Bioinformatic Sequence and Structural Analysis for Amyloidogenicity in Prions and Other Proteins

Deena Mohamad Ameen Gendoo

Doctor of Philosophy

Department of Biology
McGill University
Montreal, Quebec, Canada
April 2012

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

DEDICATION

To my wonderful parents:

A constant source of inspiration,
The object of my love and admiration,
And the driving force behind my motivation.

No words can express my extreme gratitude
for your unconditional love, unwavering support,
and infinite patience.

Thank you for everything you have done for me,
you are the role models I aspire to be.

ACKNOWLEDGEMENTS

Studying with Professor Paul Harrison has been a true learning experience, and I hope to continue applying the skills I learned from him in my future career. I am deeply grateful Paul that you have introduced me to the world of prions and amyloid biology, and opened my eyes to the bioinformatic research potential in these fields. Thank you for your continuous encouragement and guidance as I explored different projects and learned new skills. Your patience, dedication, and willingness to support me, scientifically and financially, have helped me grow as a bioinformatician. I am deeply indebted to your mentorship.

I would like to thank members of my lab, both past and present, especially Djamel Harbi and David de Lima Morais, for always offering advice and help when it was needed, but most importantly, for your friendship.

I am grateful to members of my supervisory committee, Dr. Andrea LeBlanc and Dr. Reza Salavati, for your support and guidance and stimulating discussions during supervisory committee meetings.

I would like to extend my sincerest thanks to Dr. Jerome Waldispuhl, for his fruitful discussions and suggestions on different aspects of my work, and for including me within his computational structural biology group. It has been a pleasure to interact and exchange ideas with different members of this group, including Mohamed Smaoui and Vladimir Reinharz, as well as James Wagner from the McGill Centre for Bioinformatics. Thank you Jerome for helping broaden my bioinformatics family.

Completion of this thesis would not have been possible without financial support from several funding sources, including fellowships from McGill University and the Biology department, the CIHR Systems Biology Training Program, and PrioNet Canada and NSERC funding through Paul's grants. I am particularly indebted to PrioNet Canada for supporting me to attend several PrPCanada and PRION conferences over the years, which have opened my mind to a number of research possibilities and allowed me to interact with prominent professionals in prion and amyloid biology. I am also thankful to PrioNet Canada and the CIHR Systems Biology Training Program for supporting me to

attend a number of workshops that have been instrumental in both my research and career development.

Last but not least, I would like to express my deep appreciation to a number of individuals that have always been there to assuage my concerns and fix my administrative woes. I would like to especially thank Susan Bocti and Lynda Bray for their support in the Biology Department and the MCB & Systems Biology programs, respectively. My thanks to Scott Bunnel, Martin Fleming, Ron Simpson, and Vladimir Timochevski for granting me access to several McGill servers that were of paramount importance in speeding my analysis, and well as their patience in answering my questions and requests.

ABSTRACT

Detection of amyloidogenic peptides or domains in proteins is of paramount importance towards understanding their role in amyloidosis in conformational diseases. This thesis explores different methods towards detection and prediction of amyloidogenic peptides using a variety of bioinformatic analytical methods. Bioinformatic analysis of secondary structural changes is employed to determine whether classes of structurally ambivalent peptides, mainly discordant and chameleon sequences, are efficient predictors of amyloidogenic segments. This analysis elucidates statistical relationships between discordance, chameleonism, and amyloidogenicity across a database of protein domains (SCOP), a subset of amyloid-forming proteins, and the prion family. The presented results stress upon the limitations of these peptides as predictors of amyloidogenicity, and raise issues on the predictive power that can be reaped from secondary structure prediction methods. In another bioinformatic approach, detection of conformationally variable segments in tertiary structures of PrP globular domains has been performed using Principal Component Analysis. This technique succeeded in identifying five conformationally variable domains within PrP, and ranking these subdomains by their ability to differentiate PrPs based on non-local structural response to pathogenic mutation and prion disease susceptibility. The presented results are corroborated by previous observations from experimental methods and molecular dynamic simulations, suggesting that this approach serves as a fast and reliable method for detection of potential amyloidogenic segments in amyloid-forming proteins. Finally, a structural, functional, and evolutionary bioinformatic analysis is conducted to assess the prevalence of the first experimentally verified amyloid fibril fold in nature, and whether this fold can serve as a prototype for other amyloid-forming proteins. The results indicate a limited scope of this fold in amyloid-forming proteins and across the protein universe, and have implications on future identification of amyloid-forming proteins that share this fold. Collectively, the presented thesis compares these different methods and discusses their efficacy in detection of amyloidogenic segments.

ABRÉGÉ

La détection de peptides ou de domaines amyloïdogéniques dans les protéines est d'une importance primordiale dans la compréhension de leur rôle dans l'amylose dans les maladies conformationnelles. Cette thèse explore différentes méthodes en vue de la détection et la prédiction des peptides amyloïdogéniques utilisant une variété de méthodes d'analyse bio-informatique. L'analyse bio-informatique des changements structurels secondaires est employé afin de déterminer si les classes des peptides structurellement ambivalentes, principalement des séquences discordantes et caméléons, sont des prédicteurs efficaces de segments amyloïdogéniques. Cette analyse élucide des relations statistiques entre la discordance, la chameleonism et l'amyloïdogénicité à travers une base de données de domaines protéiques (SCOP), un sous-ensemble de protéines formées d'amyloïdes, et de la famille prion. Les résultats présentés soulignent les limites de ces peptides en tant que prédicteurs d'amyloïdogénicité, et soulèvent des questions sur le pouvoir prédictif qui peut être récolté de méthodes de prédiction de structure secondaire. Dans une autre approche bio-informatique, la détection de segments de conformation variables dans les structures tertiaires de domaines globulaires PrP a été effectuée utilisant « Principal Component Analysis ». Cette technique a réussi à identifier cinq domaines de conformation variables au sein de la protéine PrP, et à classer ces sous-domaines par leur capacité à différencier les PrP fondés sur des réponses structurelles non-locales à la mutation pathogène et la susceptibilité aux maladies prion. Les résultats présentés sont corroborés par des observations antérieures à partir de méthodes expérimentales et de simulations de dynamique moléculaire, ce qui suggère que cette approche sert comme une méthode rapide et fiable pour la détection de segments amyloïdogéniques potentiels dans les protéines formées d'amyloïdes. Finalement, une analyse structurelle, fonctionnelle et évolutive bio-informatique est menée afin d'évaluer la prévalence du premier pli de fibrille amyloïde dans la nature vérifié expérimentalement, et si ce pli peut servir de prototype pour d'autres protéines formées d'amyloïdes. Les résultats indiquent une portée limitée de ce pli dans les protéines formées d'amyloïdes et à travers l'univers des protéines, et ont des répercussions sur l'identification future de protéines formées d'amyloïdes qui partagent ce pli. Collectivement, la thèse présentée compare ces différentes méthodes et discute leur efficacité dans la détection de segments amyloïdogéniques.

TABLE OF CONTENTS

LIST OF FIGURES

SUPPLEMENTARY FIGURES

LIST OF TABLES

SUPPLEMENTARY TABLES

LIST OF ABBREVIATIONS

(By order of appearance)

TSE	Transmissible Spongiform Encephalopathy
CJD	Creutzfeldt-Jakob Disease
BSE	Bovine Spongiform Encephalopathies
PrP	Prion Protein
PrP^{C}	Cellular Prion Protein
PrP^{Sc}	Pathological PrP (also referred to as Scrapie PrP)
H1	PrP Helix 1
H2	PrP Helix 2
H3	PrP Helix 3
S1 or β1	PrP Strand 1
S2 or β2	PrP Strand 2
huPrP or hPrP	Human PrP
mPrP	Mouse PrP
chPrP	Chicken PrP
tPrP	Turtle PrP
GO	Gene Ontology
BP	Biological Process
MF	Molecular Function
CC	Cellular Component
DSSP	Dictionary of Secondary Structure of Proteins
ET	Evolutionary Trace
MSA	Multiple Sequence Alignment
NMR	Nuclear Magnetic Resonance
EM	Electron Microscopy
PDB	Protein Data Bank
PCA	Principal Component Analysis
PC	Principal Component
MD	Molecular Dynamics
GSS	Gerstmann-Straussler-Scheinker
FFI	Fatal Familial Insomnia
WT	Wildtype
PFD	Prion-Forming Domain
HMM	Hidden Markov Model

CONTRIBUTION OF AUTHORS

This thesis is presented in manuscript-based format. Three manuscripts are presented that correspond to Chapters II, III, and IV. All manuscripts are first-author publications, two of which have been already published (Chapters II and IV), and one which has been submitted (Chapter III).

Chapter II:

Gendoo, D. M. and Harrison, P. M. (2011), Discordant and chameleon sequences: Their distribution and implications for amyloidogenicity. Protein Science, 20: 567–579. doi: 10.1002/pro.590

Chapter III:

Gendoo, D. M. and Harrison, P. M. (2011) The landscape of the Prion Protein structural response to mutation revealed by PCA analysis of multiple NMR ensembles. *Submitted*.

Chapter IV:

Gendoo DMA, Harrison PM (2011) Origins and Evolution of the HET-s Prion-Forming Protein: Searching for Other Amyloid-Forming Solenoids. PLoS ONE 6(11): e27342. doi:10.1371/journal.pone.0027342

Under the supervision of Professor Paul Harrison, I collected the data, conducted the experiments, interpreted the results, executed statistical tests, and wrote the manuscripts for the work presented in Chapters II, III, and IV.

Dr. Harrison and I designed all experiments for Chapters II and IV, and I conceived and executed the analysis presented in Chapter III.

Initial drafts of all of the manuscripts have been written by me, and Professor Harrison edited later drafts of the manuscripts.

CLAIMS TO ORIGINALITY

Data analysis and interpretation of the research presented in Chapters II, III, and IV was conducted by me, under the supervision of Dr. Paul Harrison, and results of these analyses have been prepared and submitted in peer-reviewed publications with a strong bioinformatics and computational biology component.

Chapter II:

The presented work is the first study that attempts to derive statistical relationships between amyloidogenicity and two classes of structurally ambivalent peptides in protein secondary structures, mainly discordant and chameleon segments. This analysis is the first of its kind on chameleon segments, and refutes an existing assumption about a strong association between discordance and amyloidogenicity. Through this study I also identify a new class of discordant-chameleonic segments that may be prone to amyloidogenicity.

Chapter III:

The presented work is a comprehensive principal component analysis (PCA) study to assess conformational variation and discern the landscape of the PrP structural response to sequence mutation. This is to my knowledge the first such large-scale analysis of multiple NMR ensembles of protein structures, and the first study of its kind for PrPs. From this analysis, five domains have been identified as conformationally variable subdomains within PrP, and PCA succeeds in raking these subdomains by their ability to differentiate PrPs based on a non-local structural response to pathogenic mutation and prion disease susceptibility. This PC-based domain ranking is a novel approach towards understanding the role these subdomains might play in the PrP conversion process.

Chapter IV:

The presented work uses structural, functional, and evolutionary analysis to assess the prevalence of the HET-s β-solenoid fold in nature, and determine the role this fold plays in amyloidosis. This is the first study that indicates a limited distribution of the

HET-s fold outside of the fungal kingdom, and its limitations for amyloidosis across the wider protein universe. The challenges of other HET-s homologs in adopting the solenoid shape are demonstrated. The implications of this study in the future identification of other amyloid-forming proteins that share the solenoid fold are also discussed.

CHAPTER I

General Introduction

1.1 The Prion Protein

1.1.1 Structure of the Prion Protein

The unique common molecular trait underlying Transmissible Spongiform Encephalopathies (TSEs) is the misfolding of a host-encoded cellular prion protein, PrP^{C}, into a pathogenic scrapie form coined PrP^{Sc} **[1]**. The structural change behind this conversion is a drastic alteration in secondary structure in which the PrP^{Sc} acquires a much higher β-sheet content (43% β-sheet vs. 30% α-helix) than the predominantly helical PrP^{C} (42% α-helix vs. 3% β-sheet), despite having the same amino acid sequence **[1-3]**.

PrP^{C} is a monomeric, GPI-linked glycoprotein attached to the outer leaflet of the plasma membrane of the cell. Following cleavage of the N-terminal signal peptide (22 residues) and cleavage of a C-terminal peptide (23 residues) on addition of the GPI anchor, PrP^{C}, in its mature form, is a 208-209 residue protein that consists of an unstructured N-terminal region (around 100 residues), and a globular C-terminal made of three α-helices and a two-strand antiparallel β-sheet **(Figure 1.1, sections A-B) [2, 4-6]**.

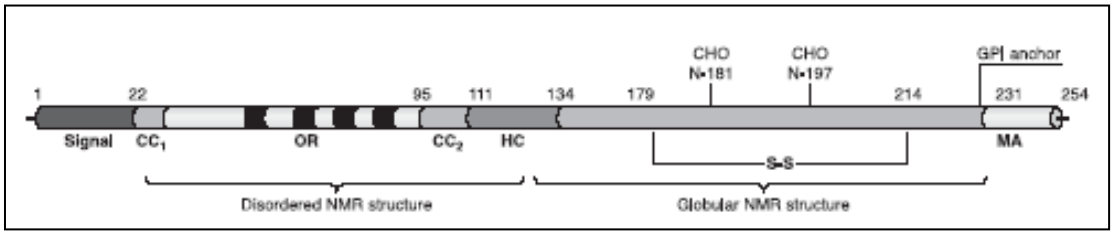
The structurally less-defined N-terminal (residues 23-124, hPrP numbering) contains several distinguishing features, including a variable number of octapeptide PHGGSWGQ repeats (OR) **[4]**. In mammals, five octarepeats are flanked by two positively charged clusters CC1 and CC2 **[2, 4]**. While the exact repeat sequences differ from organism to organism, this region is generally an unstructured, but likely helical, copper binding domain rich in glycine **[7, 8]**. Binding of copper induces α-helix formation of the peptides and is also involved in prion pathogenesis **[2, 8]**. Downstream of the octapeptide repeats is a highly hydrophobic and conserved alanine-rich profile (HC) that may form a transmembrane region in some disease-associated products **[2, 4, 9]**.

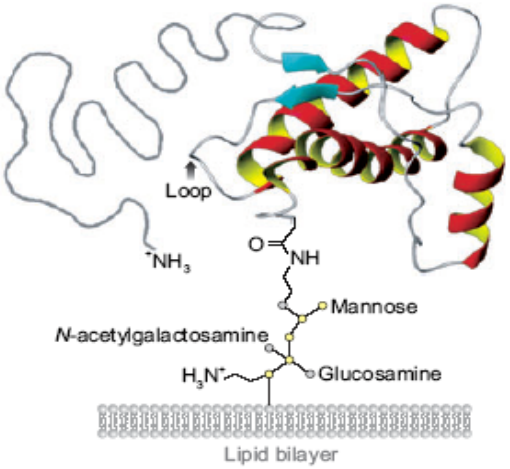
Globular PrP^{C} exhibits a high degree of sequence and structural identity within mammals **[10]**. Interestingly, despite low sequence identity between the mammalian isoforms and chicken, frog, or turtle PrP^{C}, the major structural features of PrP^{C} are preserved in these nonmammalian species **[11]**. The domain is arranged (hPrP numbering) in three helices, H1: 144-154, H2: 173-194, and H3: 200-208, with an

antiparallel β-pleated sheet flanking H1 (residues 128-131 and 161-164) [4]. Helices H2 and H3 are connected by a disulfide bond (Cys179-Cys214) that stabilizes the covalent homodimer [1, 2, 6, 9]. Full length PrP^C is found in non-, mono-, or di-glycosylated forms, depending on occupancy of the two N-linked glycosylation sites at residues Asn181 and Asn197 [4, 6], but the physiological role of PrP^C glycosylation remains unknown [4]. Notably, studies on covalent posttranslational modifications have not shown consistent differences between PrP^C and PrP^{Sc} [5].

Figure 1.1: Structure of the Prion Protein

(A) Linear representation of the N-terminal and globular domains of PrP. Within the N-terminal, Octapeptide repeats (OR) and charge clusters (CC1 and CC2) are highlighted. Glycosylation sites and the disulfide bridge within the globular domain are also indicated. Figure from Aguzzi and Calella, 2009 [4].





(B) Three dimensional structure of the prion protein, showing the unstructured N-terminal and the globular domain.
The tertiary structure of the globular domain is shown with the anti-parallel β-sheets colored in turquoise and helices in red. The GPI-anchor is also demonstrated. Figure from Aguzzi *et al.*, 2008 [2].

1.1.2 Physiological Functions of PrP^C

Developing an understanding of the physiological role(s) of PrP^C has been proposed to help in understanding the pathophysiological properties of prions. A number of functions have been attributed to PrP^C, and some of these functions are highlighted here.

One of the most interesting functions attributed to PrP^C - in stark contrast against PrP^{Sc} pathology - is a neuroprotective, or cytoprotective role for PrP^C. PrP^C decreases the rate of apoptosis after induction of apoptotic stimuli such as Bax or TNF-α; co-expression of PrP^C can reverse Bax-mediated induction of apoptosis in human neuronal cells **[12, 13]**. In addition to an anti-apoptotic function, PrP^C also plays a role in anti-oxidative stress and resistance to copper toxicity. The N-terminal octarepeats of PrP have been proposed to play a role in copper binding **[2]**, and studies on rat pheochromocytoma cells indicate that cells resistant to copper toxicity or oxidative stress showed higher PrP^C levels **[14]**. PrP also plays a role in signal transduction and growth. PrP^C is reported to play a role in cell-signaling pathways **[14, 15]**. For example, the ERK1/2 and MAP kinases are activated by binding of PrP^C binds to the adaptor protein Grb2 (growth factor receptor binding protein) **[15]**.

Despite the plethora of functions that have been attributed to PrP^C, interestingly, the only well-defined phenotype of Prnp knockout mice is their resistance to prion inoculation **[4]**. Postnatal depletion of PrP^C in neurons does not result in neurodegradation **[4]**. The uncertainty of the exact PrP^C role in physiology, and by extension pathophysiology, raises the intriguing and yet unanswered argument of whether TSE pathology is a result of PrP^C loss of function, PrP^{Sc} gain of function, or both.

1.1.3 Models of Prion Replication

Stanley Prusiner coined the term ‘Prion’ to represent a “proteinaceous and infectious particle that lacks nucleic acid” **[1]**. Indeed, increasing acceptance of the “protein-only hypothesis” by which a prion protein replicates and spreads within its host defies the standard dogma that protein production and replication are mediated by nucleic acids **[1, 2]**. The protein-only hypothesis proposes that prion replication involves a self-propagating conversion of PrP^C to its pathogenic isoform PrP^{Sc} **[16]**. Two models for this

process of conformational conversion have been proposed: a "template-assisted" or "refolding" model, and a "nuclear polymerization" or "seeding" model **(Figure 1.2)**.

The template-directed refolding model suggests a mechanism whereby PrP^{Sc} induces a catalytic cascade using PrP^{C} or a partially folded intermediate (PrP^{C*}) as a substrate to produce more PrP^{Sc} molecules. In this model, a PrP^{Sc} monomer binds to PrP^{C} or a partially unfolded intermediate, PrP^{C*}, that arises from fluctuations in PrP^{C} conformation **[2, 4, 16, 17]**. This dimerization lowers the activation-energy barrier for PrP^{C}/PrP^{C*} to convert into PrP^{SC}. As such, the exogenous PrP^{Sc} acts a template for conversion of endogenous PrP^{C}; according to the model, conformational change is kinetically controlled, as a high energy barrier would prevent spontaneous conversion of PrP^{C} to PrP^{Sc} **[2, 4, 16, 17]**.

The seeded model, as opposed to the template-mediated model, is based on the assumption that PrP^{C} to PrP^{Sc} conversion process is thermodynamically controlled. In this model, both PrP^{C} and PrP^{Sc} molecules are in equilibrium **[2, 4, 16, 17]**. In a non-disease state, PrP^{C} is strongly favored, and minute amounts of PrP^{Sc} would coexist with PrP^{C}. The infectious agent, according to this hypothesis, is a highly ordered aggregate of PrP^{Sc} molecules. This aggregate is formed by the recruitment and addition of PrP^{Sc} monomers onto an existing crystal-like "seed" of PrP^{Sc} aggregates **[2, 4, 16, 17]**. Accordingly to the model, monomeric PrP^{Sc} would be harmless, but might be prone to incorporate nascent PrP^{Sc} aggregates to generate oligomeric PrP^{Sc} in the diseased state **[2, 4]** .

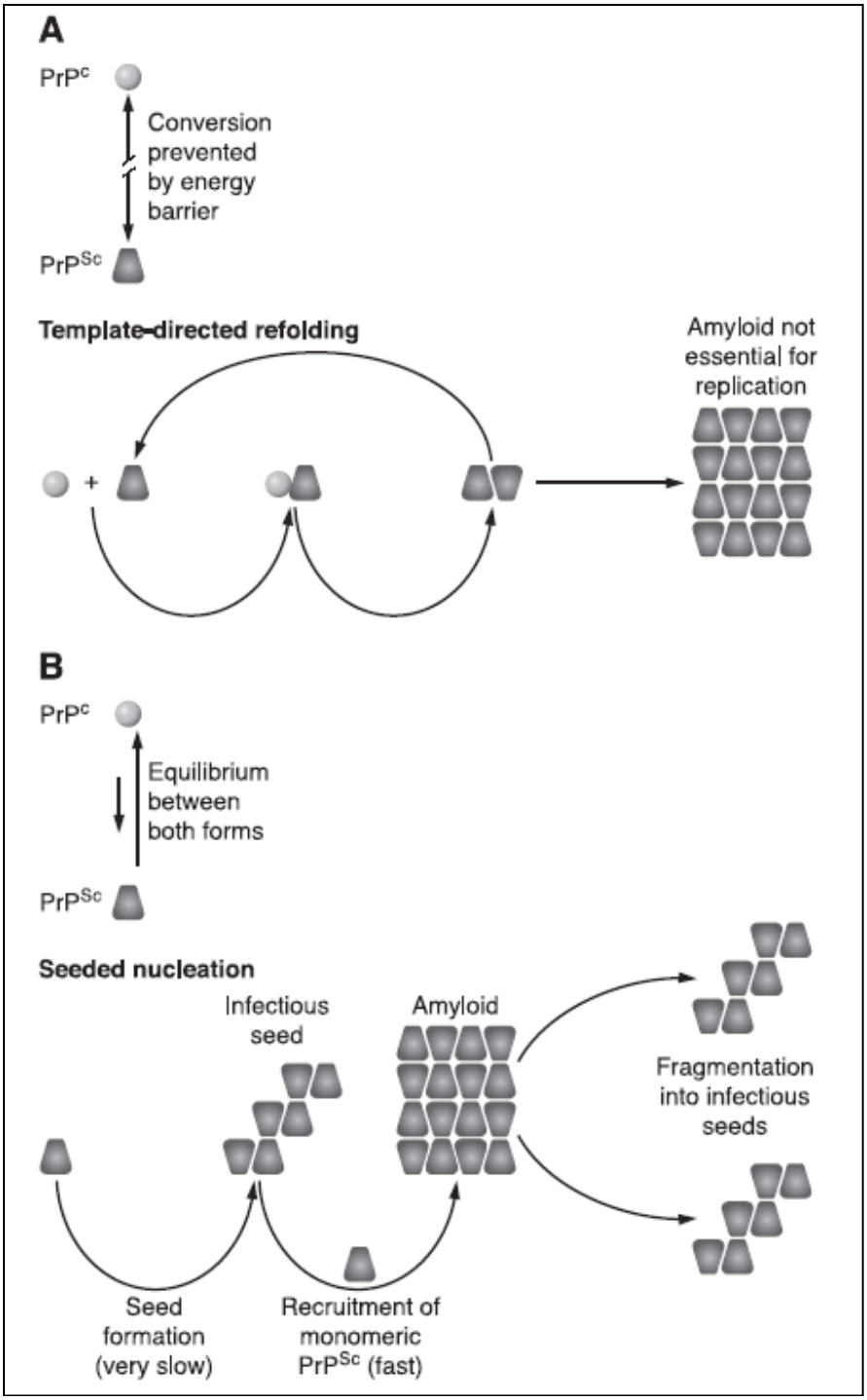


Figure 1.2: Overview of the Template-directed refolding and Seeded nucleation models

(A) Template-refolding model **(B)** Seeded nucleation model

Figure from Aguzzi and Calella, 2009 **[4]**.

