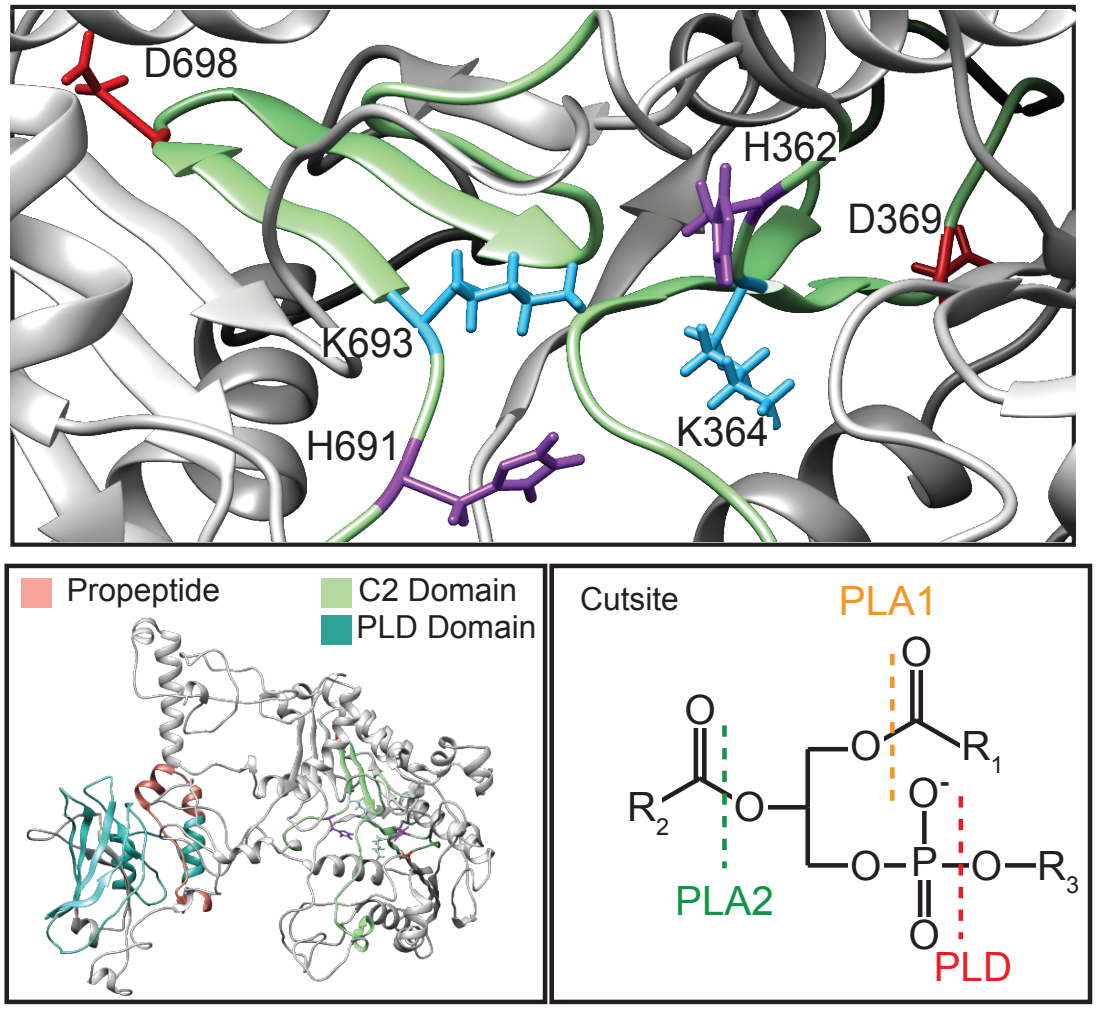

Figure 3.12: An example of the phospholipases found in *D. capensis* is seen in Figure 3.12 where the active site residues are highlighted and lablelled, and the propeptide, C2 and a PLD domain are highlighted in salmon, green and aqua colors respectively. The figure also shows the cut sites of different enzymes on a phospholipid.

Using the target selection pipeline described in Chapter 1, we have the equilibrated enzyme structures using the *in silico* maturation. Figure 3.13 shows cluster PLA_1 representative DCAP_7326 as an example in (a) and (b). The active site residues are histidine, serine and as-

partic acid (shown in purple, aqua and red respectively). PLA_2 representative DCAP_2905 as an example in (c) and (d). The active site residue is histidine (highlighted in red) while the disulfide bonds are shown in yellow. The structures will be available to download when the paper is published.

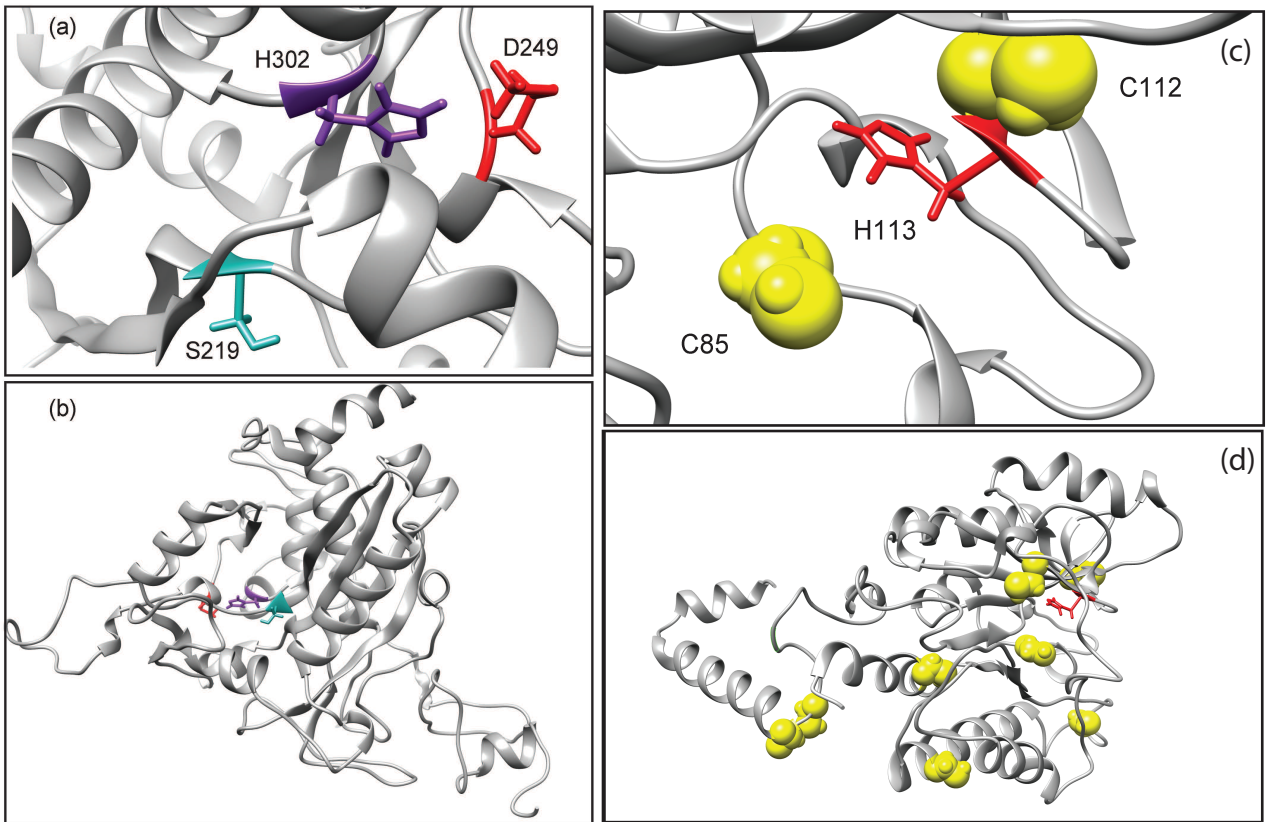

Figure 3.13: An example of the phospholipases found in *D. capensis* is seen in Figure 3.12 where the active site residues are highlighted and labelled, and the propeptide, C2 and a PLD domain are highlighted in salmon, green and aqua colors respectively. The figure also shows the cut sites of different enzymes on a phospholipid.

3.3 Nucleases from *Drosera capensis*

Nucleases are enzymes capable of cleaving the phosphodiester bonds between nucleotides of nucleic acids. We found 37 novel nucleases in *D. capensis*. An example of the nucleases found in *D. capensis* is seen in Figure 3.16 where the active site residues are highlighted and lablelled, and the metal ions are highlighted.

With Prof. Rachel Martin and Prof. Carter Butts, I chose the protein set and generated the predicted structures. Manuscript in preparation; next steps include *in silico* maturation, molecular modelling and analysis of the results.

Figure 3.14: The current chosen set for PLA1 from *D. capensis* with PLA16_ARATH, PLA20_ARATH and DESL_ARATH from *Arabidopsis thaliana* as references sequences.

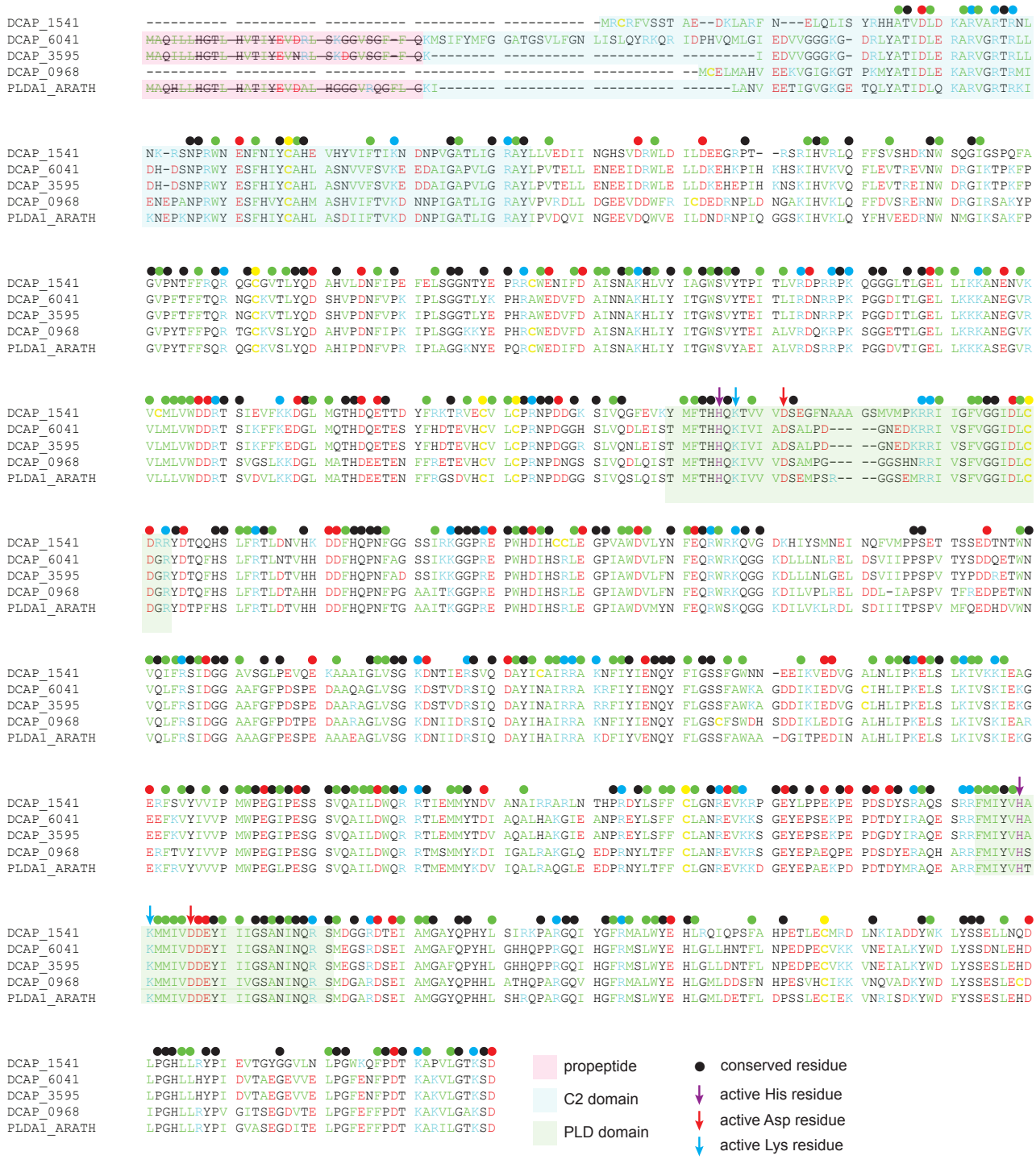

Figure 3.15: The current chosen set for PLDA from *D. capensis* with PLDA1_ARATH from *Arabidopsis thaliana* as references sequences.

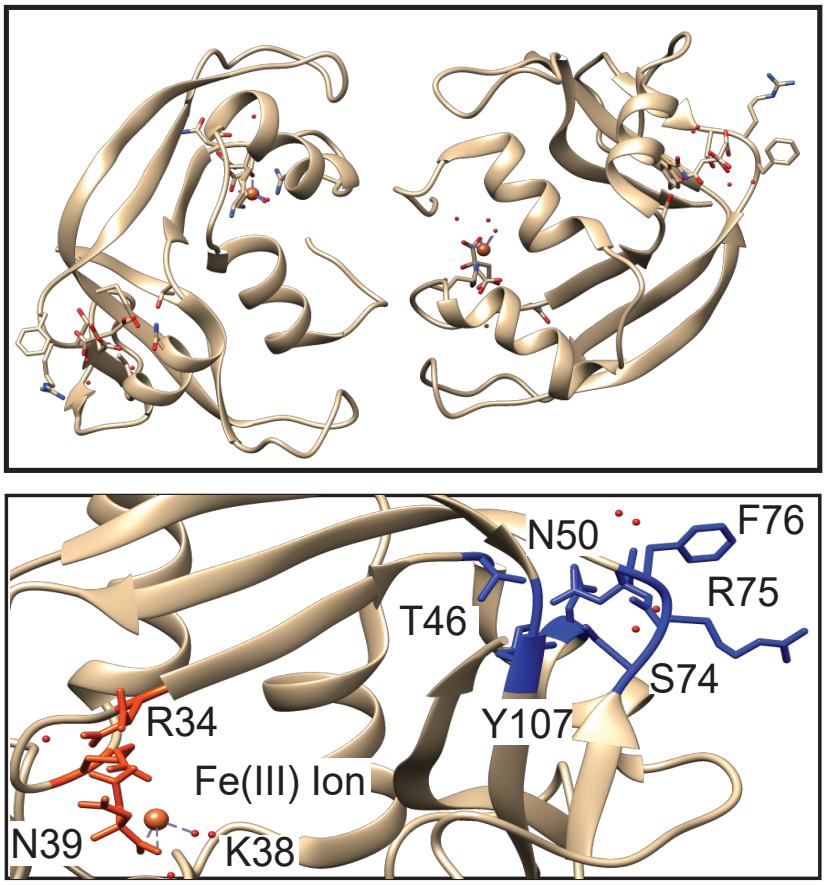

Figure 3.16: An example of the nucleases found in *D. capensis* where the active site residues are highlighted and labelled, and the metal ions are highlighted.

Chapter 4

Sequence comparison, molecular modeling, and network analysis predict structural diversity in cysteine proteases from the Cape sundew, *Drosera capensis*

Characterization of carnivorous plant digestive enzymes could lead to their use in a variety of laboratory and applications contexts, including analytical use in proteomics studies as well as preventing fouling on the surface of medical devices that cannot be treated under harsh conditions. The proteases of carnivorous plants present attractive targets for exploitation in chemical biology and biotechnology contexts. New proteases may also prove useful for cleaving amyloid fibrils, such as those responsible for the transmission of prion diseases or the formation of biofilms by pathogenic bacteria. The characterization of aspartic proteases from the tropical pitcher plants (*Nepenthes* sp.) [107, 108, 109], has already led to useful

advances in mass spectrometry-based proteomics applications, where the ability to digest proteins using a variety of cut sites is essential for identifying proteins and peptides from complex mixtures. Proteases from plant and animal sources are also important components of pharmaceutical preparations for gluten intolerance, arthritis, and pancreatic disease [110]. Therefore, characterizing proteases from *D. capensis* has the potential to diversify the toolbox of proteases with different functional properties that are available for these and other applications.

I contributed to choosing the protein set, determining the functional regions of interest, generating the predicted structures, analyzing sequence and structure data, performing sequence annotation and comparisons and wrote the manuscript with my co-authors on this paper. A portion of the paper [2] is reproduced in this chapter to understand the importance of the results found in this study. For more details and in-depth analysis, please refer the paper [2]

4.1 Background

Plant cysteine proteases form a large and diverse family of proteins that perform cellular housekeeping tasks, fulfill defensive functions, and, in carnivorous plants, digest proteins from prey. It is typical for plants to contain many different cysteine protease isoforms; for instance, in the case of tobacco (*Nicotiana tabacum*), more than 60 cysteine protease genes have been identified [111]. Many of the cysteine proteases of interest are classified by the MEROPS database as family C1 [112], a broad class of enzymes including cathepsins and viral proteases as well as plant enzymes that function to deter herbivory. C1 proteases can operate as endopeptidases, dipeptidyl peptidases, and aminopeptidases [113]. In plants, many C1 enzymes are used to degrade proteins in the vacuole, playing many of the same roles as their lysosomal counterparts in animals [114]. They are also found in fruits, particularly unripe ones; this protease activity impedes insect feeding and also serves to cleave

endogenous proteins during fruit ripening. Some families of cysteine proteases in plants have been subject to diversifying selection due to a molecular arms race between these plants and their pathogens; as plants produce proteases that suppress fungal growth, fungi evolve inhibitors specific to these proteases, driving the diversification of plant proteases involved in the immune response [64].

The plethora of paralogs found in a typical plant is indicative of the need for a range of different substrate specificities; this is particularly important in the case of carnivorous plants, which must digest prey proteins to their component amino acids. Aspartic proteases have long been implicated in *Nepenthes* pitcher plant digestion [107, 115], and more recently the cysteine protease dionain 1 has been confirmed as a major digestive enzyme in the Venus flytrap (*Dionaea muscipula*) [116]. In *Drosera capensis*, proteins from prey consititute the major nitrogen source for producing new plant tissue [40]. Given that plant carnivory appears to have evolved from defensive systems in general [117], and that the feeding responses are triggered by the same signaling pathway as is implicated in response to wounding [118], one would expect cysteine proteases to play a major role; here we investigate some of the many cysteine protease genes in *D. capensis* with the objective of adding to the portfolio of cleavage activities available for chemical biology applications. The *D. capensis* enzymes are particularly appealing for mass spectrometry-based proteomics applications, due to their ability to operate under relatively mild conditions, i.e. at room temperature and pH 5.

This study focuses on the C1 cysteine proteases from the Cape sundew (*D. capensis*), where it uses the pipeline from Chapter 1 to identify structurally distinct subgroups of proteins for subsequent expression and biochemical characterization. C1 cysteine proteases share a common papain-like fold, a property also predicted for the proteins studied here. Despite this conservation of the papain fold and critical active and structural residues, sequence analysis

of the *D. capensis* cysteine proteases indicates that they represent a highly diverse group of proteins, some of which appear to be specific to the Droseraceae. In particular, a large cluster of proteases containing dionains 1 and 3 as well as many homologs from *D. capensis* has particular sequence features not seen in papain or other reference enzymes. Finally, a new class of granulin domain-containing cysteine proteases is identified, based on clustering of the granulin domains themselves.

Molecular modeling was performed (Prof. Carter Butts and Dr. Xuhong Zhang with the team) in order to translate this sequence diversity into predicted structural diversity, which is more informative for guiding future experimental studies. Examination of the predicted enzyme structures potentially suggests diversity that may imply a variety of substrate preferences and cleavage patterns. Further, the study uses Rosetta [28, 27] to perform comparative modeling with all-atom refinement, described in detail in Chapter 1, combining local homology modeling based on short fragments with de novo structure prediction. The study then employs atomistic MD simulation of these initial structures in explicit solvent to produce equilibrated structures with corrected active site protonation states; these equilibrated structures serve as the starting point for further analysis.

Quality control was performed using both sequence alignment and inspection of the Rosetta structures; proteins that are missing one of the critical active residues (C158 or H292, papain numbering) were discarded, as were some lacking critical disulfide bonds or other structural features necessary for stability. After winnowing out sequences that are unlikely to produce active proteases, 44 potentially active proteases were chosen for further analysis. This methodology allows the development of hypotheses based on predicted 3D structure and activity, in contrast to focusing on the first discovered or most abundantly produced enzymes, enabling selection of the most promising targets for structural and biochemical characteri-

zation based on the priorities of technological utility rather than relative importance in the biological context.

4.2 Cysteine Protease Sequence Analysis

Multiple-sequence alignments for cysteine proteases from *D. capensis* and previously characterized plant cysteine proteases reveal diverse functionality. Annotated sequence alignments are shown for the DCAP cluster (Figure 4.1), the papain cluster, both catalytic domains (Figure 4.2) and (Figure 4.3), the vignain cluster, (Figure 4.4), the granulin domain cluster (Figure 4.5), the bromelain cluster (Figure 4.6) and the dionain cluster (Figure 4.7). The annotations highlight both specific amino acid properties and general sequence features. Followung the nomenclature of "Target Selection Pipeline" from Chapter 1, hydrophobic residues are shown in green, positively charged residues in blue, negatively charged residues in red, and cysteines in yellow. Conserved Cys residues involved in structure-stabilizing disulfide bonds are indicated with yellow asterisks, while other residues conserved across all the sequences considered are indicated with solid circles. Residues conserved within the cluster but not shared with papain are indicated with open circles. The residues of the catalytic dyad are indicated with colored arrows, yellow for Cys and purple for His. The position of the stabilizing Asn residue is indicated with a pink asterisk, although this residue is not conserved in all sequences. Strikethrough text indicates parts of the sequence that are expressed but removed during post-translational processing; for these proteins, this constitutes an N-terminal region comprised of the signal peptide and the pro-sequence. The presence and position of a signal sequence targeting the protein for secretion was predicted using SignalP, and is indicated in the figures by highlighting in light orange, with the predicted cut site indicated by underlining the residues on either side of the cleavage point. The

position of the pro-sequences was predicted by sequence similarity to the reference sequences as well as comparison of the predicted structures to the crystal structure of the mature form of papain. The pro-seqeunces of many of the sequences studied here contain the ERFNIN motif ($EX_3RX_3FX_2NX_3I/VX_3N$) common to C1-family cysteine proteases. When present, this sequence is shown above the relevant residues in the alignments. The presence of localization tags, when present, is indicated by purple highlighting. Granulin domains, when present, are highlighted in blue.

Sequence consensus analysis for the sequences within each cluster defined in Figure 1.1, are mapped onto the structure of a representative member of the class in Figure 4.8. Percent conservation at each position is color coded (red = more conserved, white = intermediate values, blue = less conserved). These plots demonstrate that the degree of sequence conservation varies greatly among different clusters, i.e. the DCAP cluster has much less sequence conservation overall than the vignain cluster. In all the clusters, conserved residues are concentrated in the important secondary structure elements and near the active site cleft, whereas residues in loops and linkers away from the core of the protein are less likely to be conserved. All the predicted and equlibrated structures are available to download in the paper [30].

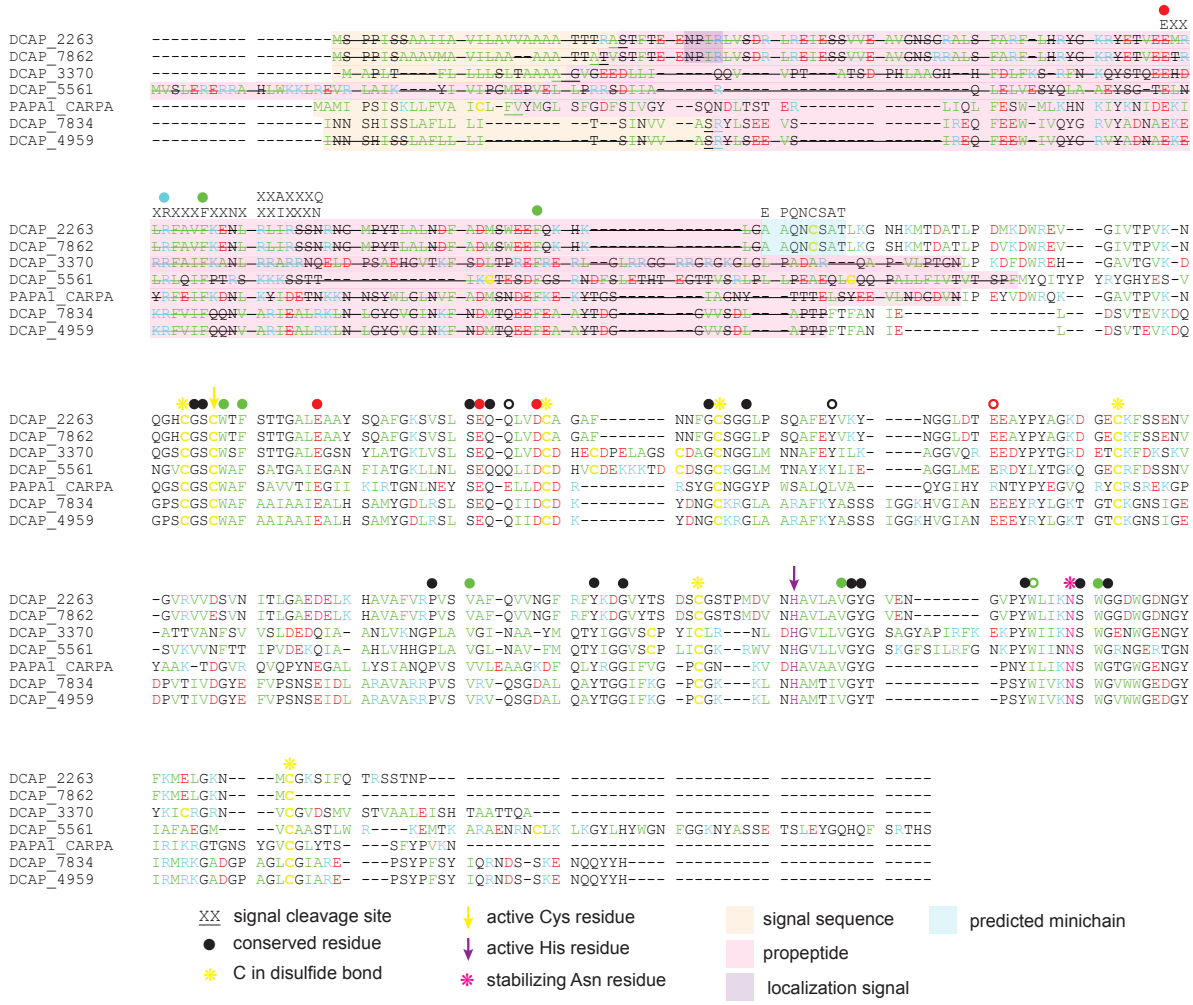

Figure 4.1: The DCAP cluster contains sequences that are more closely related to other *D. capensis* sequences than to any of the references. Several have insertions not found in other sequences, potentially indicating specific functionalities. DCAP_2263 and DCAP_7862 contain the localization tag NPIR in their N-terminal pro-domain regions, indicating targeting to the vacuole.

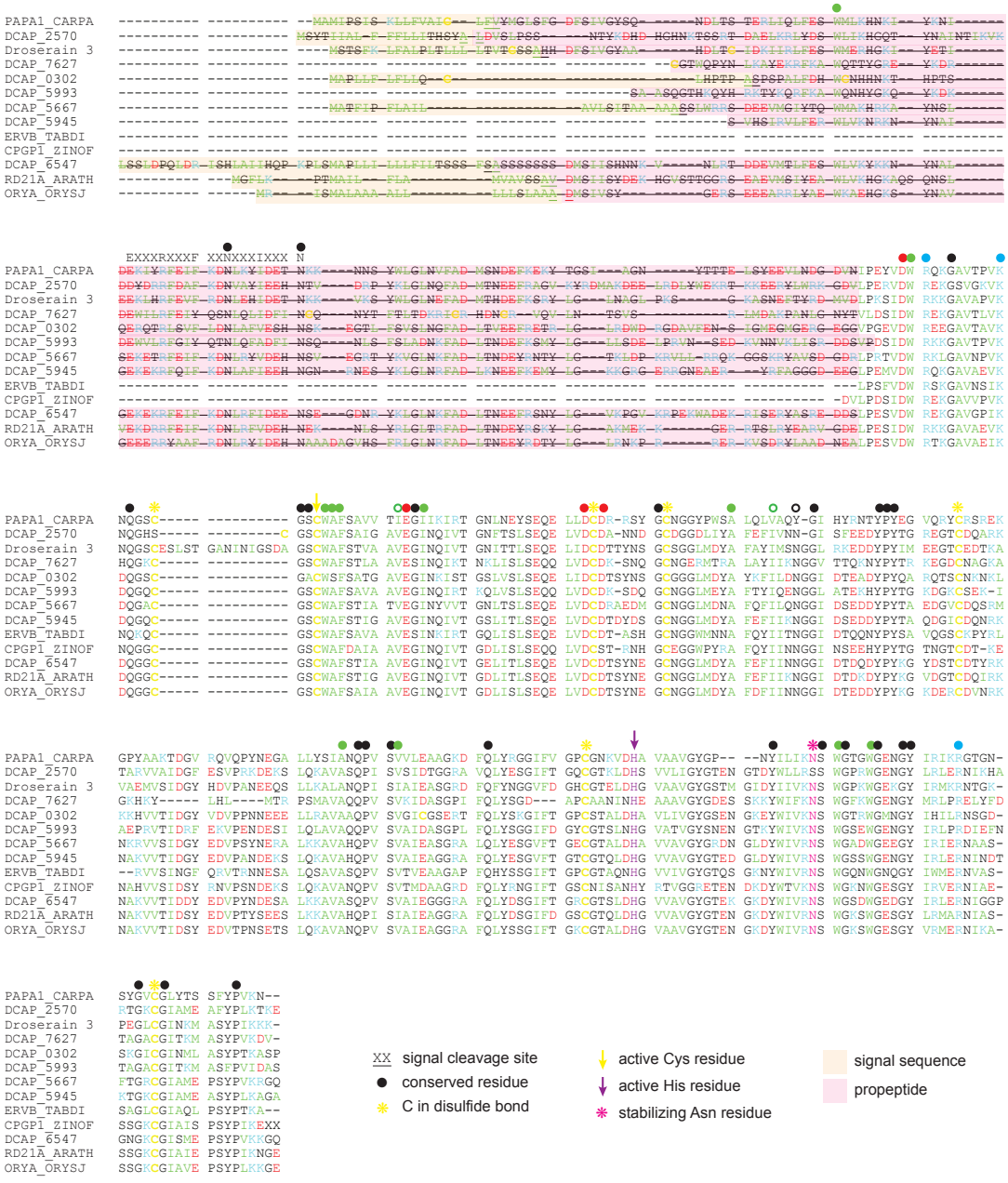

Figure 4.2: Many of the reference sequences belong to the papain cluster despite the diversity of their sources.. Several proteins in cluster also have C-terminal granulin domains, which are shown in Fig. S4.3.