1.1.4 Proposed Models of PrP^{Sc}

In contrast to high resolution data for the PrP^{C} monomer, structures of pathogenic PrP^{Sc} as well as fibrillization intermediates (PrP^{Sc}-like aggregates) responsible for infectivity and neurodegenration remain elusive. Several models of the PrP^{Sc} structure have been proposed, many of which are theoretical models based on molecular modeling and dynamics simulations **[18, 19]**, as well as more recent models based on experimental data **[20]**. Some of these models, and their implications for PrP^{Sc} pathogenesis, are highlighted here.

1.1.4.1 Features of the left-handed helix (LβH) fold

The parallel β-helix fold is a repetitive protein fold with a β-helical coil formed by segments of β-strands as its repeating unit **[21-24]**. Each rung of the β-helix is composed of 2-3 β-strands interrupted by turns or loops, and the rungs are aligned such that elongated β-sheets connected by hydrogen bonds lie parallel to the helical axis. Structural repetition of these coils creates a cylindrical hydrophobic core characterized by buried stacks of similar side chains **[21, 22]**. The left-handed beta helix (LβH) fold, as opposed to the right-handed beta-helix (RβH), is more rigid and repetitive, with each β-helical turn made of three β-strands that are connected by three loops of 1-2 residues **(Figure 1.3)** **[21, 23]**. Several models have been proposed for PrP^{Sc} based on the left-handed beta helix (LβH) fold **[18, 19]**, given experimental observations that amyloid fibrils are protease-resistant filaments with dominant β-sheet structures organized in a cross-β spine arrangement **[25]**.

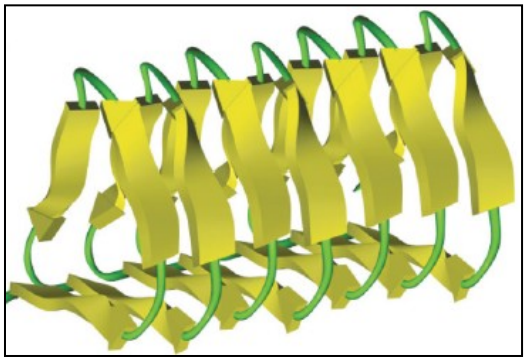


Figure 1.3: Structural representation of the LβH fold. B-sheets are represented as yellow ribbons and turns are in green. Figure from Choi *et. al.*, 2008 **[21]**

1.1.4.2 Theoretical models based on the LBH fold

"Beta-helix" [19] and "spiral" [18] models have been proposed for PrP^{Sc} using molecular modeling and dynamic simulation techniques.

The 'spiral' model proposed by De Marco and Daggett [18] is derived from molecular dynamic simulations, based on the idea that simulations with the required environment for PrP conversion, namely the presence of mutations and low pH levels, should be able to model the conversion process and allow for the analysis of prefibrillar aggregates. Simulations performed on the Syrian hamster PrP (with D147N mutation, hamster numbering) in a low pH trajectory indicate a radical conformational change involving the extension of β-structures within residues 116-164. These residues were argued to be the β-core of PrP^{Sc}, consisting of parallel and antiparallel β-strands, while the remaining helices of the globular protein retained their native conformation [18]. The β-structure adopted by this N-terminal core is composed of a three-stranded β-sheet, E1-E3, as well as an isolated strand, E4 **(Figure 1.4)**. Using this model of extended secondary structure, the authors modeled a protein aggregate that agrees with electron microscopy. In the aggregate (protofibril) form, PrP^{Sc} molecules are docked together such that the N-terminal of E1 is docked to the hydrophobic E4 sheet, forming a continuous four-stranded sheet that is aligned by interstrand backbone hydrogen bonding [18]. Propagation of this bonding forms a spiraling protofibril of PrP **(Figure 1.4)**.

The 'β-helix' model proposed by Govaerts *et. al.* [19] is obtained by threading part of the PrP sequence through a left-handed β-helical protein, based on increasing evidence arguing for a parallel β-sheet organization in amyloids structures. Using this approach, the authors threaded the amyloid core of PrP27-30, residues 89-175, against a left-handed beta helix protein, while the C-terminal H2 and H3 helices retain their α-helix conformation. By comparing potential threading of the amyloid core against the Right-handed beta helix, the authors contend that PrP is more compatible with the LβH fold, from which a trimeric β-helical model has been postulated based on low resolution 3D structures of PrP^{Sc} derived from electron crystallographic data [19]. The monomer of this PrP27-30 model is composed of the threaded β-helix amyloid core, to which the C-terminal of the PrP is connected. In the trimeric assembly, the C-termini are located outside of the trimer, with glycosylated asparagines pointing away from the center, which

is composed of the packed β-helices **(Figure 1.5)**. An oriented fibril can be produced by a head-to-tail arrangement of β-helices, linked by hydrogen bonds between the molecules.

Collectively, both models suggest that the amyloid core of PrP^{Sc} is based on the unstructured N terminus adopting the β-structure, while major helices within the globular structure (H2 and H3) remain intact.

1.1.4.3 Experimentally-derived models

In contrast to largely theoretical models, a recent model based on experimental data suggests that PrP^{C} conversion involves refolding of the C-terminal α-helical region. By analyzing amyloid fibrils of recombinant human PrP90-231 using site-directed spin labelling (SDSL) and EPR spectroscopy, Cobb *et. al.* **[20]** proposed that the amyloid core of PrP maps to the C-terminal residues 160-220 of PrP, which stack on top of one another in a parallel in-register alignment of β-strands **(Figure 1.6)**. In the native PrP structure, these residues encompass helix 2, part of helix 3, and the loop between both helices. To account for the native disulfide bridge between Cys179 and Cys214, limiting the loop regions between them implies the existence of bulges. Accordingly, to satisfy this requirement and ensure that the structure is thermodynamically stable by reducing the number of charged residues in the dry intersheet interface, the authors proposed a model containing a pair of bulges introduced within the disulfide bridged loop. The model succeeds in positioning glycosylation sites of the PrP (N181 and N197) on the outside of the intercysteine loop, making them unrestrained and compatible with a glycosylated PrP amyloid. Accordingly, the proposed experimental model differs from its *in silico* predecessors with respect to location of the β-core, as well as the folding motif of the amyloid core region **[20]**.

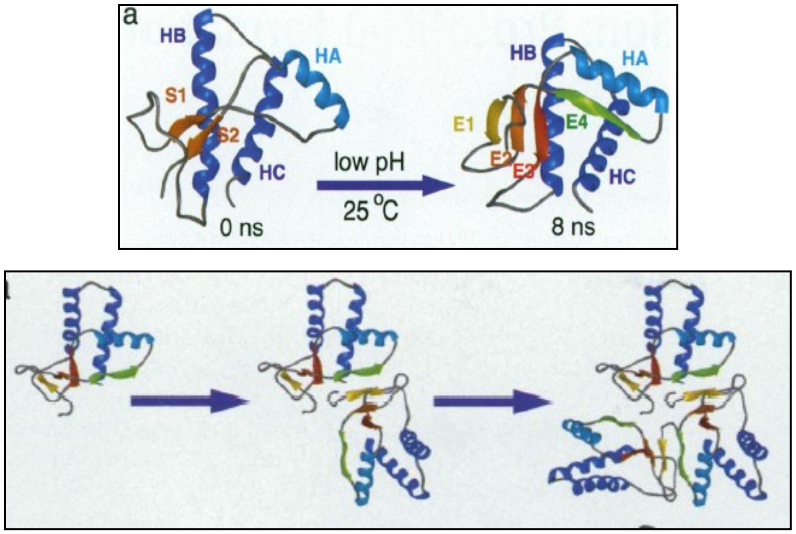


Figure 1.4: Overview of the spiral model of PrP conversion. Top panel: 4-sheet PrP^{SC} model under low pH. Bottom panel: Formation of a spiraling protofibril based on the 4-stranded sheet. Figure from DeMarco and Daggett, 2004 **[18]**.

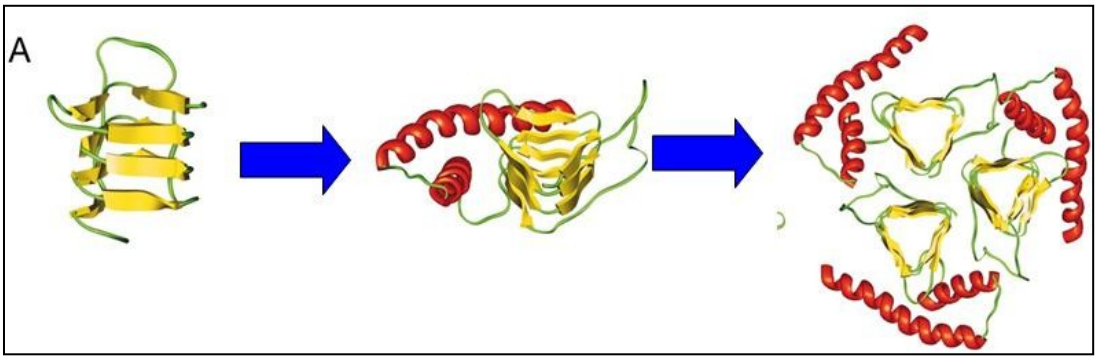


Figure 1.5: Trimeric B-helical model for PrP^{Sc} based on threading against the LβH fold. Adapted from Govaerts *et. al.*, 2004 **[19].**

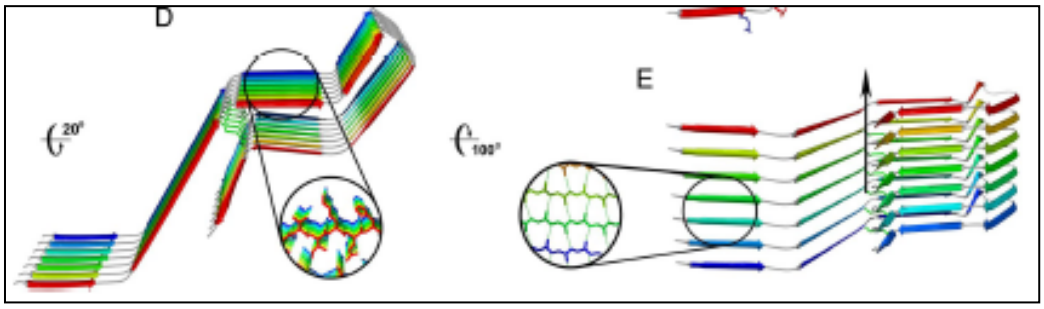


Figure 1.6: Parallel in-register arrangement of PrP^{Sc}. Figure from Cobb *et. al.*, 2007 **[20]**.

1.1.5 Key Proposed Areas of PrP Involved in Conversion and Disease

While several models of PrP^{Sc} fibrils have been suggested using modeling methods **[See Section 1.1.4.2 for a summary of PrP^{Sc} models]**, experimental methods, molecular modeling and dynamic simulations have also been used to determine the effect of pathogenic mutations on PrP conversion, that stability of PrP mutants, and conformational changes in PrP during the conversion process.

One of the key regions of PrP under heavy scrutiny as a candidate site for TSE transmissibility studies is the S2-H2 loop. Comparative studies on the flexibility of this loop indicate a difference between disease-prone and disease-resistant species, with greater degrees of flexibility in disease-resistant species, as evidenced when comparing MD trajectories from species such as elk and hamster to frog, turtle, and chicken **[26, 27]**. Similar observations have been previously described in structural comparisons of human, chicken, turtle, and frog NMR structures, insinuating that this region could serve as a "structural signature" for different evolutionary groups **[11]**. Interestingly, this loop also exhibits varying flexibility within mammalian species **[26-28]**. These structural variations make the S2-H2 loop, in conjunction with structural changes of the H2 and H3 helices, a candidate site for epitope recognition by potential chaperones. Using 50ns molecular dynamic simulations on wildtype human PrP and hPrP mutants, Rosetti *et. al.* **[29]**demonstrated that differentiating factors in the mutant structures include the increased flexibility of the H3 helix, increased solvent exposure of Y169, loss of the salt-bridge network in the H2-H3 region, and increased proximity between the S2-H2 loop and the C-terminal of H3; these findings concur with previous literature that suggested that Y169 along with the S2-H2 loop and the C-terminal of H3 form a disease-linked epitope for a monoclonal antibody **[29]**. **This is further discussed under Section 3.5.2 (Chapter III)**.

The involvement of H1 in the conversion process remains a topic of debate. H1 had been suggested as a primary interaction site with PrP^{Sc} **[1]**, but conserved sequence and structural identity between TSE-prone and susceptible species seem to negate this claim **[10]**. Studies on the unusual hydrophilic sequence pattern of H1 indicates that its charge distribution, and possible formation of salt bridges between charged residue pairs, would

ultimately stabilize the helix and prevent its conformational change during the PrP conversion process [8, 30-32]. While the PrP^{Sc} spiral model of De Marco and Dagget [18] proposes that the H1 helix converts into β-sheet to form a continuous 4 strand β-sheet, experimental analysis on C-terminal truncated forms of PrP have shown that H1 is not converted into β-sheet [32]. Notably however, these results do not exclude the possibility that the H1 helix might be shortened in PrP^{Sc} [32].

1.1.6 Evolution of the Prion Concept

The prion family has expanded to include other PrP-like molecules of neuropathological relevance, including the paralogs Doppel and Shadoo, as well as "functional" yeast prions that are beneficial to their host.

1.1.6.1 PrP Mammalian Paralogs: Doppel & Shadoo

The Doppel (Dpl) protein is a paralog of PrP that shares a similar native fold to PrP, despite its low sequence identity (about 20%) [33]. Doppel contains a globular domain with a PrP-type fold, in addition to a flexibly disordered N-terminal tail 26 residues long [33, 34]. The Doppel globular fold is composed of four helices and short two-stranded antiparallel β-sheets [33, 34]. Like PrP, the functions of Dpl in the healthy organism are unknown, but mice with Dpl and devoid of PrP have been observed to develop signs of ataxia and degeneration of Purkinje neurons, causing a different type of neurological disease [33, 35]. Interestingly however, Doppel does not convert to a scrapie form [33, 35]. Studies on this protein may help elucidate differences in stabilization and unfolding/misfolding rearrangements that are that the basis of neurodegenerative disease [35].

The Shadoo (Sho) protein, another PrP paralog, was mainly discovered in zebrafish but is also found in mammals (human, rat, mouse) and *Fugu* [36]. Unlike PrP and Dpl, the Shadoo protein lacks a defined globular structure, as the C-terminal prediction indicates mostly a coil conformation with weak prediction of secondary structure elements [36]. Interestingly, Shadoo does share structural similarities with the N-terminal of PrP, as it is the only known PrP homolog containing a conserved middle

hydrophobic region, and an unusual composition of aliphatic residues as for PrP and PrP-like proteins **[36, 37]**. Despite a lack of structural similarity to PrP, Sho and PrP demonstrate a functional and pathogenic linkage, as they both can counteract Doppel neurotoxicity in a similar way **[38]**. The ability to for Sho to convert to amyloid-like forms under native conditions has also been demonstrated **[39]**.

1.1.6.2 Yeast Prions

Recent identification of functional prions in yeast challenges the notion of the 'prion' as detrimental, infectious proteins to their hosts, and suggests that prions share a greater biological role than previously thought. Yeast prions are structurally and functionally different from their mammalian namesakes, and are not homologous to vertebrate PrPs at the level of amino acid sequence identity **[16]**. The [URE3], [PSI^+], [PIN^+], and Het-s prions are self-propagating amyloids of the Ure2p, Sup35p, Rnq1p proteins identified in *Saccharomyces cerevisae,* as well as the HET-s protein of the fungus *Podospora anserina*. The discovery of these proteins has helped shed light on the mechanism of prion conversion and propagation, whose underlying molecular basis had not been fully understood within the mammalian system **[40]**. Compared to mammalian prion 'infectivity', yeast prions mimic mammalian non-mendelian inheritance, such that they are able to transmit disease without requiring nucleic acid **[16, 41, 42]**. This behavior explained 'cytoplasmic inheritance' in yeast and other fungi, as each of these proteins forms cytoplasmic amyloid that leads to a particular phenotype, in some cases beneficial, that can be transmitted to offspring of the 'mutated' cell upon division **[16, 43-45]**. Sup35 for example, is a translation termination factor of *S. cerevisae* that ensures cessation of protein production **[43, 46]**. Its prion, [PSI+], results in increased read-through of termination codons, which generates phenotypic diversity by creating an altered proteome **[43, 46]**. Cells containing [URE3] and [PSI+], the prion determinants of Ure2p and Sup35, respectively, can be distinguished from prion-free cells by phenotypic differences at the cellular level, as well as biochemical differences in relation to solubility of the underlying prion protein **[44]**. [PSI+] variants, for example, can be differentiated by simple screens involving colony color and protein analysis that reflect conformational differences and the extent of protein aggregation **[40, 44]**.

1.2 The Nature of Amyloid

1.2.1 Unifying Events and Patterns in Conformational Diseases

A growing number of diseases, including Systemic Amyloidosis, Alzheimer's disease (AZ), Parkinson's disease, and Prion diseases are characterized as protein misfolding and protein aggregation diseases. Despite their varying pathologies, there is increasing substantial evidence of common structural and pathogenic features that underlie their protein aggregation and amyloidosis, including the accumulation of aberrant or misfolded proteins, protofibril or amyloid formation, and altered states of neurotransmission and excitotoxicity **[47]**.

A large group of protein misfolding diseases are associated with conversion of specific peptides or proteins from their soluble functional states into highly organized fibrillar aggregates that accumulate into amyloid fibrils or plaques **[48]**. These amyloid fibrils exhibit common characteristics under the microscope, mainly their ability to bind to the dyes Congo red and thioflavin T and exhibit characteristic green-yellow birefringence of Congo red **[49]**. Amyloid fibrils exhibit an elongated, unbranching morphology suggestive of an ordered arrangement of subunits **[49]**. This is a multilevel arrangement, such that multiple protofibrils (or protofilaments) are twisted together to form a rope-like fibril or long ribbons of amyloid structure **[48, 50]**. The assembly of multiple protofibrils (typically 2-6 fibrils, each 2-5 nm in diameter) creates an unbranched protein fibril with a diameter ranging from 3-10 nm **[48, 50]**. X-ray fiber diffraction data have shown that the underlying conformation behind these β-sheet rich protofilaments is a cross-β arrangement **(Explained in more detail in Section 1.2.3.1) [48, 50-52]**.

1.2.2 Functional Amyloids

The observation that amyloid fibrils can be exploited by living systems in a functional, non-pathogenic manner challenges the notorious association amyloid fibrils share with disease. 'Functional amyloids' that are unrelated to protein aggregation diseases have been observed to natively form fibrillar aggregates that share many of the biophysical qualities of amyloids, including morphological, structural, and tinctorial

properties **[43, 48, 53-55]**. Structural proteins, such as myoglobin, form *in vitro* β-rich myoglobin fibrils that share a cross-β amyloid fold and similar reactions in tinctorial assays involving thioflavin and Congo red, making them indistinguishable from pathological amyloid fibrils, despite that the native state of the protein is predominantly α-helical and lacks β-sheet structure **[56, 57]**. Another critical protein, human Pmel17, is involved in biosynthesis of the pigment melanin. Pmel17 is responsible for formation of fibril structures found in melanosomes, highly specialized secretory lysosome organelles found in melanocyte cells of the skins and eyes **[58]**. Melanosomes with these fibers enhance the rate of melanin formation by accelerating polymerization of melanogenic precursors and functioning as a receptor that templates their polymerization **[43]**. Within the past decade, functional amyloids have also been identified in lower, non-mammalian organisms, including bacteria **[59-63]**, insects **[64, 65]**, and fungi **[66-70]**. Bacteria such as *E.coli* and *Salmonella spp.* utilize extracellular proteinaceous fibrils formed by the protein curlin for cell adhesion to host proteins, as well as colony formation **[59-61, 63]**. The filamentous, soil-dwelling bacteria *Streptomyces coilecolor* secretes Chaplins, a family of 8 proteins whose β-sheet rich fibrils support aerial hyphae and allow for spore formation **[62]**. Similar structures and functions to curli are also observed in fungal Hydrophobins, amphipathic proteins that assemble into fibers at hydrophobic-hydrophilic interfaces, enabling surface attachment in pathogenesis, as well as formation of aerial structures such as spores and fruiting bodies **[66-70]**. In the silk moth (*Antheraea polyphemus*) as well as other egg-laying creatures in insects and fish, eggshells are composed of arrays of fibers of Chorion proteins, that serve as a protective barrier against proteases, microorganisms, physical stress, or other hazards **[64, 65]**. All of the above-mentioned examples and numerous others in the literature highlight the wide range of organisms that exploit the physical properties of amyloid.

1.2.3 Experimentally verified structures of Amyloid fibrils

1.2.3.1 The Cross-β spine motif

Despite a lack of sequence and structural similarity among fibril-forming proteins, X-ray fiber diffraction studies have particularly shown that amyloid-like fibrils of different proteins have a common "cross-β spine" pattern **(Figure 1.7) [48, 49, 52, 71]**. Electron microscope examinations by Cohen and Calkins (1959) of amyloid deposits in diseased tissue revealed elongated, unbranched fibrils **[72]** whose nature was further demonstrated by Eanes and Glenner (1968) in X-ray fiber diffraction experiments **[71]**. Aligned amyloid fibrils give a cross-β diffraction pattern, with a meriodional reflection at ~4.7°A (along the fiber axis) and an equatorial reflection around ~8-11°A **[49, 73]**. The 4.7°A reflection corresponds to the 4.7°A packing of β-sheet strands that run perpendicular to the fibril axis, while the ~8-11°A reflection corresponds to spacing of β-sheets that are parallel to the fibril axis **[49, 73]**. The notion of a common molecular core structure was supported by the finding that amyloid fibrils from 6 different proteins, each with its own clinical syndrome, show common reflections similar reflections in addition to those at 4.7°A and 10°A **[74]**.