Figure 4.3: The papain cluster granulin domains contain several examples homologous to the reference proteins RD21_ARATH and ORYA_ORYSJ. Papain itself lacks a C-terminal granulin domain, so it is not included in the alignment. DCAP_2570 and DCAP_5667 are truncated, and therefore do not contain both disulfide bonds stabilizing the granulin domains. DCAP_ 5945 contains an extra C-terminal extension not found in the reference sequences. The conserved sequence region characterizing animal granulin domains is shown above the corresponding sequences for comparison. The plant granulin sequences have two distinguishing features; an additional conserved Cys residue is present immediately after the first conserved CC pair in the animal sequence, and a 6-residue insertion containing another conserved C is present between the first and second CC pairs.

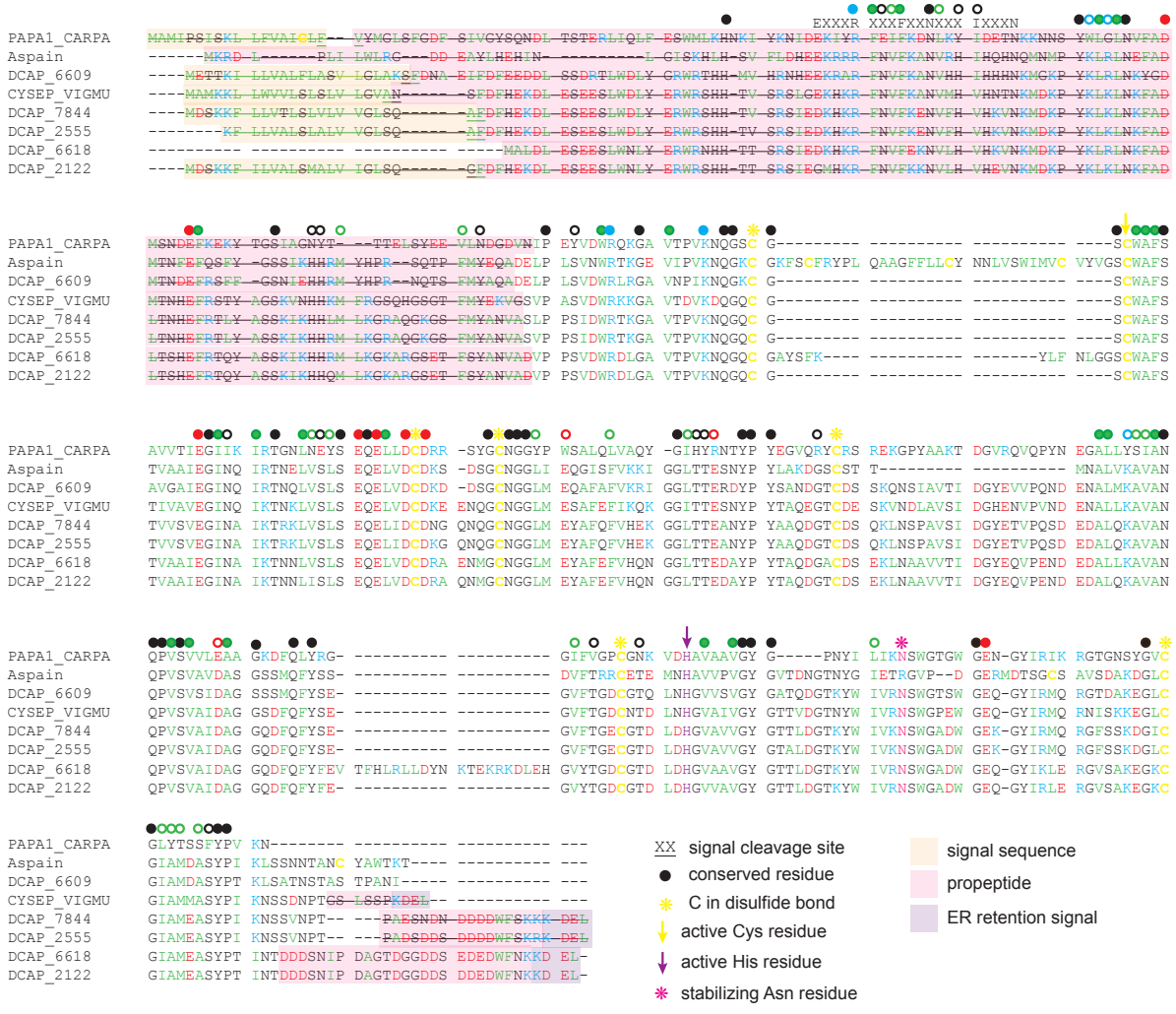

Figure 4.4: Many proteins in the vignain cluster, including vignain itself, are characterized by the localization tag KDEL at the C-terminus. This sequence element indicates that the protein is marked for retention in the endoplasmic reticulum.

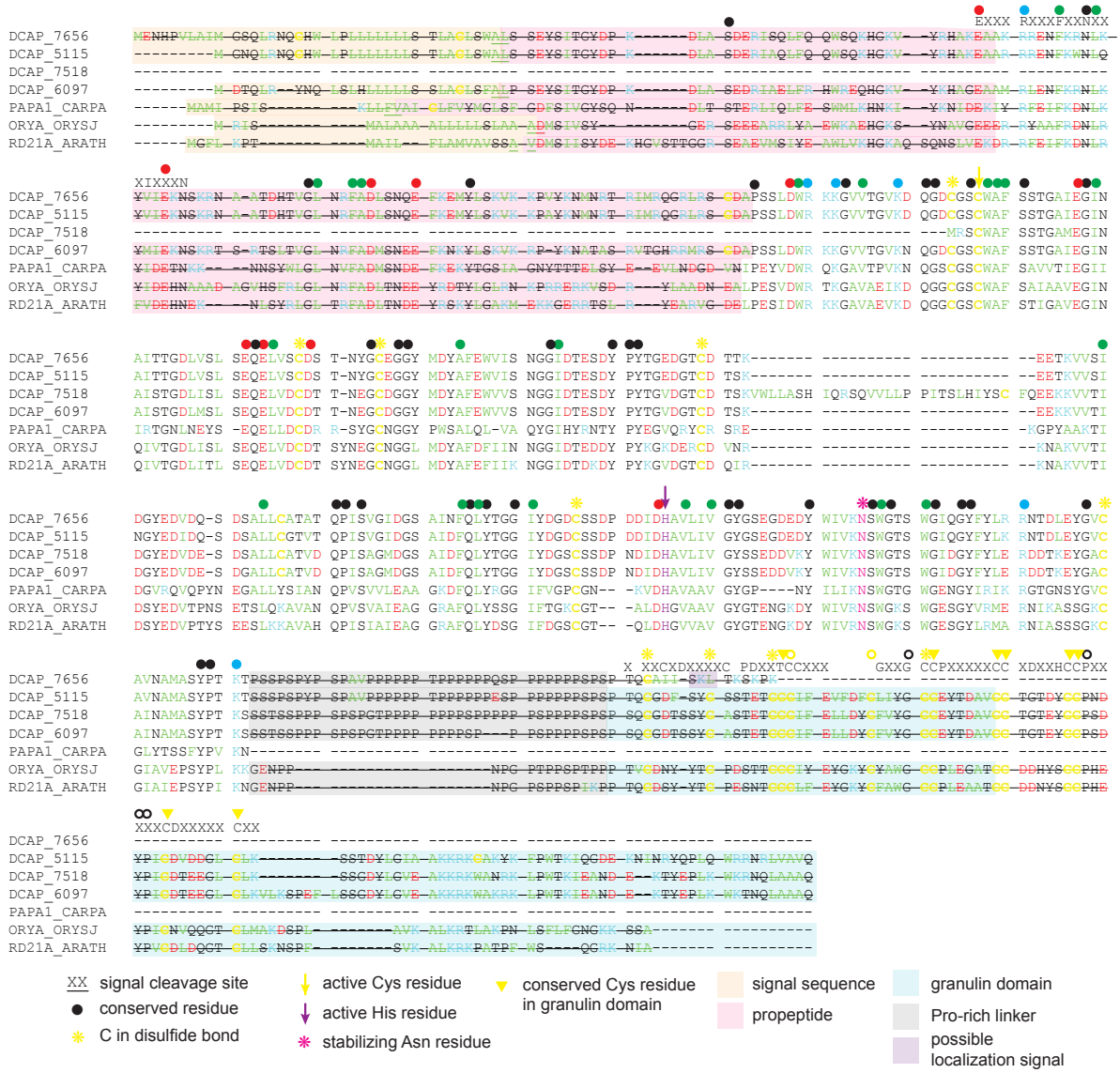

Figure 4.5: The granulin domain cluster contains proteins with C-terminal granulin domains. Although they are not closely related to any of the reference sequences, RD21_ARATH and ORYA_ORYSJ are shown in the alignment in order to compare sequence features among the granulin domains. As shown for the papain cluster granulin domains, the conserved sequence region characterizing animal granulin domains placed above the corresponding sequences. As in the papain case, there are two additional conserved Cs and a 6-residue insertion between the first and second CC pairs. In these sequence, a deletion of one residue relative to the animal sequence also occurs between the first and second conserved Cys residues in the granulin domain. DCAP_7656 is missing most of the granulin domain, and instead contains the localization tag SKL near the C-terminus, marking it for transport to the peroxisome.

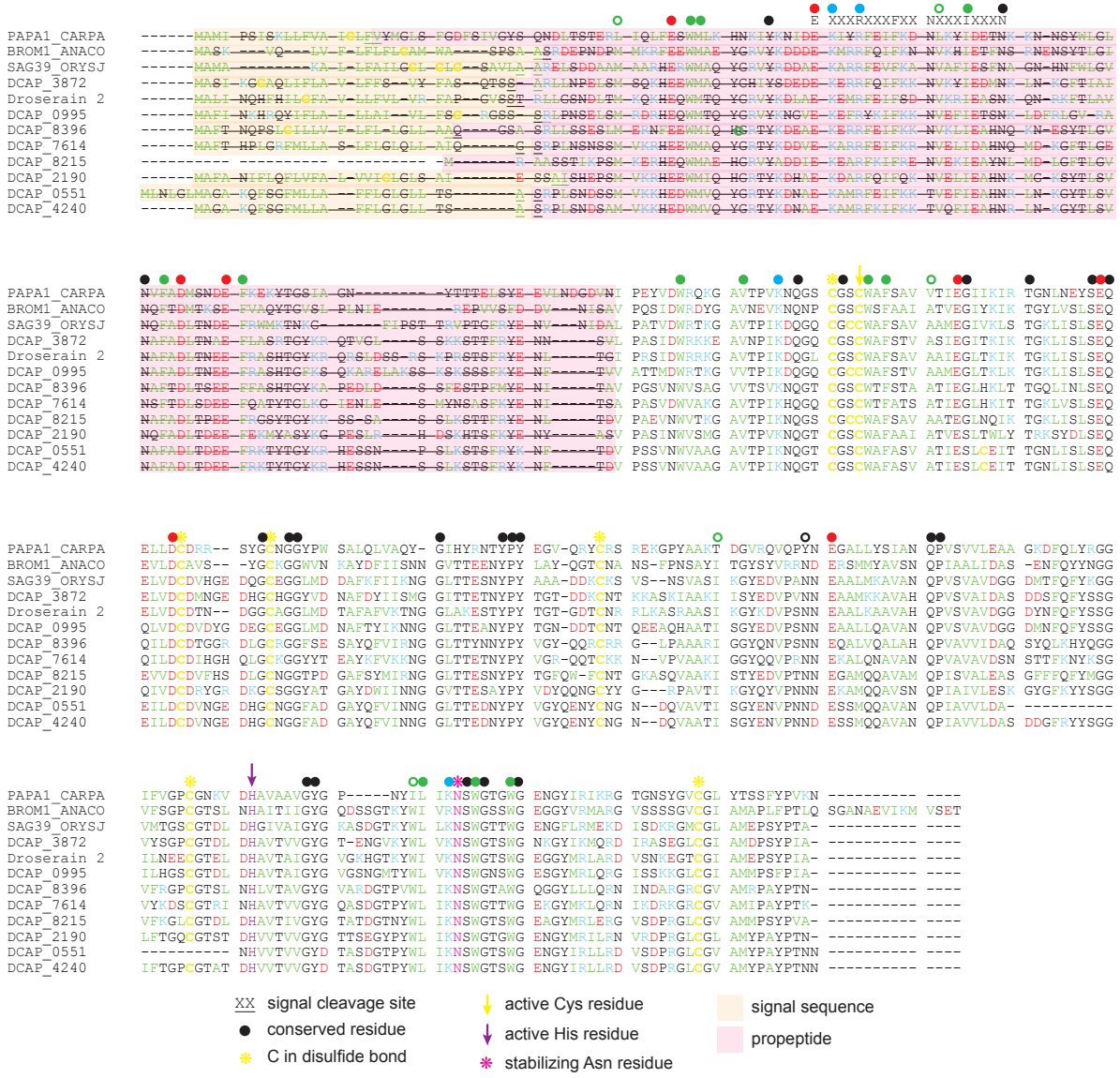

Figure 4.6: The bromelain cluster is characterized by strong sequence identity with pineapple fruit bromelain.

Figure 4.7: The dionain cluster contains many cysteine proteases that appear to be specific to Caryophylalles carnivorous plants; this cluster contains the dionains from *D. muscipula* as well as several proteins from *D. capensis*, but none of the reference sequences from other sources.

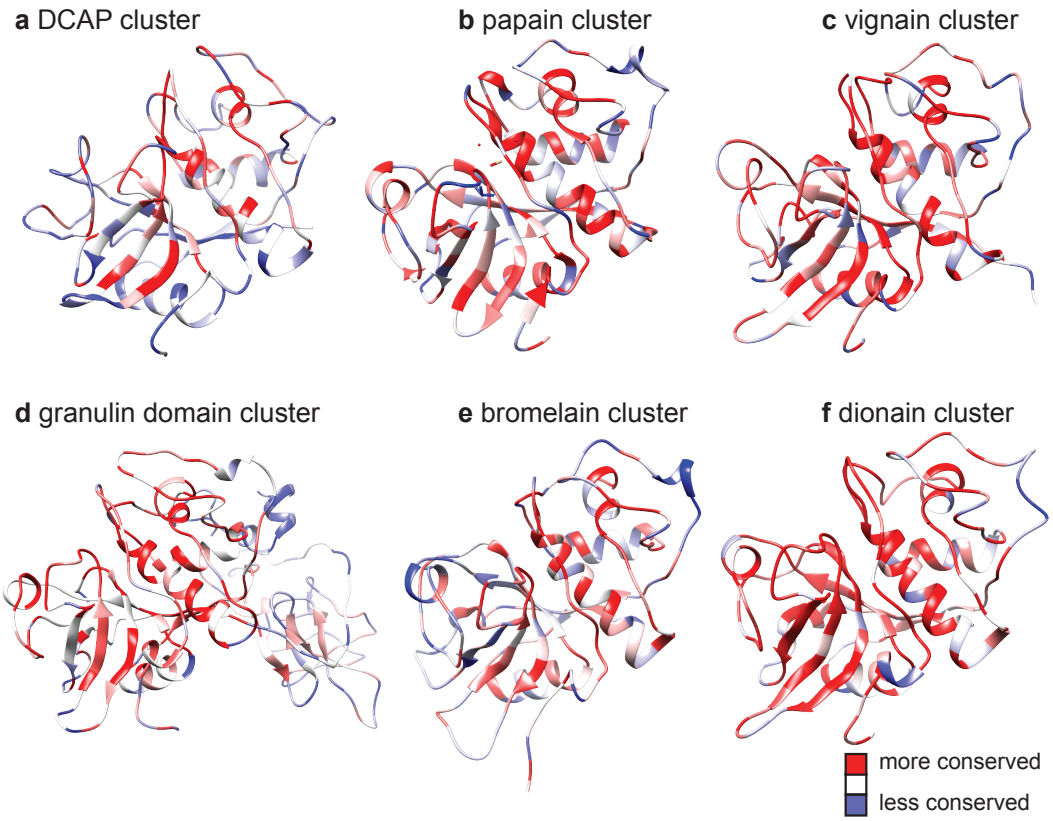

Figure 4.8: The percent conservation of each residue in the consensus sequence for each cluster is shown mapped onto a representative member of the cluster. The color scale ranges from red (more conserved) to blue (less conserved). a. DCAP cluster (DCAP_2263) b. papain cluster (papain) c. vignain cluster ((DCAP_2122) d. granulin domain cluster (DCAP_5115) e. bromelain cluster (droserain 2) and f. dionain cluster (DCAP_0624).

4.3 Results and Discussion

I discuss some important and relevant results from the paper [30] in this section- author contributions can be found in the paper. Molecular modeling, protein network analysis and all the predicted and equlibrated structures are available to download from our paper [30].

4.3.1 *D. capensis* Cysteine Proteases Cluster Into Distinct Families Based on Resemblance to Known Homologs

All *D. capensis* sequences previously annotated as coding for MEROPS C1 cysteine proteases using the MAKER-P (v2.31.8) pipeline [49] and a BLAST search against SwissProt (downloaded 8/30/15) and InterProScan [50] were clustered by sequence similarity. Several previously-characterized cysteine proteases that have been identified from other plants are also included as reference sequences. Clustering of the *D. capensis* cysteine protease sequences reveals a broad range of cysteine protease types, some of which are homologous to known plant proteases (Figure 4.9). Three of the six clusters contain only proteins from *D. capensis* or the related Venus flytrap *Dionaea muscipula*, while many of the reference sequences cluster together despite coming from a variety of different plant species from diverse orders including both monocots and eudicots. The general types of plant protease features found correlate well with previous surveys of cysteine proteases in *Arabidopsis thaliana* [119], *Populus sp.* [104], and more recently, soybeans [120] and a broader group of plant proteases from a variety of species [7].

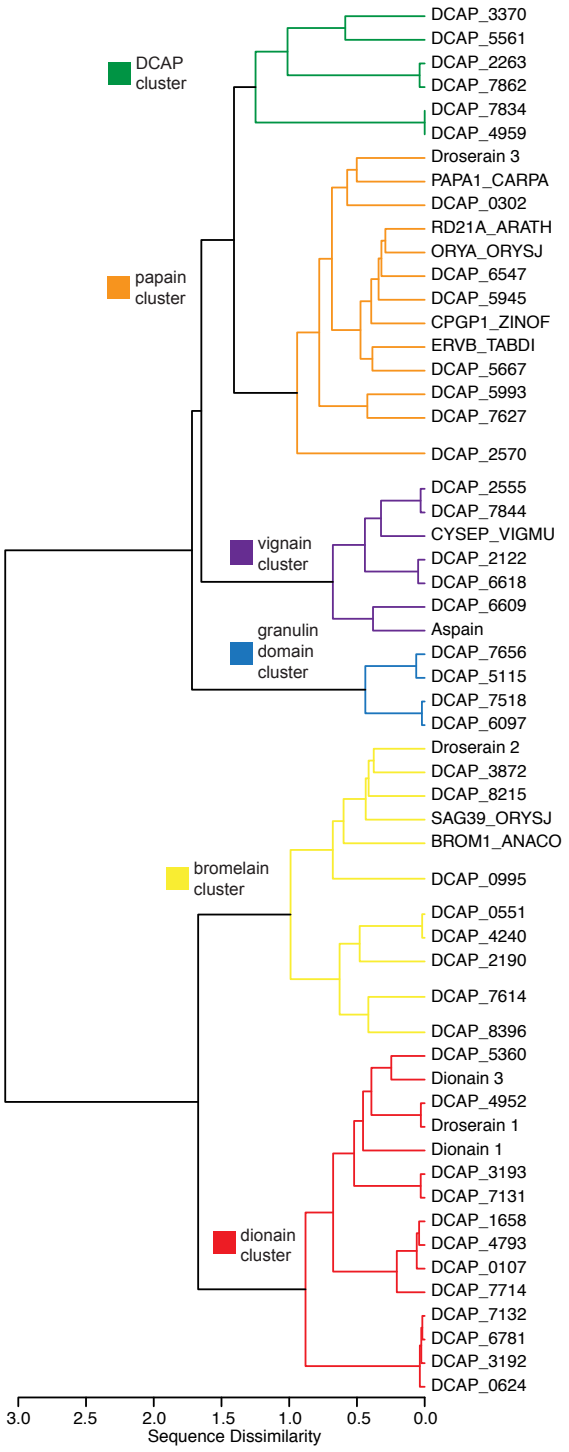

Figure 4.9: Clustering of cysteine protease sequences identified from the *D. capensis* genome. Many are homologous to known plant cysteine proteases, including dionain 1 and dionain 3 from the Venus flytrap, *Dionaea muscipula*. Dissimilarity between clusters is defined by the e-distance metric of [4] (with $\alpha = 1$), which is a weighted function of within-cluster similarities and between-cluster differences with respect to a user-specified reference metric. The underlying input metric employed here is the raw sequence dissmilarity $(1 - (\%\text{identity})/100)$.

4.3.2 Residues Conserved in *D. capensis* Cysteine Proteases Include Active Sites and Important Sequence Features

A defining feature of C1A cysteine proteases is the Cys-His catalytic dyad, which is often accompanied by an Asn residue that stabilizes the protonated catalytic His [121, 122]. The mechanism of these enzymes requires using the thiolate group on the deprotonated cysteine as a nucleophile to attack a carbonyl carbon in the backbone of the substrate. Preliminary sequence alignments comparing putative cysteine proteases from *D. capensis* were used to discard sequences lacking the conserved Cys and His residues of the catalytic dyad due to either substitution or truncation. Other conserved features were observed in many of the sequences, but were not treated as necessarily essential for activity. Reference sequences used include zingipain 1 from *Zingiber officianale* (UniProt P82473), pineapple fruit bromelain (*Ananas comosus*, UniProt O23791), RD21 from *Arabidopsis thaliana* (UniProt P43297), oryzain alpha chain (UniProt P25776) and SAG39 (UniProt Q7XWK5) from *Oryza sativa* subsp. japonica, ervatamin b from *Tabernaemontana divaricata* (UniProt P60994), and dionains 1 and 3 from the related *Dionaea muscipula* (UniProt A0A0E3GLN3, and A0A0E3M338, respectively). Several of the reference sequences, e.g. zingipain-1 [123], were characterized by mass spectrometric analysis of the mature enzyme; these sequences therefore lack the signal peptide and pro-sequence found in the initially transcribed sequence. Sequence alignments for the individual clusters (seen above) are annotated to highlight individual amino acid properties, residues conserved within the cluster and/or shared with papain, as well as functional sequence features, as described above. In addition to the cluster-specific reference sequences, all clusters include papain (*Carica papaya*, UniProt P00784) in order to have a common reference for all the C1A proteases discussed in this work.

Most of the clusters are named after a reference sequence or a distinguishing feature of its members. The DCAP cluster is highly diverse, yet it contains only sequences from *D.*

capensis. The papain cluster contains many of the reference sequences, as well as several *D. capensis* proteases, some of which have granulin domains (Figures 2 and 3), a feature that is peculiar to plant cysteine proteases. The vignain cluster (Figure 4) contains vignain from *Vigna mungo* (UniProt P12412) as well as *D. capensis* homologs. Many of the proteins in the vignain cluster have C-terminal KDEL tags, indicating retention in the ER lumen, suggesting that they are involved in germination and/or senescence. In the granulin domain cluster (Figure 5), every sequence but one contains a granulin domain connected to the catalytic domain by a proline-rich linker of about 40 residues; the one exception is truncated after the proline-rich region. Several sequences in the papain cluster also contain granulin domains, however the Pro-rich linkers in those sequences contain only about 16 residues and the sequence identity between the two types of granulin domains themselves is not high. The bromelain cluster (Figure 6) contains homologs of both defensive and senescence-related enzymes. Every sequence in the dionain cluster (Figure S7) contains an extra Cys residue immediately prior to the active site Cys. This CCWAF structural motif has been previously observed in the Arabidobsis protein SAG12 and homologs [7]; however, the function of the double Cys in unknown. It may have cataytic relevance, perhaps providing a second nucleophilic thiolate or operating as a redox switch.

Like many other proteases, the papain-family enzymes are expressed with an N-terminal pro-sequence blocking the active site. This sequence is cleaved during enzyme maturation, often upon the protein's entering a low-pH environment. This pro-sequence was found in most of the C1A proteases from *D. capensis* (highlighted with pink boxes in figures 1-7 in the above subsection). Plant C1A protease pro-sequences are often bioactive in their own right, acting as inhibitors of exogenous cysteine proteases. This enables them to deter herbivory by insects [124], nematodes [125], and spider mites [126], protecting the plants from damage. This can be technologically exploited by producing transgenic crop varieties with protective cysteine proteases they would otherwise lack [127]. This approach has proven useful in protecting

crops from Bt-resistant pests [128]. Despite some variation in the lengths of the C-terminal and N-terminal regions, all the cysteine proteases investigated here show substantial similarity in the pro-sequences; in particular, the ERFNIN motif ($EX_3RX_3FX_2NX_3IX_3N$) often found in the pro-sequence of C1A proteases [129] is conserved in many sequences spanning all the clusters. Interestingly, the alternative sequence $EX_3RX_3FX_2NX_3AX_3Q$, which is characteristic of the RD19 family of plant cysteine proteases, is found in only one of the *D. capensis* proteases, DCAP_3370 in the DCAP cluster. For all previously uncharacterized sequences, SignalP 4.1 [26] was used to predict the location of the signal sequences, if any, while the pro-sequences were predicted by sequence similarity and structural homology to papain. These sequence annotations were then used as the basis for further structure prediction and functional analysis.

In addition to the common sequence features in the N-terminal pro-region, other variations are observed, such as the presence of C-terminal granulin domains in some sequences and extra insertions that may be responsible for specific activities in others. Examples of organelle-specific targeting sequences are observed; several sequences have a C-terminal KDEL sequence targeting them for retention in the endoplasmic reticulum, while others have targeting sequences indicating their destination in the cells, including signals indicating transport to the vacuole (NPIR, but not FAEAI or LVAE) or the peroxisome (SSM at the C-terminus). The level of sequence conservation among the members of each cluster varies dramatically, as can be seen in Figure 8, where sequence conservation is mapped onto the structure of a representative member of each cluster. The sequences in the DCAP cluster are less closely related to each other than the members of any of the other clusters, and some are homologous to reference sequences used by Richau et al. [7].

Another interesting result was of DCAP_2263 and DCAP_7862, which belong to the Richau

aleurain (cathepsin H) cluster. In humans, cathepsin H is an aminopeptidase that processes neuropeptides in the brain [130], as well as acting as a lysosomal protein in other tissues. Its barley (*Hordeum vulgare*) homolog, aleurain, has both aminopeptidase and endopeptidase activity [131], suggesting that DCAP_2263 and DCAP_7862 may have both types of activity as well. This hypothesis is supported by the presence of the Cathepsin H minichain sequence in its plant orthologs, as discussed in the section devoted to these proteins. DCAP_3370 is related to the Richau RD19 (cathepsin F) cluster, and is the only protease in this set that contains the characteristic pro-sequence motif ($EX_3RX_3FX_2NX_3AX_3Q$), of the RD19 (cathepsin F) family. Human cathepsin F is distinguished by its unusually long pro-domain, which is approximately 100 residues longer than that of other cysteine proteases and adopts a cystatin fold [132]. In contrast, the pro-sequence of DCAP_3370 is about 140 residues, typical for a plant cysteine protease. The last enzyme in the DCAP cluster, DCAP_5561 is not closely related to anything in either reference set. A BLAST search yields numerous matches to uncharacterized predicted cysteine proteases from a variety of plant genomes, however, the specific function of this enzyme remains enigmatic.

4.3.3 Some Cysteine Proteases Are Targeted to Specific Locations

Several of the cysteine proteases identified from *D. capensis* contain known targeting signals that mark the protein for delivery to specific cellular locations. The most common such signal is the N-terminal signal peptide targeting the protein for secretion. As expected, the majority of proteins in this set contain such a secretion signal. In plants, the secretory pathway delivers proteins to the vacuole, the vacuolar membrane, the cell wall, and the the plasma membrane. In *D. capensis*, digestive enzymes are also expected to be secreted into the mucilage. In addition to the N-terminal signal sequences, tri- or tetrapeptides indicating that the protein is destined for a particular subcellular compartment are also found in many cases.

Figure 4.10 shows the structures predicted by Rosetta for three full-length cysteine proteases containing targeting signals, DCAP_2263, DCAP_5667, and DCAP_2122. Ribbon diagrams are shown for all three enzymes; a surface is also shown for DCAP_2122 in order to assist with visualization of the relationship of the pro-sequence, N-terminal signal peptide, and C-terminal localization sequence to the rest of the protein. The positioning of the pro-sequences (pink) and signal peptides (light orange) is highly variable, although in each example the pro-sequence blocks the active site and the signal sequences and other localization tags (light purple) are in highly exposed positions as expected based on their function.

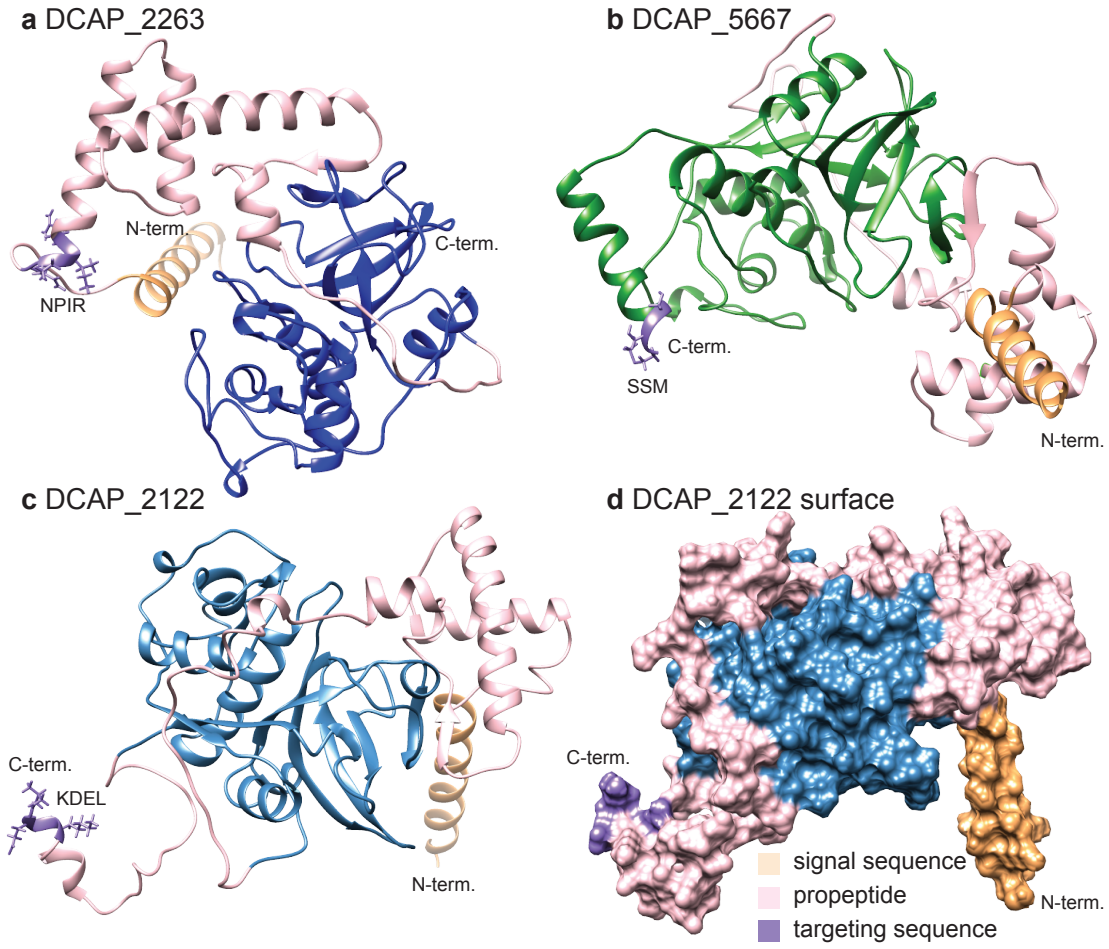

Figure 4.10: Predicted structures for three full-length cysteine proteases. The secretion signals are highlighted in light orange, the pro-sequences in pink, and the localization tags in light purple. a. DCAP_2263 contains the target sequence NPIR, indicating localization to the vacuole. b. DCAP_5667 ends in the tripeptide SSM at the extreme C-terminus, indicating transport to the peroxisome c. and d. DCAP_2122 ribbon diagram and surface model, respectively. DCAP_2122 ends in the ER-retention signal KDEL, indicating that it is retained in the ER lumen.

In plants, the subsequence NPIR in the N-terminal region of a protein indicates targeting to the vacuole, a large acidic compartment that is specific to plant cells and serves the same function as the lysozome in animal cells. These compartments, which often occupy most of the volume of the cell, contain a variety of hydrolases, including both aspartic and cysteine proteases, which normally act to recycle damaged or unneeded cellular components. Upon infection by viruses or fungal pathogens, the vacuole can also fuse with the plasma membrane to release defensive proteases into the extracellular space. Two putative vacuolar proteases, (DCAP_2263 and DCAP_7862) are found in the DCAP cluster. The NPIR tag is located in an exposed position between the secretion signal and the beginning of the N-terminal pro-sequence, as shown for DCAP_7862 in Figure 4.11. These proteases display sequence homology to mammalian cathepsin H, a lysozomal protein that is important in development and also implicated in cancer proliferation [133, 134].

In human cathepsin H, aminopeptidase activity is modulated by the minichain sequence (EPQNCSAT). DCAP_2263 and DCAP_7862 (and aleurain, but no others in this set) contain the sequence AAQNCSAT, which may have a similar function. The hypothesis that this plant-specific minichain serves a similar role in modulating the substrate specificity is supported by comparing the predicted structures with the crystal structure of porcine cathepsin H (PDBID: 8PCH) [135]. Figure 4.11 shows the predicted structures of mature DCAP_2263 (blue) and DCAP_7862 (green) overlaid with the crystal structure of porcine cathepsin H (gray). The predicted structures of the plant proteins coincide with the porcine protein in the major secondary structure elements, albeit with substantial variation in loops and linkers. The minichain sequence (EPQNCSAT in the porcine protein and AAQNCSAT in the *D. capensis* proteins) occupies a similar position in all three structures, allowing substrate approach to the active site cleft from one side (Figure 4.11a), but not the other (Figure 4.11b). Biochemical characterization of human cathepsin H has shown that deletion of the minichain abolishes aminopeptidase activity [136], making this protein a standard

endopeptidase. Based on sequence homology and examination of the predicted structures, we hypothesize that this sequence plays a similar role in modulating the substrate specificity and activity patterns of DCAP_2263 and DCAP_7862.

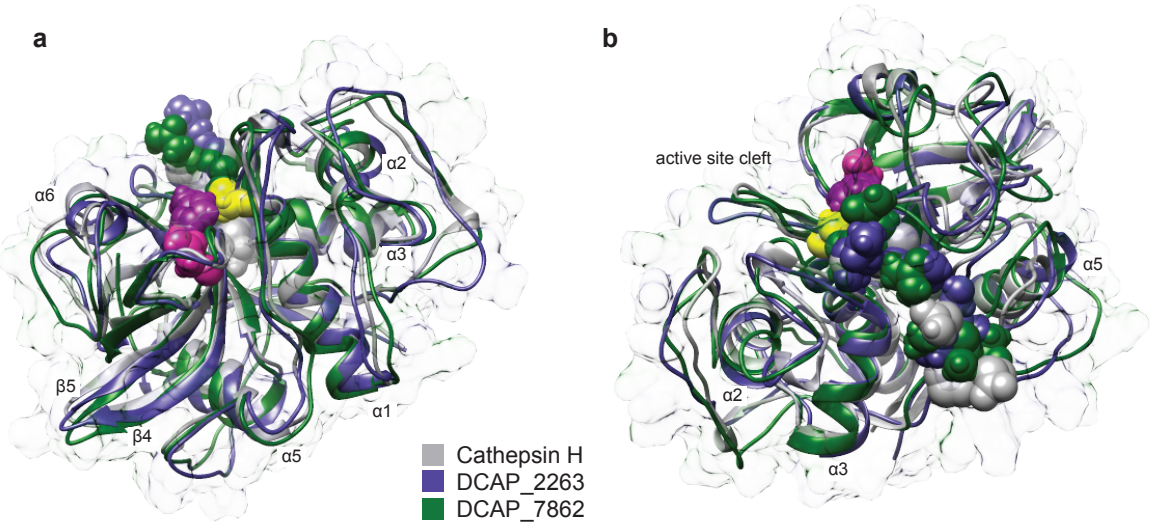

Figure 4.11: Predicted structures for two vacuolar cysteine proteases (DCAP_2263, blue and DCAP_7862, green) with sequence homology to cathepsin H (PDBID: 8PCH gray). The active site residues and the minichain are shown as space-filling models. a. One side of the active site cleft is open and accessible to substrate. b. The other side of the active site cleft is blocked by the minichain. In cathepsin H, this partial occlusion of the active site confers aminopeptidase specificity.

Other proteases are targeted to the peroxisomes, organelles that bud from the ER membrane and primarily break down long-chain fatty acids, but are also involved in the synthesis of functional small molecules, such as isoprenoids, polyamines, and benzoic acid [137]. Some proteases in the peroxisome are involved in the maturation of other enzymes imported to this organelle, as well as disposal of oxidized proteins that build up in this challenging redox environment [138]. Others are active during different developmental stages, such as differentiation of seed glyoxysomes to mature leaf peroxisomes [139]. The most common type of targeting signal for transport to the peroxisome is one of several C-terminal tripeptides. The canonical example is SKL, but others have been discovered in a variety of plant proteins [140]. DCAP_5667, which is in the papain cluster (Figure 3), has the tripeptide SSM at its

extreme C-terminal end, indicating targeting to the peroxisome. DCAP_7656, which is in the granulin domain cluster (Figure 5), contains the SKL sequence not at the C-terminal end, but at a highly exposed position near the C-terminus, suggesting possible peroxisome targeting for this protein also. DCAP_7656 contains the proline-rich linker common to this cluster, but its granulin domain is truncated. Another possibility is that the short sequence region following the SKL tripeptide may be cleaved under some circumstances, acting as a switch that determines whether this enzyme is sent to the peroxisome or elsewhere. Peroxisome-targeted proteases represent attractive targets for biotechnological studies, because they are optimized to remain stable and maintain their activity under harshly oxidizing conditions.

Proteins with the sequence KDEL at the C-terminus are retained in the lumen of the endoplasmic reticulum, enabling them to be stored in specialized vesicles as zymogens and released to mediate programmed cell death in response to a stressor or during a particular developmental phase. KDEL-tailed proteases such as vignain from *Vigna mungo* and CysEP from *Ricinus communis* play an important role during germination, when proteins stored stored in endosperm tissue are degraded for use as the cotelydons develop. A C-terminal propeptide including the KDEL tag is removed along with the N-terminal pro-sequence during maturation, to yield the soluble, active enzyme [141]. The crystal structure and biochemical characterization of a homologous KDEL-tailed protein from the castor bean indicates that this enzyme has a strong preference for large, neutral amino acids in the substrate peptides, and has an unusually large and possibly flexible substrate-binding pocket that can accommodate a variety of sidechains, including proline [142].

4.3.4 Several Discovered Proteases Possess Novel Granulin Domains

Cysteine proteases with a C-terminal granulin domain are specific to plants, where they are involved in response to dessication or infection by pathogenic fungi [143]. This type of domain is found in two of the *D. capensis* protease clusters, the papain cluster Figure 3 and the granulin domain cluster Figure 5. The reference sequences RD21A (RD21A_ARATH) from arabidopsis and oryzain (ORYA_ORYSJ) from rice both contain granulin domains, as do three proteins in the papain cluster (Figure 3) and three in the granulin domain cluster (Figure 5). An additional two sequences in the papain cluster and one in the granulin domain cluster contain truncated versions that do not contain all four cysteine residues necessary to form the two disulfide bonds stabilizing the granulin domains. The granulin domain is separated from the catalytic domain by a proline-rich linker region. In RD21A, which is found in both the vacuole and the ER bodies [144], the granulin domain is removed from the mature enzyme. Maturation within the vacuole is relatively slow and involves accumulation of an intermediate where the N-terminal pro-sequence is removed and the C-terminal granulin domain remains attached [145]. This intermediate species forms aggregates that slowly release active enzyme following cleavage of the granulin domain, which is performed by RD21 itself [146]. This suggests that aggregation mediated by the granulin domain provides a mechanism for regulating protease activity during leaf senescence.

The granulin domain is attached to the catalytic domain by a proline-rich linker of variable length, as illustrated in Fig. 4.12. Granulins in animals act as growth factors, and contain distinct sequence and structural features: the characteristic sequence motif consists of four pairs of cysteine residues, with single conserved cysteines on both sides, and the resulting fold consists of β hairpins held together by disulfide bonds [147]. In plants, the granulin domain has two additional cysteines and an insertion of 6 residues between the first two

Cys pairs, slightly modifying the structure (Fig 4.13a). Clustering of the granulin domains themselves, separately from the catalytic domains, yields three clusters (Fig 4.13b), two of which contain proteins from the *D. capensis* papain cluster and one of which is made up entirely of proteins from the *D. capensis* granulin domain cluster. The cluster analysis of Richau, et al [7] identified two subfamilies of granulin domain-containing cysteine proteases; comparison with those results places DCAP_0302 in their XBCP3 cluster, while DCAP_5945 and DCAP_6547 are in their RD21A cluster. Notably, the *D. capensis* granulin domain cluster represents a new subfamily of plant cysteine proteases that is not closely related to either of the previously described subfamilies.

The key sequence region of the canonical animal granulin motif is shown above the sequence alignment for comparison (Fig 4.13c). The plant granulin sequences have two distinguishing features; an additional conserved Cys residue is present immediately after the first conserved CC pair in the animal sequence, and a 6-residue insertion containing another conserved C is present between the first and second CC pairs. In the granulin domain cluster, there is also a one-residue deletion between the first two conserved Cys residues. The first conserved glycine in the animal sequence is not conserved in the plant granulin domains, and in fact all of the examples shown here contain a bulky residue (F, Y, or L) at that position.

4.4 Conclusion

In summary, 44 cysteine proteases were identified directly from the genomic DNA of *D. capensis*, and sorted into clusters based on sequence homology to known plant cysteine proteases in our paper [30]. Molecular modeling and network analysis indicate that these proteases have distinct structural properties suggesting potential diversity in functional characteristics (e.g., thermal stability, substrate affinity). One particularly attractive potential

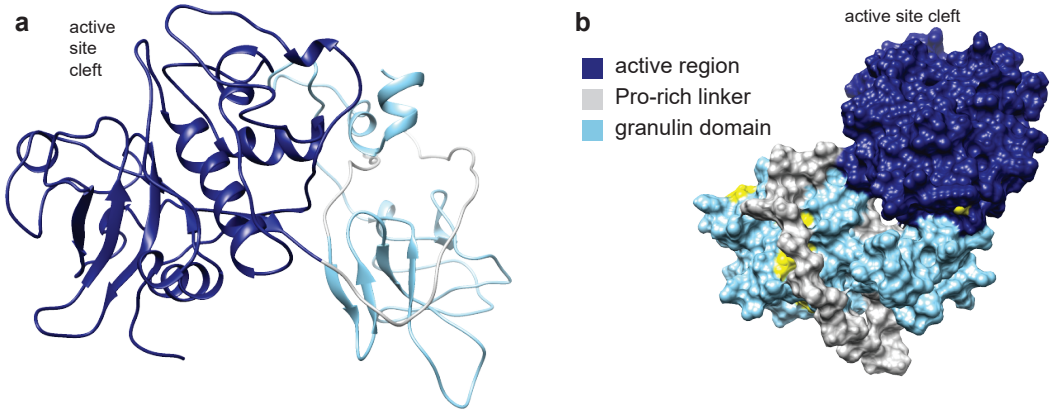

Figure 4.12: a. Ribbon diagram for the predicted structure for a representative member of the granulin domain cluster (DCAP_5115), showing the catalytic domain (dark blue), the proline-rich linker (gray) and the granulin domain (light blue). b. Surface representation of the same structure rotated to show how the proline-rich linker interacts with the granulin domain.

application for these proteases is in mass spectrometry-based proteomics. Identification and characterization of new proteases from diverse sources, including carnivorous plants, adds to the repertoire of cleavage patterns that can be used in proteomics research. The diverse properties make this class of proteins an attractive target for further characterization studies, with rich potential for biotechnology applications. Please look at the paper for detailed results, supplementary data, downloadable structures and author contributions.

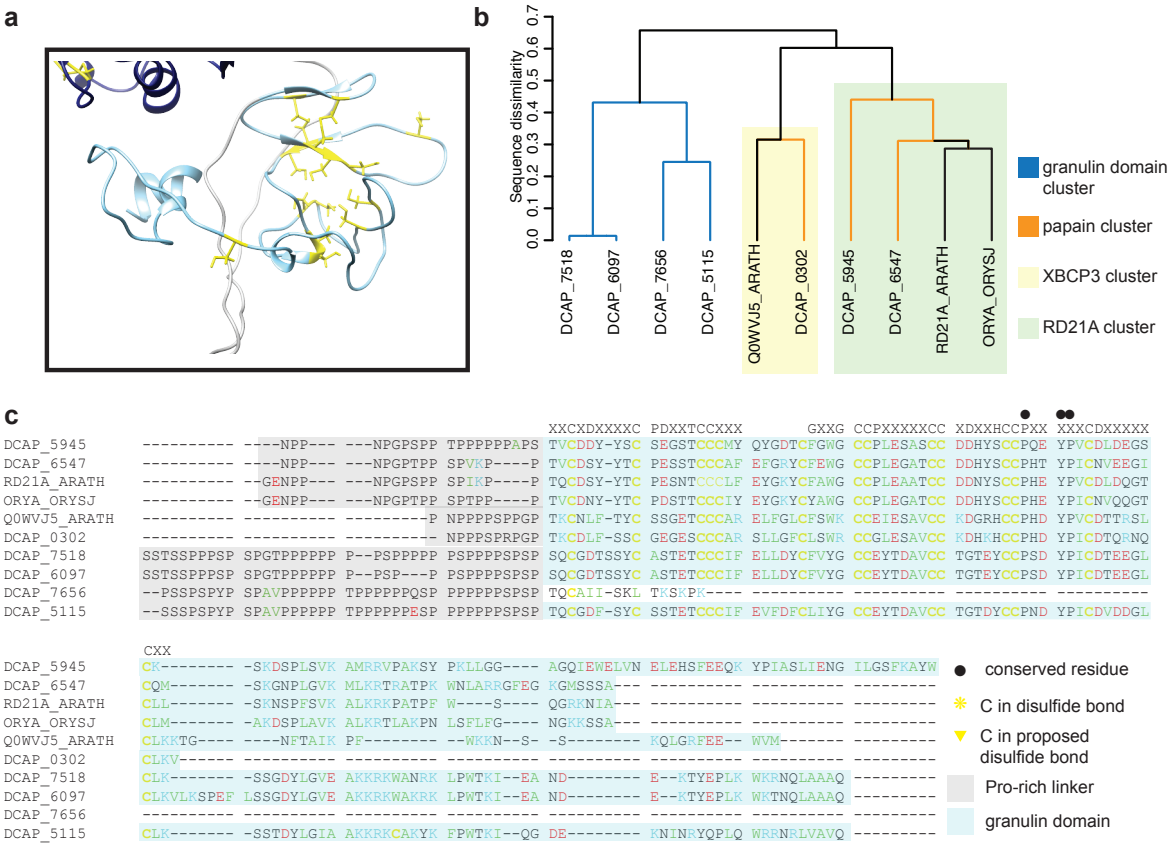

Figure 4.13: a. Ribbon diagram of the DCAP_5115 granulin domain, with cysteine residues highlighted in yellow. b. Cluster analysis of granulin domains from *D. capensis* cysteine proteases and reference sequences. Solid colors denote membership in the clusters of Fig 4.9, while the transparent boxes correspond to the clusters previously identified by Richau et al. [7]. Notably, the *D. capensis* granulin domain cluster appears to represent a new type of plant cysteine protease granulin domain. c. Sequence alignment of all the granulin domains found in the *D. capensis* cysteine proteases with reference sequences.

Chapter 5

Leveraging molecular modeling, experimental chemistry and bioinformatics to study amyloid fibril kinetics

5.1 Introduction

Amyloid fibrils are locally ordered protein aggregates that self-assemble under a variety of physiological and *in vitro* conditions. Their formation is of fundamental interest as a physical chemistry problem and plays a central role in Alzheimer's disease, Type II diabetes, and other human diseases. In fact, more than 40 different human diseases, many of which are both fatal and incurable, are associated with the formation of amyloid fibrils [8, 9, 148, 149]. These locally ordered protein aggregates are characterized by a cross-β structure, in which β-strands composed of adjacent monomers stack in a repeating linear pattern, analogous to

a crystal that grows along a single dimension rather than in all three [150]. This characteristic structure is conserved across amyloid fibrils, independent of the secondary structure of the protein monomer prior to fibrillization [151]. Despite the universality of this overall arrangement, the patterns of connectivity among monomers that form the repeating subunits of amyloid fibrils vary widely and are incompletely characterized. Indeed, the same fibrillizing monomer has been found to form different periodic structures in different experiments, ruling out simple sequence-based explanations: for example, the PDB structures 5KK3 [9] and 2BEG [152] exhibit markedly different periodic patterns along their respective growth axes, yet both fibrils are comprised of the same Aβ_{1-42} peptide. High-resolution structures have shown a diverse set of fibril subunit geometries displaying subtle but distinctive differences, e.g., linear vs. annular structures and parallel vs. antiparallel β-sheet arrangements [8, 153, 154, 155]. These structural differences have demonstrated clinical relevance: for instance, they have been shown to directly correlate with toxicity and disease progression for strains of both β-amyloid [156] and α-synuclein [157] fibrils. The key feature differentiating these subunit geometries from each other is the periodic pattern of non-covalent bonding between monomers. This study refers to such motifs of non-covalently bonded connectivity among the protein monomers making up each fibril type as *fibril topology*, the characterization and modeling of which are the focus of the present work.

As an initial illustration of how fibril topology can be extracted from high-resolution structures and specified using graph theoretic formalisms, Figure 5.1 compares the patterns of non-covalent connectivity for human Iowa mutant β-amyloid fibrils, associated with an early-onset hereditary form of Alzheimer's disease (PDBID:2LNQ [8]), with that of a fibril formed by wild-type Aβ_{1-42} (PDBID:5KK3 [9]). As detailed below and in our paper [14], the study represents the topology of each fibril by a network in which each node represents a single protein monomer, with ties indicating monomers that are non-covalently bound to one another. Application of this coarse-grained representation to solved amyloid structures demonstrates that it is sufficient to distinguish a wide range of fibril topologies while also

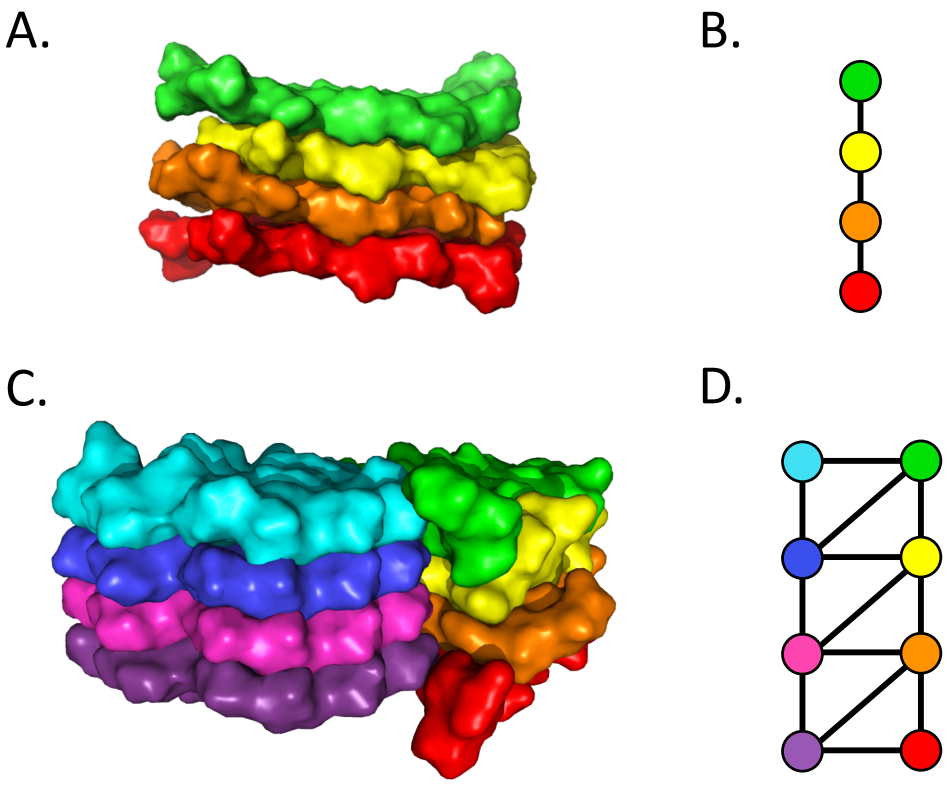

Figure 5.1: Two examples of the mapping of 3-dimensional fibril structures into their equivalent graph representations, where the color coding indicates different protein monomers. Each node in panels B and D corresponds to a protein monomer, with ties between nodes whose monomers are non-covalently bound. Panels A. and B. show the molecular structure and graph representations, respectively, of a fibril segment formed from β-amyloid D23N (PDBID:2LNQ [8]). Using the typology developed in this paper, this fibril structure is classified as a 1-ribbon. Panels C. and D. show the molecular structure and its corresponding graph representation for a segment of wild-type Aβ_{1-42} (PDBID:5KK3 [9]). In our typology, this structure is classified as a 1,2 2-ribbon.

being compact enough to serve as the basis for scalable models of fibrillization kinetics that are able to simulate the fibrillization of hundreds or thousands of protein monomers. The study's examination of experimentally solved amyloid fibril structures reveals several distinct and previously unrecognized topological classes of fibrils; to describe them, this study introduces a systematic nomenclature for fibril structure akin to the naming conventions for organic polymers or protein secondary structure. This system, demonstrated in Figure 5.5, describes all presently known fibril topologies, and is extensible to others not yet found.

A second motivation driving the present work is the need for models of fibril formation that

are able to represent the diversity of observed structures and to capture fibrillization kinetics on experimentally relevant time scales (many hours) and system sizes (hundreds to thousands of peptide monomers). Because both the number of degrees of freedom and the timescales of these events extend far beyond the reach of current atomistic methods, this study developes a mathematical formalism for a *network* Hamiltonian describing the physics of interactions between fibrillizing protein monomers in graph theoretic, or *connectivity* terms rather than in atomistic detail; our approach exploits a formal connection between the statistical mechanics of the fibrillization process and a framework for network modeling (Exponential family Random Graph Models, or ERGMs) originally developed for studying social networks. This study also describes simulation methodology that generates representative fibril topologies under this Hamiltonian, recapitulating the topologies of all currently known amyloid fibril structures reported in the Protein Data Bank (PDB [158]).

In the remainder of this paper [14], we: 1) introduce a general topological formalism for protein aggregates (fibrillar or otherwise), including a quantitative mapping from atomic-resolution structures to their topological representations; 2) introduce a typology and nomenclature for amyloid fibrils that encompasses all fibril structures found in the Protein Data Bank; 3) introduce a specific family of Exponential family Random Graph Models (ERGMs) that can reproduce all currently known amyloid fibril topologies (as well as some not yet experimentally observed); 4) employ Markov-chain Monte Carlo (MCMC) simulations to examine the equilibrium behavior of our models; and 5) employ a dynamic extension of ERGMs based on local energy differences to probe the kinetics of fibril formation.

Our work has resulted in two papers [14] and [159]. I analyzed every amyloid fibril structure from the PDB, was involved in topological classification and wrote the manuscript with my coauthors. Dr. Gianmarc Grazioli and Dr. Yue Yu developed and implemented the quantitative topological classification method and simulation studies with Prof. Carter Butts and Prof. Rachel Martin. A portion of the papers [14, 159] is reproduced in this chapter

to understand the importance of the results found in this study. For more details, in-depth analysis and author contributions, please refer the papers [14, 159].

5.2 Network-based Classification and Modeling of Amyloid Fibrils

5.2.1 Methods

Defining Protein Aggregation States with Graphs

To capture the process of fibril formation, we require a simple representation that can accommodate protein monomers, small oligomers, larger unstructured aggregates, and the repeating units of fibrils themselves. Graph theory provides a natural language for this purpose, and enables use of the ERGM formalism that has been extensively developed for modeling graphs in social network and neuroscience applications to simulate realizations of the aggregation process.[160] A *graph* is composed of a set of *nodes* or *vertices*, together with a set of *ties* or *edges* representing pairwise interactions between nodes. Here, the nodes represent individual protein monomers, with two nodes being joined by an edge if their corresponding monomers are non-covalently bonded. We refer to such structures as *aggregation graphs*. Because our focus is on fibril topology, we do not distinguish among types of intermolecular interactions (e.g. hydrogen bonds, salt bridges, non-polar interactions, etc.); in practice, protein monomers are held together by a combination of different interactions, whose combined effects are of primary interest. As Figure 5.1 illustrates, a topological representation is rich enough to distinguish among different fibril forms. Additional forms are shown in the examples throughout this work. In the case of fibril structure A. from Figure 5.1, we note that the structure displays a motif in which each protein monomer is bonded exclusively to

its two immediate neighbors along the fibril growth axis; the minimal repeating subunit is a single protein monomer. In contrast to this, the latter structure is characterized by two chains of linear growth, whereby the minimal repeating subunit is a pair of monomers (one from each chain, and of equal fibril growth axis index) that are bonded to: 1) each other; 2) both of their neighbors along the same chain; and 3) the monomer on the opposite chain with a fibril growth axis index offset by one and opposite in sign. Although our representation is by intention coarse-grained, it has some advantages over (and can be used in conjunction with) more fine-grained approaches.

Directly modeling aggregation states via topology dramatically reduces the degrees of freedom that must be explicitly represented, allowing substantial gains in computational efficiency versus geometric representations. Modeling aggregation states in this fashion, however, requires a different approach from e.g. atomistic methods. The above-mentioned ERGM framework provides a parsimonious means of parameterizing and simulating draws from probability distributions on graphs, in this case representing patterns of connectivity between protein monomers. Here, we exploit a property of this framework that allows us to relate the ERGM specification to a Boltzmann distribution over aggregation states (Eq. 5.1).

$$\Pr(G = g|\theta, t) = \frac{\exp\left(\theta^T t(g)\right)}{\sum_{g' \in \mathcal{G}} \exp\left(\theta^T t(g')\right) h(g')} h(g), \tag{5.1}$$

This correspondence also provides a basis for defining Hamiltonian functions that describe the energy of formation for all graph theoretic features sufficient for recapitulating a particular

aggregation state (Eq. 5.2).

$$\begin{aligned}\mathcal{H}(g) = (\phi_e + k_B T)\, t_e(g) + \phi_{2s} t_{2s}(g) + \phi_{NSP1} t_{NSP1}(g) + \phi_{NSP2} t_{NSP2}(g) \\ + \phi_{ESP0} t_{ESP0}(g) + \phi_{ESP1} t_{ESP1}(g) + \phi_{C5} t_{C5}(g) + \phi_{C6} t_{C6}(g) + \phi_{C7} t_{C7}(g).\end{aligned} \tag{5.2}$$

Finally, we employ the graph Hamiltonian to define a family of kinetic models whose equilibrium distributions correspond to a target ERGM distribution, and that we show to be able to qualitatively recapitulate behaviors seen in experimental studies. Further details regarding the mapping of aggregation states to graphs and the statistics describing their formation to ERGMs can be found in the paper [14]. Although the ERGM portion was performed by Giannmarc, Yue and Prof. Butts, I had to mention it here to understand the importance of the results in the next section. For more details, please look at the paper [14].