Although these data gave insight into the arrangement of the amyloid fibril core, the exact molecular structure of this core was determined in a key study involving X-ray diffraction on a 7-residue peptide from the protein Sup35, a yeast prion **[52]**. The peptide GNNQQNY is a fibril-forming segment at the amino terminus of the prion-determining domain of Sup35. In the initial X-ray structure determined from microcrystals of GNNQQNY, GNNQQNY β-strands form in-register β-sheets that are parallel to the long axis of the microcrystal, while the individual sheets are perpendicular to the long axis of the microcrystal **(Figure 1.8) [52, 73]**. For each pair of sheets, strands in one sheet are antiparallel to those in the mating sheet, with a shift between the strands of 4.87°A such that side chains from a strand in a sheet nestle between side chains of two strands from the mating sheet **[52]**. The β-sheets are packed in the crystal with two distinct interfaces between them, coined the "dry" and "wet" interfaces **(Figure 1.8)**. The wet interface is composed of water molecules that separate GNNQQNY molecules (aside from a contact between Tyr7 residues in neighboring sheets) about 15°A, suggesting that the stable

structural unit of the microcrystals is composed of a pair-of-sheets organization in cross-β motif **[52, 73]**.

The analysis of the dry interface in this newly introduced pair-of-sheets organization also introduced the concept of the "steric zipper" motif **[52]**. The dry interface between B-sheets is devoid of water (aside two molecules at the end of the peptide segments), and is primarily composed of interdigitation of complementary side chains by van der Waals interactions between Asn and Gln ladders (Asn2, Gln4, Asn6) **[52, 73]**. These interdigitating side chains form a "steric zipper". Subsequent analysis has revealed a prevalence of the cross-β spine with steric zipper side chain interactions in the oligomerization of a variety of fibril-forming proteins, including Alzheimer's amyloid-β and tau proteins, Insulin, and PrP, suggesting that steric zippers are a general principal of protein complementation in amyloid structures **[75]**. Variations of the basic steric-zipper structure indicate that there are theoretically 8 possible classes of steric zippers, based on distinguishing characteristics including i) whether β-strands in sheets are parallel or anti-parallel, ii) whether sheets pack with "face-to-face" or "face-to-back" surfaces adjacent to one another, and iii) whether sheets are oriented parallel or antiparallel with respect to one another **(Figure 1.9) [75]**. Of these 8 classes, five classes were identified in the 13 atomic-resolution structures for peptide segments of fibril forming proteins **(Figure 1.9) [75]**. Collectively, the studies on cross-β spine and steric zipper motifs suggest that these motifs are fundamental units of amyloid-like fibrils, with the possibility of more complicated geometries in full-length amyloid fibrils.

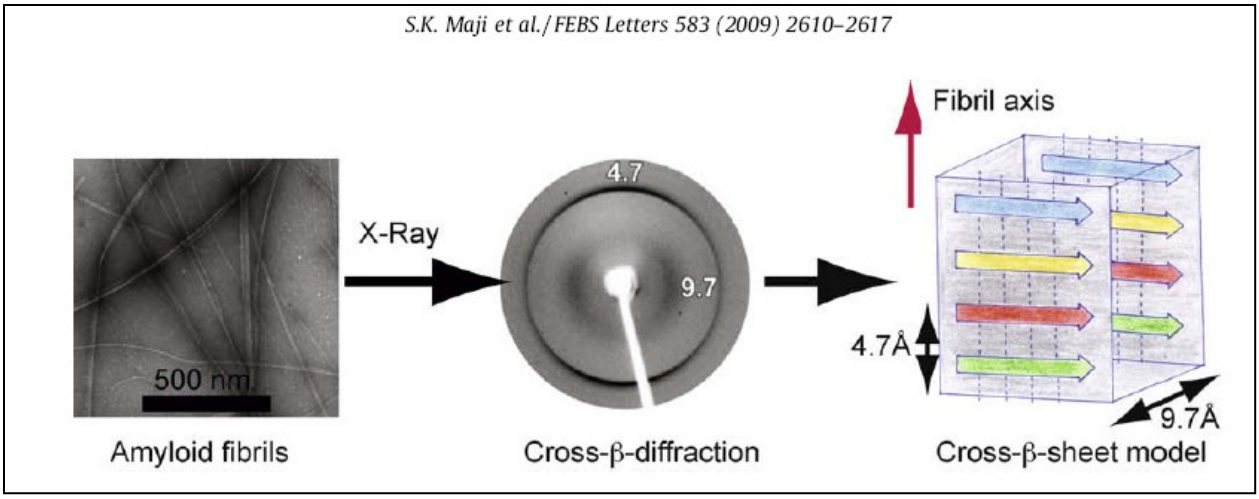


Figure 1.7: Schematic representation of the pairs of sheets in the cross-β spine motif, as determined from analysis of amyloid fibrils and cross-β-diffraction studies.

Figure from Maji et. al., 2009 **[73]**.

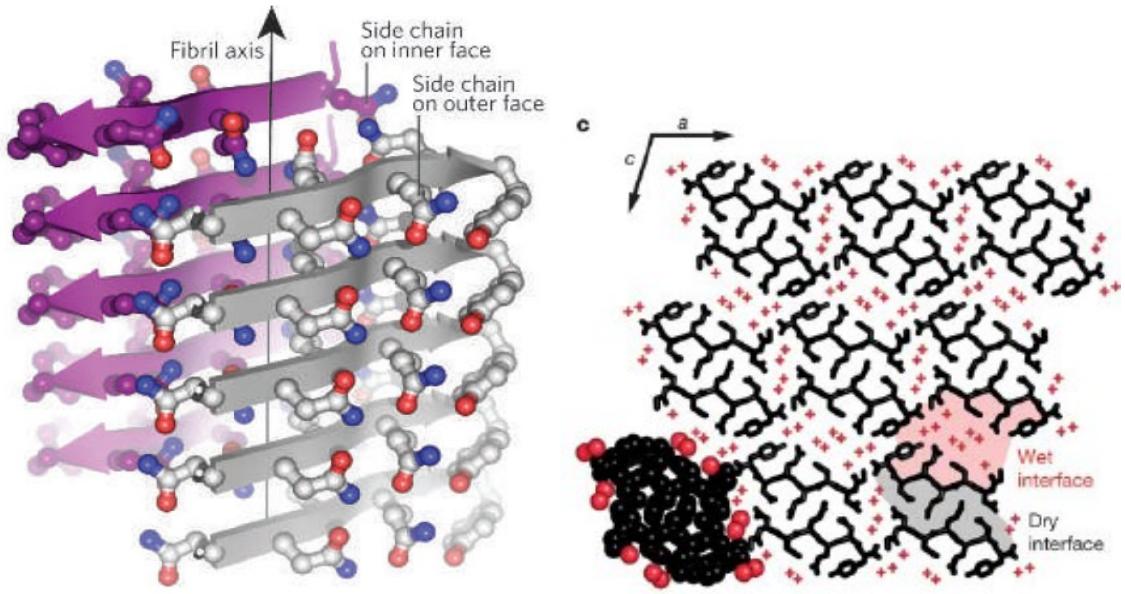


Figure 1.8: Atomic structure of the cross-β spine from Sup35.

Left panel: Opposing sheets of the fibril, showing parallel β-sheets with β-strands perpendicular to the fibril axis.

Right panel: Overview of the wet and dry interface within and between pairs of B-sheets.

Figures from Nelson *et. al.*, 2005 **[52]** and Maji *et. al.*,2009 **[73]** .

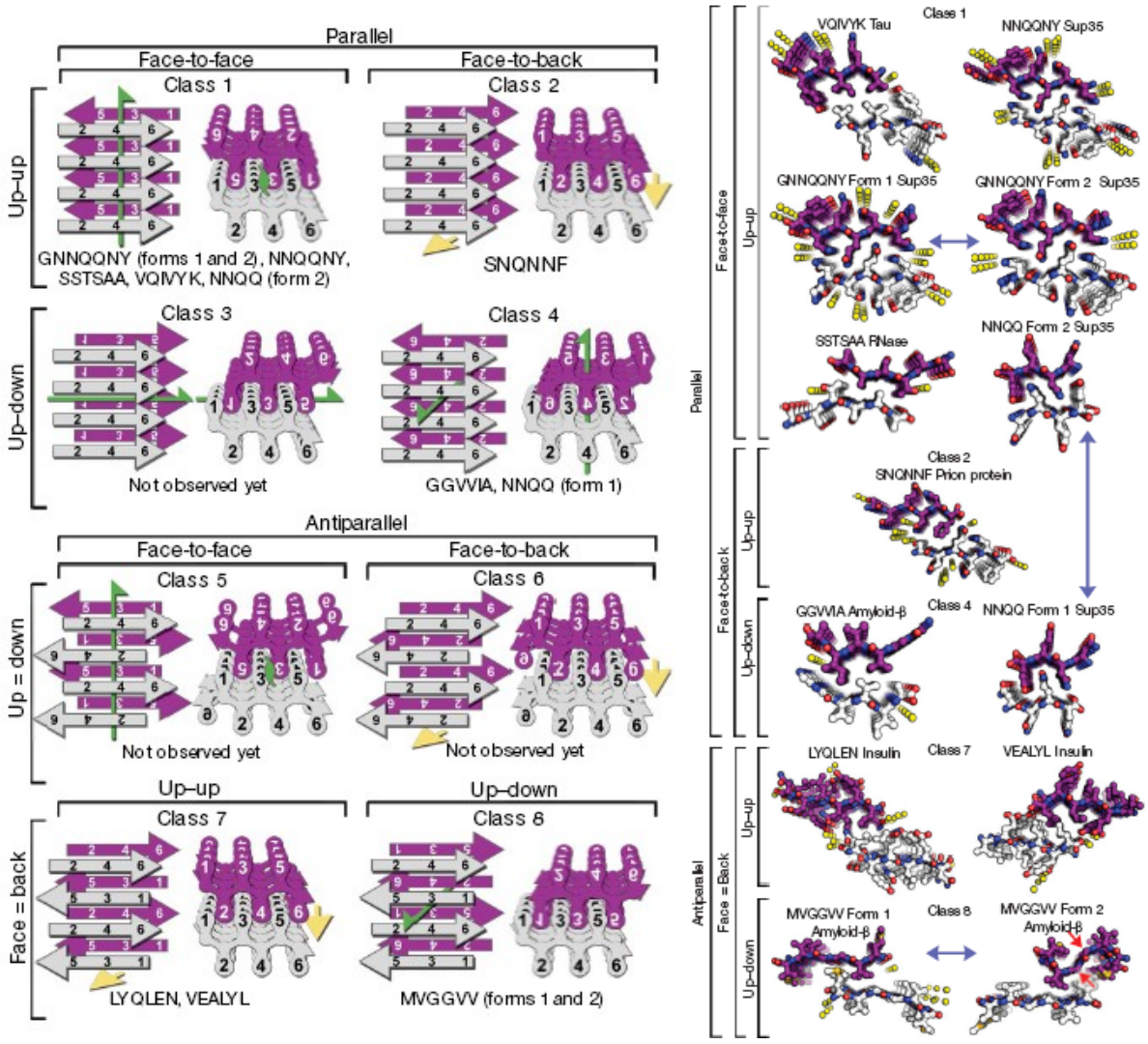


Figure 1.9: Variations of the steric-zipper structure and existing examples in nature. Left panel: Schematic representation of the 8 possible classes of steric zippers. Right panel: Five classes of steric zippers identified in 13 structures of peptides segments of fibril forming proteins. Figures from Sawaya *et. al.*, 2007 **[75]**.

1.2.3.2 The β-Solenoid Fold

The 3D structure of the HET-s (218-289) amyloid fibrils is the first well defined structure of an amyloid fibril, a left-handed β-solenoid with each 72-residue peptide forming two helical windings **[76, 77]**. The four β-strands compromising residues 226-234 (β1), 237-245 (β2), 262-270 (β3), and 273-282 (β4) are parallel and in-register, with

pseudorepeats between β1- β3 and β2- β4. The structure is stabilized by a dense hydrophobic core, intra- and inter-molecular hydrogen bonds between the pseudorepeats, three salt bridges (K229-E265, E234-K270, R236-E272), and two asparagines ladders (N226 and N262) **(Figure 1.10) [76, 77]**. Because the charged residues are stacked parallel to one another such that charge is compensated, the formation of the salt bridge and charge compensation could be both intra-molecular (within the same molecule) or inter-molecular (between the different molecules) **[76, 77]**. Collectively, these structural characteristics result in a zipper-like structure with an overall β-helix characteristic fold **[73, 77] [See section 1.1.4.1 for description of LβH fold]**, and give insights into a higher structural complexity for amyloid fibrils than for short peptide fibrils.

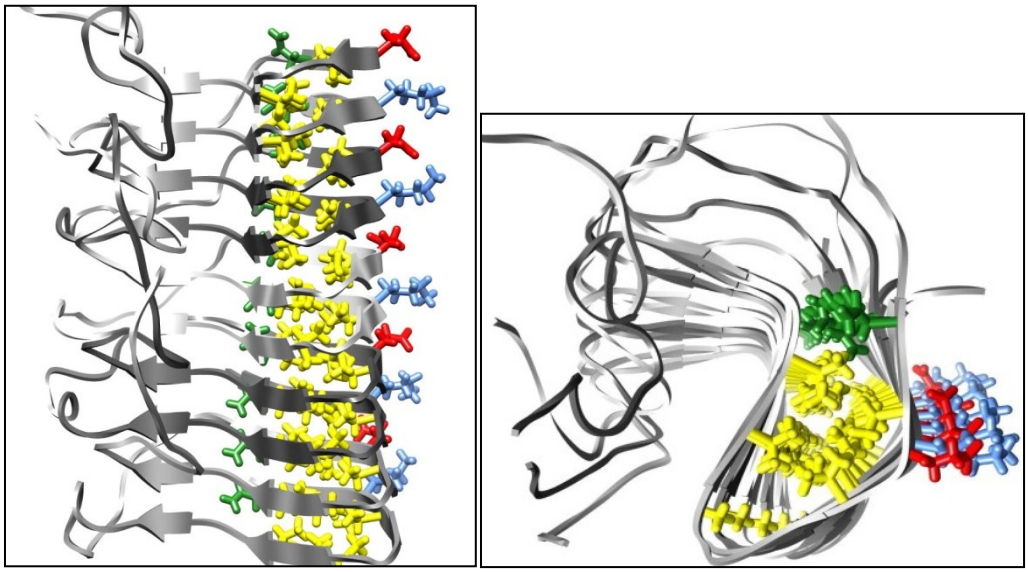


Figure 1.10: Schematic representation of the β-solenoid fold.

Top and side views are presented. The secondary structure is represented in white ribbons, salt bridges are in red and blue. The hydrophobic core (yellow) and asparagines ladder (green) are also indicated. Side and top views are shown. Figures from Maji *et. al.*, 2009 **[73]**.

1.2.4 Models of Amyloid fibrils

A number of amyloid fibril models have been developed for a variety of prions and amyloid-forming proteins, some of which are based on the cross-β spine motif and LβH fold, and others which are devoid of these structural elements. A "parallel superpleated sheet model" has been proposed for the N-terminal Asn-rich "prion" domain of Ure2p

yeast prion **(Figure 1.11) [42, 78]**. In this model, the prion domain is divided into 9, seven-residue segments whose β-strands are parallel to one another. Each of the 7-residue segments is composed of a four-residue strand and three-residue turn, that zig-zag in a planar serpentine arrangement. The sheets are packed parallel to each other in an in-register orientation that is maintained by favorable interactions between aligned amino side chains **[49, 78]**. The parallel in-register arrangement indicates that each residue of the prion domain is in contact with the same residue of adjacent molecules in the filament **[46]**. The 'polar zipper' structure is another such model whereby a parallel in-register arrangement which allows glutamine side chains hydrogen bonds with one other in amyloids fibrils **[79, 80]**.

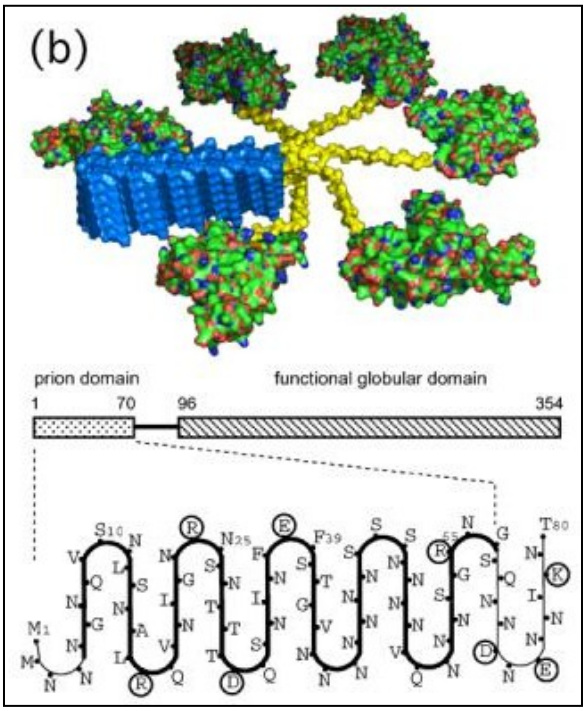


Figure 1.11: Space filling and linear representation of the Ure2p serpentine model. Figure from Baxa, 2008 **[50]**.

A variety of amyloid models have been developed for amyloid-forming proteins, based on the left-handed beta helix fold (LβH) **[See section 1.1.4.1 for description of the LβH fold]**. Indeed the LβH fold forms the basis for several models of PrP^{Sc} **[See section 1.1.4.2]**, as well as other amyloid-forming proteins such as insulin and Aβ **[See section 1.3.2 for examples]**.

1.3 Computational Techniques towards Identification and Prediction of Amyloidogenic Segments

1.3.1 Predicting β-structures and aggregation based on physicochemical properties of proteins

Given strong evidence for β-structure in amyloids, a variety of algorithms have been designed to predict β-structures, and to calculate aggregation propensities of proteins and protein segments. This section addresses tools that make such predictions based on an analysis of physicochemical properties and motif recognition of β-sheet proteins. Such properties, for example, include a general and repetitive packing pattern of buried core residues dominated by hydrophobic side chains, as well an amphipathic mosaic surface of polar and hydrophobic side chains **[21, 81]**.

Several computational methods to predict globular β-structures have been developed based on long-range pairwise interactions and the prediction of potential strand-pairs in a protein sequence **[82-84]**. Statistically, this seems to be a reasonable approach, supported by evidence that residues involved in β-sheet formation which are in close spatial proximity exhibit strong statistical biases **[85]**. However, to exploit this information requires the use of structures with topological regularity, such as β-sheets and β-trefoil structures. The BETAWRAP program was first developed to predict strand-pairs based on structure-specific knowledge derived from templates of the right-handed beta-helix (RβH) SCOP class **[82, 84]**. Statistical correlations between pairs of residues in adjacent β-strands were calculated to determine if a query protein aligns well to the structural template of the RβH motifs **[84]**. However, this limited the program to the role of a fold recognizer for only one-sub family of beta-helices, while falling short of producing a single global alignment of the putative structure of query proteins to those templates **[84, 86]**. Subsequently, an extension of this algorithm, BETAWRAPPRO, has been developed to perform fold recognition for the β-helix and β-trefoil motifs, coupled with evolutionary information from sequence profiles that lend information about residue conservation and substitutions **[84]**. Yet another algorithm, BETAPRO, relies on neural networks to predict pairwise probabilities based on profiles, secondary structure, and solvent accessibility information, and uses these probabilities in subsequent dynamic

programming and graph matching techniques to predict the β-sheet architecture of the protein [83]. A recently developed program, AmyloidMutants, predicts amyloid structures and amyloid fibril conformations based on mutational landscapes extracted from the cross-β sheet and β-solenoid folds [87]. Generally, tailoring these algorithms to β-structural folds remains an improvement over the shortcomings of 'generic' secondary structure prediction algorithms, such as PSIPRED and PHD, which are hindered by the low sequence identity and lack of sequence commonalities in amyloid-folding proteins [88].

A variety of other approaches have been developed for reliable detection and prediction of aggregation propensities in proteins, i.e., identifying aggregation-prone regions in proteins [89]. Seminal work by Chiti *et. al.* [89] indicated that physicochemical properties can be used to determine changes in aggregation rates that arise from amino acid mutations. Subsequent algorithms that have been developed to harness this predictive power include empirical tools that try to predict such regions based on amino-acid properties, or structure-based tools that rely on 3D structures of known fibril-forming peptides to predict determinants of amyloid aggregation [90]. Empirical algorithms rely on amino acid properties observed from experimental results, such as hydrophobicity, β-propensity, and solubility [90]. The TANGO algorithm, for example, uses a statistical mechanics approach to identify β-aggregating regions of a protein based on the assumption that all residues of an aggregate are hydrophobically buried and will satisfy their hydrogen-bonding potential [91]. Yet another algorithm, ZYGGREGATOR, models aggregation propensity per residue based on secondary structure propensity, hydrophobicity and charge, and the pattern of alternating hydrophobic and hydrophilic residues over a sliding window [92]. Structure-based algorithms, unlike empirical tools, are based on the study and observation of the spatial structures of peptides as well as native proteins belonging to structural classes of interest [90]. The PASTA algorithm, for example, calculates a singleton and pairwise energy functions for individual residues and residue pairs in a β-sheet, according to a Boltzmann energy function calculated from a database of 500 annotated structures [93, 94]. Similarly, AMYLOIDFOLD uses a scaling approach based on calculations of the observed packing density and statistics of hydrogen bonds in a database of 3769 proteins from the four main SCOP classes (all-α, all-β, α/β,

α+β) [95, 96]. The BETASCAN program determines strands and strand pairs based on likelihood scores that have been calculated from correlations observed in parallel β-sheet structures [86]. To date, approximately 14 computational tools have been published that include equal numbers of empirical and structure-based models [90], and this list is continually growing.

1.3.2 Structural modeling of protein segments and protein fibrils

Sequence analysis algorithms were among the first methods developed towards analysis of determinants of amyloid structure. One of the barriers to effective sequence-based computational analysis, however, is the relatively small amount and diversity of experimentally validated amyloidogenic sequences, making it difficult to gather enough information that can effectively distinguish amyloids from amorphous aggregates [86], [97]. Amyloid- and prion-forming domains are identified as unstructured random coils in secondary structure prediction algorithms, or, in the case of amyloid-forming domains, are excluded by low-complexity filters of sequence alignment tools such as BLAST [86]. Structural methods, including threading and molecular dynamic simulations, have capitalized on the structural characteristics of amyloid fibrils while overcoming the barrier of poor sequence homology. Some of the studies that utilize these approaches are highlighted here.

One of the first pioneering studies in this field involved threading six-residue peptides of proteins through the microcrystal structure of the Sup35 NNQQNY peptide, to identify new fibril-forming segments [97]. Each hexapeptide of the query protein is mapped onto an ensemble of templates that have been generated by translation of one or two β-sheets relative to other along three orthogonal directions. Threading of hexapeptides against this '3D profile' and evaluation of the energetic fit between the query and templates allowed for the identification of fibril-forming segments that could adopt the cross-β spine fold [97]. Inclusion of parameters such as apolar interactions, hydrogen bonds, and steric overlaps in the energetic calculation improves selection of fibril-forming segments, as opposed to selection of segments based on simple residue properties such as hydrophobicity or β-strand propensity. In that study, Thompson *et. al.* demonstrated that

the 3D profile template method can localize predictions to experimentally determined fibril-forming segments [97]. Variations to this method have been developed [98] with comparable results.