Identifying Amyloid Fibril Structures from the Protein Data Bank

To examine the diversity of fibril topologies observed to date, I performed an exhaustive search of the PDB for all amyloid fibril structures. Although the PDB is not a random sample from nature, it is a reasonable census of the known fibril structures so far discovered by the scientific community and their classification into topological classes is useful for showing what fibril forms have been found. The fibril structures were downloaded from the Protein Data Bank (PDB; http://www.rcsb.org/pdb/) [158] after running multiple advanced searches on the website. The search criteria included finding fibrils of sufficient size that the repeating subunit could be clearly identified. Search terms used included "Fibril," "A-beta," "Protein Fibril," and "Lysozyme." Any fibril attached to a ligand, metal or macro-scaffold was discarded. To be retained for analysis, a putative fibril structure was required to display 1) the cross-beta sheet structure definitive of amyloid fibrils (distances between adjacent

β-strands of 4.7 +/- .4 Å with 10 +/- 3 Å between β-sheets) as seen in Figure 5.3(a) [161], and 2) a periodic pattern of connectivity between monomers along a single fibril growth axis as seen in Figure 5.3(b,c). Structures not satisfying these conditions were discarded. Then, the structures were inspected individually in VMD [162] and Chimera [163]. Using PDB ID 4RIK [164] for representation in Figure 5.3 (b) and (c), each structure was color coded and mapped on the basis of topology using simple ball and stick model. The structure was also checked for its correctness, any structure that was repeating as a crystal, or did not conform to the definition of an amyloid fibril, was discarded. Figure 5.4 highlights the criteria of acceptance. Figure 5.4 panel (1a) shows the accepted amyloid fibril structure, PDB ID 4RIK, panel (1b) shows the supercell constructed in Pymol [12] and panel (1c) shows the amyloid fibrils distances between adjacent β-strands of 4.8 Å with 9.4 Å between β-sheets [11]. Figure 5.4 panel (2a) shows the rejected fibril structure, PDB ID 4RXFO, panel (2b) shows the supercell constructed in Pymol [12] and panel (2c) shows the fibrils distances between adjacent β-strands of 4.4 Å with 7.8 Å between β-sheets [13]. In case of 4XFO, the cross-layer contacts look really weak and 4XFO did not pass energy calculations performed by Dr. Gianmarc Grazioli, details can be found in the paper [14]. Figure 5.4 panel (3a) shows the rejected fibril structure, PDB ID 3FOD, panel (3b) shows the supercell constructed in Pymol [12] and panel (3c) shows the fibrils distances between adjacent β-strands of 5.3 Å with 10-11.4 Å between β-sheets [15], which does not conform to the amyloid fibril criteria of the amyloid fibrils distances between adjacent β-strands of 4.8 Å with 9.4 Å between β-sheets [11]. 3FOD did not pass energy calculations performed by Dr. Gianmarc Grazioli, details can be found in the paper [14].

This individual inspection not only help collect the accurate data samples to study but also formed the basis of GUI and a traditional interface developed by Dr. Gianmarc Grazioli and Dr. Yue Yu that automated the process of identifying and defining amyloid fibrils found in the PDB and can be found in our paper [14]. In addition to the summary of our classification results shown in Figure 5.2, a detailed list of structures satisfying our criteria is given in Table

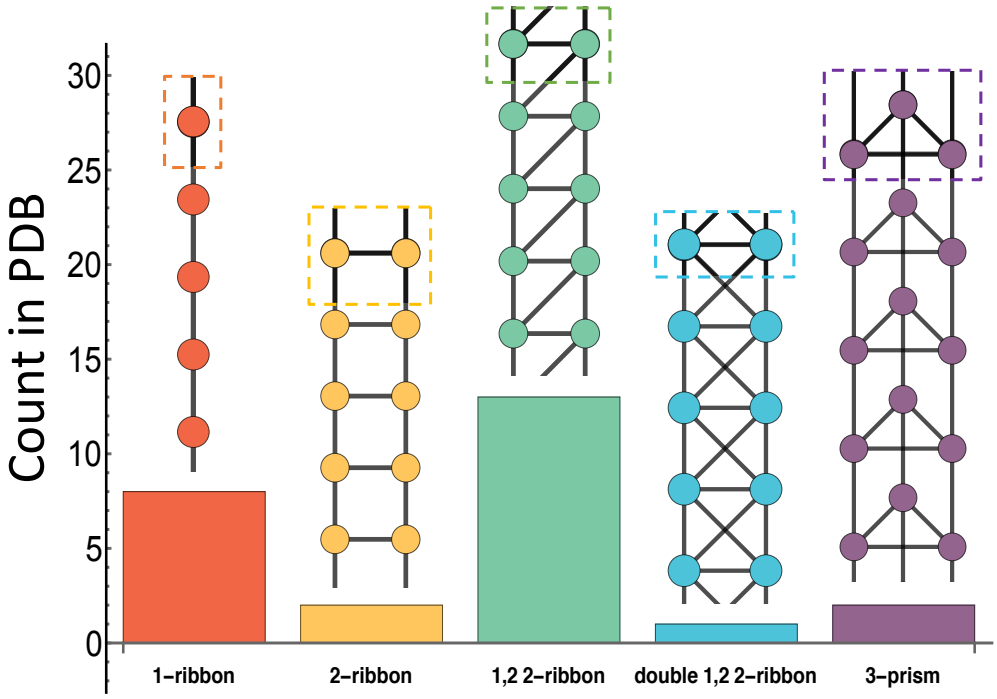

Figure 5.2: The five unique amyloid fibril topologies found in the PDB, sorted by relative complexity. Bar heights indicate the number of structures in the PDB with the indicated topology.

5.1.

Supplementary Video

We have provided a narrated video that features an animated visualization of one of our kinetic simulations of fibrillization, as well as a brief introduction to this body of work. The file name for the video is FibrilTopologyMovie.mp4. Additionally, a high resolution version of the same video can be viewed via the following link: Fibril Topology Movie.

5.2.2 Fibril Nomenclature Rules for Chords

By default, it is assumed that each monomer within a repeating unit is adjacent to the corresponding monomer in the next unit along the fibril axis. Where additional adjacencies are present, these are indicated via the specification of cross-cutting ties that we refer to as

Table 5.1: The following table is a comprehensive list of all fibrillar structures found in the PDB at the time this analysis was done, as well as their topological classification.

PDB ID	**Fibril Type**	**Description**	**Reference**
2BEG	1-ribbon	Aβ 1-42	[152]
2E8D	1-ribbon	β2-microglobulin	[165]
2LMQ	1-ribbon	Aβ 1-40	[166]
2LMP	1-ribbon	Aβ 1-40	[166]
2LNQ	1-ribbon	Aβ D23N 1-40	[8]
2MXU	1-ribbon	Aβ 1-40	[167]
2NNT	1-ribbon	TTCERG1 Y22F CA150	[168]
5OQV	1-ribbon	Aβ 1-42	[169]
2LMN	2-ribbon	Aβ 1-40	[166]
2LMO	2-ribbon	Aβ 1-40	[166]
2M5N	1,2 2-ribbon	TTR 105-115	[11]
2M5K	1,2 2-ribbon	TTR 105-115	[11]
2M5M	1,2 2-ribbon	TTR 105-115	[11]
2OMQ	1,2 2-ribbon	Insulin 12-17	[170]
2OMM	1,2 2-ribbon	Yeast Prion sup35 7-13	[170]
3FTK	1,2 2-ribbon	IAPP 64-70	[171]
3FVA	1,2 2-ribbon	Elk Prion	[171]
3ZPK	1,2 2-ribbon	TTR 105-115	[11]
4R0U	1,2 2-ribbon	α-synuclein 72-78	[172]
4RIL	1,2 2-ribbon	α-synuclein 68-78	[164]
4ZNN	1,2 2-ribbon	α-synuclein 47-56	[164]
5K2H	1,2 2-ribbon	Yeast Prion sup35 7-13	[170]
5KK3	1,2 2-ribbon	Aβ 1-42	[9]
2MVX	double 1,2 2-ribbon	Aβ 1-40	[173]
2MPZ	3-prism	Aβ D23N 1-40	[174]
2M4J	3-prism	Aβ 1-40	[156]

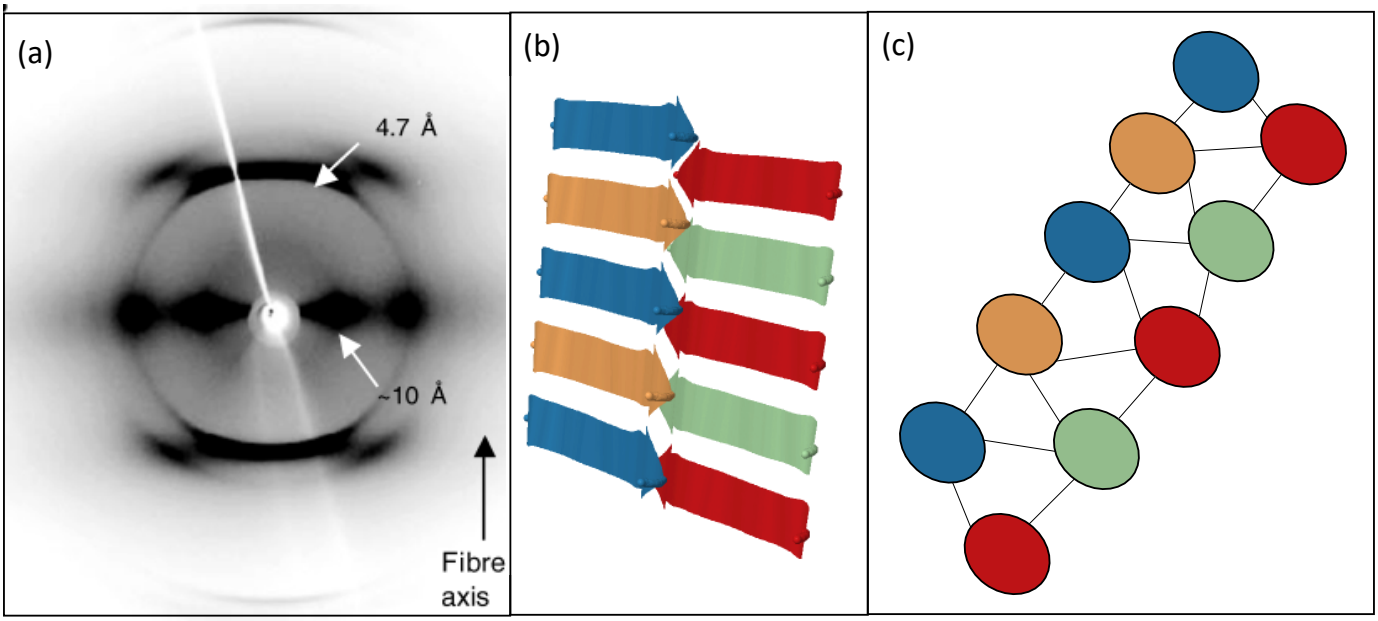

Figure 5.3: Panel (a) represents the characteristics of amyloid fibril appearance. X-ray fiber diffraction pattern from aligned IAPP amyloid fibrils, showing the positions of the 4.7 Å meridional and 9.8 Å equatorial reflections in a cross- pattern. The figure is taken from the paper [10]. Using PDB ID 4RIK [11], (c) is the color coded topology representation of (b).

chords. A chord is indicated by a pair of numbers, indicating the repeating units joined by the chord in question. The first number is always 1, referring to the focal unit; succeeding units are numbered as 2, 3, etc. Thus, a chord from the focal unit to the next unit is a 1,2 chord. A chord from the focal unit to the unit after next is a 1,3 chord, and so on. By default, a chord is assumed to connect each element in the focal unit not having an incoming chord of the same type from a previous unit to the next vertex along the ribbon or prism in the unit to which it connects. Where specified explicitly, these are called *trans* chords. When a chord connects to the corresponding element in a subsequent unit, it is known as a *cis* chord. (Fig. 5.5 gives examples of both trans and cis chords.) Hence a 2-ribbon in which the first vertex in the ribbon in a focal unit is adjacent to the second ribbon vertex in the next unit is called a trans-1,2 2-ribbon, or simply a 1,2 2-ribbon (a cis-1,2 chord would be a reference to an inherent edge between subunits, thus all all 1,2 chords must be trans, ergo trans is a redundant qualifier for a 1,2 chord). Where chords connect both previous and subsequent elements in the ribbon or prism structure, the chords are said to be *doubled*.

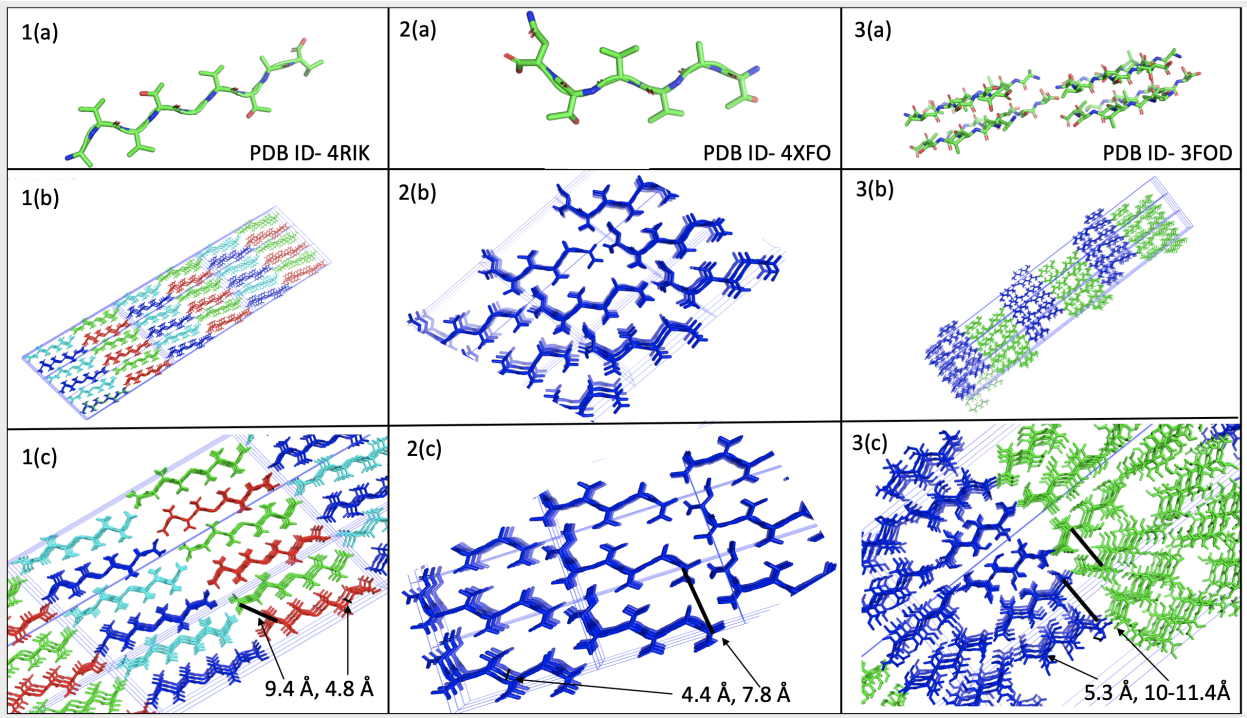

Figure 5.4: Panel (1a) shows the accepted amyloid fibril structure, PDB ID 4RIK, panel (1b) shows the supercell constructed in Pymol [12] and panel (1c) shows the amyloid fibrils distances between adjacent β-strands of 4.8 Å with 9.4 Å between β-sheets [11]. Panel (2a) shows the rejected fibril structure, PDB ID 4RXFO, panel (2b) shows the supercell constructed in Pymol [12] and panel (2c) shows the fibrils distances between adjacent β-strands of 4.4 Å with 7.8 Å between β-sheets [13]. In case of 4XFO, the cross-layer contacts look really weak and 4XFO did not pass energy calculations performed by Dr. Gianmarc Grazioli, details can be found in the paper [14]. Panel (3a) shows the rejected fibril structure, PDB ID 3FOD, panel (3b) shows the supercell constructed in Pymol [12] and panel (3c) shows the fibrils distances between adjacent β-strands of 5.3 Å with 10-11.4 Å between β-sheets [15], which does not conform to the amyloid fibril criteria of the amyloid fibrils distances between adjacent β-strands of 4.8 Å with 9.4 Å between β-sheets [11]. 3FOD did not pass energy calculations performed by Dr. Gianmarc Grazioli, details can be found in the paper [14].

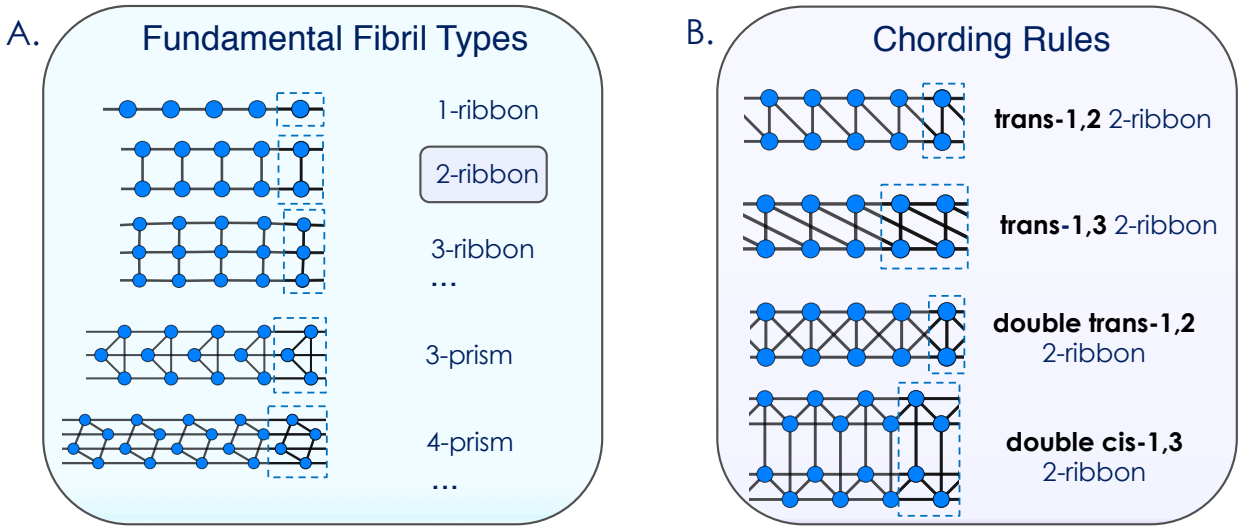

Figure 5.5: The topology of all amyloid structures can be described using a simple network framework. Shown in A are the fundamental fibril forms: the n-ribbon and the n-prism. These fundamental forms are the basis for describing any fibril by either adding (chording) or deleting (nulling) edges between nodes in a repeating pattern. In B, we demonstrate various chording operations to the 2-ribbon. Chords are indexed by the subunits they connect, e.g. consecutive chorded subunits are labeled 1,2, while subunits two positions apart are labeled 1,3. Cis- and trans- indicate whether chords are between subunits occupying equivalent or different embedded 1-ribbon "backbones," respectively.

A double trans-1,2 2-ribbon hence has chords from both vertices within each repeating unit to the opposite vertex in the next repeating unit. When multiple chords types are present, they are listed sequentially from longest to shortest; one may thus have e.g. a cis-1,3 double trans-1,2 2-ribbon, or any other combination that forms a valid aggregation graph.

Extracting Topology from Atomistic Protein Structures

After identifying an amyloid fibril structure, we then extract its underlying topology (i.e., the pattern of bound interactions among monomers). For structures derived from x-ray crystallography, this requires distinguishing between crystal contacts and the much stronger interactions that define the fibril structure itself; in the case of NMR and EM structures, the corresponding problem is distinguishing ephemeral and dynamically unstable contacts from structural ones. In both cases, we employ an energy scoring protocol to remove spurious

contacts.

Although we use the same scoring scheme for all structures, the definition of crystal structures in terms of a repeating asymmetric unit requires special processing. This is performed as follows. To ensure that we are working with a collection of fibril segments, we first generate supercells consisting of multiple repeats of the reported crystal unit cell [12]. We then calculate the score for the total interaction energy between each spatially adjacent monomer in the supercell (using the approach described below), using a -10 kcal/mol cutoff to filter out spurious contacts. If this yields a periodic fibril structure along a single axis (and not, e.g. a three-dimensional, sheet-like, or non-periodic structure), the resulting ties are taken to define the fibril topology. If this does not yield such a structure, then we lower the energy cutoff until either such a structure is observed, or until the structure decomposes into independent monomers; in the latter case, the structure is considered not to meet our criteria of being composed of a repeating one-dimensional pattern of bound monomers, and is rejected. Although the focus of the current work is fibril units held together by strong interactions, we note that this methodology can be extended to more complex treatments in which valued graphs are used to represent strong and weak interactions. In this case, multiple scoring thresholds would be used to classify edges into their respective classes.

Since the NMR and EM structures employed here were not obtained from crystallized protein, we do not need to build supercells prior to scoring inter-monomeric interactions. Instead, we simply score contacts among spatially adjacent monomers within the structure, as before beginning with a -10 kcal/mol cutoff and lowering the energy threshold until a periodic structure is obtained. For the NMR and EM structures examined here, it was not necessary to lower the default cutoff in order to obtain periodic structures. Most of this work was done by Gianmarc and Yue, and further details can be found in the paper [14].

Computational Experiments

Other experiments perfomred in this paper, mainly by Gianmarc and Yue include the following and further details can be found in [14]- Binding Energy Scoring Protocol for Edge Determination, Kinetic Extension of the Fibrillization ERGM, Statistical Mechanics of the Aggregation Graph, Aggregation Model Reference Measure, Aggregation Model Behavior at Extreme Temperatures, Extended Stability Testing of Fibrils, Definition of Fibril Epochs and Sample Code for 2-Ribbon Simulation.

5.2.3 Results and Discussion

A Systematic Nomenclature for Fibril Topology

In the study of individual protein structures, the paradigm of secondary structure has provided researchers with a tool for concisely describing structural details of proteins based on commonly observed hydrogen bonding patterns. Here we present an analogous, systematic, and standardized nomenclature system for fibril topologies that encompasses all of the diverse amyloid fibril forms currently represented in the PDB, and which can straightforwardly be extended to describe forms yet to be discovered (Figure 5.5). The goal of the present study is to develop a formalism for describing amyloid fibrils purely in network terms, using connectivity among monomers. Although it differs in the details, the general approach is part of the rich tradition in computational chemistry of adopting coarse-grained models that capture the most salient features of the system of interest, while discarding other details that are judged to add complexity but not essential information [175, 176]. Such coarse-grained representations of protein systems often necessitate use of a specialized force field, whereby the coarse-grained mapping is performed in the most literal sense, i.e. atomistic degrees of freedom deemed to be unnecessary for capturing the phenomena of interest are fused

together to create the "coarse grains" that compose less computationally expensive models with fewer degrees of freedom.[177, 178] We address this objective below.

Amyloid fibrils are characterized by interesting structural features at multiple scales, from the atomistic details of the molecular structures all the way to the distribution of plaques in the brain. [179] The former have been the subject of considerable experimental investigation, leading to details about the shapes of individual monomers within both parallel and antiparallel assemblies. Types of fibrils include S- [9] (or tilde) [180] and double-horseshoe [181] shapes, along with more complicated supramolecular assemblies. Our particular coarse-grained model focuses on fibril topology at the level of interacting monomers, thus opening the door to similar investigations at a higher level of structure, including potential relationships with disease etiology (to which other features are hypothesized to be related [182]) and connections between atomistic and topological features. Because all fibrils, by definition, have a unique axis along which fibril growth occurs, we first posit that all fibrils are chains of repeating subunits that are non-covalently bonded end to end. Each unique fibril topology is identified by this one-dimensional repeating pattern of linked subgraphs, as summarized in Figure 5.5. Our convention for visually indicating the axis of fibril growth in the graphs is to depict "stubs" (i.e., endpoints of cross-unit edges) at both ends of the fibril segments shown. In all examples, the repeating pattern of nodes, edges, and stubs that defines the repeating unit of the fibril form is enclosed with dotted lines.

The subunit topologies observed to date can be divided into two categories: toroidal (i.e., cyclic) and linear. If the subunit is linear, we call it an n-ribbon, and if it is toroidal we call it an n-prism (see Figure 5.5). For example, if a subunit comprises 3 monomers i, j, and k, and the edges they share within the subunit can be represented as $\{i \leftrightarrow j, j \leftrightarrow k\}$, the fibril type is a 3-ribbon. On the other hand, if the set of edges between monomers i, j, and k in a single subunit are $\{i \leftrightarrow j, j \leftrightarrow k, k \leftrightarrow i\}$, the fibril type is a 3-prism. The simplest case, where the repeating subunit is a single monomer, is called a 1-ribbon.

We can further extend the naming convention to describe ties between monomers from different unit cells. We introduce a convention for describing these additional ties of the form i, j, where i and j give the indices of the subunits between which the ties occur. Additionally, if the ties between unit cells connect nodes along the same embedded ("backbone") 1-ribbon, we add the term *cis*, while the term *trans* indicates that the tie is between nodes on different ribbon segments. For example, if we have a 2-ribbon, where an additional tie connects each unit cell to its immediate neighbor (a 1,2 tie), and the tie is between nodes that belong to the opposite constituent 1-ribbons that make up the 2-ribbon (trans), the 2-ribbon becomes a trans-1,2 2-ribbon (see Figure 5.5).

Characterizing Fibril Topologies Observed in Nature

In order to demonstrate that our system of nomenclature is general enough to be useful, we have categorized the complete set of amyloid fibrils in the Protein Data Bank by topology (Fig. 5.2). The amyloid fibril type classification follows a two-step protocol (see Methods). First, we confirm that the PDB entry possesses the characteristic cross-beta sheet structure definitive of amyloid fibrils (i.e. distances of 4.7 +/- .4 Å between strands within β-sheets and 10 +/- 3 Å across β-sheets [161]). We carry out this step using the distance measurement tool in PyMOL [12]. Second, we apply an energy scoring criterion to identify monomer pairs whose interactions are sufficiently strong to constitute an edge in the aggregation graph. The interaction energy score between two monomers is calculated in a style similar to free energy difference scoring commonly practiced in molecular docking simulations. A conservative minimum energetic threshold of -10 kcal/mol is employed to ensure that the inter-monomeric interactions included in the graph representation of the fibril are at least an order of magnitude more enthalpically favorable than a typical hydrogen bond between water molecules. For the 28 amyloid fibril structures solved to date that meet our inclusion criteria, we find 5 unique topological classes (Figure 5.2). We have identified simulation

parameters for our mathematical model of fibril formation that are capable of recapitulating all 5 of these known amyloid fibril topologies, as described in the following sections.

Computational Experiments and Simulation Results

Prof. Carter T. Butts, Dr. Gianmarc Grazioli and Dr. Yue Yu were successful in recapitulating fibril structure in equilibrium by adapting exponential family random graph (ERGM) techniques originally developed for the study of social networks. Details on how we employed the ERGMs using a Hamiltonian function and Markov Chain Monte Carlo (MCMC) can be found in the paper [14] [183, 160]. We were also successful in constructing the fibrilization kinetics simulations [14]. Our followup study to the above paper [14] deals with the mechanisms of amyloid fibril formation, a problem of considerable interest in terms of both basic science and medical applications [159]. Our distinctive approach is to employ a graph-theoretic formalism to represent the underlying topology of fibril structure, giving us a parsimonious way of describing fibrils that is nevertheless flexible enough to accommodate a wide range of aggregation states [159]. Further details can be found in the papers [14, 159]

5.2.4 Experimental Validation of the Predicted Statistical Model: *in vitro* Amyloid Fibril Kinetics

Network statistical characterization of the pathways to fibrillization predicted by network Hamiltonian models [14, 159] offer a uniquely capable approach toward identifying potential intermediates in the process of amyloid fibril formation for experimental validation or invalidation. Most papers used two-step nucleation induced mechanism to study amyloid fibril kinetics [184, 185]. To understand the disease and treatment better, we need to understand the actual process, including formation of monomers, to better understand pathology, process and disease treatment and management As I want to study the process of amyloid fibril

kinetics, I propose to use unseeded, no nucleation method to understand the kinetics better.

Materials and Methods

Hen egg white lysozyme (HEWL), Enzyme Commission (EC) Number 3.2.1.17, CAS molecular weight 14.4 kDa, was obtained from Millipore Sigma. γ-crystallin was expressed in the lab [186].

Sample Preparation- 1 mg/ml, 2 mg/ml or 2.5 mg/ml of HEWL or γS-crystallin in different buffers were made by vortexing in a test tube. Tested buffers included phosphate buffer (pH 2.15, 6.8, 12) and carbonate buffer (3.2, 9.2) as amyloid fibrils tend to form generally at extreme pH, temperatures and agitation [184, 185, 187, 188]. Aliquots of 120 μL of sample were taken in an eppendorf tube, placed in the Digital Heating Shaking Drybath (Thermofisher) with or without shaking at temperatures between 70-76 degree celcius.

Thioflavin Binding Experiments- Thioflavin binding-dependent fluorescence was routinely assessed on kinetic time points as described in literature [184, 185, 187, 188]. ThT was sourced from Sigma Aldrich. 10 μL aliquot of a 2.5 mM stock of ThT was added to 90 μL to the aliquot of the amyloid fibril sample in fluorescence 96-well plate. The fluorometer was set to an excitation wavelength of 450 nm and an emission wavelength of 489 nm. Calculations were performed using excel and blank and negative controls were used.