Molecular modeling and simulations can be used to sample the conformational space of a given system and assess structural changes that arise at the atomic level over a given period of time [99]. Several amyloid-forming proteins have been modeled against β-helix folds in an attempt to determine the potential structure and stability of their ensuing amyloid fibrils. Choi *et. al.* attempted to model the monomeric subunits of the insulin amyloid fibrils against the β-helix and β-roll folds, and ascertained which of those folds exhibits greater stability using molecular dynamics simulations [100]. Analysis of the physicochemical properties of the rung structures for both models indicated that both models satisfy the sequence and structural features of the β-roll and β-helix folds, but molecular simulations suggested that the β-roll subunit model exhibits greater stability as possible subunits of fibrillar insulin. Construction of polymeric fibrils based on the β-roll and β-helix subunits also suggested that fibers composed of 6 twisted β-roll protofilaments are the most reasonable fit supported by previous experimental data [100]. Previous and ongoing studies of similar nature were conducted on different sizes of Aβ peptides [101, 102]. A structural model for amyloid fibrils of the Aβ protein was achieved by successful modeling of the Aβ(15-36) peptide against parallel β-helical proteins, thus supporting experimental evidence on full-length Aβ fibrils that suggested an in-register, parallel arrangement for the fibril core [102]. Collectively, these studies and many others suggest that LβH-like structures serve as models of misfolded human proteins associated with disease, an intriguing concept since there are no known human or mammalian proteins that have been documented to incorporate these folds to date [21].

In addition to focused studies on the globular PrP structure, several studies have focused on amyloid-forming peptides, such the Sup35 peptide for example, to obtain deeper understanding about the amyloid core and the oligomerization process [103-105]. Molecular dynamic simulations on the Sup35 hexapeptide GNNQQNY indicated that the hexapeptide β-strands stack in a parallel in-register arrangement during aggregation, and that this arrangement is favored over anti-parallel stacking because side-chain hydrogen bonds and aromatic stacking stabilize the aggregates [103]. Studies on the aggregation

and polymerization of different oligomer sizes (3-mer, 12-mers, 20-mers) for this hexapeptide also indicated that 20-mer oligomers adopt elongated structure reminiscent of the zipper-spine of fibril microcrystals, but would need to be stabilized by establishing contacts with multiple copies of the same structure to evolve into a full fibril **[105]**. As such, the use of molecular dynamics in these studies has provided valuable information about structural changes that arise at the atomic level upon amyloid fibril formation.

1.3.3 Benefits of predicting amyloidogenic segments in proteins

The development of computational tools and algorithms to predict amyloidogenic segments in proteins promises significant predictive power that can be reaped in a variety of applications. Some of these multi-faceted benefits are discussed in the following sections.

1.3.3.1 Metascale Analysis of Aggregation Propensity in Proteomes

In addition to protein-specific identification of potential fibril forming segments, and analysis of the effect of mutations on aggregation propensity in these proteins, the availability of computational tools has allowed for rapid and systemic analysis of full proteomes **[106, 107]**. This, in turn, sheds light on the distribution of aggregation-prone proteins in different proteomes, as well as functional and structural characteristics that may be associated with these proteins. Using an analytical model to predict β-aggregation rates and aggregation-prone segments based on physicochemical properties, Tartaglia and Caflisch **[107]** analyzed the *Saccharaomyces cerevisae* proteome and demonstrated links between amyloidogenic propensity and certain biological functions, as well as preferred localization of β-aggregation prone proteins. The authors observed that β-aggregation prone proteins are "accrued in molecular transport, protein biosynthesis, and cell wall organization processes, while they are underrepresented in ribosome biogenesis, RNA metabolism, and vitamin metabolism" **[107]**. Yeast transporters, for example, demonstrated the highest level of aggregation potential, and as several of these proteins demonstrate significant sequence matches with known amyloidogenic proteins to the human proteome, this suggests that transport proteins have the highest β-potential in a

proteome [107]. Another study on the aggregation propensity in the human proteome revealed a discrepancy between aggregation propensities of proteins taking the secretory pathway, versus proteins operating in intracellular compartments [106]. Interestingly, that study also demonstrated that while different subpopulations of the proteome have different aggregation propensities, the aggregation propensity of proteins involved in protein deposition diseases does not differ extensively from the human proteome as a whole [106]. This finding supports that idea of amyloid fibril formation as a generic property of proteins.

1.3.3.2 Design of Beta Breakers & Inhibiting β-helix aggregation

One of the approaches to prospective treatment of prion disease, as well as other amyloid-forming proteins, is preventing amyloid formation using aggregation inhibitors. In the case of PrP, one possible approach is to disrupt the interaction between PrP^{C} and PrP^{SC}; this has been attempted through the design of beta-breakers that specifically interact with prion protein conversion and slow disease progression [108, 109]. Beta-sheet breaker peptides consist of sequences of the target protein with extra proline residues inserted, which inhibit its formation into the desired β-sheet [109]. The 13-residue β-sheet breaker peptide (iPrP13) designed by Soto *et al* [110], corresponding to residues 115-122 of PrP, was demonstrated to reduce PrP^{SC} in an *in vitro* experiment with scrapie infected brain homogenate. Similar compounds have been developed to inhibit Aβ fibrillogenesis in Alzheimer's disease [111], and α-synuclein fibrillogenesis in Parkinson's disease [112]. An example of such an extensively studied peptide is the five-residue peptide, iAB5, designed against hydrophobic region of the N-terminal domain of the AB protein. This peptide succeeded in inhibiting amyloid formation of the Aβ 1-40 and Aβ 1-42 peptides, as well as inhibiting Aβ neurotoxicity [111].

1.3.3.3 Design of Therapies against Amyloid-forming proteins

Identification of aggregation-prone regions in amyloid-forming proteins facilitates the development of a variety of therapeutic methods, including the development of monoclonal antibodies (mABs) or chemical drugs [113]. For the sake of simplicity, the

examples listed here focus on PrP, although these techniques are equally applicable to other amyloid forming proteins.

Given knowledge of key regions in a protein that adopt conformational changes, structure-based drug design and molecular docking can be used to design drugs and identify drugs against these potential target sites. Perrier *et al.* **[114]** designed drugs to mimic a four-residue epitope on PrP^{C} that represents the Protein X binding site. These residues were demonstrated to confer and mimic dominant negative inhibition of the prion protein. A notable strategy by Kuwata *et al* **[115]** uses a dynamics-based drug discovery approach to identify hot spots of pathogenic conversion in PrP^{C}, based on the observation of residues that undergo conformational changes in the high-energy PrP* states. Identification of these 'hot spots' allowed for the identification and development of novel anti-prion drugs **[115]**. There are numerous studies of similar nature that rely on drug discovery, with different variations on how the target sites are determined, what molecular docking tools are used, or how the experimental assay is undergone.

1.4 Objectives of the Research

The analysis of secondary and tertiary structure conformational changes in proteins to detect potential amyloidogenic segments can be achieved via bioinformatic sequence and structural analysis of prions, amyloid-forming proteins, and the protein universe. Accordingly, the presented thesis is a bioinformatics-based thesis that specifically aims to:

1. Analyze conformational changes in secondary and tertiary structures of native prions and other amyloid-forming proteins, and discuss the ramifications of conformationally-variable segments on the prediction of amyloidogenic segments.

2. Analyze the distribution of the first atomic fibril structures in proteins, and determine the extent to which prions and amyloid-forming proteins can adopt that structure.

Chapters II and III of this thesis address the first aim, while Chapter IV addresses the second aim.

Chapter II is a bioinformatic sequence analysis study that aims to analyze the distribution of conformationally variable segments observed in secondary structures, mainly discordant and chameleon sequences, and test their efficacy as predictors of amyloidogenic segments. To this end, a meta-scale statistical analysis of the distribution of these segments has been conducted in a database of protein domains, a subset of amyloid-forming proteins, and the prion family.

Chapter III is a bioinformatic structural analysis that attempts to identify conformationally-variable segments in prions via an analysis of the tertiary structures of the globular domains of prion proteins. Dominant motions within PrP are determined using a multidimensionality reduction technique, Principal Component Analysis (PCA), on the backbone of prion structures. PCA transforms the high-dimensional representation of correlated variables of protein motion into a lower-dimensional representation, called principal components, which can highlight structural differences between proteins and allow for identification and characterization of interconformer relationships. Residues

and domains that contribute the most to the variation along the principal components (PCs), and ultimately separate structures based on their conformational differences, are considered 'conformationally variable'. The study conducted on PrP aims to identify conformationally variable domains that may be involved in the conversion process between $PrP^{C} \rightarrow PrP^{Sc}$, and ultimately, amyloidogenesis in this protein.

Chapter IV combines both bioinformatic sequence and structural based analyses, as well as evolutionary analysis, to identify homologs to the HET-s prion-forming domain (PFD), which is the first atomic structure of a functional amyloid fibril to date. Searching for structural homologs to the PFD aims to determine potential fibril-forming proteins that are amenable to adopting this specific, highly-structured B-aggregate. This search also aims to determine the prevalence of this fold in nature, and whether such as fold can serve as a common prototype of an amyloid fibril in amyloid-forming proteins.

1.5 REFERENCES

1. Prusiner SB: **Prions**. *Proceedings of the National Academy of Sciences* 1998, **95**(23):13363-13383.
2. Aguzzi A, Sigurdson C, Heikenwaelder M: **Molecular Mechanisms of Prion Pathogenesis**. *Annual Review of Pathology: Mechanisms of Disease* 2008, **3**(1):11-40.
3. Corsaro A, Thellung S, Bucciarelli T, Scotti L, Chiovitti K, Villa V, D'Arrigo C, Aceto A, Florio T: **High hydrophobic amino acid exposure is responsible of the neurotoxic effects induced by E200K or D202N disease-related mutations of the human prion protein**. *The International Journal of Biochemistry & Cell Biology* 2010, **43**(3):372-382.
4. Aguzzi A, Calella AM: **Prions: Protein Aggregation and Infectious Diseases**. *Physiological Reviews* 2009, **89**(4):1105-1152.
5. Collinge J: **Prion diseases of humans and animals: Their causes and molecular basis**. *Annual Review of Neuroscience* 2001, **24**:519-550.
6. Pastore A, Zagari A: **A Structural Overview of the Vertebrate Prion Proteins**. *Prion* 2007, **1**(3):185-197.
7. Pinotsis N, Wilmanns M: **Protein assemblies with palindromic structure motifs**. *Cellular and Molecular Life Sciences* 2008, **65**(19):2953-2956.
8. Dima RI, Thirumalai D: **Probing the instabilities in the dynamics of helical fragments from mouse PrPC**. *Proc Natl Acad Sci U S A* 2004, **101**(43):15335-15340.
9. Luisa Ronga BT, Pasquale Palladino, Raffaele Ragone, Emanuela Urso, Michele Maffia, Menotti Ruvo, Ettore Benedetti, Filomena Rossi: **The prion protein: structural features and related toxic peptides**. *Chemical Biology & Drug Design* 2006, **68**(3):139-147.
10. Lysek DA, Schorn C, Nivon LG, Esteve-Moya V, Christen B, Calzolai L, von Schroetter C, Fiorito F, Herrmann T, Guntert P *et al*: **Prion protein NMR structures of cats, dogs, pigs, and sheep**. *Proc Natl Acad Sci U S A* 2005, **102**(3):640-645.
11. Calzolai L, Lysek DA, Perez DR, Guntert P, Wuthrich K: **Prion protein NMR structures of chickens, turtles, and frogs**. *Proc Natl Acad Sci U S A* 2005, **102**(3):651-655.
12. Jodoin J, Misiewicz M, Makhijani P, Giannopoulos PN, Hammond J, Goodyer CG, LeBlanc AC: **Loss of Anti-Bax Function in Gerstmann-Straussler-Scheinker Syndrome-Associated Prion Protein Mutants**. *PLoS ONE* 2009, **4**(8).
13. Laroche-Pierre S, Jodoin J, LeBlanc AC: **Helix 3 is necessary and sufficient for prion protein's anti-Bax function**. *Journal of Neurochemistry* 2009, **108**(4):1019-1031.
14. Aguzzi A, Baumann F, Bremer J: **The prion's elusive reason for being**. In: *Annual Review of Neuroscience.* vol. 31; 2008: 439-477.
15. Gains MJ, LeBlanc AC: **Canadian Association of Neurosciences Review: Prion protein and prion diseases: The good and the bad**. *Canadian Journal of Neurological Sciences* 2007, **34**(2):126-145.

16. Weissmann C: **The state of the prion**. *Nat Rev Micro* 2004, **2**(11):861-871.
17. Harrison PM, Bamborough P, Daggett V, Prusiner SB, Cohen FE: **The prion folding problem**. *Current Opinion in Structural Biology* 1997, **7**(1):53-59.
18. DeMarco ML, Daggett V: **From conversion to aggregation: Protofibril formation of the prion protein**. *Proc Natl Acad Sci U S A* 2004, **101**(8):2293-2298.
19. Govaerts C, Wille H, Prusiner SB, Cohen FE: **Evidence for assembly of prions with left-handed beta-helices into trimers**. *Proc Natl Acad Sci USA* 2004, **101**:8342 - 8347.
20. Cobb NJ, Soennichsen FD, McHaourab H, Surewicz WK: **Molecular architecture of human prion protein amyloid: A parallel, in-register beta-structure**. *Proc Natl Acad Sci U S A* 2007, **104**(48):18946-18951.
21. Choi JH, Govaerts C, May BCH, Cohen FE: **Analysis of the sequence and structural features of the left-handed β-helical fold**. *Proteins: Structure, Function, and Bioinformatics* 2008, **73**(1):150-160.
22. Choi JH, May BCH, Govaerts C, Cohen FE: **Site-Directed Mutagenesis Demonstrates the Plasticity of the beta Helix: Implications for the Structure of the Misfolded Prion Protein**. *Structure* 2009, **17**(7):1014-1023.
23. Iengar P, Josh NV, Balaram P: **Conformational and sequence signatures in beta helix proteins**. *Structure* 2006, **14**(3):529-542.
24. Jenkins J, Pickersgill R: **The architecture of parallel beta-helices and related folds**. *Progress in Biophysics & Molecular Biology* 2001, **77**(2):111-175.
25. Kajava AV, Squire JM, Parry DAD, Andrey Kajava JMS, David ADP: **β-Structures in Fibrous Proteins**. In: *Advances in Protein Chemistry.* vol. Volume 73: Academic Press; 2006: 1-15.
26. Blinov N, Berjanskii M, Wishart DS, Stepanova M: **Structural Domains and Main-Chain Flexibility in Prion Proteins**. *Biochemistry* 2009, **48**(7):1488-1497.
27. Santo KP, Berjanskii M, Wishart DS, Stepanova M: **Comparative analysis of essential collective dynamics and NMR-derived flexibility profiles in evolutionarily diverse prion proteins**. *Prion* 2011, **5**(3).
28. Gossert AD, Bonjour S, Lysek DA, Fiorito F, Wuthrich K: **Prion protein NMR structures of elk and of mouse/elk hybrids**. *Proc Natl Acad Sci U S A* 2005, **102**(3):646-650.
29. Rossetti G, Cong X, Caliandro R, Legname G, Carloni P: **Common Structural Traits across Pathogenic Mutants of the Human Prion Protein and Their Implications for Familial Prion Diseases**. *Journal of Molecular Biology* 2011, **411**(3):700-712.
30. Aizhuo Liu RRRZSHRGKW: **Peptides and proteins in neurodegenerative disease: Helix propensity of a polypeptide containing helix 1 of the mouse prion protein studied by NMR and CD spectroscopy**. *Peptide Science* 1999, **51**(2):145-152.
31. Speare JO, Rush TS, III, Bloom ME, Caughey B: **The Role of Helix 1 Aspartates and Salt Bridges in the Stability and Conversion of Prion Protein**. *J Biol Chem* 2003, **278**(14):12522-12529.

32. Watzlawik J, Skora L, Frense D, Griesinger C, Zweckstetter M, Schulz-Schaeffer WJ, Kramer ML: **Prion Protein Helix1 Promotes Aggregation but Is Not Converted into beta-Sheet**. *J Biol Chem* 2006, **281**(40):30242-30250.
33. Riek R, Luhrs T: **Three-dimensional structures of the prion protein and its doppel**. *Clinics in Laboratory Medicine* 2003, **23**(1):209-+.
34. Lührs T, Riek R, Güntert P, Wüthrich K: **NMR Structure of the Human Doppel Protein**. *Journal of Molecular Biology* 2003, **326**(5):1549-1557.
35. Colacino S, Tiana G, Colombo G: **Similar folds with different stabilization mechanisms: the cases of prion and doppel proteins**. *Bmc Structural Biology* 2006, **6**.
36. Premzl M, Sangiorgio L, Strumbo B, Marshall Graves JA, Simonic T, Gready JE: **Shadoo, a new protein highly conserved from fish to mammals and with similarity to prion protein**. *Gene* 2003, **314**:89-102.
37. Premzl M, Gamulin V: **Comparative genomic analysis of prion genes**. *BMC Genomics* 2007, **8**(1):1.
38. Watts JC, Drisaldi B, Ng V, Yang J, Strome B, Horne P, Sy M-S, Yoong L, Young R, Mastrangelo P *et al*: **The CNS glycoprotein Shadoo has PrPC-like protective properties and displays reduced levels in prion infections**. *Embo J* 2007, **26**(17):4038-4050.
39. Daude N, Ng V, Watts JC, Genovesi S, Glaves JP, Wohlgemuth S, Schmitt-Ulms G, Young H, McLaurin J, Fraser PE *et al*: **Wild-type Shadoo proteins convert to amyloid-like forms under native conditions**. *Journal of Neurochemistry* 2010, **113**(1):92-104.
40. Tuite MF: **Cell biology - The strain of being a prion**. *Nature* 2004, **428**(6980):265-+.
41. Brockes JP: **Topics in prion cell biology**. *Current Opinion in Neurobiology* 1999, **9**(5):571-577.
42. Wickner RB, Edskes HK, Kryndushkin D, McGlinchey R, Bateman D, Kelly A: **Prion diseases of yeast: Amyloid structure and biology**. *Seminars in Cell & Developmental Biology* 2011, **22**(5):469-475.
43. Fowler DM, Koulov AV, Balch WE, Kelly JW: **Functional amyloid from bacteria to humans**. *Trends in Biochemical Sciences* 2007, **32**(5):217-224.
44. Tuite MF, Cox BS: **Propagation of yeast prions**. *Nat Rev Mol Cell Biol* 2003, **4**(11):878-890.
45. Wickner RB: **[URE3] AS AN ALTERED URE2 PROTEIN - EVIDENCE FOR A PRION ANALOG IN SACCHAROMYCES-CEREVISIAE**. *Science* 1994, **264**(5158):566-569.
46. Wickner RB, Edskes HK, Shewmaker F, Nakayashiki T: **Prions of fungi: inherited structures and biological roles**. *Nat Rev Micro* 2007, **5**(8):611-618.
47. Bossy-Wetzel E, Schwarzenbacher R, Lipton SA: **Molecular pathways to neurodegeneration**. *Nature Medicine* 2004, **10**(7):S2-S9.
48. Chiti F, Dobson CM: **Protein Misfolding, Functional Amyloid, and Human Disease**. *Annual Review of Biochemistry* 2006, **75**(1):333-366.
49. Nelson R, Eisenberg D: **Structural models of amyloid-like fibrils**. In: *Fibrous Proteins: Amyloids, Prions and Beta Proteins.* Edited by Kajava ASJMPDAD, vol. 73; 2006: 235-+.

50. Baxa U: **Structural basis of infectious and non-infectious amyloids**. *Curr Alzheimer Res* 2008, **5**(3):308-318.
51. Makin OS, Serpell LC: **Structures for amyloid fibrils**. *Febs J* 2005, **272**(23):5950-5961.
52. Nelson R, Sawaya MR, Balbirnie M, Madsen AO, Riekel C, Grothe R, Eisenberg D: **Structure of the cross-β spine of amyloid-like fibrils**. *Nature* 2005, **435**(7043):773-778.
53. Greenwald J, Riek R: **Biology of Amyloid: Structure, Function, and Regulation**. *Structure (London, England : 1993)*, **18**(10):1244-1260.
54. Shewmaker F, McGlinchey RP, Wickner RB: **Structural Insights into Functional and Pathological Amyloid**. *Journal of Biological Chemistry* 2011.
55. Uversky VN, Fink AL: **Conformational constraints for amyloid fibrillation: the importance of being unfolded**. *BBA-Proteins Proteomics* 2004, **1698**(2):131-153.
56. Fandrich M, Fletcher MA, Dobson CM: **Amyloid fibrils from muscle myoglobin**. *Nature* 2001, **410**(6825):165-166.
57. Fändrich M, Forge V, Buder K, Kittler M, Dobson CM, Diekmann S: **Myoglobin forms amyloid fibrils by association of unfolded polypeptide segments**. *Proceedings of the National Academy of Sciences* 2003, **100**(26):15463-15468.
58. Berson JF, Theos AC, Harper DC, Tenza D, Raposo G, Marks MS: **Proprotein convertase cleavage liberates a fibrillogenic fragment of a resident glycoprotein to initiate melanosome biogenesis**. *J Cell Biol* 2003, **161**(3):521-533.
59. Barnhart MM, Chapman MR: **Curli biogenesis and function**. In: *Annual Review of Microbiology*. vol. 60. Palo Alto: Annual Reviews; 2006: 131-147.
60. Chapman MR, Robinson LS, Pinkner JS, Roth R, Heuser J, Hammar Mr, Normark S, Hultgren SJ: **Role of Escherichia coli Curli Operons in Directing Amyloid Fiber Formation**. *Science* 2002, **295**(5556):851-855.
61. Cherny I, Rockah L, Levy-Nissenbaum O, Gophna U, Ron EZ, Gazit E: **The Formation of Escherichia coli Curli Amyloid Fibrils is Mediated by Prion-like Peptide Repeats**. *Journal of Molecular Biology* 2005, **352**(2):245-252.
62. Claessen D, Rink R, de Jong W, Siebring J, de Vreugd P, Boersma FGH, Dijkhuizen L, Wosten HAB: **A novel class of secreted hydrophobic proteins is involved in aerial hyphae formation in Streptomyces coelicolor by forming amyloid-like fibrils**. *Genes Dev* 2003, **17**(14):1714-1726.
63. Shewmaker F, McGlinchey RP, Thurber KR, McPhie P, Dyda F, Tycko R, Wickner RB: **The functional curli amyloid is not based on in-register parallel beta-sheet structure**. *Journal of Biological Chemistry* 2009.
64. Iconomidou VA, Chryssikos GD, Gionis V, Galanis AS, Cordopatis P, Hoenger A, Hamodrakas SJ: **Amyloid fibril formation propensity is inherent into the hexapeptide tandemly repeating sequence of the central domain of silkmoth chorion proteins of the A-family**. *J Struct Biol* 2006, **156**(3):480-488.
65. Iconomidou VA, Vriend G, Hamodrakas SJ: **Amyloids protect the silkmoth oocyte and embryo**. *FEBS Letters* 2000, **479**(3):141-145.