Results

Although incomplete, the results seen below are promising and the project will continue on the experimental path. Some results are presented in Figure 5.6, 5.7 , 5.8 , 5.9 and 5.10 General trend shows a minimum concentration 2 mg/ml at pH 3.7 with agitation (150 RPM) at 75 degree Celsius shows propmising results. In Figure 5.6, we see ThT assay results of

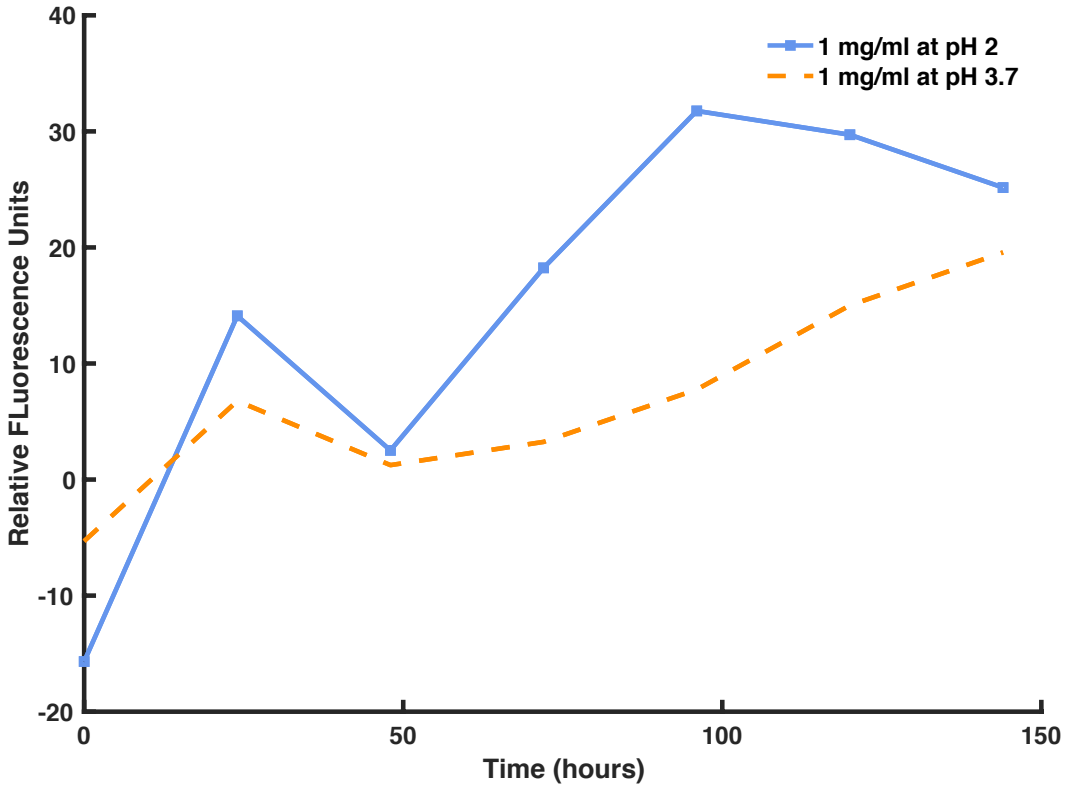

Figure 5.6: ThT assay results of 1 mg/ml HEWL using phosphate buffer at pH 2 and pH 3.7 with agitation (150 RPM) at 75 degree celcius.

1 mg/ml HEWL using phosphate buffer at pH 2 and pH 3.7 with agitation (150 RPM) at 75 degree Celsius. Some amyloid fibril growth is observed however my hypothesis is the fibrils crashes out of the solution. In figure 5.7, ThT assay results of 1 mg/ml and 2 mg/ml γS-crystallin using phosphate buffer at pH 1 and pH 2 with agitation (150 RPM) at 75 degree Celsius. In Figure 5.8, ThT assay results of 1 mg/ml and 2 mg/ml γS-crystallin using phosphate buffer at pH 2 and pH 3.7 with agitation (150 RPM) at 75 degree Celsius. In Figure 5.9, ThT assay results of 0.5 mg/ml and 1 mg/ml HEWL using carbonate buffer at pH 2 and pH 3.7 with agitation (150 RPM) at 75 degree Celsius. In Figure 5.10, ThT assay results of 1 mg/ml γS-crystallin using phosphate buffer at pH 3.7 with agitation (150 RPM) and without agitation at 75 degree Celsius.

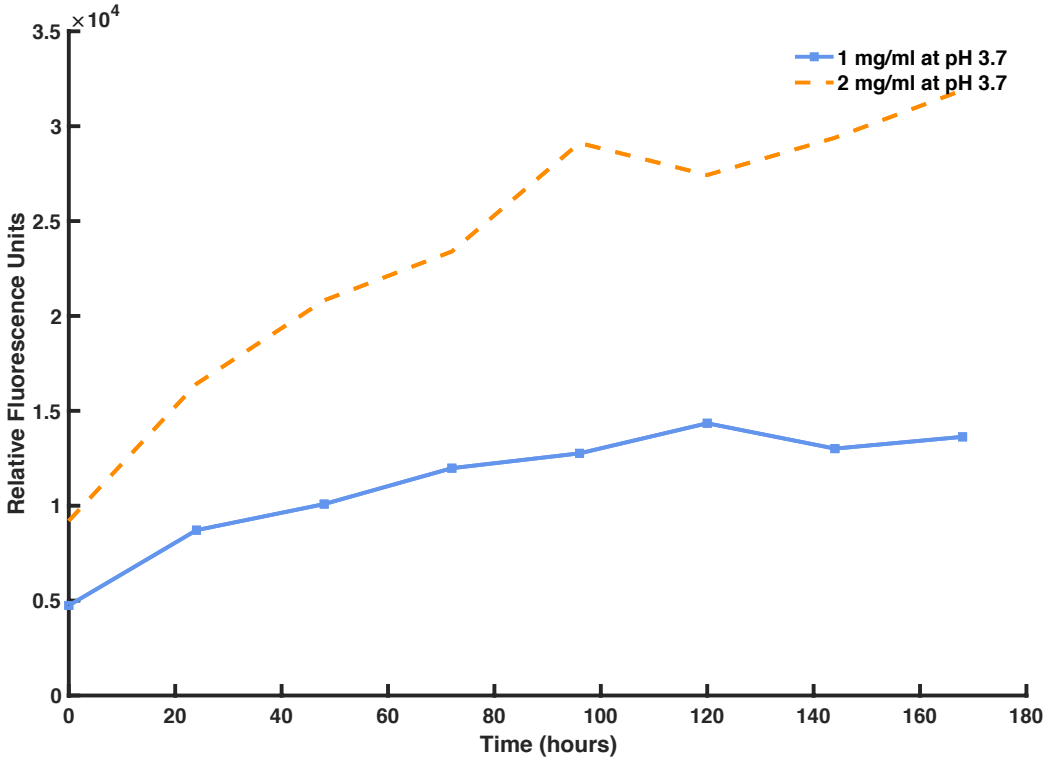

Figure 5.7: ThT assay results of 1 mg/ml and 2 mg/ml γS-crystallin using phosphate buffer at pH 1 and pH 2 with agitation (150 RPM) at 75 degree celcius.

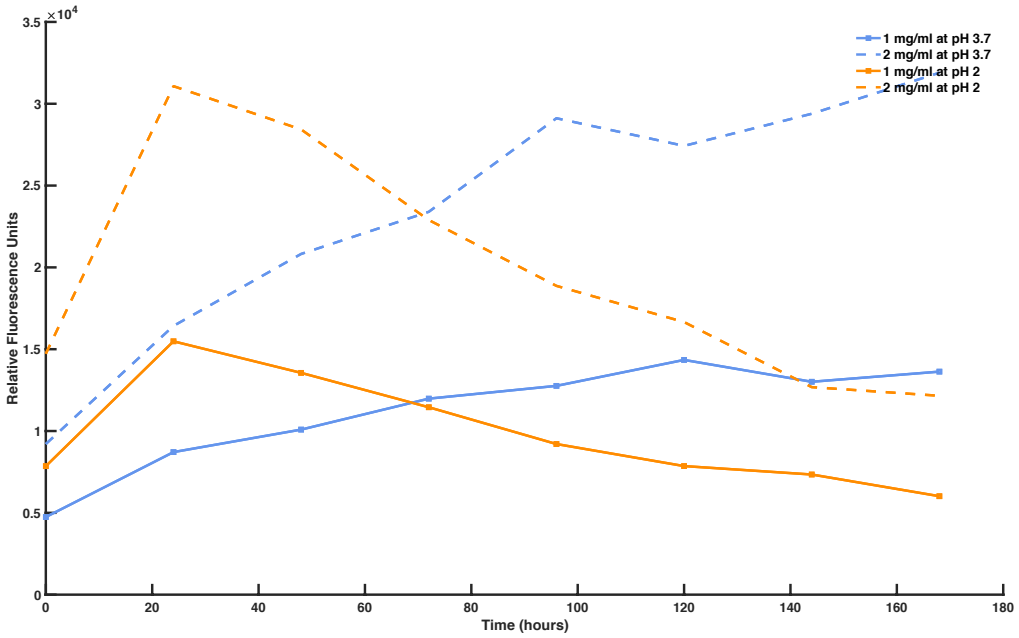

Figure 5.8: ThT assay results of 1 mg/ml and 2 mg/ml γS-crystallin using phosphate buffer at pH 2 and pH 3.7 with agitation (150 RPM) at 75 degree celcius.

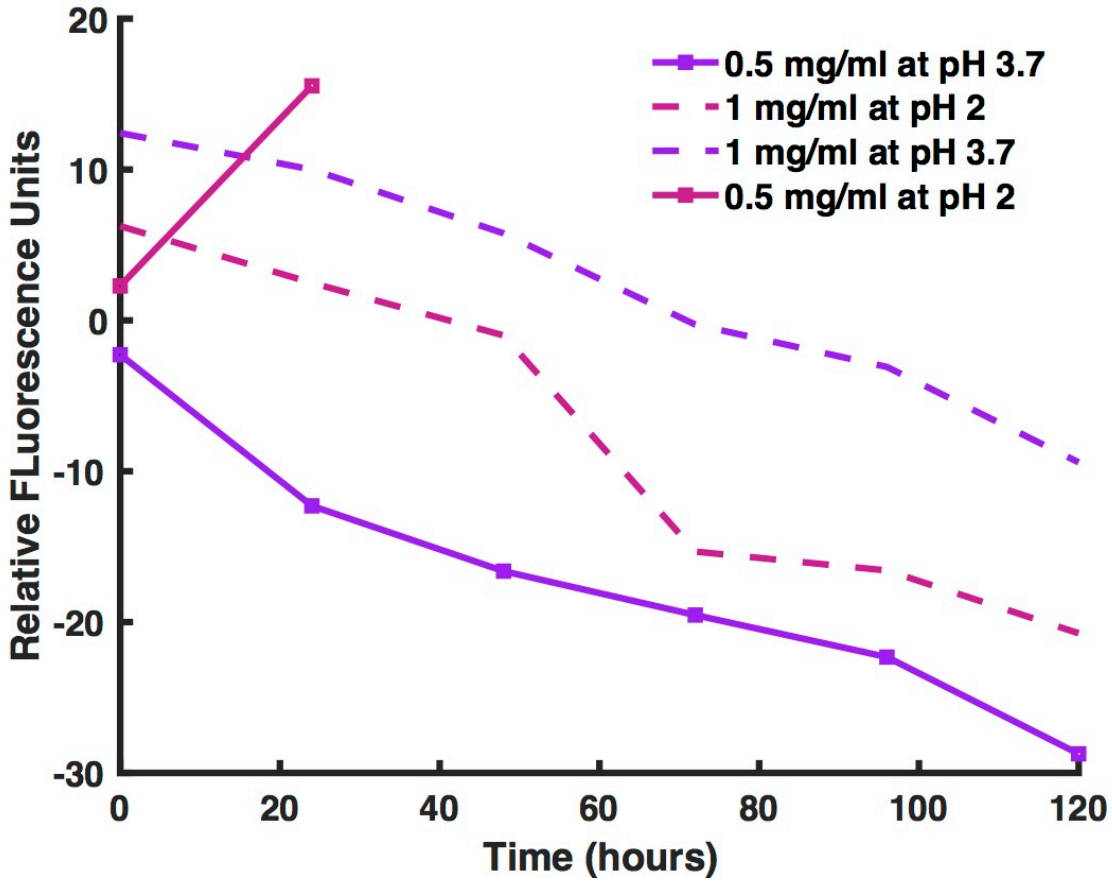

Figure 5.9: ThT assay results of 0.5 mg/ml and 1 mg/ml HEWL using carbonate buffer at pH 2 and pH 3.7 with agitation (150 RPM) at 75 degree celcius.

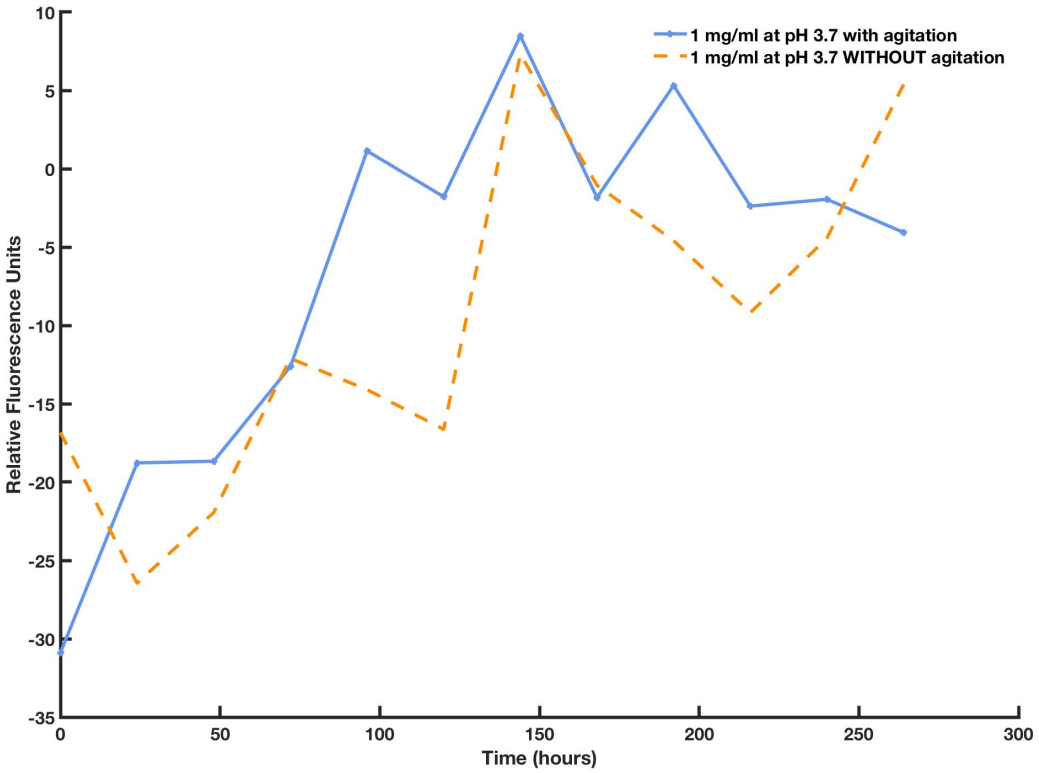

Figure 5.10: ThT assay results of 1 mg/ml γS-crystallin using phosphate buffer at pH 3.7 with agitation (150 RPM) and without agitation at 75 degree celcius.

Chapter 6

Biophysical characterization and solution-state NMR assignments of J2 crystallin: Novel eye lens protein from the box jellyfish

In this chapter, I describe the preparation, biophysical characterization, and solution-state NMR structure determination of J2-crystallin, a previously uncharacterized eye lens protein from *Tripedalia cystophora* (box jellyfish).

6.1 Background

The crystallins of the eye lens are extremely stable, soluble proteins that are responsible for the transparency of this highly specialized tissue. The high refractivity of the eye lens results from two major contributions; the high protein concentration (up to 1000 mg/mL in some

fish species) [189, 190] and the high refractive indices of the crystallin proteins [191, 192]. Many crystallins are derived from either metabolic enzymes or physiological stress proteins and appear to have undergone selection for increased refractive index after gene duplication [193]. Mammalian lenses mostly consist of two strongly conserved classes of crystallins, the chaperone α-crystallins, and the structural $\beta\gamma$-crystallins [194, 195, 196, 193, 197, 198].

Taxon-specific crystallins with diverse structural properties are found in other species, including the ϵ-crystallin in avians and reptiles and the S-crystallins in cephalopods [194, 196]. Unlike other non-cephalopod invertebrates, in which typical visual systems consist of simple ocelli and/or compound eyes made up of an array of ommatidia [199, 200], the box jellyfish, *Tripedalia cystophora*, has camera-type eyes similar to those of vertebrates and cephalopods [199]. Although the lenses are capable of forming sharp images, the eyes are used for navigation rather than detection of detailed objects [201], as the distance between the lens and the retina is not optimized for maximum visual acuity. *T. cystophora* has a total of 24 eyes split among four rhopalia (specialized structures that sense light), around the bell of the medusa, each containing two camera-type eyes as well as an array of simpler pigment cup eyes [202, 200]. The camera-type eye lenses are composed of the J1-, J2- and J3-crystallin proteins [202]. J1- and J3-crystallins are homologous to known enzymes, (ADP-ribohydroglycosylase and saposins, respectively) [202], whereas J2-crystallin (J2) has no known homologs [201, 203]; a BLAST search of the Protein Data Bank (PDB) found no proteins above 37% similarity [204].

To my knowledge, J2 has not previously been expressed and characterized. Here I focus on the box jellyfish eye lens protein J2-crystallin. This 157-amino acid protein has a molecular weight of 18.2 kDa [194] and a theoretical isoelectric point (pI) of 9.25 [205]. As expected for an eye lens protein, the thermal stability of J2 crystallin is high, with an unfolding temperature of 75.2 C. The project was initiated by Dr. Domarin Khago [192] in the lab. She pioneered the expression and optimization of J2 crystallin [192] and was also involved

in biophysical characterization and taking initial HSQC and NMR data. Please refer to her thesis [192] for more details on expression and optimization. I will be focusing on the biophysical studies and 3D NMR data that I collected for J2 crystallin [206, 207].

As seen in Figure 6.1, Rosetta and iTasser servers were used to predict J2 crystallin protein models.

Rosetta, of Robetta server [206, 207], is an automated protein structure prediction software developed by the Baker laboratory for *ab initio* and comparative modeling. On submission of the fasta sequence, in this case J2 crystallin, Rosetta first searches for structural homologs using BLAST, PSI-BLAST, and 3D-Jury, by breaking down the target sequence in small portions of individual domains, or independently folding units of proteins. It then matches the sequence to structural families in the Pfam database. Domains with structural homologs then follow a "template-based model" (i.e., homology modeling) protocol [206, 207]. The lowest energy model as determined by Rosetta energy function is then selected as the final predicted structure [206, 207].

iTASSER server [208, 209] is an on-line software that implements the iTASSER based algorithms for protein structure and function predictions from amino acid (fasta) sequences. When user submits an amino acid sequence, the server first tries to retrieve template proteins of similar folds (or super-secondary structures) from the PDB library by LOMETS, a locally installed meta-threading approach developed Zhang Lab [208, 209]. Next, the continuous fragments excised from the PDB templates are reassembled into full-length models using Monte Carlo simulations with the threading unaligned regions built by *ab initio* modeling [208, 209]. In the following step, the fragment assembly simulation is performed again starting from the SPICKER cluster centroids, where the spatial restrains collected from the LOMETS templates and the PDB structures are used to fasten the simulations [208, 209]. The final structure is predicted from the consensus of top structural matches between model and templates as evaluated by TM-score, and the sequence identity in the structurally aligned

regions [208, 209].

As seen in Firgure 6.1, there is variability and high dissimilarity in the models predicted by iTasser [208, 209] and Rosetta [206, 207]. As these both tools use homology modelling of the known protein structures, I hypothesize J2 crystallin has a unique protein fold (structure).

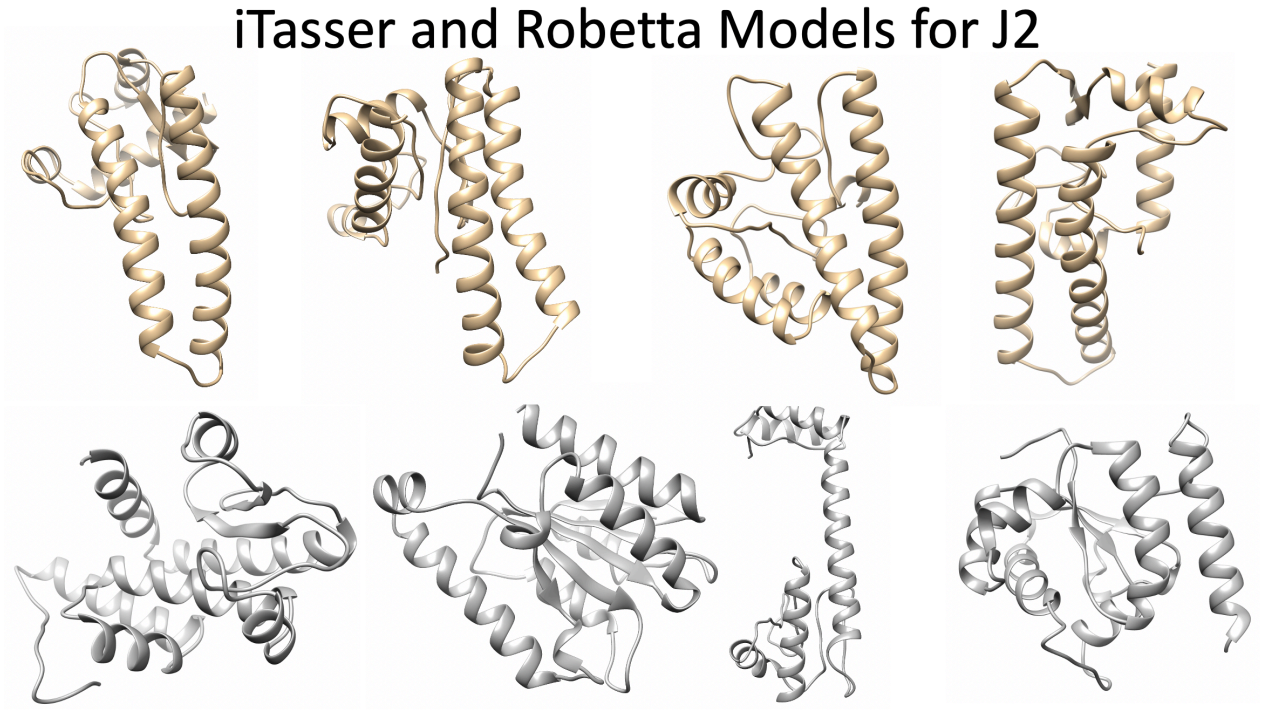

Figure 6.1: Rosetta (gold) and iTasser (silver) servers were used to predict J2 crystallin protein model. Even within the server, there is variability and high dissimilarity of the predicted models. As these both tools use homology modelling of the known protein structures, I hypothesize J2 crystallin has a unique protein fold (structure).

6.2 Materials and Methods

6.2.1 Expression and purification of ^{15}N-labeled and ^{13}C-labeled J2-crystallin

For the starter culture of J2 crystallin, 50 mL of LB media was inoculated with a single colony of Rosetta (DE3) *Escherichia coli* cells containing a pET28(+)a vector with J2-crystallin gene inserts was grown at 37 °C for 16 hours with shaking at 225 RPM [192]. Optical density measurement (OD), defined as a logarithmic intensity ratio of the light falling upon the material, to the light transmitted through the material, measures the concentration of cells in the solution [210]. Until an OD of 0.60 was reached, the cultures were grown at 37 °C with shaking at 225 RPM. The cells were then collected in 500 mL batches by centrifugation at 3000 RPM for 30 minutes and each 500 mL batch was resuspended in 1 L ^{15}N-labeled ^{13}C-labeled minimal media cultures [192]. The 1 L minimal media cultures were grown for an additional 2 hours at 37 °C at 225 RPM. Protein overexpression was induced using IPTG (Gold Biotechnology) at a final concentration of 0.10 mM at 25 °C for 30 hours [192]. Cells were collected via centrifugation at 6000 RPM, and pellet was resuspended in 40 mL of 50 mM sodium phosphate buffer with 300 mM sodium chloride, 10 mM imidazole, and 0.05% sodium azide at pH 7.4 [192]. Cells were lysed by sonication in 30 second intervals for a total of 10 minutes, followed by centrifugation at 15000 RPM for 90 minutes. The supernatant was filtered with through a 0.22 μm filter (Millipore) before being loaded onto a Bio-Rad Duo-Inject FPLC system (Hercules, CA) [192]. The His-tagged crystallin was purified and cut by a His-tagged TEV protease, using a Ni-IDA (Bio-Rad) and a second application to a Ni-IDA column was done to remove TEV protease and His-tag [192]. The purified labeled crystallin was dialyzed into 10 mM sodium phosphate buffer with 0.05% sodium azide at pH 6.0 [192]. The final 1.8 mM concentrated J2-crystallin sample was prepared in 10 % D_2O and 2 mM TMSP [192].

6.2.2 NMR experiments

A Varian $^{\text{Unity}}$INOVA spectrometer (Agilent Technologies, Santa Clara, CA) operating at 800 MHz coupled with a ^{1}H–^{13}C–^{15}N 5 mm tri-axis PFG triple-resonance probe, using an 18.8 Tesla superconducting electromagnet (Oxford Instruments) was used to perform most of the 2D and 3D experiments (HNCA and HN(CO)CA were performed on a Bruker Spectrometer operating at 800 MHz at University of California, San Diego NMR Facility). ^{1}H chemical shifts were referenced to TMSP, and ^{15}N and ^{13}C shifts were referenced indirectly to TMSP. NMR data were processed using NMRPipe [211] and analyzed using CcpNMR Analysis [212]. Center operating frequencies and (unless otherwise stated) center frequency offsets were as follows:

Center	^{1}H: 799.7988955 MHz	^{15}N: 81.04252684 MHz
Offset	^{1}H: 4.81 ppm	^{15}N: 118.70 ppm

6.3 Results and Discussion

6.3.1 Biophysical characterization of J2-crystallin reveals a stable protein

Circular dichroism (CD) spectroscopy and aggregation propensity were used for biophysical characterization of J2-crystallin. As expected of a structural protein, J2 crystallin is a stable protein, aggregating at 73.5 degree Celcius and unfolding at 76.5 degree Celcius. A part of biophyical characterization was performed by Dr. Domarin Khago and the results can be found in her thesis [192].

Aggregation under thermal stress measured as a function of temperature using Dynamic Light Scattering (DLS)

Dynamic Light Scattering is robust and sensitive technique that can be used to characterize protein aggregates in solution, because of its ability to resolve molecular or particle sizes ranging from sub-nanometer to several microns [213]. DLS measurements to understand the aggregation propensity of J2 crystallin were obtained on Zetasizer Nano ZS (Malvern Instruments, Malvern, U.K.). The sample was prepared at the concentration of 1.0 mg/ml in 10 mM phosphate buffer at pH 6.9. At each temperature, the sample was allowed to equilibrate for 2 min before measurements were obtained, after which scattering measureazments were performed in triplicate, resulting in a heating rate of 0.5 C/min. The average apparent particle size is plotted as a function of temperature. J2 aggregates at 73.5° Celsius as seen in Figure 6.2.

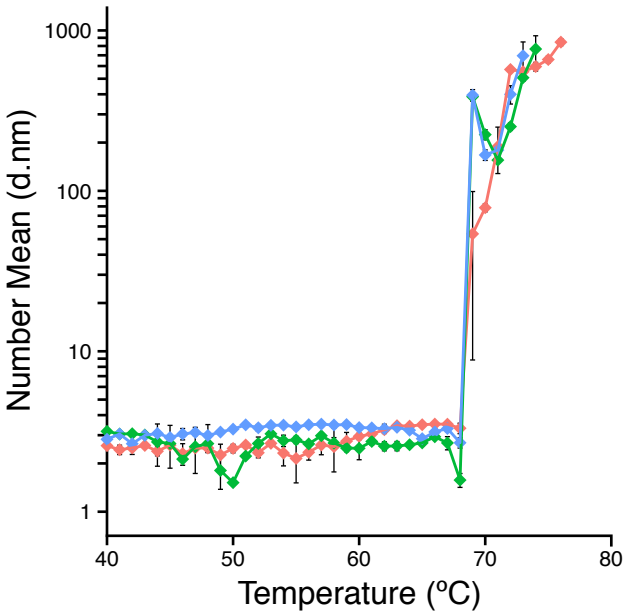

Figure 6.2: DLS measurements of thermally induced aggregation of J2 crystallin over a range of 40-80 degree celcius. Salmon, green and blue colors represent the data taken in triplicate. The average apparent particle size is plotted as a function of temperature. J2 aggregates at 73.5 degree Celsius.

Circular Dichroism (CD) to study the thermal unfolding curves of J2 crystallin

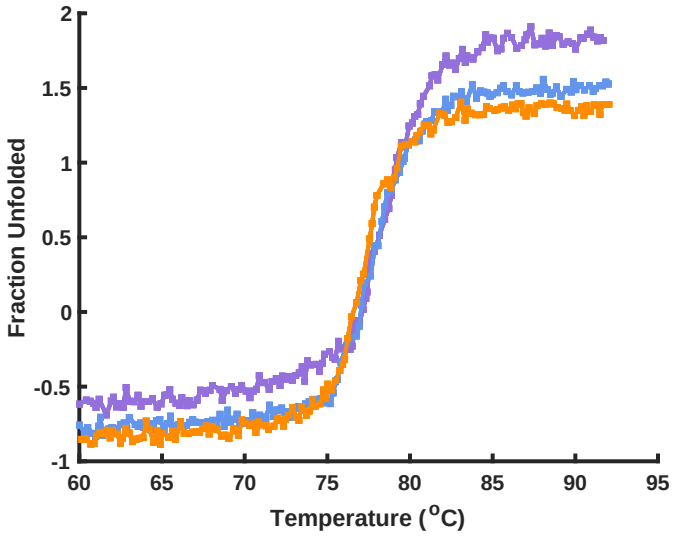

Figure 6.3: Thermal unfolding curve of J2 crystallin measured by monitoring the circular dichroism signal at 218 nm with the Tm value for J2 is 76.5° celcius.

Thermal denaturation provides complementary information regarding protein stability and aggregation propensity. Thermal unfolding curves of J2 are measured by monitoring the circular dichroism signal at 218 nm on a J-810 spectropolarimeter (JASCO, Easton, MD) equipped with a thermal controller. For unfolding measurments, the samples were heated at a rate of 2° C/ minute. The midpoint of the unfolding transition (Tm) is itself a useful measure of protein stability. Tm value for J2 is 76.5° Celcius exhibiting characteristics of a stable protein as seen in the Figure 6.3.

For the sake of comparison, Figure 6.4, shows the Tm comparison of J2 crystallin with other crystallins studied in the lab. Figure from the paper [16].

6.3.2 Backbone assignment of J2 crystallin using Nuclear Magnetic Resonance

Solution state NMR was used to collect the data for J2 crystallin structure determination.

Table 1. Thermal Unfolding Temperatures for γS-Crystallin Variants

protein variant	T_m (°C)
γS-WT	72.0 ± 0.1
γS-G18V	63.6 ± 0.1
γS-G106V	58.9 ± 0.1
γS-G18V/R20A	67.1 ± 0.1
γS-G18V/R20M	63.1 ± 0.1
γS-G106V/M108R	62.7 ± 0.1

Figure 6.4: Tm comparison of J2 crystallin with other crystallins studied in the lab. Figure from the paper [16].

^{15}N HSQC

To assess the suitability of J2-crystallin for structural studies, a ^{1}H–^{15}N heteronuclear single-quantum correlation experiment (HSQC) was collected using samples of ^{15}N-labelled J2-crystallin. The first of many solution-state NMR experiments in a structure determination effort, this experiment determines how well-folded the protein is by the distribution of cross peaks [192]. The cross peaks are representative of the amide N-H pairs of the protein backbone and sidechains. Clean and separated cross peaks allow for distinction among the 157 amino acid residues that should be seen in the spectrum. An HSQC also can indicate the signal-to-noise ratio at the particular pH and concentration of the sample [192]. J2-crystallin HSQC, as seen in Figure 6.5, shows well-separated cross peaks in both spectral dimensions, indicating that the protein is folded and monomeric.

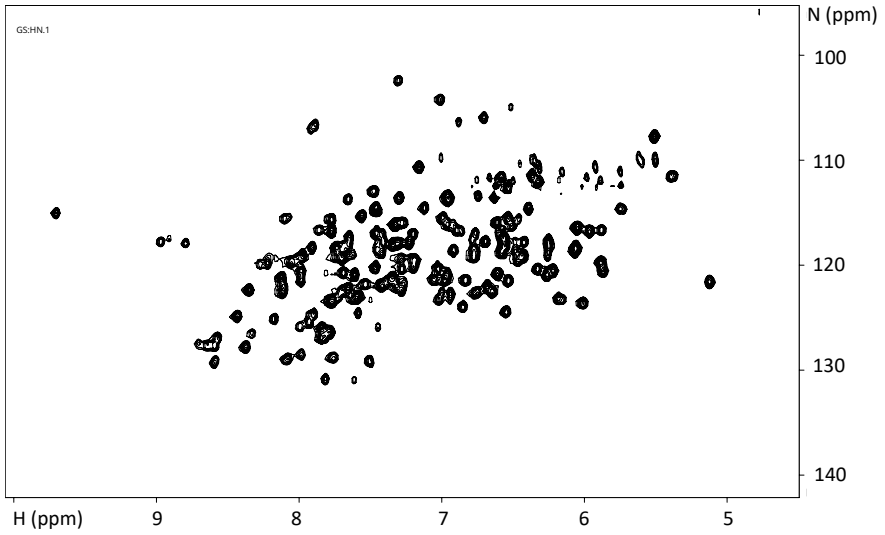

Figure 6.5: ^{1}H–^{15}N HSQC spectrum of ^{15}N-labelled J2-crystallin acquired at 25 °C, indicating that the protein is folded and monomeric. The crystallin sample was prepared in 10 % D_2O and 2 mM TMSP at a final concentration of 1.8 mM.

^{15}N-Temperature dependent HSQC

^{1}H–^{15}N HSQC spectrum of ^{15}N-labelled J2-crystallin acquired at 20, 25, 30, 35 and 40 degree Celcius. indicating that the protein is folded, monomeric and stable. The crystallin sample was prepared in 10 % D_2O and 2 mM TMSP at a final concentration of 1.8 mM. Temperature dependent HSQC does not reveal a major shift in the peaks showing that J2 is a stable protein as seen in Figure 6.6.

^{15}N-^{13}C HNCA

HNCA experiment is useful for backbone assignment when used in conjunction with the HN(CO)CA [214]. This is a 3D experiment because the chemical shifts are evolved for ^{1}HN, ^{15}NH and $^{13}C\alpha$ [214]. The magnetisation is passed in the following order- ^{1}H - ^{15}N - N-Cα J-coupling - $^{13}C\alpha$ - ^{15}N - ^{1}H hydrogen, where it is detected [214]. Since the amide nitrogen is coupled both to the Cα of its own residue and that of the preceding residue, both these transfers occur and peaks for both Cαs are visible in the spectrum [214]. HNCA spectrum

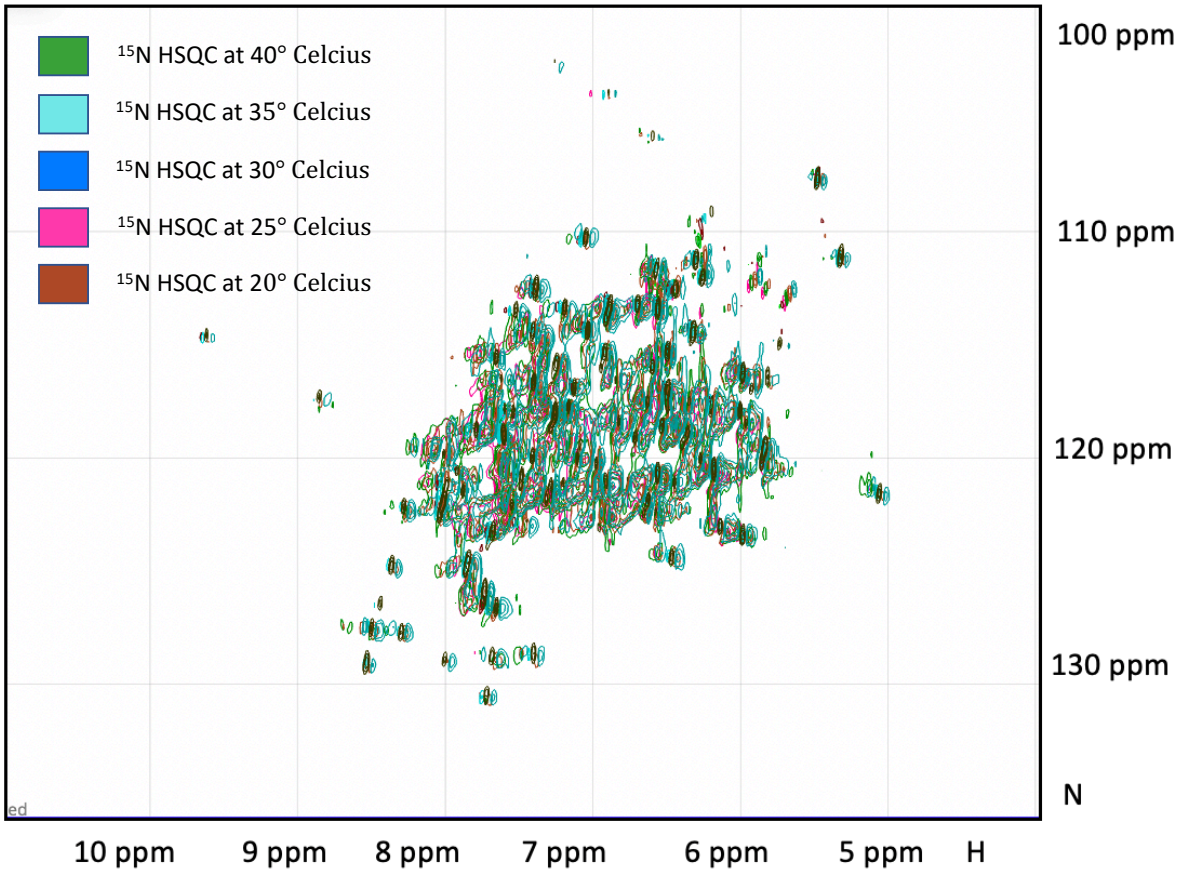

Figure 6.6: Temperature dependent HSQC does not reveal a major shift in the peaks showing that J2 is a stable protein.

for J2 crystallin is seen in Figure 6.7.

^{15}N-^{13}C HN(CO)CA

HN(CO)CA experiment is useful for backbone assignment when used in conjunction with the HNCA [214]. The magnetisation is passed as follows- ^{1}H to ^{15}N- to ^{13}CO, then transferred to ^{13}C where the chemical shift is evolved [214]. The magnetisation is then transferred back via ^{13}CO to ^{15}N and ^{1}H for detection. The chemical shift is only evolved for the ^{1}HN, the ^{15}N and the ^{13}C, but not for the ^{13}COα [214]. HNCA spectrum for J2 crystallin is seen in Figure 6.8.

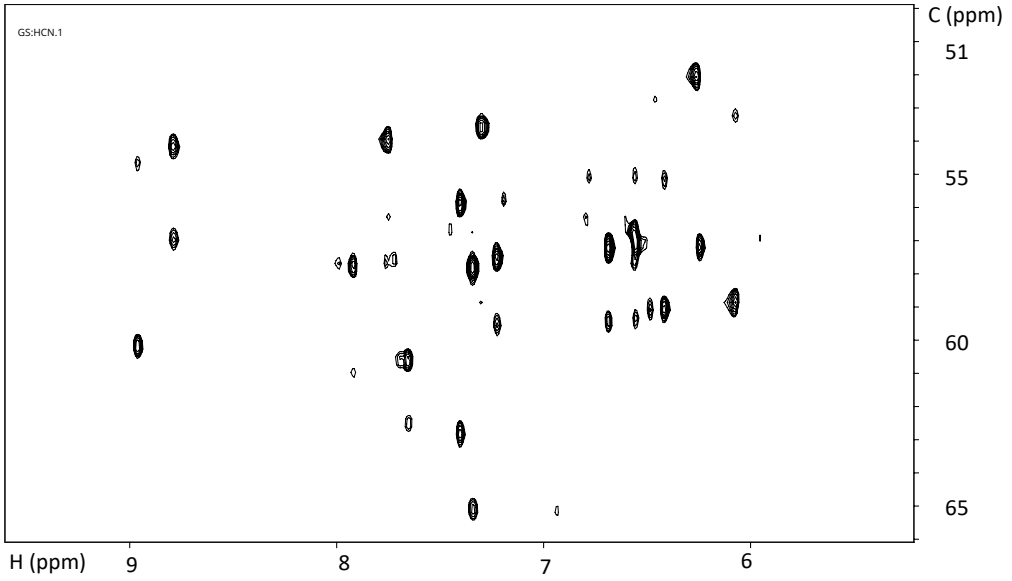

Figure 6.7: Slice of ^{1}H–^{15}N HNCA spectrum of ^{15}N ^{13}C labelled J2-crystallin acquired at 25 °C. The J2 crystallin sample was prepared in 10 % D_2O and 2 mM TMSP at a final concentration of 1.8 mM.

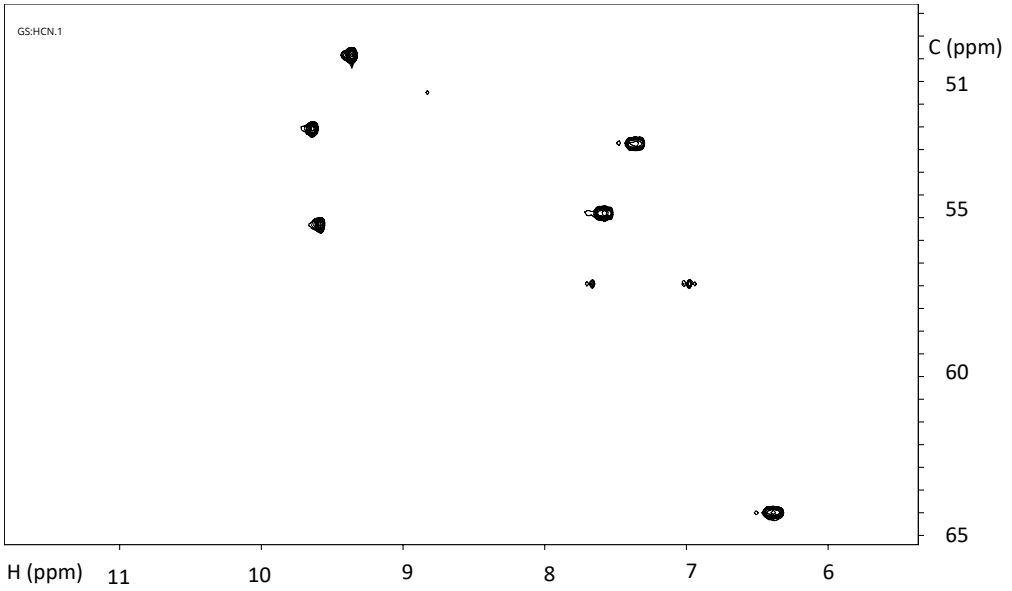

Figure 6.8: Slice of ^{1}H–^{15}N HNCOCA spectrum of ^{15}N ^{13}C labelled J2-crystallin acquired at 25 °C. The crystallin sample was prepared in 10 % D_2O and 2 mM TMSP at a final concentration of 1.8 mM.

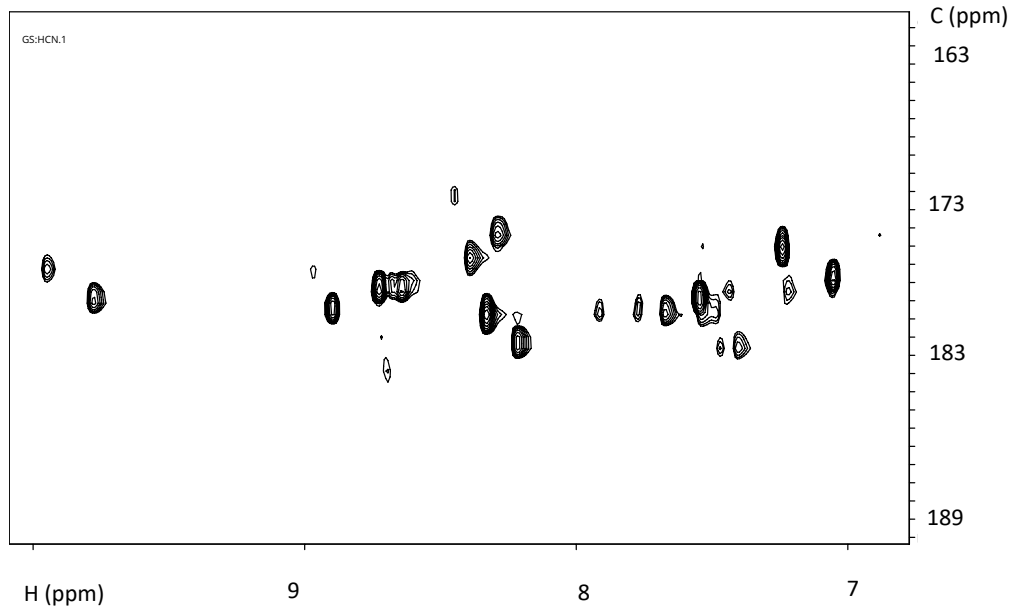

Figure 6.9: Slice of ^{1}H–^{15}N HNCO spectrum of ^{15}N ^{13}C ^{1}H labelled J2-crystallin acquired at 25 °C. The crystallin sample was prepared in 10 % D_2O and 2 mM TMSP at a final concentration of 1.8 mM.

^{15}N-^{13}C HNCO

This is the most sensitive triple-resonance experiment as in addition to the backbone CO-N-HN correlations, asparginine and glutamine are also visible [214]. Magnetisation is passed in the following order ^{1}H to ^{15}N to the carbonyl ^{13}C through ^{15}NH–^{13}C J-coupling to ^{15}N to ^{1}H where it is detected [214]. HNCO is used in conjunction with the HN(CA)CO [214]. HNCO spectrum for J2 crystallin is seen in Figure 6.9.

^{15}N-^{13}C HNCACO

Magnetisation is transferred in the following order ^{1}H to ^{15}N and then via the N-C J-coupling to the ^{13}Cα to ^{13}CO via the ^{13}C α-^{13}CO J-coupling [214] and then detected on H. HNCACO spectrum for J2 crystallin is seen in Figure 6.10 which can be used in conjunction with HNCO.

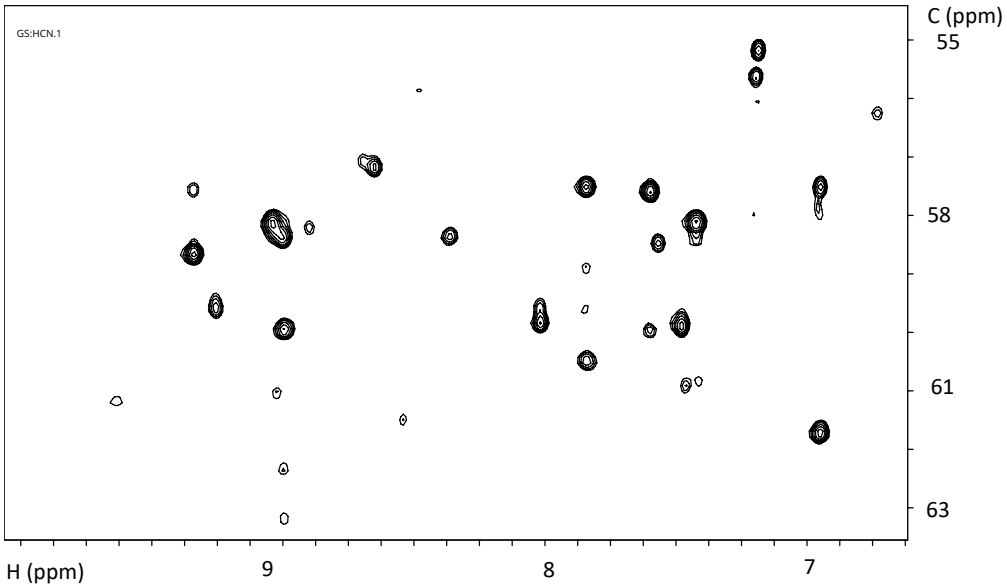

Figure 6.10: Slice of ^{1}H–^{15}N HNCACO spectrum of ^{15}N ^{13}C labelled J2-crystallin acquired at 25 °C. The crystallin sample was prepared in 10 % D_2O and 2 mM TMSP at a final concentration of 1.8 mM.

^{15}N-^{13}C CBCACONH, ^{15}N-^{13}C CBCANH and ^{15}N-^{13}C HNCACB

^{15}N-^{13}C CBCACONH, ^{15}N-^{13}C CBCANH and ^{15}N-^{13}C HNCACB, as seen in Figures 6.11, 6.12 and 6.13, are a standard set of experiments needed for backbone assignment especially for large proteins which tend to have a big signal to noise ratio [214].

HCCH-COSY

HCCH-COSY is used for side-chain assignment. Magnetisation transfer is from the side-chain hydrogen nuclei to their attached ^{13}C nuclei which is then exchanged between neighbouring ^{13}C nuclei via the J-coupling and finally transferred back to the side-chain hydrogen atoms for detection [214]. Figure 6.14 shows the HCCH-COSY spectrum of J2 crystallin.

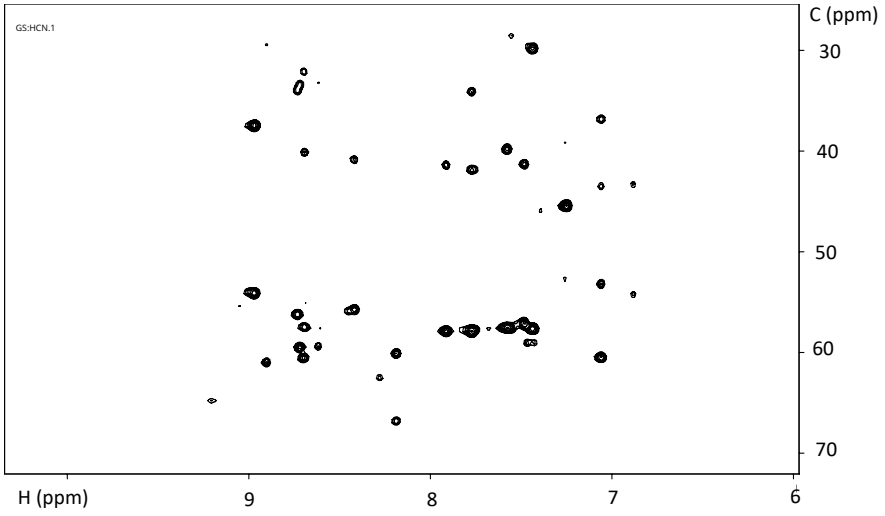

Figure 6.11: Slice of ^{1}H–^{15}N ^{15}C CBCACONH spectrum of ^{15}N-labelled J2-crystallin acquired at 25 °C. The crystallin sample was prepared in 10 % D_2O and 2 mM TMSP at a final concentration of 1.8 mM.

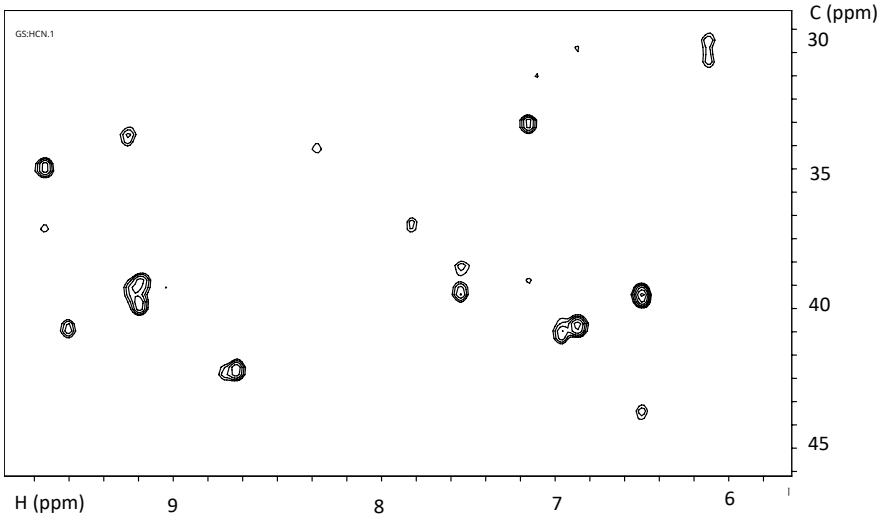

Figure 6.12: Slice of ^{1}H–^{15}N CBCANH spectrum of ^{15}N ^{13}C labelled J2-crystallin acquired at 25 °C. The crystallin sample was prepared in 10 % D_2O and 2 mM TMSP at a final concentration of 1.8 mM.

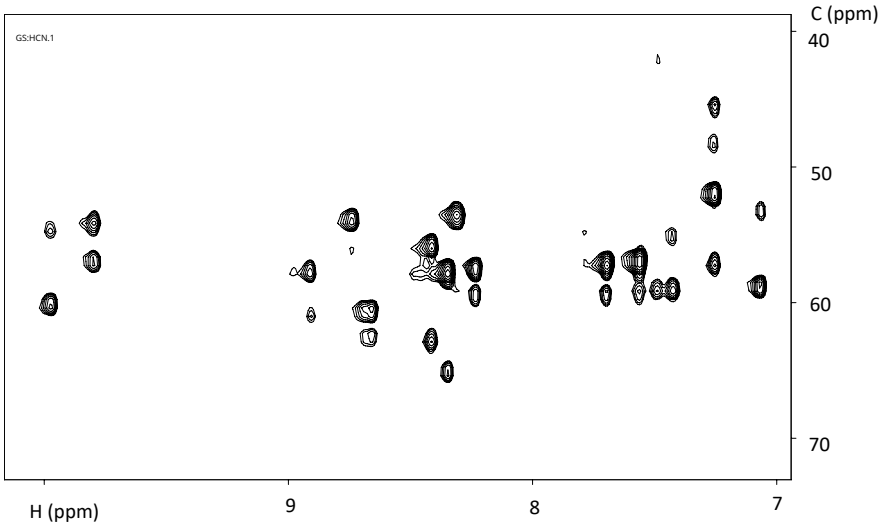

Figure 6.13: Slice of ^{1}H–^{15}N HNCACB spectrum of ^{15}N ^{13}C labelled J2-crystallin acquired at 25 °C. The crystallin sample was prepared in 10 % D_2O and 2 mM TMSP at a final concentration of 1.8 mM.

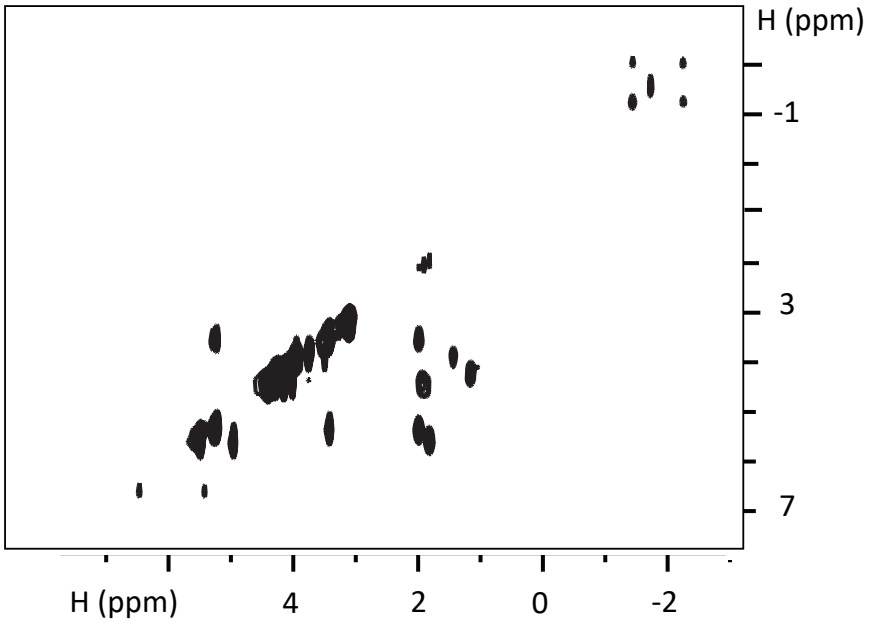

Figure 6.14: Slice of HCCH-COSY spectrum of ^{15}N ^{13}C labelled J2-crystallin acquired at 25 °C. The crystallin sample was prepared in 10 % D_2O and 2 mM TMSP at a final concentration of 1.8 mM.

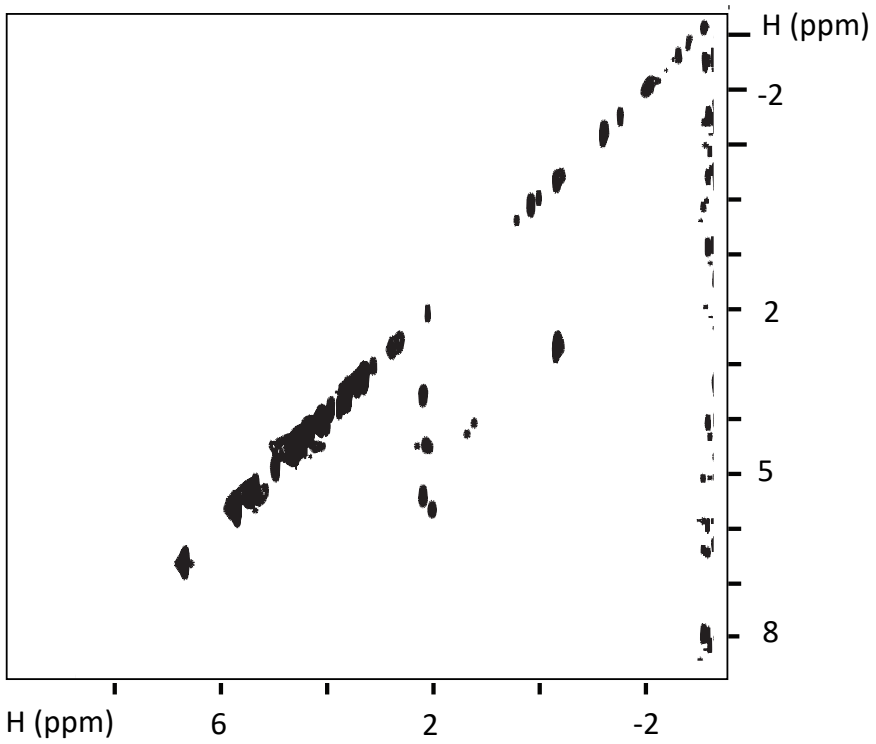

Figure 6.15: Slice of HCCH- TOCSY spectrum of ^{15}N ^{13}C labelled J2-crystallin acquired at 25 °C. The crystallin sample was prepared in 10 % D_2O and 2 mM TMSP at a final concentration of 1.8 mM.

HCCH-TOCSY

HCCH-TOCSY is used for side-chain assignment. FIgure 6.15 shows the HCCH- TOCSY spectrum of J2 crystallin. Magnetisation transfer begins from the side-chain hydrogen nuclei to their attached ^{13}C nuclei, followed by isotropic ^{13}C mixing and back to the side-chain hydrogen atoms where it is detected [214].

^{13}C HSQC Aliphatic and ^{13}C HSQC Aromatic

^{13}C HSQC Aliphatic and ^{13}C HSQC Aromatic experiments (Seen in Figure 6.14, provide correlations between carbons and its attached protons and helps circumvent the issue of splitting of signal due to homonuclear ^{13}C—^{13}C J couplings making the spectral resolution

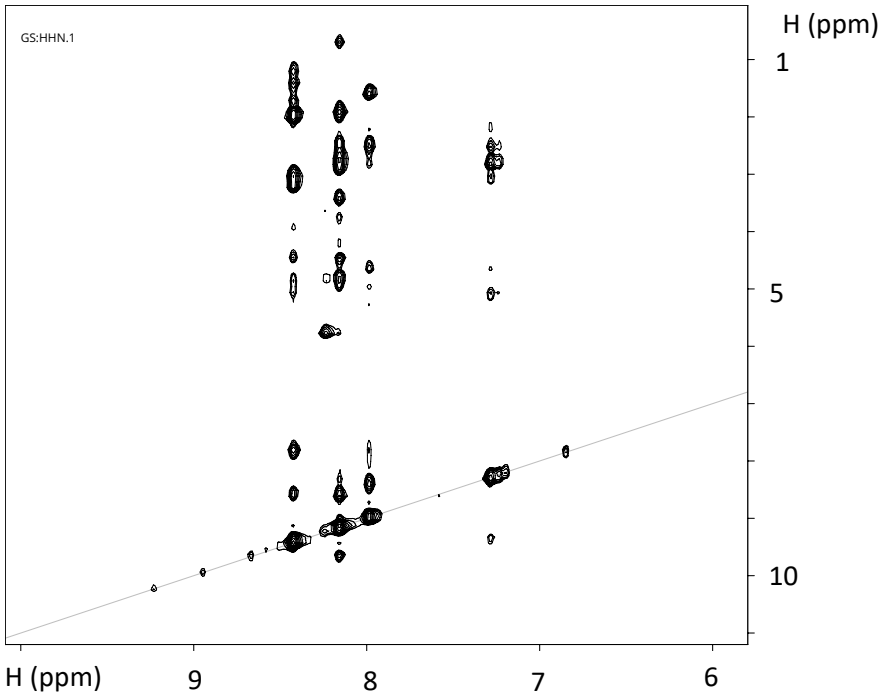

Figure 6.16: Slice of ^{1}H–^{15}N NOESY spectrum of ^{15}N ^{13}C labelled J2-crystallin acquired at 25 °C. The crystallin sample was prepared in 10 % D_2O and 2 mM TMSP at a final concentration of 1.8 mM.

better.