66. Butko P, Buford JP, Goodwin JS, Stroud PA, McCormick CL, Cannon GC: **Spectroscopic evidence for amyloid-like interfacial self-assembly of hydrophobin Sc3**. *Biochem Biophys Res Commun* 2001, **280**(1):212-215.
67. de Vocht ML, Reviakine I, Wosten HAB, Brisson A, Wessels JGH, Robillard GT: **Structural and functional role of the disulfide bridges in the hydrophobin SC3**. *Journal of Biological Chemistry* 2000, **275**(37):28428-28432.
68. Gebbink M, Claessen D, Bouma B, Dijkhuizen L, Wosten HAB: **Amyloids - A functional coat for microorganisms**. *Nat Rev Microbiol* 2005, **3**(4):333-341.
69. Mackay JP, Matthews JM, Winefield RD, Mackay LG, Haverkamp RG, Templeton MD: **The hydrophobin EAS is largely unstructured in solution and functions by forming amyloid-like structures**. *Structure* 2001, **9**(2):83-91.
70. Wosten HAB, de Vocht ML: **Hydrophobins, the fungal coat unravelled**. *Biochim Biophys Acta-Rev Biomembr* 2000, **1469**(2):79-86.
71. Eanes ED, Glenner GG: **X-RAY DIFFRACTION STUDIES ON AMYLOID FILAMENTS**. *Journal of Histochemistry & Cytochemistry* 1968, **16**(11):673-&.
72. Cohen AS, Calkins E: **ELECTRON MICROSCOPIC OBSERVATIONS ON A FIBROUS COMPONENT IN AMYLOID OF DIVERSE ORIGINS**. *Nature* 1959, **183**(4669):1202-1203.
73. Maji SK, Wang L, Greenwald J, Riek R: **Structure-activity relationship of amyloid fibrils**. *FEBS Letters* 2009, **583**(16):2610-2617.
74. Sunde M, Serpell LC, Bartlam M, Fraser PE, Pepys MB, Blake CCF: **Common core structure of amyloid fibrils by synchrotron X-ray diffraction**. *Journal of Molecular Biology* 1997, **273**(3):729-739.
75. Sawaya MR, Sambashivan S, Nelson R, Ivanova MI, Sievers SA, Apostol MI, Thompson MJ, Balbirnie M, Wiltzius JJW, McFarlane HT *et al*: **Atomic structures of amyloid cross-beta spines reveal varied steric zippers**. *Nature* 2007, **447**(7143):453-457.
76. Adam L, Zrinka G, Hélène Van M, Christian W, Alice S, Wilfred FvG, Beat HM: **A Combined Solid-State NMR and MD Characterization of the Stability and Dynamics of the HET-s(218-289) Prion in its Amyloid Conformation**. *ChemBioChem* 2009, **10**(10):1657-1665.
77. Wasmer C, Lange A, Van Melckebeke H, Siemer AB, Riek R, Meier BH: **Amyloid Fibrils of the HET-s(218-289) Prion Form a {beta} Solenoid with a Triangular Hydrophobic Core**. *Science* 2008, **319**(5869):1523-1526.
78. Kajava AV, Baxa U, Wickner RB, Steven AC: **A model for Ure2p prion filaments and other amyloids: The parallel superpleated Î²-structure**. *Proc Natl Acad Sci U S A* 2004, **101**(21):7885-7890.
79. Chan JCC, Oyler NA, Yau WM, Tycko R: **Parallel beta-sheets and polar zippers in amyloid fibrils formed by residues 10-39 of the yeast prion protein Ure2p**. *Biochemistry* 2005, **44**(31):10669-10680.
80. Perutz MF, Johnson T, Suzuki M, Finch JT: **GLUTAMINE REPEATS AS POLAR ZIPPERS - THEIR POSSIBLE ROLE IN INHERITED NEURODEGENERATIVE DISEASES**. *Proc Natl Acad Sci U S A* 1994, **91**(12):5355-5358.

81. Kajava AV, Steven AC, Andrey Kajava JMS, David ADP: **β-Rolls, β-Helices, and Other β-Solenoid Proteins**. In: *Advances in Protein Chemistry.* vol. Volume 73: Academic Press; 2006: 55-96.
82. Bradley P, Cowen L, Menke M, King J, Berger B: **BETAWRAP: Successful prediction of parallel beta-helices from primary sequence reveals an association with many microbial pathogens**. *Proc Natl Acad Sci U S A* 2001, **98**(26):14819-14824.
83. Cheng JL, Baldi P: **Three-stage prediction of protein beta-sheets by neural networks, alignments and graph algorithms**. *Bioinformatics* 2005, **21**:I75-I84.
84. McDonnell AV, Menke M, Palmer N, King J, Cowen L, Berger B: **Fold recognition and accurate sequence-structure alignment of sequences directing beta-sheet proteins**. *Proteins* 2006, **63**(4):976-985.
85. Steward RE, Thornton JM: **Prediction of strand pairing in antiparallel and parallel beta-sheets using information theory**. *Proteins* 2002, **48**(2):178-191.
86. Bryan AW, Jr., Menke M, Cowen LJ, Lindquist SL, Berger B: **BETASCAN: Probable β-amyloids Identified by Pairwise Probabilistic Analysis**. *PLoS Comput Biol* 2009, **5**(3):e1000333.
87. O'Donnell CW, Waldispühl J, Lis M, Halfmann R, Devadas S, Lindquist S, Berger B: **A method for probing the mutational landscape of amyloid structure**. *Bioinformatics* 2011, **27**(13):i34-i42.
88. Bryan AW, Menke M, Cowen LJ, Lindquist SL, Berger B: **BETASCAN: Probable beta-amyloids Identified by Pairwise Probabilistic Analysis**. *PLoS Comput Biol* 2009, **5**(3):11.
89. Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM: **Rationalization of the effects of mutations on peptide and protein aggregation rates**. *Nature* 2003, **424**(6950):805-808.
90. Belli M, Ramazzotti M, Chiti F: **Prediction of amyloid aggregation in vivo**. *EMBO Rep* 2011, **12**(7):657-663.
91. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L: **Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins**. *Nat Biotechnol* 2004, **22**(10):1302-1306.
92. Tartaglia GG, Pawar AP, Campioni S, Dobson CM, Chiti F, Vendruscolo M: **Prediction of aggregation-prone regions in structured proteins**. *Journal of Molecular Biology* 2008, **380**(2):425-436.
93. Trovato A, Chiti F, Maritan A, Seno F: **Insight into the structure of amyloid fibrils from the analysis of globular proteins**. *PLoS Comput Biol* 2006, **2**(12):1608-1618.
94. Trovato A, Seno F, Tosatto SCE: **The PASTA server for protein aggregation prediction**. *Protein Eng Des Sel* 2007, **20**(10):521-523.
95. Galzitskaya OV, Garbuzynskiy SO, Lobanov MY: **Prediction of amyloidogenic and disordered regions in protein chains**. *PLoS Comput Biol* 2006, **2**(12):1639-1648.
96. Garbuzynskiy SO, Lobanov MY, Galzitskaya OV: **FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence**. *Bioinformatics* 2010, **26**(3):326-332.

97. Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D, Eisenberg D: **The 3D profile method for identifying fibril-forming segments of proteins**. *Proc Natl Acad Sci U S A* 2006, **103**(11):4074-4078.
98. Zhang Z, Chen H, Lai L: **Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential**. *Bioinformatics* 2007, **23**(17):2218-2225.
99. Karplus M, McCammon JA: **Molecular dynamics simulations of biomolecules**. *Nat Struct Biol* 2002, **9**(9):646-652.
100. Choi JH, May BCH, Wille H, Cohen FE: **Molecular Modeling of the Misfolded Insulin Subunit and Amyloid Fibril**. *Biophysical Journal* 2009, **97**(12):3187-3195.
101. Chaney MO, Webster SD, Kuo YM, Roher AE: **Molecular modeling of the Aβ1-42 peptide from Alzheimer's disease.** *Protein Engineering* 1998, **11**(9):761-767.
102. Guo J-t, Wetzel R, Xu Y: **Molecular modeling of the core of Aβ amyloid fibrils**. *Proteins: Structure, Function, and Bioinformatics* 2004, **57**(2):357-364.
103. Gsponer J, Haberthur U, Caflisch A: **The role of side-chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid-forming peptide from the yeast prion Sup35**. *Proc Natl Acad Sci U S A* 2003, **100**(9):5154-5159.
104. Matthes D, Gapsys V, Daebel V, de Groot BL: **Mapping the Conformational Dynamics and Pathways of Spontaneous Steric Zipper Peptide Oligomerization**. *PLoS ONE* 2011, **6**(5).
105. Nasica-Labouze J, Meli M, Derreumaux P, Colombo G, Mousseau N: **A Multiscale Approach to Characterize the Early Aggregation Steps of the Amyloid-Forming Peptide GNNQQNY from the Yeast Prion Sup-35**. *PLoS Comput Biol* 2011, **7**(5).
106. Monsellier E, Ramazzotti M, Taddei N, Chiti F: **Aggregation Propensity of the Human Proteome**. *PLoS Comput Biol* 2008, **4**(10).
107. Tartaglia GG, Caflisch A: **Computational analysis of the S-cerevisiae Proteome reveals the function and cellular localization of the least and most amyloidogenic proteins**. *Proteins* 2007, **68**(1):273-278.
108. Aguzzi A, O'Connor T: **Protein aggregation diseases: pathogenicity and therapeutic perspectives**. *Nature Reviews Drug Discovery* 2010, **9**(3):237-248.
109. Trevitt CR, Collinge J: **A systematic review of prion therapeutics in experimental models**. *Brain* 2006, **129**:2241-2265.
110. Soto C, Kascsak RJ, Saborio GP, Aucouturier P, Wisniewski T, Prelli F, Kascsak R, Mendez E, Harris DA, Ironside J *et al*: **Reversion of prion protein conformational changes by synthetic beta-sheet breaker peptides**. *Lancet* 2000, **355**(9199):192-197.
111. Soto C, Sigurdsson EM, Morelli L, Kumar RA, Castano EM, Frangione B: **beta-sheet breaker peptides inhibit fibrillogenesis in a rat brain model of amyloidosis: Implications for Alzheimer's therapy**. *Nature Medicine* 1998, **4**(7):822-826.
112. Kim YS, Lim D, Kim JY, Kang SJ, Kim Y-H, Im H: **beta-Sheet-breaking peptides inhibit the fibrillation of human alpha-synuclein**. *Biochem Biophys Res Commun* 2009, **387**(4):682-687.

113. Agrawal NJ, Kumar S, Wang X, Helk B, Singh SK, Trout BL: **Aggregation in Protein-Based Biotherapeutics: Computational Studies and Tools to Identify Aggregation-Prone Regions**. *Journal of Pharmaceutical Sciences* 2011, **100**(12):5081-5095.
114. Perrier Vr, Wallace AC, Kaneko K, Safar J, Prusiner SB, Cohen FE: **Mimicking dominant negative inhibition of prion replication through structure-based drug design**. *Proc Natl Acad Sci U S A* 2000, **97**(11):6073-6078.
115. Kuwata K, Nishida N, Matsumoto T, Kamatari YO, Hosokawa-Muto J, Kodama K, Nakamura HK, Kimura K, Kawasaki M, Takakura Y *et al*: **Hot spots in prion protein for pathogenic conversion**. *Proceedings of the National Academy of Sciences* 2007, **104**(29):11921-11926.

CHAPTER II

Discordant and chameleon sequences: Their distribution and implications for amyloidogenicity

Amyloid fibrils are characterized as highly ordered and distinct β-sheet-rich aggregates of many copies of a peptide or protein. This molecular aggregation is contingent on the ability of a native protein or peptide to undergo a secondary structure change such that it acquires substantial β-sheet content. Arguably, identification of these conformationally variable regions in a protein may be indicative of potentially amyloidogenic segments. Chapter II analyzes two classes of structurally ambivalent peptides observed in secondary structures, mainly discordant and chameleon sequences, and tests their efficacy as predictors of amyloidogenic segments. A meta-scale distribution of these segments is conducted on a database of protein domains and several cohorts representing amyloid-forming proteins as well as the prion family. Through this sequence-based analysis, statistical relationships are derived between each class of segments and amyloidogenicity, and the ramifications of these relationships on future prediction of amyloidogenic proteins is discussed.

A version of this chapter is originally published as:
Gendoo, D. M. and Harrison, P. M. (2011), Discordant and chameleon sequences: Their distribution and implications for amyloidogenicity. Protein Science, 20: 567–579. doi: 10.1002/pro.590

2.1 ABSTRACT

Identification of ambiguous encoding in protein secondary structure is paramount to develop an understanding of key protein segments underlying amyloid diseases. We investigate two types of structurally ambivalent peptides, which were hypothesized in the literature as indicators of amyloidogenic proteins: discordant α-helices and chameleon sequences. Chameleon sequences are peptides discovered experimentally in different secondary-structure types. Discordant α-helices are α-helical stretches with strong β-strand propensity or prediction. To assess the distribution of these features in known protein structures, and their potential role in amyloidogenesis, we analyzed the occurrence of discordant α-helices and chameleon sequences in nonredundant sets of protein domains (n = 4263) and amyloidogenic proteins extracted from the literature (n = 77). Discordant α-helices were identified if discordance was observed between known secondary structures and secondary-structure predictions from the GOR-IV and PSIPRED algorithms. Chameleon sequences were extracted by searching for identical sequence words in α-helices and β-strands. We defined frustrated chameleons and very frustrated chameleons based on varying degrees of total β propensity ≥α propensity. To our knowledge, this is the first study to discern statistical relationships between discordance, chameleons, and amyloidogenicity. We observed varying enrichment levels for some categories of discordant and chameleon sequences in amyloidogenic sequences. Chameleon sequences are also significantly enriched in proteins that have discordant helices, indicating a clear link between both phenomena. We identified the first set of discordant-chameleonic protein segments we predict may be involved in amyloidosis. We present a detailed analysis of discordant and chameleons segments in the family of one of the amyloidogenic proteins, the Prion Protein.

2.2 INTRODUCTION

Identification of ambiguous encoding in protein secondary structure is paramount to develop an understanding of key protein segments underlying many conformational diseases. Amyloid diseases, such as prion disease, for example, are characterized by major conformational changes whereby native proteins stretches can adopt β-sheet

conformations and stabilize into multimeric assemblies that are highly pathogenic **(1-3)**. This behavior is linked to many human diseases, including Transmissible Spongiform Encephalopathies (TSEs), and has been experimentally observed in dozens of other proteins **(2; 3)**. The resultant amyloids are fibrillar assemblies of β–sheets with a characteristic 'cross–β' configuration, *i.e.*, with β–strand axes orthogonal to the major long axis of the fibrils **(4-6)**. Accordingly, a key task is to understand how the protein sequence can encrypt these alternative conformations and configurations of protein chains, in additional to their normal, cellular forms. Two sequence feature types that evidence ambiguous encoding of secondary structures, and which are hypothesized in the literature as indicators of amyloidogenic sequences, are "discordant" and "chameleon" sequences.

The phenomenon of *discordance* **(7; 8)** refers to sequences of a native α–helix which may sometimes have a high intrinsic β propensity; this may arise because of specific long-range side-chain interactions causing the conservation of amino acids that would otherwise prefer to be in β–strands. Kallberg *et al*. detected 37 incidences of 'discordant' α–helices in >1300 protein structures, of ≥7 residues in length. These 'discordant' α–helices were predicted to form β-strands by multiple orthogonal secondary-structure algorithms, even though they had been determined to be α–helices in known experimental three-dimensional structures **(7; 8)**. These α–β discordant stretches were hypothesized to be associated with amyloid fibril formation **(8)**, although no statistical relationship between discordance and amyloidogenicity was discerned.

Another aspect of ambiguity in encoding secondary structure is the existence of *chameleon* sequences. Chameleon sequences are short peptide stretches experimentally shown to adopt multiple secondary structure conformations in different proteins **(9; 10)**. Chameleons are thus structurally ambivalent peptides, because they assume different secondary structures in different contexts **(11-13).** Short 5-mer chameleon sequences in the PDB were first reported in a study of the use of sequence homology for protein structure prediction **(14),** and later studies reported hexapeptide chameleon sequences in a larger PDB database **(9)**. The longest 'chameleon-HE' (*i.e.*, Helix (H) *vs.* Sheet (E)) sequence reported to date is seven residues long **(11; 12; 15)**. Chameleon sequences that

adopt both α-helical and β-sheet conformation are of particular interest in the analysis of the sequence determinants of amyloidogenicity.

The discordant α-helix and chameleon sequence phenomena are possibly strong indicators of conformational plasticity, and prime candidates for causation of amyloidogenicity.

In this work, we decided to rigorously test the hypotheses that discordant α-helices and chameleon sequences have a causative link to amyloid formation. To this end, we have performed a meta-scale statistical analysis of the distribution of discordant and chameleon sequences in a database of protein domains (SCOP), as well as in amyloidogenic proteins and their determinants. From our analysis of discordant stretches in protein domains, we suggest protein functions where structurally-ambivalent peptides may be of importance. We also discuss the enrichment we have observed for various definitions of discordant α-helices and chameleon sequences, in amyloidogenic determinants. We introduce the first set of identified discordant-chameleonic segments that may be involved in amyloidosis. Finally, we explore in detail the specific important case of chameleon and discordant segments in the PrP protein family.

2.3 RESULTS

2.3.1 Distribution of discordant α–helices

The distribution of discordant α–helices of length ≥5 residues was analyzed in a non-redundant data set of 4263 SCOP protein domains (see *Methods* for details). Using the consensus of the GOR-IV and PSIPRED secondary-structure prediction algorithms, we identified 119 discordant α–helices **(Table 2.1)**. The complete list is in **(Supplementary Table S2.1)**. α–β Discordances or 'mispredictions' by the individual secondary-structure prediction algorithms are also tabulated.

Table 2.1: Occurrences of Discordant Stretches in Amyloidogenic Proteins and SCOP Domains.

	Amyloidogenic protein chains (n=77)	**SCOP domains (n=4263)**	**Significance ✱**	**Comments**
Total Number of Residues in the dataset	11939 [3810 residues in α–helices ≥5 residues long]	800218 [273330 residues in α–helices ≥5 residues long]	---------	---------
Discordant proteins	5 [1300 total residues]	111 [28348 total residues]	---------	Consensus ✱✱✱
Discordant helices ✱✱	5 [30]	119 [667]	0.027	Consensus ✱✱✱
Discordant helices ✱✱	40 [266]	907 [6352]	<0.000001	GOR ✱✱✱
Discordant helices ✱✱	9 [59]	280 [1689]	0.018	Psipred ✱✱✱

✱ Binomial probability (one-tailed test, using a Poisson approximation) of obtaining the observed number (or a larger number) of *initial* residue positions in discordant alpha-helical stretches in the alpha-helices ≥5 residues in length in the amyloidogenic protein set, compared to the same statistics for the SCOP domain set. Subtracting the maximum possible disallowed positions (*i.e.,* from other positions within the discordant alpha-helical stretches, and the residue positions immediately adjacent to the start and end of them) has a negligible effect on the calculations.

✱✱ Uninterrupted stretches of discordant residues with a length of 5 or greater constitute a discordant helix. The number of discordant residues is indicated in brackets.

✱✱✱ As comparison to our consensus approach, the number of discordant helices obtained using one secondary structure prediction tool is demonstrated.

The discordant α–helices occur in 111 protein domains, with eight domains having two discordant stretches each **(Figure 2.1)**. These 111 domains are dubbed 'discordant protein domains'. Most discordant stretches were 5 residues long, with stretches greater than 7 being exceedingly rare **(Figure 2.1)**. The list of discordant protein domains includes known amyloidogenic proteins, including pilin, alpha-lactalbumin (pdb 1b9o), chicken

lysozyme (pdb 3lzt), triacylglycerol lipase (pdb 1tca), cytochrome c (pdb 1gu2), and human PrP (pdb 1i4m) **(8; 16-18)**. The list also includes the PrP paralog Doppel (which has not been shown experimentally to make amyloid), and a yeast prion candidate, MCM1, which ranked amongst the top 10 yeast prion candidates in an experimental screen **(19)**. Interestingly, almost 10% of the discordant protein domains are viral proteins; conformational changes of proteins in viral envelopes facilitate host membrane fusion and entry of viruses. The discordant protein domains do not have a general tendency for increased β propensity, compared with protein domains in general **(Figure 2.2, section A)**; this is further demonstrated by the discordant stretches clearly having greater β than α propensity, compared with random samplings from these protein domains of α–helical stretches of the same length (**Figure 2.2, section B**; P-values for two-tailed *t*-tests $\leq 10^{-3}$ for all stretch lengths).

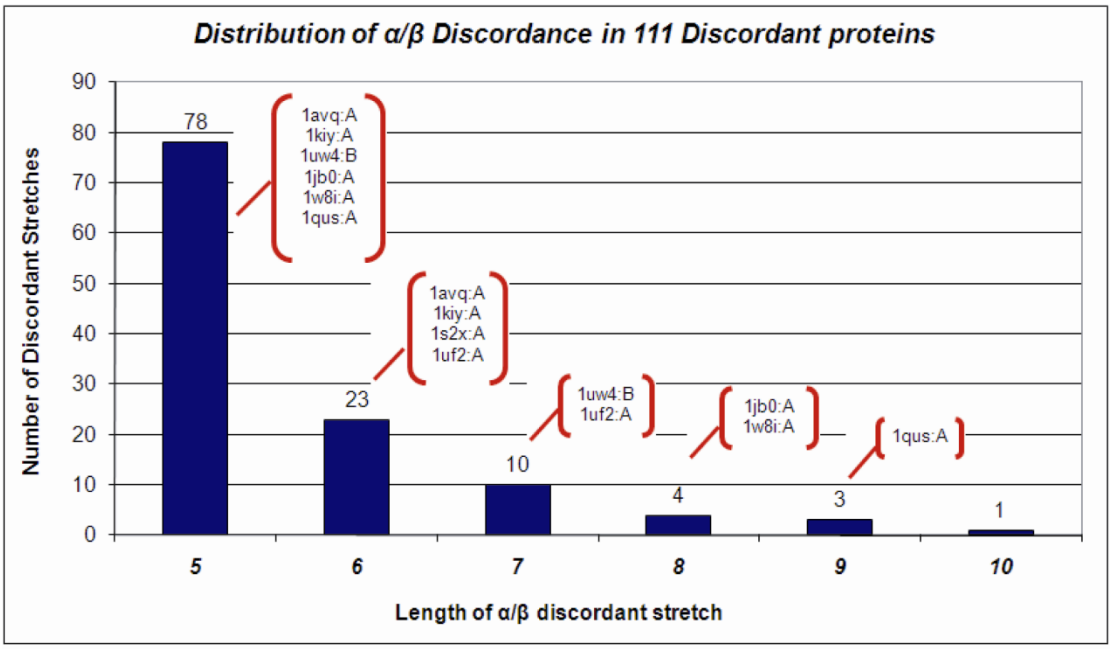
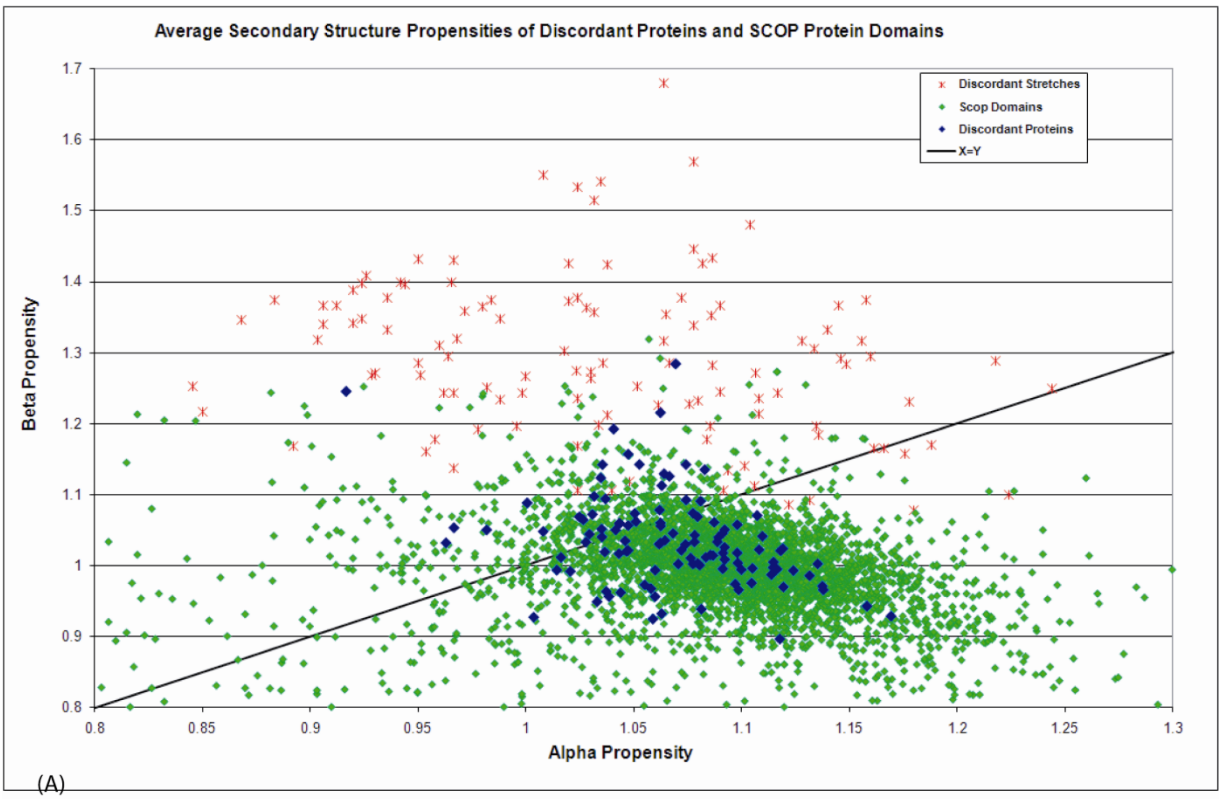
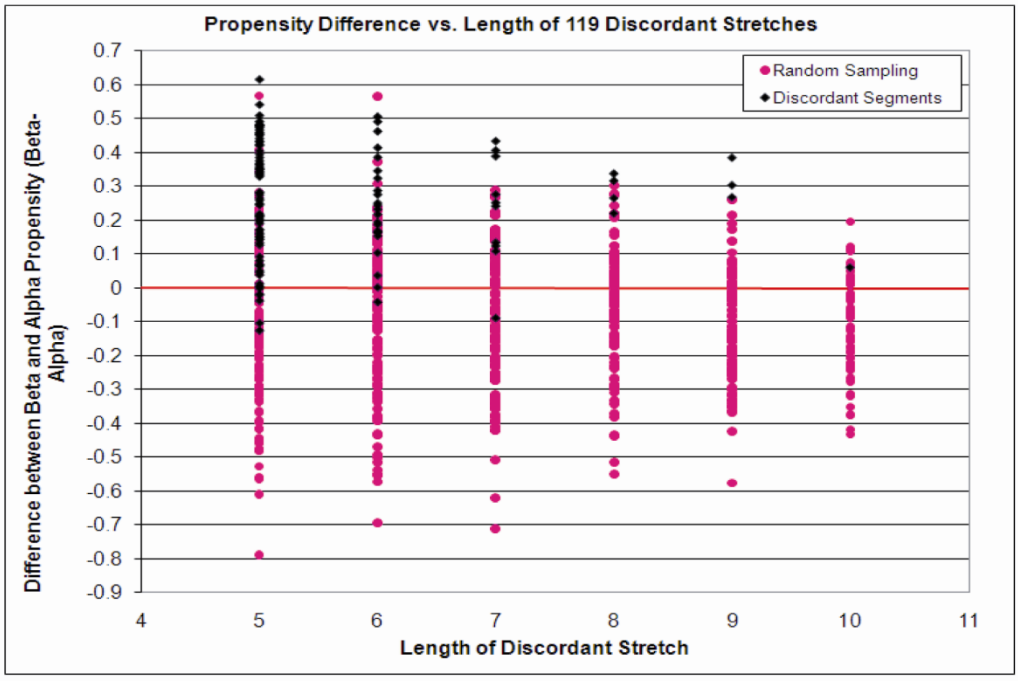


Figure 2.1: Distribution of 119 discordant stretches by length. There are 119 discordant stretches in 111 discordant proteins. Protein chains with multiple discordant stretches are identified in red brackets.


Average Secondary Structure Propensities of Discordant Proteins and SCOP Protein Domains
Discordant Stretches
Scop Domains
Discordant Proteins
X=Y
Beta Propensity
Alpha Propensity
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.8
0.85
0.9
0.95
1
1.05
1.1
1.15
1.2
1.25
1.3

(A)


Propensity Difference vs. Length of 119 Discordant Stretches
Random Sampling
Discordant Segments
Difference between Beta and Alpha Propensity (Beta-Alpha)
Length of Discordant Stretch
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
4
5
6
7
8
9
10
11

(B)

Figure 2.2 (previous page): (A) Comparison of average secondary structure propensities of the discordant proteins and the SCOP database. For the discordant proteins, propensities of the discordant stretches as well as the average of helices for each of the discordant proteins are shown. For the SCOP domain dataset, the average of entire helices for each protein is shown.

(B) Net gain in secondary structure propensity of discordant segments. The difference between beta and alpha propensities of discordant segments (black) compared to 600 random helical segments of the same length (pink) is shown. P-values for comparing the discordant stretches to the random helical samples are 8.36529E-26, 3.29408E-10, 3.5661E-06, 2.173E-03, 8.11975E-06 for 5-mer, 6-mer, 7-mer, 8-mer, and 9-mer fragments, respectively, based on a two-tailed t-test assuming equal variance.

2.3.2 Distribution of chameleon sequences

Chameleons were defined as peptides of five or six residues in length that occur in both α–helix and β–strand secondary assignments made by the DSSP algorithm **(10-12)** (see *Methods* for details). In the ASTRALSCOP protein domain data set analyzed, we observed that a sizeable fraction (~14%) of all 5-mer α–helical peptides are chameleons **(Table 2.2)**; however, a much smaller fraction of 6-mers α–helical peptides are also observed in β–strands (0.6%). *Very Frustrated chameleons* were defined as the subset of these chameleon sequences that have β propensity ≥1.5× their α propensity **(Figure 2.3, section C)**. These sequences are thus predicted to be more 'frustrated' when in an α–helical state, *i.e.*, the specific local side-chain environment of the α–helix is 'frustrating' the propensity for the sequence to adopt a β conformation. These sequences are very unusual, occurring at the rate of only 1 in ~890 α–helical 5-mers, and almost non-existent for 6-mers (just 2 cases).

Table 2.2: The number of helical segments, chameleons, and frustrated chameleons for each cohort.

	SCOP domains with <40% Identity (n=4263 chains)		**Discordant Proteins (n=119 chains)**		**Amyloidogenic Proteins with <40% Identity (n=77 chains) ✱**		**Amyloidogenic Determinants (n=45) ✱✱**	
	5-mer	**6-mer**	**5-mer**	**6-mer**	**5-mer**	**6-mer**	**5-mer**	**6-mer**
Total Helical Fragments	213854	160843	9035	7341	2786	2436	120	99
Chameleons in Helices	29645 (13.86%)	986 (0.613%)	1329 (14.70%)	49 (0.66%)	364 (13.06%)	15 (0.615%)	20 (16.7%)	2 (3.36%)
Frustrated Chameleons in Helices✱✱✱	16283 (7.614%)	498 (0.310%)	846 (9.363%)	29 (0.395%)	197 (7.071%)	6 (0.246%)	16 (13.3%)	2 (3.36%)
Very Frustrated Chameleons in Helices✱✱✱	240 (0.1122%)	2 (0.0012%)	14 (0.15%)	0	6 (0.215%)	0	2 (7.09%)	0 (1.68%)
% Frustrated Chameleons in Chameleon Fragments	0.8095	0.2028	1.053	0	1.648	0	18.18	50
$P_{Chameleon}$ ✱✱✱✱	----------	----------	**0.01**	0.29	0.89	0.53	0.22	0.12
$P_{FrustratedChameleon}$ ✱✱✱✱	----------	----------	**2.9×10^{-10}**	0.11	0.15	0.37	**0.02**	**0.04**
$P_{VeryFrustratedChameleon}$ ✱✱✱✱	----------	----------	0.15	0.92	0.10	0.97	**0.008**	----------

(continued from Table 2.2, previous page)

Significance of identified chameleons (See $P_{Chameleon}$ and $P_{FrustratedChameleon}$ and $P_{VeryFrustratedChameleon}$) is calculated using a hypergeometric test against the number of chameleonic fragments identified in SCOP (for all other cohorts than SCOP).

✱ Single-chain domains of 77 Amyloidogenic protein structures are selected for the analysis.

✱✱ From the selected determinants (n=45), determinants with experimentally verified secondary structure were selected for analysis. Of these, 17 determinants had helical structures ≥5 residues.

✱✱✱ *Frustrated Chameleons* and *Very Frustrated Chameleons* are calculated as described in the *Methods* section. We defined *frustrated chameleons* as chameleons with higher β propensity than α propensity, and *very frustrated chameleons* as chameleon sequences with very high β propensity values (operationally, with total β propensity ≥1.5× α propensity).

✱✱✱✱ This is the length-specific binomial probability of finding chameleon or frustrated chameleon sequences in each of the cohorts in comparison to the numbers of chameleon and frustrated chameleons observed in all helices in the SCOP database. A Poisson approximation is used for expected values <0.1. Significant P-values ($P<0.05$) are in **bold**. All counts in categories with $P<0.05$ are also significant for tests of significance against non-redundant sets of whole protein chains, derived from the DSSP database with a 40% sequence identity threshold, using BLASTCLUST **(52)**.

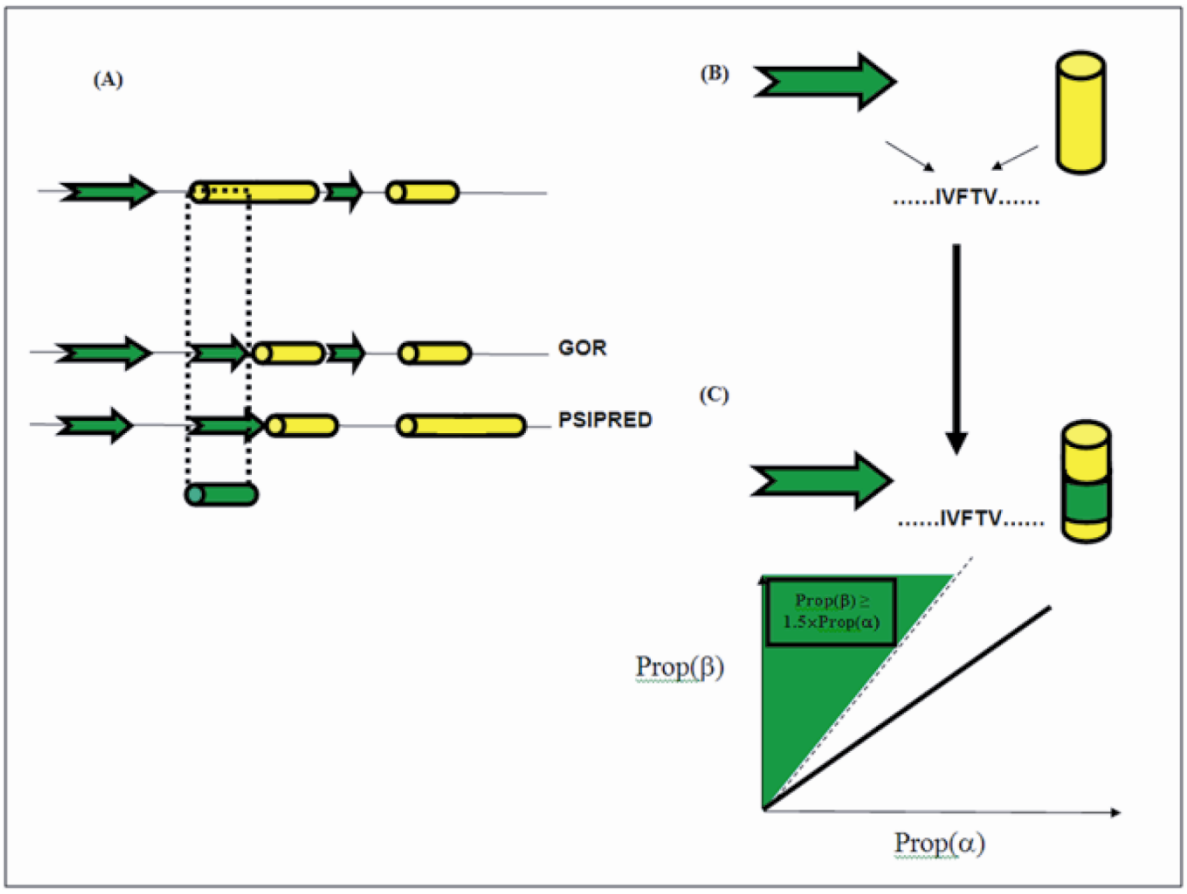


Figure 2.3: Schematic of definition of discordance, chameleon and *very frustrated chameleon*. (A) A discordant α-helix is any stretch in a known α-helix ≥5 residues long that is predicted as β-strand. These are annotated using the GOR and PSIPRED algorithms **(43; 44)**. The list of cases that are assigned by either or both algorithms have been analyzed. **(B)** A chameleon is a protein sequence word that is observed in both α-helices and β-strands in known protein structures. **(C)** A *very frustrated chameleon* is a chameleon with β propensity [Prop(β)] greater than or equal to 1.5× the α propensity [Prop(α)].

One might expect that chameleon sequences would have a tendency to low sequence complexity, *i.e.,* sequences that fit into both α and β secondary structure might have an enrichment of amino-acid runs in them. To test this hypothesis, we analyzed the distribution of amino-acid runs in the 5-mer and 6-mer chameleon sequences from the ASTRALSCOP domain data set **(Table 2.3)**. We examined runs of size 4, 5 and 6 for all twenty amino-acid residue types. We found that a small number of amino-acid run types are significantly over-represented in chameleon sequences (most frequently runs of alanine, histidine, valine and leucine) **(Table 2.3)**.

Table 2.3: Analysis of Sequence Complexity in Pentameric and Hexameric Chameleons of the SCOP domain dataset.

Sequence and Length of Run		**Count Observed in Pentameric Chameleons of non-redundant SCOP data set (n=29,645)**	**Count Observed in Hexameric Chameleons of non-redundant SCOP data set (n=986)**
Length (4X,5X,6X)	**Sequence Observed**		
4	AAAA	47†	7†
4	HHHH	4†	2†
4	LLLL	45†	0
4	TTTT	1	0
4	VVVV	15†	0
5	AAAAA	8†	0
5	HHHHH	3	1

Identified same-residue runs of lengths 4-6 are shown, and their corresponding counts in the chameleon sets.

† $P<0.01$ for a Poisson distribution given observed frequency of each run in all α–helices in the SCOP database. There are no significantly underrepresented runs in the chameleons.

2.3.3 Are discordant, chameleon, and frustrated chameleon sequences over-represented in amyloidogenic sequences?

To assess whether these ambiguous encoding segments are enriched in amyloidogenic determinants and their proteins, a set of amyloidogenic proteins from the current literature **(16-18; 20)** were reduced for sequence redundancy (with a 40% sequence identity threshold) using the PISCES tool and manual curation **(21)**. We identified five cases of discordant α–helices (identified as the consensus of mispredictions by both the GOR-IV and PSIPRED algorithms) in amyloidogenic proteins **(Tables 2.1 and 2.4)**. This is a moderately significant enrichment **(Table 2.1)**. All but one of these discordant stretches are not in amyloidogenic determinants of these proteins **(Table 2.4)**; the lone exception being in an amyloidogenic determinant of the Prion Protein PrP. Interestingly, comparison of discordance using only one prediction tool indicates a highly significant enrichment of discordant α–helices that were identified through GOR-IV β mispredictions alone (**Table 2.1**, $P<0.000001$).

Table 2.4: Identified Discordant Segments in Amyloidogenic Proteins.

Protein	PDB	Discordant Region	Discordant Segment	Chameleon Sequence	Amyloidogenic Determinant?
Coagulation factor XIII [*H. sapiens*]	1ex0:A	239-244	IKVSRV	NONE	NO
Lysozyme [*H. sapiens*]	1jsf:A	28-33	WMCLAK	NONE	NO
Cytochrome c [*B.taurus*]	1ppj:T	52-58	VAFYLVY	NONE	NO
Prion Protein [*H. sapiens*]	3hak:A	56-60	VNITI	1e4k:C, 1e4j:A, 1fnl:A, 1hfl, 1op8:F, 1op8:C, 1op8:D, 2vov:A, 2vow:A, 2vox:A	YES
Triacylglycerol lipase [*C. antartica*]	3icv:A	255-260	FSYVVG	NONE	NO

The discordant segments and corresponding chameleon proteins are identified.

We investigated whether chameleon and frustrated chameleon sequences are enriched in amyloidogenic proteins, and more specifically, in their experimentally-defined amyloidogenic determinants. We compared these chameleon occurrences to those observed for the non-redundant SCOP protein domain sets **(Table 2.2)**. As one would expect, in all cohorts analyzed, there is an over-abundance of pentameric chameleon sequences over hexameric ones. Counting up chameleon sequences simply, we find marginal significant enrichments in amyloidogenic determinants of frustrated and very frustrated chameleons (**Table 2.2**, binomial P-values ≤ 0.04). However, arguably, one should remove over-lapping cases of chameleon 5-mers and 6-mers in protein sequences. After doing this, we only observe a significant enrichment of 6-mer frustrated chameleons in amyloidogenic determinants **(Table 2.5)**. This indicates a marginal link of $\alpha-\beta$ chameleons in known $\alpha-$helices to amyloidogenicity.

Table 2.5: N*on-overlapping Counts* of Chameleons and Frustrated Chameleons for each Cohort.

	SCOP domains with <40% Identity (n=4263 chains)		**Discordant Proteins (n=119 chains)**		**Amyloidogenic Proteins with <40% Identity (n=77 chains) ✱**		**Amyloidogenic Determinants (n=45) ✱✱ †**	
	5-mer	**6-mer**	**5-mer**	**6-mer**	**5-mer**	**6-mer**	**5-mer**	**6-mer**
Total Helical Fragments	46199	36255	2045	1627	632	506	33	25
Chameleons in Helices	15577 (33.7%)	888 (2.5%)	661 (32.3%)	42 (2.6%)	208 (32.9%)	12 (2.4%)	11 (33.3%)	2 (8%)
Frustrated Chameleons in Helices✱✱✱	8056 (17.4%)	443 (1.2%)	393 (19.2%)	24 (1.5%)	106 (16.77%)	5 (0.988%)	8 (24.2%)	2 (8%)
$P_{Chameleon}$ ✱✱✱✱	----------	----------	0.09	0.38	0.35	0.53	0.56	0.12
$P_{FrustratedChameleon}$ ✱✱✱✱	----------	----------	**0.02**	0.20	0.35	0.41	0.21	**0.04**

✱ Single-chain domains of 77 Amyloidogenic protein structures are selected for the analysis.

✱✱ From the selected determinants (n=45), determinants with experimentally verified secondary structure were selected for analysis. Of these, 17 determinants had helical structures ≥5 residues.
† Adding the small transmembrane protein 1SFP amyloidogenic determinant does not change the significances, except for making an enrichment of 6-mer chameleons generally (P=0.03).

✱✱✱ *Frustrated Chameleons* and *Very Frustrated Chameleons* are calculated as described in the *Methods* section. We defined *frustrated chameleons* as chameleons with higher β propensity than α propensity, and *very frustrated chameleons* as chameleon sequences with very high β propensity values (operationally, with total β propensity ≥1.5× α propensity).

✱✱✱✱ This is the length-specific binomial probability of finding chameleon or frustrated chameleon sequences in each of the cohorts in comparison to the numbers of chameleon and frustrated chameleons observed in all helices in the SCOP database. A Poisson approximation is used for expected values <0.1. Significant P-values (P<0.05) are in **bold**. All counts in categories with P<0.05 are also significant for tests of significance against non-redundant sets of whole protein chains, derived from the DSSP database with a 40% sequence identity threshold, using BLASTCLUST **(52)**.

2.3.4 Segments that are both chameleon and discordant

Although we have demonstrated that both discordant and chameleon segments are moderately or marginally linked to amyloidogenicity, respectively, we discovered that 5-mer chameleons are significantly enriched in the discordant protein domains we had identified **(Table 2.2)**. This indicates a clear link between the two phenomena, and any sequence segment can occur as both phenomena. We wished to discern whether combining both phenomena may improve predictions of amyloidogenic segments. We identified 28 discordant α–helical segments that also contain chameleon sequences that occur in at least one further protein **(Supplementary Table S2.2)**, as described in *Methods*. The subset of these that are also predicted by the algorithm Pafig **(22)** to be amyloidogenic are listed in **(Table 2.6)**. Notable examples of discordant α–helical segments exhibiting chameleon conformational behavior include the discordant stretch of Human PrP helix 2 (which to date is the only example of a known amyloidogenic protein with a combined discordant and chameleonic sequence segment), and the only identified chameleon sequence of the PrP paralog Doppel, in its most N-terminal α–helix. The list includes other interesting candidates, such as HSV Glycoprotein D; HSV is proposed to be linked with amyloidogenicity in Alzheimer's disease.

Table 2.6 (next page): Conformationally-flexible protein segments from SCOP that are both discordant and chameleonic, and additionally predicted to be amyloidogenic by the Pafig algorithm.

Proteins are sorted in descending order by the reliability score (RS) of the Pafig fibril-forming hexapeptide segment.

** The reliability score is not shown for some proteins which were part of the training set of the Pafig support vector model.