NOESY

This spectrum can be used to obtain restraints for structure calculations and can also help in assigninng the backbone [214]. J2 crystallin NOESY is seen in Figure 6.16.

6.3.3 J2 Crystallin triple resonance backbone assignment

Sequential protein backbone assignments for J2 crystallin using triple-resonance experiments is seen in figure 6.17. Most spectra used for triple resonance backbone assignment have a ^{1}H, ^{15}N and ^{13}C dimension each. Assigning the backbone is step one on the path to strcuture from NMR spectra. In Figure 6.17, Alanine 54 and 55 and Serine 56 residue are assigned

using the CBCA(CO)NH and HNCACB 3D experiments.

6.4 Conclusion

As this is a structural protein, the structure is of paramount importance in determining its function. J2-crystallin is a novel eye lens protein that has been expressed recombinantly and purified in high yield [192]. In agreement with the secondary structure prediction and CD spectroscopy, J2-crystallin has primarily α-helical character and thermal studies have concluded a melting temperature of 75.2 °C [192]. A ^{1}H–^{15}N HSQC shows the protein is well-folded, allowing for further NMR experiments to be performed. Next steps include spin assignments and residual dipolar coupling.

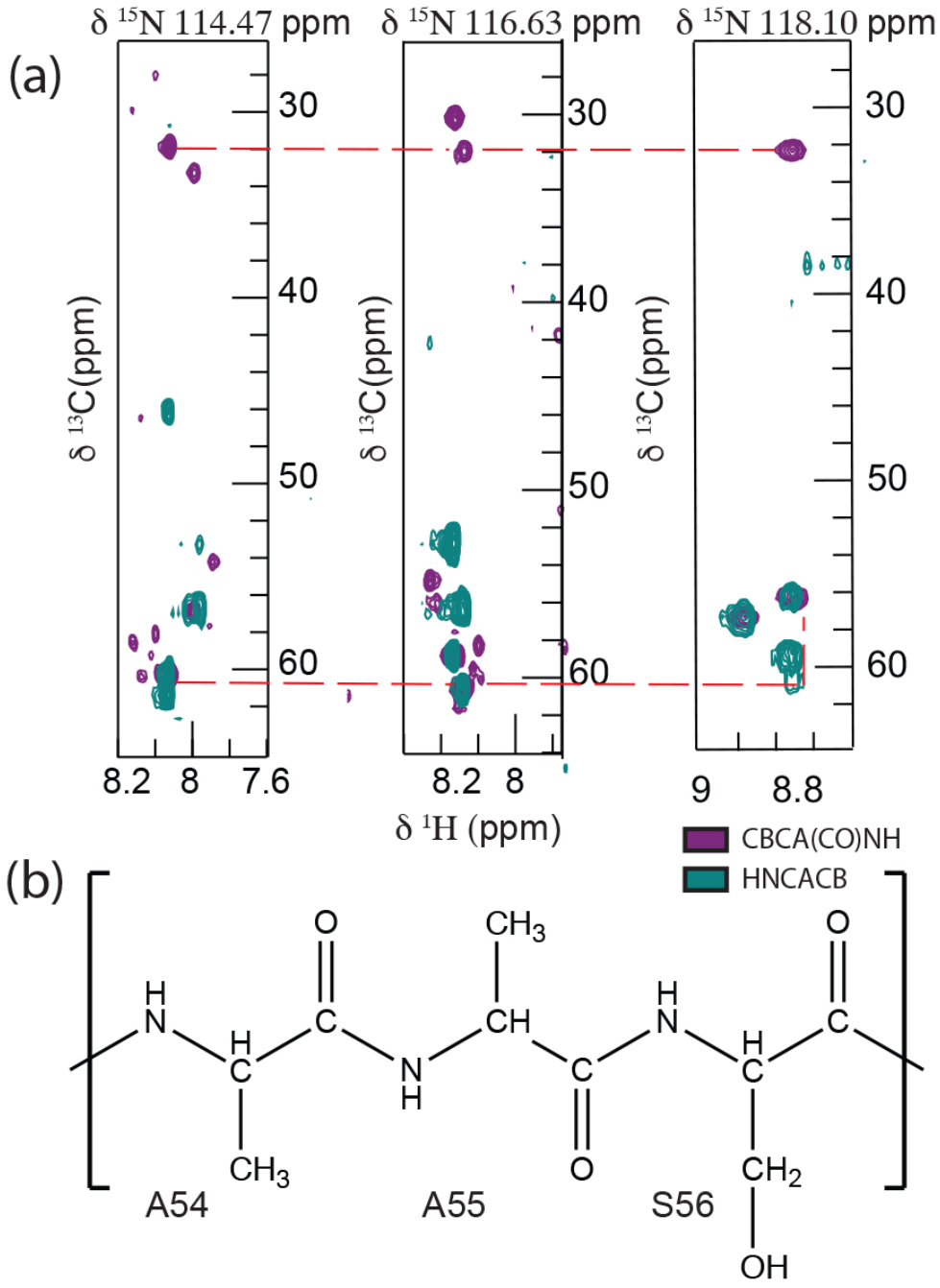

Figure 6.17: Sequential protein backbone assignments for J2 crystallin using triple-resonance experiments where the Alanine 54 and 55 and Serine 56 residue are assigned using the CBCA(CO)NH and HNCACB 3D experiments.

Bibliography

[1] Megha H Unhelkar, Vy T Duong, Kaosoluchi N Enendu, John E Kelly, Seemal Tahir, Carter T Butts, and Rachel W Martin. Structure prediction and network analysis of chitinases from the cape sundew, drosera capensis. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1861(3):636–643, 2017.

[2] Vy T Duong, Megha H Unhelkar, John E Kelly, Suhn H Kim, Carter T Butts, and Rachel W Martin. Protein structure networks provide insight into active site flexibility in esterase/lipases from the carnivorous plant drosera capensis. *Integrative biology*, 10(12):768–779, 2018.

[3] Carter T. Butts, Domarin Khago, and Rachel W. Martin. Bayesian analysis of static light scattering data for globular proteins. In Preperation, 2016.

[4] Gábor J. Székely and Maria L. Rizzo. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of Classification*, 22(2):151–183, 2005.

[5] Jing-Na Si, Ren-Xiang Yan, Chuan Wang, Ziding Zhang, and Xiao-Dong Su. TIM-Finder: A new method for identifying TIM-barrel proteins. *BMC Structural Biology*, 9(73):doi:10.1186/1472–6807–9–73, 2009.

[6] Paulina Paszota, Maria Escalante-Perez, Line R. Thomsen, Michael W. Risør, Alicja Dembski, Laura Sanglas, Tania A. Nielsen, Henrik Karring, Ida B. Thøgersen, Rainer Hedrich, Jan J. Enghild, Ines Kreuzer, and Kristian W. Sanggaard. Secreted major Venus flytrap chitinase enables digestion of arthropod prey. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1844(2):374 – 383, 2014.

[7] Kerstin H. Richau, Farnusch Kaschani, Martijn Verdoes, Twinkal C. Pansuriya, Sherry Niessen, Kurt Stüber, Tom Colby, Hermen S. Overkleeft, Matthew Bogyo, and Renier A. L. van der Hoorn. Subclassification and biochemical analysis of plant papain-like cysteine proteases displays subfamily-specific characteristics. *Plant Physiology*, 158:1583–1599, 2012.

[8] Wei Qiang, Wai-Ming Yau, Yongquan Luo, Mark P Mattson, and Robert Tycko. Antiparallel β-sheet architecture in Iowa-mutant β-amyloid fibrils. *Proceedings of the National Academy of Sciences*, 109(12):4443–4448, 2012.

[9] Michael T Colvin, Robert Silvers, Qing Zhe Ni, Thach V Can, Ivan Sergeyev, Melanie Rosay, Kevin J Donovan, Brian Michael, Joseph Wall, Sara Linse, et al. Atomic resolution structure of monomorphic Aβ42 amyloid fibrils. *Journal of the American Chemical Society*, 138(30):9663–9674, 2016.

[10] O Sumner Makin and Louise C Serpell. Structures for amyloid fibrils. *The FEBS journal*, 272(23):5950–5961, 2005.

[11] Anthony WP Fitzpatrick, Galia T Debelouchina, Marvin J Bayro, Daniel K Clare, Marc A Caporini, Vikram S Bajaj, Christopher P Jaroniec, Luchun Wang, Vladimir Ladizhansky, Shirley A Müller, et al. Atomic structure and hierarchical assembly of a cross-β amyloid fibril. *Proceedings of the National Academy of Sciences*, 110(14):5468–5473, 2013.

[12] Schrödinger, LLC. The *PyMOL* molecular graphics system, version 1.8, November 2015.

[13] Lorena Saelices, Lisa M Johnson, Wilson Y Liang, Michael R Sawaya, Duilio Cascio, Piotr Ruchala, Julian Whitelegge, Lin Jiang, Roland Riek, and David S Eisenberg. Uncovering the mechanism of aggregation of human transthyretin. *Journal of Biological Chemistry*, 290(48):28932–28943, 2015.

[14] Gianmarc Grazioli, Yue Yu, Megha H Unhelkar, Rachel W Martin, and Carter T Butts. Network-based classification and modeling of amyloid fibrils. *The Journal of Physical Chemistry B*, 123(26):5452–5462, 2019.

[15] Jed JW Wiltzius, Meytal Landau, Rebecca Nelson, Michael R Sawaya, Marcin I Apostol, Lukasz Goldschmidt, Angela B Soriaga, Duilio Cascio, Kanagalaghatta Rajashankar, and David Eisenberg. Molecular mechanisms for protein-encoded inheritance. *Nature structural & molecular biology*, 16(9):973, 2009.

[16] Kyle W Roskamp, Carolyn N Paulson, William D Brubaker, and Rachel W Martin. Function and aggregation in structural eye lens crystallins. *Accounts of Chemical Research*, 53(4):863–874, 2020.

[17] Michael Krogh Jensen, Josef Korbinian Vogt, Simon Bressendorff, Andaine Seguin-Orlando, Morten Petersen, Thomas Sicheritz-PontÃ©n, and John Mundy. Transcriptome and genome size analysis of the venus flytrap. *PLoS ONE*, 10(4):e0123887, 04 2015.

[18] Carter T Butts, Jan C Bierma, and Rachel W Martin. Novel proteases from the genome of the carnivorous plant drosera capensis: structural prediction and comparative analysis. *Proteins: Structure, Function, and Bioinformatics*, 84(10):1517–1533, 2016.

[19] E. V. Leushkin, R. A. Sutormin, E. R. Nabieva, A. A. Penin, A. S. Kondrashov, and M. D. Logacheva. The miniature genome of a carnivorous plant Genlisea aurea contains a low number of genes and short non-coding sequences. *BMC Genomics*, 14(476), 2013.

[20] Enrique Ibarra-Laclette, Eric Lyons, Gustavo Hernandez-Guzman, Claudia Anahi Perez-Torres, Lorenzo Carretero-Paulet, Tien-Hao Chang, Tianying Lan, Andreanna J. Welch, Maria Jazmin Abraham Juarez, June Simpson, Araceli Fernandez-Cortes, Mario Arteaga-Vazquez, Elsa Gongora-Castillo, Gustavo Acevedo-Hernandez, Stephan C. Schuster, Heinz Himmelbauer, Andre E. Minoche, Sen Xu, Michael Lynch, Araceli Oropeza-Aburto, Sergio Alan Cervantes-Perez, Maria de Jesus Ortega-Estrada, Jacob Israel Cervantes-Luevano, Todd P. Michael, Todd Mockler, Douglas Bryant, Alfredo Herrera-Estrella, Victor A. Albert, and Luis Herrera-Estrella. Architecture and evolution of a minute plant genome. *Nature*, 498(7452):94–98, 2013.

[21] Megha H. Unhelkar, Vy T. Duong, Kaosoluchi N. Enendu, John E. Kelly, Seemal Tahir, Carter T. Butts, and Rachel W. Martin. Structure prediction and network analysis of chitinases from the Cape sundew, *Drosera capensis*. *Biochimica et Biophysica Acta*, 1861:636–643, 2017.

[22] UniProt Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2019.

[23] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular systems biology*, 7(1):539, 2011.

[24] Thomas Nordahl Petersen, Søren Brunak, Gunnar Von Heijne, and Henrik Nielsen. Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10):785–786, 2011.

[25] Olof Emanuelsson, Søren Brunak, Gunnar Von Heijne, and Henrik Nielsen. Locating proteins in the cell using targetp, signalp and related tools. *Nature protocols*, 2(4):953, 2007.

[26] T.N. Petersen, S. Brunak, G. von Heijne, and H. Henrik Nielsen. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, 8:785–786, 2011.

[27] Srivatsan Raman, Robert Vernon, James Thompson, Michael Tyka, Ruslan Sadreyev, Jimin Pei, David Kim, Elizabeth Kellogg, Frank DiMaio, Oliver Lange, Lisa Kinch, Will Sheffler, Bong-Hyun Kim, Rhiju Das, Nick V. Grishin, and David Baker. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins*, 77(Suppl 9):89–99, 2009.

[28] Kim D.E., D. Chivian, and D. Baker. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, 32(Supplement 2):W526–31, 2004.

[29] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26(16):1781–1802, December 2005.

[30] Carter T Butts, Xuhong Zhang, John E Kelly, Kyle W Roskamp, Megha H Unhelkar, J Alfredo Freites, Seemal Tahir, and Rachel W Martin. Sequence comparison, molecular modeling, and network analysis predict structural diversity in cysteine proteases from the cape sundew, drosera capensis. *Computational and structural biotechnology journal*, 14:271–282, 2016.

[31] Feisal Khoushab and Montarop Yamabhai. Chitin research revisited. *Marine Drugs*, 8(7):1988–2012, 2010.

[32] Rathore A.S and Gupta R.D. Chitinases from bacteria to human: Properties, applications, and future perspectives. *Enzyme Research*, 2015(8):1 – 8, 2015.

[33] Taro Masuda, Guanghua Zhao, and Bunzo Mikami. Crystal structure of class III chitinase from pomegranate provides the insight into its metal storage capacity. *Bioscience, Biotechnology, and Biochemistry*, 79(1):45–50, 2015.

[34] Karina Vega and Markus Kalkum. Chitin, chitinase responses, and invasive fungal infections. *International Journal of Microbiology*, 2012:Article ID 920459, 2012.

[35] Ines Mack, Andreas Hector, Marlene Ballbach, Julius Kohlhäufl, Katharina J. Fuchs, Alexander Weber, Marcus A. Mall, and Dominik Hartl. The role of chitin, chitinases, and chitinase-like proteins in pediatric lung diseases. *Molecular and Cellular Pediatrics*, 2(3):DOI 10.1186/s40348–015–0014–6, 2015.

[36] Gunther Busam, Hanns-Heinz Kassemeyer, and Ulrich Matern. Differential expression of chitinases in vitis vinifera l. responding to systemic acquired resistance activators or fungal challenge. *Plant Physiol.*, 15:1029–1038, 1997.

[37] S Karmakar, K.A. Molla, P.K. Chanda, S.N. Sarkar, S.K. Datta, and K. Datta. Green tissue-specific co-expression of chitinase and oxalate oxidase 4 genes in rice for enhanced resistance against sheath blight. *Planta*, 243(1):115–130, 2016.

[38] H. Eilenberg, S. Pnini-Cohen, S. Schuster, A. Movtchan, and A. Zilberstein. Isolation and characterization of chitinase genes from pitchers of the carnivorous plant *Nepenthes khasiana*. *Journal of Experimental Botany*, 57:2775–2784, 2006.

[39] S. Rottloff, R. Stieber, H. Maischak, F. G. Turini, G. Heubl, and A. Mithöfer. Functional characterization of a class iii acid endochitinase from the traps of the carnivorous pitcher plant genus, *Nepenthes*. *Journal of Experimental Botany*, 62:4639–4647, 2011.

[40] Andrej Pavlovic, Miroslav Krausko, and Lubomír Adamec. A carnivorous sundew plant prefers protein over chitin as a source of nitrogen from its traps. *Plant Physiology and Biochemistry*, 104:11–16, 2016.

[41] L. Adamec. Leaf absorption of mineral nutrients in carnivorous plants stimulates root nutrient uptake. *New Phytolologist*, 2155:89–100, 2002.

[42] A. Pavlovic, M. Krausko, M. Libiakova, and L. Adamec. Feeding on prey increases photosynthetic efficiency in the carnivorous sundew *Drosera capensis*. *Ann. Bot.*, 113:69–78, 2014.

[43] B.E. Juniper, R.J. Robins, and D.M. Joel. *The Carnivorous Plants*. Academic Press, London, UK, 1989.

[44] I. Matusíková, J. Salaj, J. Moravcíková, L. Mlynárová, J.P. Nap, and J. Libantová. Tentacles of in vitro-grown round-leaf sundew (*Drosera rotundifolia* L.) show induction of chitinase activity upon mimicking the presence of prey. *Planta*, 222:1020–1027, 2005.

[45] N. Hatano and T. Hamada. Proteomic analysis of secreted protein induced by a component of prey in pitcher fluid of the carnivorous plant *Nepenthes alata*. *Journal of Proteomics*, 75:4844–4852, 2012.

[46] F. Buch, M. Rott, S. Rottloff, C. Paetz, I. Hilke, M. Raessler, and A. Mithöfer. Secreted pitfall-trap fluid of carnivorous nepenthes plants is unsuitable for microbial growth. *Annals of Botany*, 111(375–383), 2013.

[47] Carter T. Butts, Jan C. Bierma, and R. W. Martin. Genome sequence and comparative analysis of putative proteases from the carnivorous plant *Drosera capensis*. *submitted*, 2016.

[48] Carter T. Butts, Xuhong Zhang, John E. Kelly, Kyle W. Roskamp, Megha H. Unhelkar, J. Alfredo Freites, Seemal Tahir, and Rachel W. Martin. Sequence comparison, molecular modeling, and network analysis predict structural diversity in cysteine proteases from the Cape sundew, *Drosera capensis*. *Computational and Structural Biotechnology Journal*, in press, 2016.

[49] M. Campbell, M. Law, C. Holt, J. Stein, G. Moghe, D. Hufnagel, J. Lei, R. Achawanantakun, D. Jiao, C. J. Lawrence, D. Ware, S. H. Shiu, K. L. Childs, Y. Sun, N. Jiang, , and M Yandell. MAKER-P: A tool-kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiology*, 164:513–524, 2013.

[50] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. InterProScan: Protein domains identifier. *Nucleic Acids Research*, 33:W116–W120, 2005.

[51] T. Renner and C. D. Specht. Molecular and functional evolution of Class I chitinases for plant carnivory in the caryophyllales. *Molecular Biology and Evolution*, 29(10):2971–2985, 2012.

[52] B. L. Cantarel, P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Research*, 37:D233–D238, 2009.

[53] Vincent Lombard, Hemalatha Golaconda Ramulu, Elodie Drula, Pedro M. Coutinho, and Bernard Henrissat. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research*, 42:D490–D495, 2014.

[54] Hans Merzendorfer. The cellular basis of chitin synthesis in fungi and insects: Common principles and differences. *European Journal of Cell Biology*, 90(9):759 – 769, 2011.

[55] Simone Zach Lukas Hartl and Verena Seidl-Seiboth. Fungal chitinases: diversity, mechanistic properties and biotechnological potential. *Appl Microbiol Biotechnol*, pages 533 – 543, 2011.

[56] Fabian Sievers, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D. Thompson, and Desmond G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol Syst Biol*, 7:539–539, Oct 2011. 21988835[pmid].

[57] William L Jorgensen, Jayaraman Chandrasekhar, Jeffry D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.

[58] Robert B. Best, Xiao Zhu, Jihyun Shim, Pedro E. M. Lopes, Jeetain Mittal, Michael Feig, and Alexander D. MacKerell, Jr. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ, and side-chain χ_1 and χ_2 dihedral angles. *Journal of Chemical Theory and Computation*, 8(9):3257–3273, 2012.

[59] P. J. Hart, H. D. Pfluger, A. F. Monzingo, T Hollis, and J. D. Robertus. The refined crystal structure of an endochitinase from *Hordeum vulgare L.* seeds at 1.8åresolutiontructure of an endochitinase from *Hordeum vulgare L.* seeds at 1.8åresolution. *Journal of Molecular Biology*, 248:402, 1995.

[60] J. Michael Word, Simon C. Lovell, Jane S. Richardson, and David C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of Molecular Biology*, 285(4):1735–1747, January 1999.

[61] Fukamizo T. Chitinolytic enzymes: catalysis, substrate binding, and their application. *Current Protein and Peptide Science*, 1:105–124, 2000.

[62] B. Iseli, S. Armand, T. Boller, J.M. Neuhaus, and B. Henrissat. Plant chitinases use two different hydrolytic mechanisms. *FEBS Letters*, 382:186–188, 1996.

[63] Z. Hua, C. Zou, S. H. Shiu, and R.D. Vierstra. Phylogenetic comparison of F-Box (FBX) gene superfamily within the plant kingdom reveals divergent evolutionary histories indicative of genomic drift. *PLoS One*, 6:e16219, 2011.

[64] F. Kaschani, M. Shabab, T. Bozkurt, T. Shindo, S. Schornack, C. Gu, M. Ilyas, J. Win, S. Kamoun, and R.A. van der Hoorn. An effector-targeted protease contributes to defense against *Phytophthora infestans* and is under diversifying selection in natural hosts. *Plant Physiology*, 154(4):1794–1804, 2010.

[65] Jaap J. Beintema. Structural features of plant chitinases and chitin-binding proteins. *FEBS Letters*, 350(2):159–163, 1994.

[66] Pooja Kesari, Dipak Narhari Patil, Pramod Kumar, Shailly Tomar, Ashwani Kumar Sharma, and Pravindra Kumar. Structural and functional evolution of chitinase-like proteins from plants. *PROTEOMICS*, 15(10):1693–1705, 2015.

[67] Yoshihito Kitaoku, Naoyuki Umemoto, Takayuki Ohnuma, Tomoyuki Numata, Toki Taira, Shohei Sakuda, and Tamo Fukamizo. A class III chitinase without disulfide bonds from the fern, *Pteris ryukyuensis*: crystal structure and ligand-binding studies. *Planta*, 242:895–907, 2015.

[68] A.C. Terwisscha van Scheltinga, K.H. Kalk, J.J. Beintema, and Dijkstra B.W. Crystal structures of hevamine, a plant defence protein with chitinase and lysozyme activity, and its complex with an inhibitor. *Structure*, 2:1181–1189, 1994.

[69] Noah C. Benson and Valerie Daggett. A chemical group graph representation for efficient high-throughput analysis of atomistic protein simulations. *Journal of Bioinformatics and Computational Biology*, 10(04):1250008, July 2012.

[70] W. Humphrey, A. Dalke, and K. Schulten. VMD: visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38, 27–28, February 1996.

[71] Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of Statistical Software*, 24(1):1–11, 2008.

[72] Carter T. Butts. network: a package for managing relational data in R. *Journal of Statistical Software*, 24(2), 2008.

[73] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[74] Carter T. Butts and Kathleen M. Carley. Some simple algorithms for structural comparison. *Computational and Mathematical Organization Theory*, 11(4):291–305, 2005.

[75] Carter T. Butts. Social network analysis with sna. *Journal of Statistical Software*, 24(6), 2008.

[76] DMF Van Aalten, D Komander, B Synstad, S Gåseidnes, MG Peter, and VGH Eijsink. Structural insights into the catalytic mechanism of a family 18 exo-chitinase. *Proceedings of the National Academy of Sciences*, 98(16):8979–8984, 2001.

[77] Christina M. Payne, Jamil Baban, Svein J. Horn, Paul H. Backe, Andrew S. Arvai, Bjørn Dalhus, Magnar Bjørås, Vincent G. H. Eijsink, Morten Sørlie, Gregg T. Beckham, and Gustav Vaaje-Kolstad. Hallmarks of processivity in glycoside hydrolases from crystallographic and computational studies of the serratia marcescens chitinases. *Journal of Biological Chemistry*, 287(43):36322–36330, 2012.

[78] Jogi Madhuprakash, Avinash Singh, Sanjit Kumar, Mau Sinha, Punit Kaur, Sujata Sharma, Appa R Podile, and Tej P Singh. Structure of chitinase d from serratia

proteamaculans reveals the structural basis of its dual action of hydrolysis and transglycosylation. *International journal of biochemistry and molecular biology*, 4(4):166, 2013.

[79] Svein J. Horn, Pawel Sikorski, Jannicke B. Cederkvist, Gustav Vaaje-Kolstad, Morten Sórlie, Bjórnar Synstad, Gert Vriend, Kjell M. Vårum, and Vincent G. H. Eijsink. Costs and benefits of processivity in enzymatic degradation of recalcitrant polysaccharides. *Proceedings of the National Academy of Sciences of the United States of America*, 103(48):18089–18094, 2006.

[80] I. von Ossowski, Ståhlberg J., A. Koivula, K. Piens, D. Becker, H. Boer, R. Harle, M. Harris, C. Divne, S. Mahdi, Y. Zhao, Driguez. H., M. Claeyssens, M.L. Sinnott, and T.T. Teeri. Engineering the exo-loop of *Trichoderma reesei* cellobiohydrolase, Cel7A. a comparison with *Phanerochaete chrysosporium* Cel7D. *Journal of Molecular Biology*, 333(4):817–829, 2003.

[81] F. Meins, B. Fritig, H.J.M. Linthorst, J.D. Mikkelsen, J.M. Neuhaus, and J. Ryals. Plant chitinase genes. *Plant Molecular Biology Reporter*, 12(2):S22–S28, 1994.

[82] J.M. Neuhaus, B. Fritig, H.J.M. Linthorst, F. Meins, J.D. Mikkelsen, and J. Ryals. A revised nomenclature for chitinase genes. *Plant Mol Biol Rep*, 14:102–104, 1996.

[83] Pallinti Purushotham, P. V. Parvati Sai Arun, Jogadhenu S. S. Prakash, and Appa Rao Podile. Chitin binding proteins act synergistically with chitinases in *Serratia proteamaculans* 568. *PLoS ONE*, 7(5):e36714, 2012.

[84] Kana Ishisaki, Yuji Honda, Hajime Taniguchi, Naoya Hatano, and Tatsuro Hamada. Heterogonous expression and characterization of a plant class IV chitinase from the pitcher of the carnivorous plant *Nepenthes alata*. *Glycobiology*, 22(3):345–351, 2012.

[85] Darrian Talamantes, Nazmehr Biabini, Hoang Dang, Kenza Abdoun, and Renaud Berlemont. Natural diversity of cellulases, xylanases, and chitinases in bacteria. *Biotechnology for Biofuels*, 9(133):DOI: 10.1186/s13068–016–0538–6, 2016.

[86] Takeshi Tanaka, Shinsuke Fujiwara, Shingo Nishikori, Toshiaki Fukui, Masahiro Takagi, and Tadayuki Imanaka. A unique chitinase with dual active sites and triple substrate binding sites from the hyperthermophilic archaeon *Pyrococcus kodakaraensis* KOD1. *Applied and Environmental Microbiology*, 65(12):5338–5344, 1999.

[87] Yoshihiro Kikkawa, Masato Fukuda, Ayumi Kashiwada, Kiyomi Matsuda, Masatoshi Kanesato, Masahisa Wada, Tadayuki Imanaka, and Takeshi Tanaka. Binding ability of chitinase onto cellulose: an atomic force microscopy study. *Polymer Journal*, 43:742–744, 2011.

[88] Marcia M. Chaudet, Todd A. Naumann, Neil P.J. Price, and David R. Rose. Crystallographic structure of ChitA, a glycoside hydrolase family 19, plant class IV chitinase from *Zea mays*. *Protein Science*, 23(5):586–593, 2014.

[89] M.D. Andersen, A. Jensen, J.D. Robertus, R. Leah, and K. Skriver. Heterologous expression and characterization of wild-type and mutant forms of a 26 kda endochitinase from barley (hordeum vulgare l.). *Biochemical Journal*, 822:815–822, 1997.

[90] G. Garcia-Casado, C. Carmen, I. Allona, R. Casado, L.F. Pacios, C. Aragoncillo, and L. Gomez. Site-directed mutagenesis of active site residues in a class i endochitinase from chestnut seeds. *Glycobiology*, 8(10):1021–1028, 1998.

[91] C.M. Tang, M.L. Chye, S. Ramalingam, S.W. Ouyang, K.J. Zhao, W. Ubhayasekera, and S.L. Mowbray. Functional analyses of the chitin-binding domains and the catalytic domain of brassica juncea chitinase bjchi1. *Plant Molecular Biology*, 56:285–298, 2004.

[92] Wimal Ubhayasekera, Reetika Rawat, Sharon Wing Tak Ho, Malgorzata Wiweger, Sara Von Arnold, Mee-Len Chye, and Sherry L. Mowbray. The first crystal structures of a family 19 class iv chitinase: the enzyme from norway spruce. *Plant Molecular Biology*, 71(3):277–289, 2009.

[93] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, 1994.

[94] S. B. Seidman. Network structure and minimum degree. *Social Networks*, 5:269–287, 1983.

[95] D. B. West. *Introduction to Graph Theory*. Prentice Hall, Upper Saddle River, NJ, 1996.

[96] Olga Serra, Subhasish Chatterjee, Wenlin Huang, and Ruth E. Stark. Review: What nuclear magnetic resonance can tell us about protective tissues. *Plant Science*, 195:120–124, 2012.

[97] Subhasish Chatterjee, Antonio J. Matas, Tal Isaacson, Cindie Kehlet, Jocelyn K.C. Rose, and Ruth E. Stark. Solid-state 13C NMR delineates the architectural design of biopolymers in native and genetically altered tomato fruit cuticles. *Biomacromolecules*, 17(1):215–224, 2016.

[98] Kathleen Clauss, Alfred Baumert, Manfred Nimtz, Carsten Milkowski, and Dieter Strack. Role of a gdsl lipase-like protein as sinapine esterase in brassicaceae. *The Plant Journal*, 53(5):802–813, 2008.

[99] Yukio Kikuta, Hirokazu Ueda, Masafumi Takahashi, Tomonori Mitsumori, Gen Yamada, Koji Sakamori, Kengo Takeda, Shogo Furutani, Koji Nakayama, Yoshio Katsuda, et al. Identification and characterization of a gdsl lipase-like protein that catalyzes the ester-forming reaction for pyrethrin biosynthesis in *Tanacetum cinerariifolium*–a new target for plant protection. *The Plant Journal*, 71(2):183–193, 2012.