PDB	Discordant Segment	Protein [Organism]	Pafig RS	Chameleon Sequences
1ccw:B	306-310 GVIVT	Glutamate mutase, large subunit [Clostridium cochlearium]	9	1hzp:A, 1hzp:B, 1m1m:A, 1m1m:B, 1okk:D, 1rj9:A, 1u6e:A, 1u6e:B, 1u6s:A, 1u6s:B, 2ahb:A, 2ahb:B, 2aj9:A, 2aj9:B, 2cnw:F, 2cnw:E, 2cnw:D, 2iyl:D, 2j7p:E, 2j7p:D, 2q9a:A, 2q9a:B, 2q9b:A, 2q9b:B, 2q9c:A, 2q9c:B, 2qnx:A, 2qnx:B, 2qny:A, 2qny:B, 2qnz:A, 2qnz:B, 2qO0:A, 2qO0:B, 2qO1:A, 2qO1:B, 2qx1:A, 2qx1:B, 3dii:A, 3dii:B, 3dij:A, 3dij:B
1gxy:A	70-74 TALVA	Eukaryotic mono-ADP-ribosyltransferase ART2.2 [Rattus norvegicus]	9	1jxh:A, 1jxi:A, 1jxi:B, 2eay:A, 2eay:B, 2uzh:C, 2uzh:A, 2uzh:B, 3ddy:A, 1llj:A
1i4m:A	62-66 VNITI	Prion protein domain [Homo sapiens]	**	1e4k:C, 1e4j:A, 1fnl:A, 1hf1, 1op8:F, 1op8:C, 1op8:D, 2vov:A, 2vow:A, 2vox:A
1jma:A	233-237 VYSLK	HSV glycoprotein D [Herpes simplex virus type 1]	9	1a22:B, 1axi:B, 1hwg:B, 1hwg:C, 1hwh:B, 3hhr:B, 3hhr:C, 1kf9:F, 1kf9:E, 1kf9:C, 1kf9:B, 2aew:A, 2aew:B
1nth:A	273-277 TTIVD	Monomethylamine methyltransferase MtmB [Archaeon Methanosarci-na barkeri]	8	1nfg:C, 1nfg:A, 1nfg:B, 1nfg:D, 1nu5:A
1tca:A	232-236 FSYVV	Triacylglycerol lipase [Candida antarctica), form b]	7	1iic:A, 1iic:B, 1iid:A, 2nmt:A, 2p6e:F, 2p6e:E, 2p6e:C, 2p6e:A, 2p6e:B, 2p6e:D, 2p6f:F, 2p6f:E, 2p6f:C, 2p6f:A, 2p6f:B, 2p6f:D, 2p6g:F, 2p6g:E, 2p6g:C, 2p6g:A, 2p6g:B, 2p6g:D
1v74:A	99-103 RIYLE	Colicin D nuclease domain [Escherichia coli]	5	1s3o:A, 1s3o:B, 2dud:A, 2dud:B,3ull:A, 3ull:B
1xg7:A	149-153 IVFTV	Hypothetical protein PF0904 [Pyrococcus furiosus]	5	2hew:F, 2hey:F, 2hey;G
1muk:A	505-509 SVAIL	Reovirus polymerase lambda3 Reovirus [TaxId: 10891]}	9	1knx:C, 1knx:B, 1knx:D

2.3.5 Chameleons and discordance in the Prion Protein (PrP) family

In our data set here, PrP was the only known example of an amyloidogenic protein with a sequence segment that is both chameleonic and discordant **(Table 2.7)**. Using new sequences for echinoderms, reptiles and birds **(23)** we re-examined the phylogenetic distribution of the α–helical discordance in PrP helix 2, and found that it not only occurs in mammals, but also in birds, and is absent from amphibian and reptilian prion protein family members (**Table 2.7** and **Supplementary Table S2.3**). However, globular PrP domain structures from amphibians and reptiles do contain chameleon sequences **(Supplementary Table S2.3)**. Analysis of the PrP discordant segments using Consurf **(24)** indicates deep conservation in mammals for residues that have high beta propensity, such as Valine **(Supplementary Figure S2.1, section A)**. This conservation trend is also correlated with an increased predicted relative importance for these residues as determined by the Evolutionary Trace algorithm **(25; 26)**; 80% of the discordant stretch is found within the top 68% of important residues of the protein **(Supplementary Figure S2.1, section B)**.

The discordant stretch contains an N-glycosylation site (N-*x*-[T or S], where *x* is any residue); we checked whether this was a general phenomenon for discordance, but observed no significant association, with PrP being the only such case. However, one notable tendency is that the orthogonal bundle is the most observed protein architecture amongst the discordant proteins (20% of the protein list, **Supplementary Figure S2.2**), and is the same as that of the PrP fold **(7; 27)**.

Interestingly, the combined chameleonic and α–helix discordant region was highly conserved throughout mammals (as VNITI or VNITV) **(Supplementary Table S2.3)**. Discordant stretches that are also chameleonic, are additionally observed in the first helix of Doppel **(Table 2.7)**, thus providing a prediction for an amyloidogenic determinant in this protein.

Table 2.7: Discordant and Chameleon Segments in Representatives of the Prion Protein (PrP) family.

Discordant Chain	Protein	Discordant Segment	Discordant Sequence	Chameleon Sequences
1xyx:A	PrP Mouse (Mus musculus)	60-64	VNITI	1e4j:A, 1e4k:A, 1fnl:C, 1hf1:A
1b10:A	PrP Golden hamster (Mesocricetus auratus)	56-60	VNITI	1e4j:A, 1e4k:A, 1fnl:C, 1hf1:A
1i4m:A	PrP Human (Homo sapiens)	62-66	VNITI	1e4j:A, 1e4k:A, 1fnl:C, 1hf1:A
1y2s:A	PrP Sheep (Ovis aries)	62-66	VNITV	1iz6:A
1xyw:B	PrP American Elk (Cervus elaphus nelsoni)	60-64	VNITV	1iz6:A
1u3m:A	PrP Chicken (Gallus gallus)	64-68	ITVTE	
1xyj:A	PrP Cat (Felis silvestris catus)	60-64	VNITV	1iz6:A
1xyk:A	Prp Dog (Canis familiaris)	60-64	VNITV	1iz6:A
1xyq:A	Prp Pig (Sus scrofa)	60-64	VNITV	1iz6:A
2fj3:A	PrP Rabbit (Ornithorhynchus anatinus)	56-60	VNITV	1iz6:A
1dx0:A	PrP Bovine	57-61	VNITV	1iz6:A
2k56:A	PrP Bank Vole (Clethrionomys glareolus)	62-66	VNITI	1e4j:A, 1e4k:A, 1fnl:C, 1hf1:A
1lg4:A	Human Doppel (Homo sapiens)	25-29	RYYEA	1a6c:A
1i17:A	Mouse Doppel (Mus musculus)	27-31	RYYAA	1crf:A

2.4 DISCUSSION

We have performed an exhaustive study for sequences capable of being found in secondary structure types, either explicitly, such as chameleons, or potentially, as is the case with discordant stretches. Conformational plasticity of these sequences makes them prime candidates for amyloidogenic segments, which are largely characterized by a conformational change from an α-helix to β-sheet conformation. To test this hypothesis it was imperative to first develop an understanding of the distribution of these segments in protein domains in general, to facilitate statistical comparisons with the subset of amyloid-forming proteins.

Our meta-analysis of discordant proteins in a non-redundant dataset of protein domains suggests possible roles of discordance which have been overlooked in previous publications addressing this topic **(8)**. We have observed that discordant protein domains are enriched for specific protein-fold types and functional categories. For example, using Gene Ontology (GO) terms, we have observed that the most frequent molecular functions of discordant proteins were 'Metal-ion binding' (GO:0046872) and 'hydrolase' activity (GO:0016787), with more than a quarter of the discordant proteins exhibiting either activity **(Supplementary Table S2.4)**. Discordant proteins were found frequently in the 'Extracellular' region (GO:0005576, 16 proteins), and 'Membrane' (GO:0016020, 16 proteins) of cells, while the most frequent biological process of these proteins was 'Transport' (GO:0006810, 9 proteins). These results complement our findings that almost 10% of discordant proteins are viral proteins, where such functions are imperative for host interaction and viral replication and survival. An analysis of 3D folds using CATH also indicated significant enrichment of the orthogonal bundle ($P \leq 10^{-31}$, using hypergeometric probability) and three-layer αβα sandwich ($P \leq 10^{-11}$) architectures **(Supplementary Figure S2.2)**, suggesting that amino acid orientations in these folds may promote discordance. The effect of these folds on the discordant stretches, and their implications on the overall function of their respective discordant proteins would be an interesting point for future research. Taken collectively, our findings shed light on other protein functions – besides amyloidogenicity – where discordance may be of importance, including protein-ligand interactions and viral replication.

Testing for enrichment of discordant and chameleon segments in amyloid proteins has revealed, contrary to our expectations, that these characteristics are poor predictors of amyloidogenic segments. When 'discordance' was first proposed by *Kallberg et al*. **(8)**, the authors proposed that discordant segments are associated with amyloid fibril formation, but no significant statistical relationship between the two was discerned. Interestingly, many more publications have since emerged involving sequence analysis of discordant segments and experimental analyses of 'discordant' proteins from the Kallberg study, such as the Alzheimer Aβ peptide and lung surfactant proteins **(28-30)**. However, none of the subsequent publications had rigorously tested for a statistical relationship between amyloid proteins and discordance, despite the increasing availability of protein domains and continuous identification of new amyloid sequences. Since the pioneering study of Kallberg **(8)**, this is the first study to discern a statistical relationship between discordance and amyloidogenicity, to determine whether such segments are truly associated with amyloids. Our analysis raised a couple of important points, mainly that prediction of discordance is heavily dependent on the prediction algorithm. A decade after its publication, and with the current protein databases more than triple in size than the initial discordance study, our study indicates that discordant segments are only "moderately enriched" in amyloid proteins. Although our results, using the consensus predictions from both GOR and PSIPRED, are significant ($p = 0.027$), they are only "moderate" in comparison to the "high" enrichment observed using either of the tools separately, such as GOR ($p<0.000001$). The GOR-IV algorithm works through considering all possible residue pair frequencies in a sliding window of 17 residues in length; it thus just considers the local sequence environment in a basic way; in contrast, the neural-network based PSIPRED method is a 'black-box' machine-learning technique **(31; 32)**. We opted for a consensus approach to increase stringency of our selection criteria, and prevent bias by the use of only one prediction algorithm. Comparing discordance predictions using GOR with other state-of-the-art algorithms is an interesting point for future research.

With respect to chameleon segments, to our knowledge, this is the first study that attempts to derive a statistical relationship between chameleons and amyloidogenicity. An initial analysis of chameleons in amyloid proteins and their determinants indicated a

significant enrichment between frustrated and *very frustrated* chameleons and amyloid determinants, but a more rigorous analysis (removing overlapping chameleon segments) severely limits the classes of chameleons that still share a significant association **(Table 2.5)**. Hexapeptide frustrated chameleons are the only class of chameleons that remains enriched in these segments, but that significance is marginal (p= 0.04). Taken collectively, the paucity of chameleons and frustrated chameleons in observed amyloid proteins and their determinants, and their poor or marginal enrichment in these proteins, suggests that chameleons are not reliable predictors of amyloidogenic segments.

It was interesting to discover a significant enrichment of chameleons in discordant proteins **(Table 2.2)**, even after the removal of overlapping chameleons **(Table 2.5)**. This suggested a clear link between the phenomena and raised the question of how sequences sharing both characteristics may play a role in amyloidosis. Although our observations indicate that discordant and chameleon sequences, *taken separately*, are not reliable predictors of amyloidogenic segments, we found that 32% of the sequences with *both* phenomena may be prone to amyloidogenicity. These discordant-chameleonic segments and their proteins (identified in **Table 2.6**) include already known amyloidogenic proteins such as PrP. Prion proteins (PrP) are responsible for a variety of neurodegenerative prion diseases, including human Creutzfeldt-Jakob Disease (CJD), sheep scrapie, and Bovine Spongiform Encephalopathy (BSE) in cattle **(1; 33)**. Notably, our analysis of a conserved discordant segment in PrP helix 2 within mammals is further supported by genome-wide analysis **(34)** and MD simulations **(34; 35)** of the PrP helices, which indicated conformational instability in the second half of helix 2, and a drastic decrease of α-helical content accompanied by an increase of β-strands during transition of PrP from its cellular form (PrP^{C}) to its pathogenic, aggregated form (PrP^{Sc}). Interestingly, our analysis of ambiguous encoding in the prion family also identified a discordant-chameleonic segment in its paralog, Doppel, which may suggest an evolutionary importance for discordant-chameleonic segments. Notably, this segment in Doppel was the only identified chameleon segment for that protein, but its relationship to discordance had not been previously elucidated. The discordant-chameleonic segments we have identified in **Table 2.6** also include other proteins, such as HSV Glycoprotein D (1jma:A), whose homologs are already involved in amyloidogenicity. HSV1 is a member of the herpes

virus and is proposed as a strong risk factor for Alzheimer's disease, which is primarily characterized by Amyloid Beta (Aβ) amyloid plaque formation in the brain. Neuronal and glial cells infected with HSV1 led to the increased production and rise in intracellular levels of Aβ amyloid protein, and Aβ amyloid plaques have been observed in mouse brains after HSV1 infection **(36)**. Indeed, homology has been observed between the carboxyl-terminal region of the Aβ peptide with an internal sequence of HSV1 Glycoprotein B (gB), subsequently shown to form B-pleated sheets, self-assemble into fibrils, and accelerate Aβ fibril formation in vitro **(37)**. In HSV1, gB is responsible for attaching the HSV protein to the cell surface, whereas glycoprotein D (gD) facilitates binding of the virus to cell surface receptors. Our discovery of discordance in HSV Glycoprotein D suggests that, like Glycoprotein B, HSV may contain several discordant proteins that facilitate its viral entry and ultimately contribute to amyloidosis. Experimental analyses of HSV Glycoprotein D and other interesting candidates from **Table 2.6** would we required to verify their role in amyloidosis, and to shed light on how the combined discordant-chameleonic effect may play a role in amyloid formation.

Although this study has focused on discordant and chameleonic segments, and their potential for amyloidogenicity based on secondary structure predictions and sequence analysis, it is worth noting that energy barriers also influence their potential for amyloidogenicity. As has been demonstrated for the VGSN peptide in Aβ **(38)**, overcoming the large energy barriers of peptide interactions must first happen before aggregate structures can be formed. One aspect for future research, which is beyond the scope of this study, would be to analyze and compare the peptide interactions and energy barriers of discordant segments, chameleon segments, and discordant-chameleonic segments.

We have performed a meta-scale analysis of chameleon and discordant stretches in protein domains and amyloidogenic proteins and their determinants, to understand the extent to which these segments contribute to conformational flexibility in proteins, as well as their relationship to amyloid formation. From our analysis of discordant stretches in protein domains, we propose several protein functions where conformationally variable segments may play a strong role. Our analysis of discordant stretches in amyloidogenic proteins and their determinants indicates an enrichment of discordant

stretches in amyloid determinants, but this enrichment is dependent on the prediction algorithm. To our knowledge, this is the first study to also address the statistical relationship between chameleons and amyloids, and after alleviating sources of potential bias, we conclude that chameleons are not reliable predictors of amyloidogenic segments. We have however uncovered interesting exceptions where a combination of discordant-chameleonic segments may be heavily involved in amyloidosis, but further experimental analysis would be required to develop an understanding of how they contribute to amyloid formation.

2.5 MATERIALS AND METHODS

2.5.1 Protein Data Sets

We explored the distribution of discordant α–helices and chameleon sequences in three separate cohorts of protein sequence data: *(i)* single-chain protein domains from the SCOP database*, (ii)* known amyloid-forming proteins, and *(iii)* the Prion Protein (PrP) protein family.

(i) SCOP protein domains:

We downloaded a non-redundant set of 'genetic' single-chain domain protein sequences (n=4263) from ASTRALSCOP **(39)**, based on PDB SEQRES records (release 1.73, astral-scopdom-seqres-gd-all-1.73). This was the non-redundant set made such that all sequences in it have pairwise sequence similarity ≤40%).

Also, we derived a data of single protein chains from the entire DSSP database **(40)**, also with a 40% sequence identity threshold (n=10940).

(ii) Known amyloid-forming proteins:

We identified pathogenic and non-pathogenic amyloid-forming proteins through cross-referencing the current literature with the UniProt database **(16-18; 20; 41)**. In Uniprot, 59 identifiers were mapped to 1346 PDB structures. Of these, we selected a non-

redundant set of PDB sequences with less than 40% identity, using the PISCES procedure (n=77) **(Supplementary Table S2.5)** (21).

Amyloidogenic determinants are defined as the union of overlapping subsequences within amyloidogenic protein chains that have been experimentally determined to be amyloidogenic. We determined a non-redundant list of 45 amyloidogenic determinants from the current literature **(Supplementary Table S2.6)** **(16; 20)**; of these, 17 determinants had an α−helical sequence of ≥5 residues.

(iii) The Prion Protein (PrP) protein family:

We selected representative prion and doppel three-dimensional structures from the PDB and SUPERFAMILY databases **(42)**, ignoring mutant or engineered models. For species with multiple structures, we selected one structure per species as designated by SUPERFAMILY. A total of 16 PDB structures were selected, comprising 2 Doppels (Human: 1lg4; Mouse: 1i17) and 14 Prion Proteins (Mouse: 1xyx; Hamster: 1b10; Human: 1i4m; Sheep: 1y2s; Elk: 1xyw; Chicken: 1u3m; Turtle: 1u5l; Frog: 1xu0; Cat: 1xyj; Dog: 1xyk; Pig: 1xyq; Bovine: 1dx0; Bank Vole: 2k56; Rabbit: 2fj3). We selected from Genbank additional prion sequences for which a three-dimensional structure was not available. A total of 24 such sequences were selected (Fruit Bat: gi|27733840; Eurasian bat: gi|27733844; Silky Anteater: gi|27733872; Nine-banded Armadillo: gi|202071082; Asiatic elephant: gi|27733858; African Bush elephant: gi|182636942; Sunda flying lemur: gi|27733816; Large tree shrew: gi|27733818; Aardvark: gi|27733864; Elephant shrew: gi|27733866; Carribean manatee: gi|27733860; Platypus: gi|171473244; Hottentot Golden mole: gi|27733870; Short-tailed opossum: gi|91680539; Tammar Wallaby: gi|49618779; Sperm whale: gi|27733856; Zebrafish: gi|45387601; Zebra finch: gi|123303169; Bottle-nosed dolphin: gi|61743503; Hippopotamus: gi|27733854; Roe Deer: gi|50442322; Domestic goat: gi|119489906; Chimpanzee: gi|56122310; Orangutan: gi|474369).

2.5.2 Experimentally Determined and Predicted Secondary Structures

Secondary structures assignments of three-dimensional protein structures were extracted from the Dictionary of Secondary Structure of Proteins (DSSP) **(40)**. DSSP

defines eight classes of secondary structures based on hydrogen bond patterns: α-helix (H), 3_{10}-helix (G), π-helix(I), extended strand (E), isolated β-bridge (B), hydrogen bonded turn (T), bend (S), and coil (_). To facilitate comparison of DSSP with secondary structure prediction tools, these classes were reduced to three states (Helix, Strand, and Loop), as in previous analyses **(8)**. Secondary structures predictions on selected proteins were performed using the GOR-IV **(43)** and PSIPRED **(44)** algorithms, with default parameters. Both programs employ a three-state classification of secondary structures ('Helix', 'Strand', and 'Loop' (or 'Coil')).

2.5.3 Identification of Discordant Stretches

A schematic of the definition of discordant α-helices is illustrated in **(Figure 2.3, section A)**. Discordant α-helices were identified if discordance was observed between DSSP secondary-structure assignments and: *(i)* the GOR-IV secondary-structure prediction algorithm alone; *(ii)* the PSIPRED secondary-structure algorithm alone; *(iii)* the consensus of the GOR-IV and PSIPRED secondary structure prediction algorithms **(8)**. Discordant stretches ≥5 residues were selected for further statistical analysis.

2.5.4 Structural and Functional Analysis of Discordant Proteins

Classification of protein architectures was determined using CATH **(45)**. Over-representation of CATH architectures was calculated using hypergeometric probability. Molecular functions and biological processes were determined using the Gene Ontology (GO) **(46)** database, and functional relationships between GO terms and protein sequence sets were mapped using GOLEM **(47)** and DAVID **(48)**.

2.5.5 Identification of Chameleon Sequences

Using an in-house script, we identified α–β chameleon sequences by searching for the same 5-mer and 6-mer protein sequence words in both α–helices and β–strands from DSSP secondary structure assignments **(40)**, in SCOP protein domains **(39)** **(Figure**

2.3, section B). NetCSSP **(49)** and ChamSequence Finder **(50)** were also used to identify additional cases.

2.5.6 Calculation of secondary structure propensities

Conformational preferences for α–helix or β–strand for any protein subsequence was determined by averaging the respective amino acid propensities of the whole subsequence using the physiochemical scales of Chou and Fasman **(51)**. For chameleon sequences, we defined *'frustrated chameleons'* as chameleon sequences with higher β than α propensity, and '*very frustrated chameleons*' as chameleon sequences with very high β propensity values (operationally, with a ≥1.5-fold occurrence of beta propensity over alpha propensity) **(Figure 2.3, section C)**.

2.5.7 Prediction of amyloid fibrillogenicity

We used the algorithm Pafig **(22)** to identify sequence segments that predict as fibril-forming hexapeptides. Pafig uses a support vector machine (SVM) to identify fibril-forming hexapeptides based on a classifier of 41 physiochemical properties of amino acids, with an overall prediction accuracy (Q2) of 81% and a Matthews correlation coefficient of 0.63. Predictions generated includes a reliability index (RI) based on the output of the SVM. Fibril-forming segments are defined as having RIs ≥5 (out of 10).

2.5.8 Evolutionary Conservation

Evolutionary conservation was analyzed using the tool Consurf **(24)**. Data Residue Variety for each of the amino-acid positions in each segment was noted, as and the greatest and least occurring amino acid per position in a multiple sequence alignment. Evolutionary importance of conformationally-variable segments was also determined using the Evolutionary Trace Report Maker **(25)** and Evolutionary Trace Viewer **(26)**.