[100] Casimir C. Akoh, Guan-Chiun Lee, Yen-Chywan Liaw, Tai-Huang Huang, and Jei-Fu Shaw. GDSL family of serine esterases/lipases. *Progress in Lipid Research*, 43(534–552), 2004.

[101] Slim Abdelkafi, Hiroyuki Ogata, Nathalie Barouh, Benjamin Fouquet, Régine Lebrun, Michel Pina, Frantz Scheirlinckx, Pierre Villeneuve, and Frédéric Carrière. Identification and biochemical characterization of a GDSL-motif carboxylester hydrolase from *Carica papaya* latex. *Biochimica et Biophysica Acta*, 1791:1048–1056, 2009.

[102] A. El Moussaoui, M. Nijs, R. Paul, C.and Wintjens, J. Vincentelli, M. Azarkan, and Y. Looze. Revisiting the enzymes stored in the laticifers of *Carica papaya* in the context of their possible participation in the plant defence mechanism. *Cellular and Molecular Life Sciences*, 58:556–570, 2001.

[103] C. Upton and J.T. Buckley. A new family of lipolytic enzymes? *Trends in Biochemical Science*, 20:178–179, 1995.

[104] M. García-Lorenzo, A. Sjödin, S. Jansson, and C. Funk. Protease gene families in populus and arabidopsis. *BMC Plant Biology*, 6:30, 2006.

[105] Xuemin Wang. Plant phospholipases. *Annual review of plant biology*, 52(1):211–231, 2001.

[106] B Profotová, L Burketová, Z Novotná, J Martinec, and O Valentová. Involvement of phospholipases c and d in early response to sar and isr inducers in brassica napus plants. *Plant Physiology and Biochemistry*, 44(2-3):143–151, 2006.

[107] C. L. An, E. Fukusaki, and A. Kobayashi. Aspartic proteinases are expressed in pitchers of the carnivorous plant *Nepenthes alata* Blanco. *Planta*, 214:661–667, 2002.

[108] S.B.P. Athauda, K. Matsumoto, S. Rajapakshe, M. Kuribayashi, N. Kojima, M.and Kubomura-Yoshida, A. Iwamatsu, C. Shibata, H. Inoue, and K. Takahashi. Enzymic and structural characterization of nepenthesin, a unique member of a novel subfamily of aspartic proteinases. *Biochemical Journal*, 381:295–306, 2004.

[109] K. Takahashi, S.B. Athauda, K. Matsumoto, S. Rajapakshe, M. Kuribayashi, M. Kojima, N. Kubomura-Yoshida, A. Iwamatsu, C. Shibata, and H. Inoue. Nepenthesin, a unique member of a novel subfamily of aspartic proteinases: enzymatic and structural characteristics. *Current Protein and Peptide Science*, 6(6):513–525, 2005.

[110] Gerhard Lorkowski. Gastrointestinal absorption and biological activities of serine and cysteine proteases of animal and plant origin: review on absorption of serine and cysteine proteases. *International Journal of Physiolology, Pathophysiology and Pharmacology*, 4(1):10–27, 2012.

[111] K. Duwadi, L. Chen, R. Menassa, and S. Dhaubhadel. Identification, characterization and down-regulation of cysteine protease genes in tobacco for use in recombinant protein production. *PLoS ONE*, 10(7):e0130556, 2015.

[112] N.D. Rawlings, M. Waller, A.J. Barrett, and A. Bateman. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, 42:D503–D509, 2014.

[113] M. Novinec and B. Lenarčič. Papain-like peptidases: structure, function, and evolution. *Biomolecular Concepts*, 4(3):287–308, 2013.

[114] V. Turk, V. Stoka, O. Vasiljeva, M. Renko, T. Sun, B. Turk, and D. Turk. Cysteine cathepsins: From structure, function and regulation to new frontiers. *Biochimica et Biophysica Acta*, 1824:68–88, 2012.

[115] F. Buch, W. E. Kaman, F. J. Bikker, A. Yilamujiang, and A. Mithöfe. Nepenthesin protease activity indicates digestive fluid dynamics in carnivorous *Nepenthes* plants. *PLoS ONE*, 10(3):e0118853, 2015.

[116] M. Libiaková, K. Floková, O. Novák, L. Slováková, and A. Pavlovič. Abundance of cysteine endopeptidase dionain in digestive fluid of Venus flytrap (dionaea muscipula ellis) is regulated by different stimuli from prey through jasmonates. *PLoS ONE*, 9:e104424, 2014.

[117] Aaron M. Ellison and Nicholas J. Gotelli. Energetics and the evolution of carnivorous plants—Darwin's 'most wonderful plants in the world'. *Journal of Experimental Botany*, 60(1):19–42, 2009.

[118] Y. Nakamura, M. Reichelt, V. E. Mayer, and A. Mithöfer. Jasmonates trigger prey-induced formation of outer stomach in carnivorous sundew plants. *Proceedings of the Royal Society B*, 280(20130228), 2013.

[119] E. P. Beers, A. M. Jones, and A. W. Dickerman. The S8 serine, C1A cysteine and A1 aspartic protease families in Arabidopsis. *Phytochemistry*, 65:43–58, 2004.

[120] Stefan George van Wyk, Magdeleen Du Plessis, Christoper Ashley Cullis, Karl Josef Kunert, and Barend Juan Vorster. Cysteine protease and cystatin expression and activity during soybean nodule development and senescence. *BMC Plant Biology*, 14:294, 2014.

[121] T. Vernet, D. C. Tessier, J. Chatellier, C. Plouffe, T. S. Lee, D.Y. Thomas, and R. Storer, A.C.and M'enard. Structural and functional roles of asparagine 175 in the cysteine protease papain. *The Journal of Biological Chemistry*, 270(28):16645–16652, 1995.

[122] Robert Ménard, Céline Plouffe, Pierre Laflamme, Thierry Vemet, Daniel C. Tessier, David Y. Thomas, and Andrew C. Storer. Modification of the electrostatic environment is tolerated in the oxyanion hole of the cysteine protease papain. *Biochemistry*, 34:464–471, 1995.

[123] K.H. Choi and R.A. Laursen. Amino-acid sequence and glycan structures of cysteine proteases with proline specificity from ginger rhizome Zingiber officinale. *European Journal of Biochemistry*, 267:1516–1526, 2000.

[124] S. Visal, M.A.J. Taylor, and D. Michaud. The proregion of papaya proteinase IV inhibits Colorado potato beetle digestive cysteine proteinases. *FEBS Letters*, 434:401–405 PMID: 9742962, 1998.

[125] F.B. Silva, J.A. Batista, B.M. Marra, R.R. Fragoso, A.C. Monteiro, E.L. Figueira, and M.F. Grossi-de-S{'a. Prodomain peptide of HGCP-Iv cysteine proteinase inhibits nematode cysteine proteinases. *Genetics and Molecular Research*, 3:342–355 PMID: 15614726, 2004.

[126] M. E. Santamaria, A. Arnaiz, M. Diaz-Mendoza, and I. Martinez, M. andDiaz. Inhibitory properties of cysteine protease pro-peptides from barley confer resistance to spider mite feeding. *PLoS ONE*, 10(6):e0128323, 2015.

[127] B.M. Marra, D.S. Souza, J.N. Aguiar, A.A. Firmino, R.P. Sarto, F.B. Silva, C.D. Almeida, J.E. Cares, M.V. Continho, C. Martins-de-Sá, O.L. Franco, and M.F. Grossi-de-Sá. Protective effects of a cysteine proteinase propeptide expressed in transgenic soybean roots. *Peptides*, 30(5):825–831, 2009.

[128] G.Z. Rovenska, R. Zemek, J.E.U. Schmidt, and A. Hilbeck. Altered host plant preference of Tetranychus urticae and prey preference of its predator Phytoseiulus persimilis (Acari: Tetranychidae, Phytoseiidae) on transgenic Cry3Bb-eggplants. *Biological Control*, 33:293–300 PMID: 15781137, 2005.

[129] K.M. Karrer, S.L. Peiffer, and M.E. DiTomas. Two distinct gene subfamilies within the family of cysteine protease genes. *Proceedings of the National Academy of Sciences of the United States of America*, 90:3063–3067, 1993.

[130] W. D. Lu, L. Funkelstein, T. Toneff, T. Reinheckel, C. Peters, and V. Hook. Cathepsin H functions as an aminopeptidase in secretory vesicles for production of enkephalin and galanin peptide neurotransmitters. *Journal of Neurochemistry*, 122:512 – 522, 2012.

[131] Barry C. Holwerda and John C. Rogers. Purification and characterization of Aleurain 1: A plant thiol protease functionally homologous to mammalian Cathepsin H. *Plant Physiology*, 99:848–855, 1992.

[132] D. K Nägler, T. Sulea, and R. Ménard. Full-length cDNA of human cathepsin F predicts the presence of a cystatin domain at the N-terminus of the cysteine protease zymogen. *Biochemical and Biophysical Research Communications*, 257:313 – 318, 1999.

[133] Anuradha Waghray, Daniel Keppler, Bonnie F. Sloane, Lucia Schuger, and Yong Q. Chen. Analysis of a truncated form of cathepsin h in human prostate tumor cells. *Journal of Biological Chemistry*, 277(13):11533–11538, 2002.

[134] Zala Jevnikar, Matija Rojnik, Polona Jamnik, Bojan Doljak, Urša Pečar Fonovi/'c, and Janko Kos. Cathepsin H mediates the processing of talin and regulates migration of prostate cancer cells. *The Journal of Biological Chemistry*, 288:2201–2209, 2013.

[135] G. Guncar, M. Podobnik, J. Pungercar, B. Strukelj, V. Turk, and D. Turk. Crystal structure of porcine cathepsin H determined at 2.1 åresolution: location of the mini-chain C-terminal carboxyl group defines cathepsin H aminopeptidase function. *Structure*, 6:51–61, 1998.

[136] J. Dodt and J. Reichwein. Human cathepsin H: deletion of the mini-chain switches substrate specificity from aminopeptidase to endopeptidaseh: deletion of the mini-chain switches substrate specificity from aminopeptidase to endopeptidase. *Biol Chem.*, 384(9):1327–1332, 2003.

[137] Alison Baker and Rupesh Paudyal. The life of the peroxisome: from birth to death. *Current Opinion in Plant Biology*, 22:39–47, 2014.

[138] Klaas J. van Wijk. Protein maturation and proteolysis in plant plastids, mitochondria, and peroxisomes. *Annual Review of Plant Biology*, 66:75–111, 2015.

[139] José M. Palma, Luisa M. Sandalio, F. Javier Corpas, María C. Romero-Puertas, Iva McCarthy, and Luis A. del Río. Plant proteases, protein degradation, and oxidative stress: role of peroxisomes. *Plant Physiology and Biochemistry*, 40:521–530, 2002.

[140] Thomas Lingner, Amr R. Kataya, Gerardo E. Antonicelli, Aline Benichou, Kjersti Nilssen, Xiong-Yan Chen, Tanja Siemsen, Burkhard Morgenstern, Peter Meinicke, and Sigrun Reumann. Identification of novel plant peroxisomal targeting signals by a combination of machine learning methods and *in vivo* subcellular targeting analyses. *The Plant Cell*, 23:1556–1572, 2011.

[141] Takashi Okamoto, Hiroshi Nakayama, Kazuo Seta, Toshiaki Isobe, and Takao Minamikawa. Posttranslational processing of a carboxy-terminal propeptide containing a KDEL sequence of plant vacuolar cysteine endopeptidase (SH-EP). *FEBS Letters*, 351(1):31–34, 1994.

[142] Manuel E. Than, Michael Helm, David J. Simpson, Friedrich Lottspeich, Robert Huber, and Christine Gietl. The 2.0 åcrystal structure and substrate specificity of the KDEL-tailed cysteine endopeptidase functioning in programmed cell death of *Ricinus communis* endosperm. *Journal of Molecular Biology*, 336:1103–1116, 2004.

[143] Takayuki Shindo, Johana C. Misas-Villamil, Anja C. Hörger, Jing Song, and Renier A. L. van der Hoorn. A role in immunity for arabidopsis cysteine protease RD21, the ortholog of the tomato immune protease C14. *PLoS ONE*, 7(1):e29317, 2012.

[144] Y. Hayashi, K. Yamada, T. Shimada, R. Matsushima, N. K. Nishizawa, M. Nishimura, and I. Hara-Nishimura. A proteinase-storing body that prepares for cell death or stresses in the epidermal cells of *Arabidopsis*. *Plant and Cell Physiology*, 42:894–899, 2001.

[145] Kenji Yamada, Ryo Matsushima, Mikio Nishimura, and Ikuko Hara-Nishimura. A slow maturation of a cysteine protease with a granulin domain in the vacuoles of senescing *Arabidopsis* leaves. *Plant Physiology*, 127:1626–1634, 2001.

[146] Christian Gu, Mohammed Shabab, Richard Strasser, Pieter J. Wolters, Takayuki Shindo, Melanie Niemer, Farnusch Kaschani, Lukas Mach, and Renier A. L. van der Hoorn. Post-translational regulation and trafficking of the granulin-containing protease RD21 of *Arabidopsis thaliana*. *PLoS ONE*, 7(3):e32422, 2012.

[147] A Bateman and H. P. J. Bennett. Granulins: the structure and function of an emerging family of growth factors. *Journal of Endocrinology*, 158:145–151, 1998.

[148] Wojciech Pulawski, Umesh Ghoshdastider, Vincenza Andrisano, and Slawomir Filipek. Ubiquitous amyloids. *Applied Biochemistry and Biotechnology*, 166(7):1626–1643, 2012.

[149] Fabrizio Chiti and Christopher M Dobson. Protein misfolding, amyloid formation, and human disease: A summary of progress over the last decade. *Annual Review of Biochemistry*, 86:27–68, 2017.

[150] Fabrizio Chiti, Paul Webster, Niccolò Taddei, Anne Clark, Massimo Stefani, Giampietro Ramponi, and Christopher M Dobson. Designing conditions for in vitro formation of amyloid protofilaments and fibrils. *Proceedings of the National Academy of Sciences*, 96(7):3590–3594, 1999.

[151] Margaret Sunde, Louise C Serpell, Mark Bartlam, Paul E Fraser, Mark B Pepys, and Colin CF Blake. Common core structure of amyloid fibrils by synchrotron x-ray diffraction. *Journal of Molecular Biology*, 273(3):729–739, 1997.

[152] Thorsten Lührs, Christiane Ritter, Marc Adrian, Dominique Riek-Loher, Bernd Bohrmann, Heinz Döbeli, David Schubert, and Roland Riek. 3D structure of Alzheimer's amyloid-β (1–42) fibrils. *Proceedings of the National Academy of Sciences of the United States of America*, 102(48):17342–17347, 2005.

[153] Jakob T Nielsen, Morten Bjerring, Martin D Jeppesen, Ronnie O Pedersen, Jan M Pedersen, Kim L Hein, Thomas Vosegaard, Troels Skrydstrup, Daniel E Otzen, and Niels C Nielsen. Unique identification of supramolecular structures in amyloid fibrils by solid-state NMR spectroscopy. *Angewandte Chemie*, 121(12):2152–2155, 2009.

[154] Jed JW Wiltzius, Stuart A Sievers, Michael R Sawaya, Duilio Cascio, Dmitriy Popov, Christian Riekel, and David Eisenberg. Atomic structure of the cross-β spine of islet amyloid polypeptide (amylin). *Protein Science*, 17(9):1467–1474, 2008.

[155] Marcin I Apostol, Jed JW Wiltzius, Michael R Sawaya, Duilio Cascio, and David Eisenberg. Atomic structures suggest determinants of transmission barriers in mammalian prion disease. *Biochemistry*, 50(13):2456–2463, 2011.

[156] Jun-Xia Lu, Wei Qiang, Wai-Ming Yau, Charles D Schwieters, Stephen C Meredith, and Robert Tycko. Molecular structure of β-amyloid fibrils in Alzheimer's disease brain tissue. *Cell*, 154(6):1257–1268, 2013.

[157] Luc Bousset, Laura Pieri, Gemma Ruiz-Arlandis, Julia Gath, Poul Henning Jensen, Birgit Habenstein, Karine Madiona, Vincent Olieric, Anja Böckmann, Beat H. Meier, and Ronald Melkia. Structural and functional characterization of two alpha-synuclein strains. *Nature Communications*, 4:2575, 2013.

[158] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[159] Yue Yu, Gianmarc Grazioli, Megha H Unhelkar, Rachel W Martin, and Carter T Butts. Network hamiltonian models reveal pathways to amyloid fibril formation. *Scientific reports*, 10(1):1–11, 2020.

[160] Dean Lusher, Johan Koskinen, and Garry Robins. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Cambridge University Press, Cambridge, 2012.

[161] Louise C Serpell, Margareth Sunde, and Colin CF Blake. The molecular basis of amyloidosis. *Cellular and Molecular Life Sciences*, 53(11):871–887, 1997.

[162] William Humphrey, Andrew Dalke, Klaus Schulten, et al. Vmd: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.

[163] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.

[164] Jose A Rodriguez, Magdalena I Ivanova, Michael R Sawaya, Duilio Cascio, Francis Reyes, Dan Shi, Smriti Sangwan, Elizabeth L Guenther, Lisa M Johnson, Meng Zhang, et al. Structure of the toxic core of α-synuclein from invisible crystals. *Nature*, 525(7570):486, 2015.

[165] Kentaro Iwata, Toshimichi Fujiwara, Yoh Matsuki, Hideo Akutsu, Satoshi Takahashi, Hironobu Naiki, and Yuji Goto. 3D structure of amyloid protofilaments of β2-microglobulin fragment probed by solid-state NMR. *Proceedings of the National Academy of Sciences*, 103(48):18119–18124, 2006.

[166] Anant K Paravastu, Richard D Leapman, Wai-Ming Yau, and Robert Tycko. Molecular structural basis for polymorphism in Alzheimer's β-amyloid fibrils. *Proceedings of the National Academy of Sciences*, 105(47):18349–18354, 2008.

[167] Yiling Xiao, Buyong Ma, Dan McElheny, Sudhakar Parthasarathy, Fei Long, Minako Hoshi, Ruth Nussinov, and Yoshitaka Ishii. Aβ(1-42) fibril structure illuminates self-recognition and replication of amyloid in Alzheimer's disease. *Nature Structural & Molecular Biology*, 22(6):499–505, 2015.

[168] Neil Ferguson, Johanna Becker, Henning Tidow, Sandra Tremmel, Timothy D Sharpe, Gerd Krause, Jeremy Flinders, Miriana Petrovich, John Berriman, Hartmut Oschkinat, et al. General structural motifs of amyloid protofilaments. *Proceedings of the National Academy of Sciences*, 103(44):16248–16253, 2006.

[169] Lothar Gremer, Daniel Schölzel, Carla Schenk, Elke Reinartz, Jörg Labahn, Raimond BG Ravelli, Markus Tusche, Carmen Lopez-Iglesias, Wolfgang Hoyer, Henrike Heise, et al. Fibril structure of amyloid-β (1–42) by cryo–electron microscopy. *Science*, 358(6359):116–119, 2017.

[170] Michael R Sawaya, Shilpa Sambashivan, Rebecca Nelson, Magdalena I Ivanova, Stuart A Sievers, Marcin I Apostol, Michael J Thompson, Melinda Balbirnie, Jed JW Wiltzius, Heather T McFarlane, et al. Atomic structures of amyloid cross-β spines reveal varied steric zippers. *Nature*, 447(7143):453, 2007.

[171] Jed JW Wiltzius, Meytal Landau, Rebecca Nelson, Michael R Sawaya, Marcin I Apostol, Lukasz Goldschmidt, Angela B Soriaga, Duilio Cascio, Kanagalaghatta Rajashankar, and David Eisenberg. Molecular mechanisms for protein-encoded inheritance. *Nature Structural & Molecular Biology*, 16(9):973–978, 2009.

[172] Dan Li, Eric M Jones, Michael R Sawaya, Hiroyasu Furukawa, Fang Luo, Magdalena Ivanova, Stuart A Sievers, Wenyuan Wang, Omar M Yaghi, Cong Liu, et al. Structure-based design of functional amyloid materials. *Journal of the American Chemical Society*, 136(52):18044–18051, 2014.

[173] Anne K Schütz, Toni Vagt, Matthias Huber, Oxana Y Ovchinnikova, Riccardo Cadalbert, Joseph Wall, Peter Güntert, Anja Böckmann, Rudi Glockshuber, and Beat H Meier. Atomic-resolution three-dimensional structure of amyloid β fibrils bearing the Osaka mutation. *Angewandte Chemie International Edition*, 54(1):331–335, 2015.

[174] Nikolaos G Sgourakis, Wai-Ming Yau, and Wei Qiang. Modeling an in-register, parallel "Iowa" Aβ fibril structure using solid-state NMR data from labeled samples with Rosetta. *Structure*, 23(1):216–227, 2015.

[175] Valentina Tozzini. Coarse-grained models for proteins. *Current Opinion in Structural Biology*, 15(2):144–150, 2005.

[176] Alexander J. Pak and Gregory A. Voth. Advances in coarse-grained modeling of macromolecular complexes. *Current Opinion in Structural Biology*, 52:119–126, 2018.

[177] Ana Rojas, Adam Liwo, Dana Browne, and Harold A. Scheraga. Mechanism of fiber assembly: Treatment of Aβ peptide aggregation with a coarse-grained united-residue force field. *Journal of Molecular Biology*, 404(3):537–552, 2010.

[178] Weihua Zheng, Min-Yeh Tsai, and Peter G. Wolynes. Comparing the aggregation free energy landscapes of amyloid beta(1-42) and amyloid beta(1-40). *Journal of the American Chemical Society*, 139:16666–16676, 2017.

[179] Robin Roychaudhuri, Mingfeng Yang, Minako M. Hoshi, and David B. Teplow. Amyloid β-protein assembly and Alzheimer's disease. *The Journal of Biological Chemistry*, 8:4749–4753, 284.

[180] Matthias Schmidt, Alexis Rohou, Keren Lasker, Jay K. Yadav, Cordelia Schiene-Fischer, Marcus Fändrich, and Nikolaus Grigorieff. Peptide dimer structure in an Aβ(1–42) fibril visualized with cryo-EM. *Proceedings of the National Academy of Sciences of the United States of America*, 112(38):11858–11863, 2015.

[181] Marielle Aulikki Wälti, Francesco Ravotti, Hiromi Arai, Charles G Glabe, Joseph S Wall, Anja Böckmann, Peter Güntert, Beat H Meier, and Roland Riek. Atomic-resolution structure of a disease-relevant Aβ (1–42) amyloid fibril. *Proceedings of the National Academy of Sciences*, 113(34):E4976–E4984, 2016.

[182] David S. Eisenberg and Michael R. Sawaya. Implications for Alzheimer's disease of an atomic resolution structure of amyloid-β(1–42) fibrils. *Proceedings of the National Academy of Sciences of the United States of America*, 113(34):9398–9400, 2016.

[183] Paul W. Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, 76(373):33–50, 1981.

[184] Songming Chen, Valerie Berthelier, J Bradley Hamilton, Brian O'Nuallai, and Ronald Wetzel. Amyloid-like features of polyglutamine aggregates and their assembly kinetics. *Biochemistry*, 41(23):7391–7399, 2002.

[185] Wei Qiang, Kevin Kelley, and Robert Tycko. Polymorph-specific kinetics and thermodynamics of β-amyloid fibril growth. *Journal of the American Chemical Society*, 135(18):6860–6871, 2013.

[186] Kyle W. Roskamp, David M. Montelongo, Chelsea D. Anorma, Diana N. Bandak, Janine A. Chua, Kurtis Malecha, and Rachel W. Martin. Thermal-, pH-, and UV-induced aggregation of human γS-crystallin and its aggregation-prone G18V variant. *Investigative Ophthalmology and Visual Science*, 58:2397–2405, 2017.

[187] Dorothea Pinotsi, Alexander K Buell, Christopher M Dobson, Gabriele S Kaminski Schierle, and Clemens F Kaminski. A label-free, quantitative assay of amyloid fibril growth based on intrinsic fluorescence. *ChemBioChem*, 14(7):846, 2013.

[188] Daizo Hamada and Christopher M Dobson. A kinetic study of β-lactoglobulin amyloid fibril formation promoted by urea. *Protein Science*, 11(10):2417–2426, 2002.

[189] Zbynek Kozmik, Jana Ruzickova, Kristyna Jonasova, Yoshifumi Matsumoto, Pavel Vopalensky, Iryna Kozmikova, Hynek Strnad, Shoji Kawamura, Joram Piatigorsky, Vaclav Paces, et al. Assembly of the cnidarian camera-type eye from vertebrate-like components. *Proceedings of the National Academy of Sciences*, 105(26):8989–8993, 2008.

[190] Fon-Yi Yin, Ya-Huei Chen, Chung-Ming Yu, Yu-Chin Pon, and Hwei-Jen Lee. Kinetic refolding barrier of guanidinium chloride denatured goose δ-crystallin leads to regular aggregate formation. *Biophysical journal*, 93(4):1235–1245, 2007.

[191] Joseph Horwitz. Alpha crystallin: the quest for a homogeneous quaternary structure. *Experimental eye research*, 88(2):190–194, 2009.

[192] Domarin Khago. *Understanding the Molecular Basis of Transparency and Refraction in the Eye Lens*. PhD thesis, UC Irvine, 2016.

[193] J. Horwitz. Alpha-crystallin. *Experimental Eye Research*, 76:145—153, 2003.

[194] Z. Kozmik, K. Shivalingappa, J. Ruzickova, K. Jonasova, V. Paces, C. Vlcek, and J. Piatigorsky. Cubazoan crystallins: evidence for convergent evolution of pax regulatory sequences. *Evolution & Development*, 10:52—61, 2008.

[195] F. Yin, Y. Chen, C. Yu, Y. Pon, and H. Lee. Kinetic refolding barrier of guanidinium chloride denatured goose δ-crystallin leads to regular aggregate formation. *Biophysical Journal*, 93:1235—1245, 2007.

[196] Stanislav I. Tomarev, Sambath Chung, and Joram Piatigorsky. Glutathione s-transferase and s-crystallins of cephalopods: Evolution from active enzyme to lens-refractive proteins. *Journal of Molecular Evolution*, 41:1048—1056, 1995.

[197] T.D. Ingolia and E.A. Craig. Four small drosophila heat shock proteins are related to each other and to mammalian alpha-crystallins. *Proceedings of the National Academy of Sciences of the United States of America*, 79:2360–2364, 1982.

[198] W. W. de Jong, J. A. M. Leunissen, and C. E. M. Voorter. Evolution of the alpha-crystallin/small heat shock protein family. *Molecular Biology and Evolution*, 10:103–126, 1993.

[199] J. Piatigorsky and Z. Kozmik. Cubozoan jellyfish: An evo/devo model for eyes and other sensory systems. *International Journal of Developmetnal Biology*, 48:719—729, 2004.

[200] Z. Kozmik, J. Ruzickova, K. Jonasova, Y. Matsumoto, P. Vopalensky, I. Kozmikoza, H. Strnad, S. Kawamura, J. Piatigorsky, V. Paces, and C. Vlcek. Assembly of the cnidarian camera-type eye from vertebrate-like components. *Proceedings of the National Academy of Sciences of the United States of America*, 105:8989—8993, 2008.

[201] J. Piatigorsky, J. Horwitz, and B. Norman. J1-crystallins of the cubomedusan jellyfish lens constitute a novel family encoded in at least three intronless genes. *Journal of Biological Chemistry*, 268:11894—11901, 1993.

[202] J. Piatigorsky, B. Norman, L. Dishaw, L. Kos, J. Horwitz, P. Steinbach, and Z. Kozmik. J3-crystallin of the jellyfish lens: Similarity to saposins. *Proceedings of the National Academy of Sciences of the United States of America*, 28:12362—12367, 2001.

[203] M. Bloemendal and A. Toumadje. Bovine lens crystallins do contain helical structure. *Biochimica et Biophysica Acta*, 1432:234—238, 1999.

[204] S. Altschul, A. Madden, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped blast and psi-blast;a new generation of prtein database search programs. *Nucleic Acids Research*, 25:3389—3402, 1997.

[205] E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M.R. Wilkins, R.D. Appel, and A. Bairoch. *The Proteomics Protocols Handbook*, chapter Protein identification and analysis tools on the ExPASy Server. Humana Press, 2005.

[206] Srivatsan Raman, Robert Vernon, James Thompson, Michael Tyka, Ruslan Sadreyev, Jimin Pei, David Kim, Elizabeth Kellogg, Frank DiMaio, Oliver Lange, et al. Structure prediction for casp8 with all-atom refinement using rosetta. *Proteins: Structure, Function, and Bioinformatics*, 77(S9):89–99, 2009.

[207] Yifan Song, Frank DiMaio, Ray Yu-Ruei Wang, David Kim, Chris Miles, TJ Brunette, James Thompson, and David Baker. High-resolution comparative modeling with rosettacm. *Structure*, 21(10):1735–1742, 2013.

[208] Jianyi Yang, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. The i-tasser suite: protein structure and function prediction. *Nature methods*, 12(1):7–8, 2015.

[209] Ambrish Roy, Alper Kucukural, and Yang Zhang. I-tasser: a unified platform for automated protein structure and function prediction. *Nature protocols*, 5(4):725–738, 2010.

[210] Ronit Freeman and Itamar Willner. Optical molecular sensing with semiconductor quantum dots (qds). *Chemical Society Reviews*, 41(10):4067–4085, 2012.

[211] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfiefer, and A. Bax. Nmrpipe: A multidimensional spectral processing system based on unix pipes. *Journal of Biomolecular NMR*, 6:277—293, 1995.

[212] W. F. Vranken, W. Boucher, T. J. Stevens, R. H. Fogh, A. Pjon, M. Llinas, E. L. Ulrich, J. L. Markley, J. Ionides, and E. D. Laue. The ccpn data model for nmr spectroscopy: Development of a software pipeline. *Proteins*, 59(4):687—696, 2005.

[213] Ye Li, Vassiliy Lubchenko, and Peter G Vekilov. The use of dynamic light scattering and brownian microscopy to characterize protein aggregation. *Review of Scientific Instruments*, 82(5):053106, 2011.

[214] Stephan Grzesiek and AD Bax. Improved 3d triple-resonance nmr techniques applied to a 31 kda protein. *Journal of Magnetic Resonance (1969)*, 96(2):432–440, 1992.