2.6 REFERENCES

1. Aguzzi A, Sigurdson C, Heikenwaelder M (2008) Molecular Mechanisms of Prion Pathogenesis. Annual Review of Pathology: Mechanisms of Disease 3:11-40. PMID: 18233951 %U http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.pathmechdis.3.121806.154326 {Medline}
2. Caughey B, Lansbury PT (2003) PROTOFIBRILS, PORES, FIBRILS, AND NEURODEGENERATION: Separating the Responsible Protein Aggregates from The Innocent Bystanders*. Annual Review of Neuroscience 26:267-298.
3. Chiti F, Dobson CM (2006) Protein Misfolding, Functional Amyloid, and Human Disease. Annual Review of Biochemistry 75:333-366.
4. Kajava AV, Steven AC, Andrey Kajava JMS, David ADP. [beta][hyphen (true graphic)]Rolls, [beta][hyphen (true graphic)]Helices, and Other [beta][hyphen (true graphic)]Solenoid Proteins. (2006) Advances in Protein Chemistry. Academic Press, pp. 55-96.
5. Kajava AV, Squire JM, Parry DAD, Andrey Kajava JMS, David ADP. [beta][hyphen (true graphic)]Structures in Fibrous Proteins. (2006) Advances in Protein Chemistry. Academic Press, pp. 1-15.
6. Nelson R, Sawaya MR, Balbirnie M, Madsen AO, Riekel C, Grothe R, Eisenberg D (2005) Structure of the cross-[beta] spine of amyloid-like fibrils. Nature 435:773-778.
7. Dima RI, Thirumalai D (2002) Exploring the Propensities of Helices in PrPC to Form [beta] Sheet Using NMR Structures and Sequence Alignments. Biophysical Journal 83:1268-1280.
8. Kallberg Y, Gustafsson M, Persson B, Thyberg J, Johansson J (2000) Prediction of amyloid fibril-forming proteins. J Biol Chem:M010402200.
9. Bruce IC, Scott RP, Fred EC (1993) Origins of structural diversity within sequentially identical hexapeptides. Protein Science 2:2134-2145.
10. Igor BK, Shalom R (2004) Comparative computational analysis of prion proteins reveals two fragments with unusual structural properties and a pattern of increase in hydrophobicity associated with disease-promoting mutations. Protein Science 13:3230-3244.
11. Jun-Tao G, Jerzy WJ, Ying X (2007) Analysis of chameleon sequences and their implications in biological processes. Proteins: Structure, Function, and Bioinformatics 67:548-558.
12. Mezei M (1998) Chameleon sequences in the PDB. Protein Eng 11:411-414.
13. Minor DL, Kim PS (1996) Context-dependent secondary structure formation of a designed protein sequence. Nature 380:730-734.
14. Kabsch W, Sander C (1984) ON THE USE OF SEQUENCE HOMOLOGIES TO PREDICT PROTEIN-STRUCTURE - IDENTICAL PENTAPEPTIDES CAN HAVE COMPLETELY DIFFERENT CONFORMATIONS. Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences 81:1075-1078. PMID: ISI:A1984SH73800021 {Medline}

15. Xianghong Z, Frank A, Gerd F, Gaston HG, Gareth C (2000) An analysis of the helix-to-strand transition between peptides with identical sequence. Proteins: Structure, Function, and Genetics 41:248-256.
16. Harrison RS, Sharpe PC, Singh Y, Fairlie DP. Amyloid peptides and proteins in review. (2007) Reviews of Physiology, Biochemistry and Pharmacology, Vol 159. Springer-Verlag Berlin, Berlin, pp. 1-77.
17. Uversky VN (2008) Amyloidogenesis of natively unfolded proteins. Curr Alzheimer Res 5:260-287. PMID: ISI:000256569700005 {Medline}
18. Uversky VN, Fink AL (2004) Conformational constraints for amyloid fibrillation: the importance of being unfolded. BBA-Proteins Proteomics 1698:131-153. PMID: ISI:000221371200001 {Medline}
19. Alberti S, Halfmann R, King O, Kapila A, Lindquist S (2009) A Systematic Survey Identifies Prions and Illuminates Sequence Features of Prionogenic Proteins. Cell 137:146-158.
20. Susan T, Andrew JD Amyloidogenic sequences in native protein structures. Protein Science 19:327-348.
21. Wang G, Dunbrack RL, Jr. (2003) PISCES: a protein sequence culling server. Bioinformatics 19:1589-1591.
22. Tian J, Wu N, Guo J, Fan Y (2009) Prediction of amyloid fibril-forming segments based on a support vector machine. BMC Bioinformatics 10:S45. PMID: doi:10.1186/1471-2105-10-S1-S45 {Medline}
23. Harrison PM, Khachane A, Kumar M Genomic assessment of the evolution of the prion protein gene family in vertebrates. Genomics 95:268-277.
24. Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucl Acids Res 33:W299-302.
25. Mihalek I, Res I, Lichtarge O (2006) Evolutionary trace report_maker: a new type of service for comparative analysis of proteins. Bioinformatics 22:1656-1657. PMID: ISI:000238905700016 {Medline}
26. Morgan DH, Kristensen DM, Mittelman D, Lichtarge O (2006) ET viewer: an application for predicting and visualizing functional sites in protein structures. Bioinformatics 22:2049-2050. PMID: ISI:000239900200018 {Medline}
27. Riek R, Hornemann S, Wider G, Billeter M, Glockshuber R, Wuthrich K (1996) NMR structure of the mouse prion protein domain PrP(121-231). Nature 382:180-182.
28. Johansson J (2001) Membrane properties and amyloid fibril formation of lung surfactant protein. Biochem Soc Trans 29:601-606. PMID: ISI:000170780000046 {Medline}
29. Li J, Hosia W, Hamvas A, Thyberg J, Jornvall H, Weaver TE, Johansson J (2004) The N-terminal propeptide of lung surfactant protein C is necessary for biosynthesis and prevents unfolding of a metastable alpha-helix. Journal of Molecular Biology 338:857-862. PMID: ISI:000221305200001 {Medline}
30. Paivio A, Nordling E, Kallberg Y, Thyberg J, Johansson J (2004) Stabilization of discordant helices in amyloid fibril-forming proteins. Protein Science 13:1251-1259. PMID: ISI:000221042200008 {Medline}

31. Garnier J, Gibrat J-F, Robson B, Russell FD. [32] GOR method for predicting protein secondary structure from amino acid sequence. (1996) Methods in Enzymology. Academic Press, pp. 540-553.
32. King RD, Sternberg MJ (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Science 5:2298-2310.
33. Prusiner SB (1998) Prions. Proc Natl Acad Sci USA 95:13363 - 13383. PMID: doi:10.1073/pnas.95.23.13363 {Medline}
34. Dima RI, Thirumalai D (2004) Probing the instabilities in the dynamics of helical fragments from mouse PrPC. Proceedings of the National Academy of Sciences of the United States of America 101:15335-15340.
35. Bae S-H, Legname G, Serban A, Prusiner SB, Wright PE, Dyson HJ (2009) Prion Proteins with Pathogenic and Protective Mutations Show Similar Structure and Dynamics. Biochemistry 48:8120-8128.
36. Wozniak MA, Itzhaki RF, Shipley SJ, Dobson CB (2007) Herpes simplex virus infection causes cellular beta-amyloid accumulation and secretase upregulation. Neurosci Lett 429:95-100. PMID: ISI:000251869600005 {Medline}
37. Cribbs DH, Azizeh BY, Cotman CW, LaFerla FM (2000) Fibril formation and neurotoxicity by a herpes simplex virus glycoprotein B fragment with homology to the Alzheimer's A beta peptide. Biochemistry 39:5988-5994. PMID: ISI:000087399700008 {Medline}
38. Tarus B, Straub JE, Thirumalai D (2006) Dynamics of Asp23âˆ'Lys28 Salt-Bridge Formation in AÎ²10-35 Monomers. Journal of the American Chemical Society 128:16159-16168.
39. Chandonia J-M, Hon G, Walker NS, Lo Conte L, Koehl P, Levitt M, Brenner SE (2004) The ASTRAL Compendium in 2004. Nucl Acids Res 32:D189-192.
40. Kabsch W, Sander C (1983) DICTIONARY OF PROTEIN SECONDARY STRUCTURE - PATTERN-RECOGNITION OF HYDROGEN-BONDED AND GEOMETRICAL FEATURES. Biopolymers 22:2577-2637. PMID: ISI:A1983RV60400010 {Medline}
41. The UniProt C The Universal Protein Resource (UniProt) in 2010. Nucl Acids Res 38:D142-148.
42. Gough J, Chothia C (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucl Acids Res 30:268-272.
43. Sen TZ, Jernigan RL, Garnier J, Kloczkowski A (2005) GOR V server for protein secondary structure prediction. Bioinformatics 21:2787-2788.
44. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology 292:195-202.
45. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH - a hierarchic classification of protein domain structures. Structure 5:1093-1108. PMID: ISI:A1997XV93200010 {Medline}
46. Gene Ontology C (2004) The Gene Ontology (GO) database and informatics resource. Nucl Acids Res 32:D258-261.

47. Sealfon R, Hibbs M, Huttenhower C, Myers C, Troyanskaya O (2006) GOLEM: an interactive graph-based gene-ontology navigation and analysis tool. BMC Bioinformatics 7:443. PMID: doi:10.1186/1471-2105-7-443 {Medline}
48. Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protocols 4:44-57.
49. Kim C, Choi J, Lee SJ, Welsh WJ, Yoon S (2009) NetCSSP: web application for predicting chameleon sequences and amyloid fibril formation. Nucl Acids Res 37:W469-473.
50. Yoon S, Jung H (2006) Analysis of chameleon sequences by energy decomposition on a pairwise per-residue basis. Protein J 25:361-368. PMID: ISI:000241162900007 {Medline}
51. Chou PY, Fasman GD (1974) CONFORMATIONAL PARAMETERS FOR AMINO-ACIDS IN HELICAL, BETA-SHEET, AND RANDOM COIL REGIONS CALCULATED FROM PROTEINS. Biochemistry 13:211-222. PMID: ISI:A1974R818500001 {Medline}
52. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215:403-410.

2.7 SUPPLMENTAL DATA

Supplemental data includes 6 tables and 2 figures that can be found in **Appendix A**, as well as online with this article at: http://onlinelibrary.wiley.com/doi/10.1002/pro.590/full

CHAPTER III

The Landscape of the Prion Protein's Structural Response to Mutation Revealed by Principal Component Analysis of Multiple NMR Ensembles

The recent application of multivariate analytical methods towards computational modeling of protein folding facilitates the understanding of dominant protein motions underlying key biological functions. This is particularly advantageous for proteins involved in conformational diseases, as identification of protein peptides or domains likely to undergo conformational change during fibrillogenesis can be determined by analyzing molecular motions of protein tertiary structures. In Chapter III, computational modeling of protein flexibility in the Prion family is conducted via a Principal Component Analysis (PCA) on the backbone of prion structures. The presented analysis succeeds in identifying protein segments that are likely to undergo a conformational change during the PrP conversion process, and thus likely involved in fibril formation in these proteins. Interestingly, the identified domains facilitate the differentiation of PrP structures based on non-local structural response to pathogenic mutation and prion disease susceptibility. The novelty of this approach with respect to the prion family is discussed, as well as the potential for adapting this approach towards structural identification of conformationally-variable segments in other amyloid-forming proteins.

A version of this chapter has been submitted as:
Gendoo, D. M. and Harrison, P. M. (2011), The landscape of the Prion Protein structural response to mutation revealed by PCA analysis of multiple NMR ensembles.
Submitted to Plos Computational Biology

3.1 ABSTRACT

Prion Proteins (PrP) are among a small number of proteins for which large numbers of NMR ensembles have been resolved for sequence mutants and diverse species. Here, we perform a comprehensive principle components analysis (PCA) on the tertiary structures of PrP globular proteins to discern PrP subdomains that exhibit conformational change in response to point mutations and clade-specific evolutionary sequence mutation trends. This is to our knowledge the first such large-scale analysis of multiple NMR ensembles of protein structures, and the first study of its kind for PrPs. We conducted PCA on human (n=11), mouse (n=14), and wildtype (n=21) sets of PrP globular structures, from which we identified five conformationally variable subdomains within PrP. PCA shows that different non-local patterns and rankings of variable subdomains arise for different pathogenic mutants. These subdomains may thus be key areas for initiating PrP conversion during disease. Furthermore, we have observed the conformational clustering of divergent TSE-non-susceptible species pairs; these non-phylogenetic clusterings indicate structural solutions towards TSE resistance that do not necessarily coincide with evolutionary divergence. We discuss the novelty of our approach and the importance of PrP subdomains in structural conversion during disease.

3.2 AUTHOR SUMMARY

Prion Proteins (PrP) cause of variety of incurable TSE diseases, and are among a small number of proteins for which large numbers of NMR ensembles have been resolved for sequence mutants and diverse species. Here, we perform a comprehensive principle components analysis (PCA) study to assess conformational variation and discern the landscape of the PrP structural response to sequence mutation. This is to our knowledge the first such large-scale analysis of multiple NMR ensembles for a specific protein, and the first study to perform a multivariate PCA on the native globular structures of PrP. We conducted exhaustive PCA on three subsets of PrP, human and mouse PrP subsets that include structures of sequence mutants, and the set of wild-type PrP globular proteins (representing 16 PrP species). PCA shows that different non-local patterns of variable subdomains arise for different pathogenic mutants. These subdomains may thus be key areas for initiating PrP conversion during disease. Furthermore, we have observed that some evolutionarily divergent species that are not susceptible to prion diseases, have surprising structural similarities in their prion proteins. We discuss the novelty of our approach with respect to the prion protein, and the advantage of this analysis as a fast, reliable starting point to identify domains of interest that may warrant further experimental and computational analysis.

3.3 INTRODUCTION

The extraordinary conformational change witnessed between the normal, non-pathological prion protein, PrP^{C}, and its virulent pathological form, PrP^{SC}, in which the latter acquires substantial β-sheet content, is a significant contributor to the role this protein plays as an agent of many incurable Transmission Spongiform Encephalopathies (TSEs). Such diseases, including human Creutzfeldt-Jakob Disease (CJD) and Bovine Spongiform Encephalopathy (BSE), are caused by the misfolding and subsequent aggregation of PrP^{SC} to produce amyloid fibrils, highly ordered and distinct β-sheet-rich molecular aggregates **[1,2]**. The PrP protein is a 208 residue protein (residues 23-230, hPrP numbering) composed of a largely disordered N-terminal tail (23-124) and a C-terminal globular domain (125-231), in addition to two signal peptides (1-23, 232-253) **[3,4]**. The globular domain contains three α-helices **(H1,H2,H3)** and two anti-parallel β-sheets **(S1,S2)**. Globular domains of multiple PrP species have been resolved to develop an understanding of PrP structures in relation to TSE-susceptibility, and discern subdomains of the protein that are involved in the PrP conversion process **[4,5,6,7,8,9]**. The S2-H2 loop and H2-H3 regions, for example, demonstrate structural plasticity in pathogenic PrP and are proposed to be involved in the conversion process, making them candidate sites for transmissibility studies and potential target sites for drug design **[10,11,12,13,14,15]**. The prion protein is one of few proteins with a large number of pathogenic mutants, and the increasing availability of these structures in the protein databank (PDB) provides ample material for a multivariate analysis of structural plasticity of PrP domains.

Principal Component Analysis (PCA) **[16]** is a dimensionality reduction technique that can be used to analyze protein structures by reducing variation observed within 3D atomic coordinates of the protein structures. PCA has been used on several protein families to analyze key regions of interest, including ligand-binding sites and cavities **[17,18]**, receptor sites **[19]**, catalytic subunits **[20]**, as well as large-scale analysis of whole proteins **[21]**. Most interesting is the recent application of PCA towards modeling protein flexibility computationally, and characterizing structural variation of protein domains **[22,23]**. Identifying structural plasticity within protein domains is

especially advantageous for proteins involved in conformational diseases, such as amyloid-forming proteins.

In this work, we perform an exhaustive PCA analysis on the tertiary structures of PrP globular proteins to discern PrP subdomains that exhibit conformational plasticity in response to pathogenic point mutations and clade-specific evolutionary sequence mutation trends; these subdomains may thus be key areas for initiating the conversion of PrP^{C} to PrP^{SC}. To our knowledge, this is the first PCA study on native globular structures of PrP, using NMR ensembles, and without relying on structures generated from protein dynamics methods. We focus our analysis on three subsets of PrP, human and mouse PrP subsets that include structures of sequence mutants, and the set of wild-type PrP globular proteins (representing 16 PrP species). From this analysis, we identify five conformationally variable subdomains of PrP whose relative importance changes for different pathogenic mutations and species groupings. Also, PCA indicates that PrPs exhibit a marked non-phylogenetic clustering, with some notable divergent pairs of species that are non-susceptible to TSEs. We discuss the implications of these results for the conformational basis of TSEs.

3.4 RESULTS

3.4.1 Analysis of Human PrP Proteins

PCA was conducted on the NMR ensembles of 11 human wildtype, variant and mutant prion proteins (230 models in total), to examine major conformational changes between the structures and map them onto a lower (mostly 2-dimensional) space. The resulting eigenvalue contribution of PCA shows that 65% of the total mean-square displacement of atom positional fluctuations was captured in the first three components **(Figure 3.1, section C)**.

Plotting of the hPrP structures onto the two most significant principal components (PC1 and PC2) characterizes conformational relationships between the hPrP structures that are reflective of human prion TSEs. Four major conformational clusters have been observed, of which the largest cluster (encircled in the black oval in **Figure 3.1, section A**) corresponds to PDB structures of WT proteins, as well as hPrP artificial variant structures [PDBs 1E1G, 1E1P, 1E1U, 1H0L] that maintain a similar structure to WT PrPs (mPrP, shPrP) **[26,27]**. For each of the remaining three clusters, each cluster is composed of the models of the NMR ensemble representing the PDB structure of each of the human TSE diseases of GSS (red oval) [PDB 2KUN] **[24]**, FFI (blue oval) [PDB 2K1D], and CJD (green oval) [PDB 1FO7] **[25]** **(Figure 3.1 sections A, B)**. These four clusters, as observed by projection of the hPrP structures onto PC1 and PC2 **(Figure 3.1, section A)**, as well as PC1 and PC3 **(Figure 3.1, section B)**, indicate that these principal component projections facilitate the discrimination of key, pathogenic mutant structures that reflect PrP diseases. Interestingly, such projections also highlight variation between models within an NMR ensemble, as is clearly demonstrated for the structure 2K1D (encircled in blue in **Figure 3.1, section A**), whereby an additional hierarchical cluster is introduced for some models (model numbers 8, 14, 16, 20, encircled in a dashed brown oval in **Figure 3.1, section A**) which cluster further away from the 2K1D ensemble along PC1 **(Figure 3.1 sections A, B)**. This contrasts with other NMR ensembles whose models remain tightly clustered together along the PCs, such as 1FO7 (encircled in green in **Figure 3.1, section A**).

The contribution of each residue in hPrP to each of the first three PCs is displayed, whereby the height of each bar indicates the maximum atomic displacement of each residue for a given PC, and regions of increased displacement highlight structurally variable subdomains in the hPrP structures **(Figure 3.2, sections A, C-E)**. The mutant structure ensembles are separable on the conformer plots **(Figure 3.1)** because of distinct patterns of variable subdomains observed in the residue contribution plot **(Figure 3.2, section A)**. The variable subdomains captured by PC1 include the S2-H2 loop and the C-terminal end of H3. PC2, which contributes to the large separation between the FFI and GSS clusters on the conformer plot (**Figure 3.1, section A**), is characterized by concerted structural variability of the H2-H3 loop, the N-terminus of the globular domain, and S1. The remaining variations captured by PC3 include the S1-H1 loop, and increased displacement of the S2-H2 loop region witnessed in PC1. In total, 5 variable subdomains have been identified: the N-terminal region of the globular domain and S1, the S1-H1 loop, the S2-H2 loop, the H2-H3 loop, and the C terminus of H3 **(Figure 3.2, section B)**. Strikingly, these subdomains of structural variation are not localized to the variant or mutation spots of the protein, which reflects on the nonlocal changes in the protein that are induced by these highly localized substitutions **(Figure 3.2, section B)**.

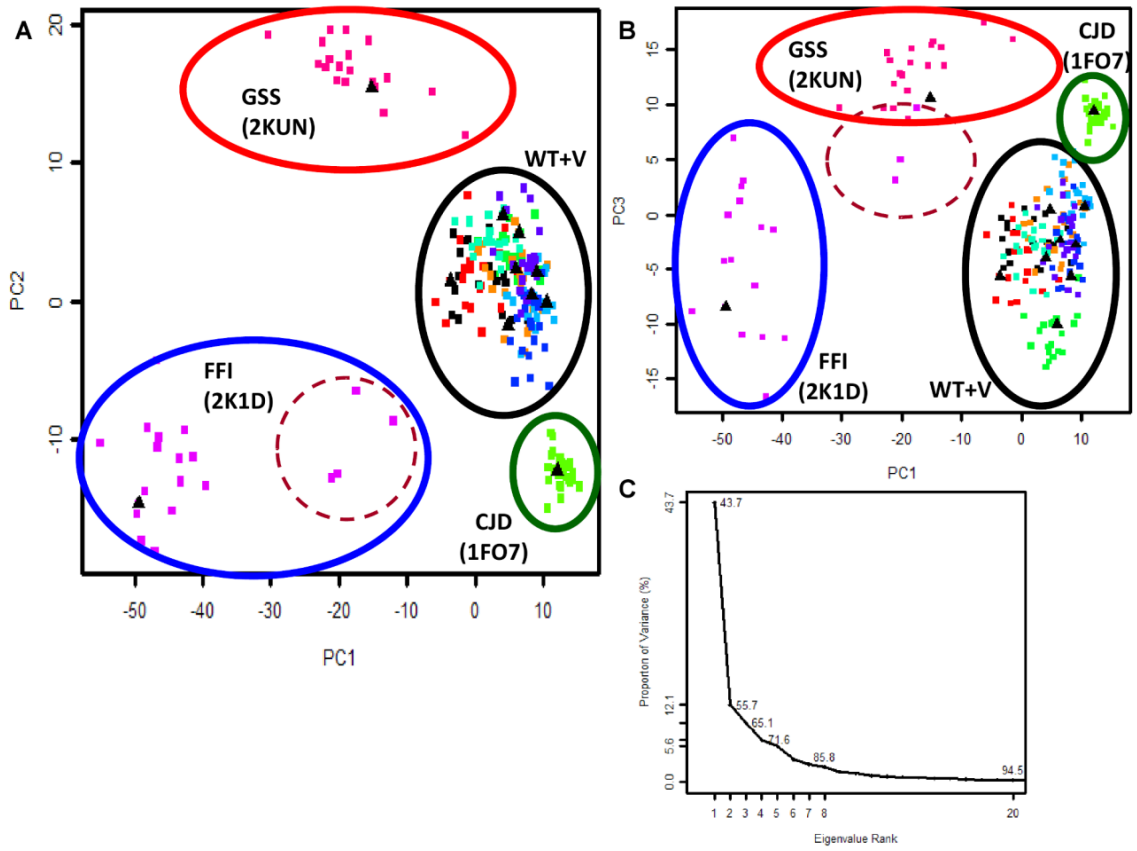


Figure 3.1: PCA analysis on 11 hPrP structures reveal structural perturbations correlated with prion disease.

(A) Projection of hPrP NMR ensembles onto PC1 and PC2.

(B) Projection of hPrP NMR ensembles onto PC1 and PC3.

For (A) and (B), each point on the conformer plot represents an NMR model, and the models are colored to reflect NMR ensembles. For each NMR ensemble, the NMR representative model that has been selected by OLDERADO **[42]** is indicated by a black triangle. Ovals indicate dominant clusters that represent the hPrP diseases of CJD (green oval), FFI (blue oval), GSS (red oval), as well as the set of WT and variant proteins (WT+V, black oval). The ovals representing hPrP disease are also labeled, with the PDB code of their corresponding NMR ensemble in brackets. 2K1D models which cluster separately from the rest of the 2K1D ensemble are circled (dashed brown oval).

(C) Eigenvalue contribution of PCs to variance of the dataset.

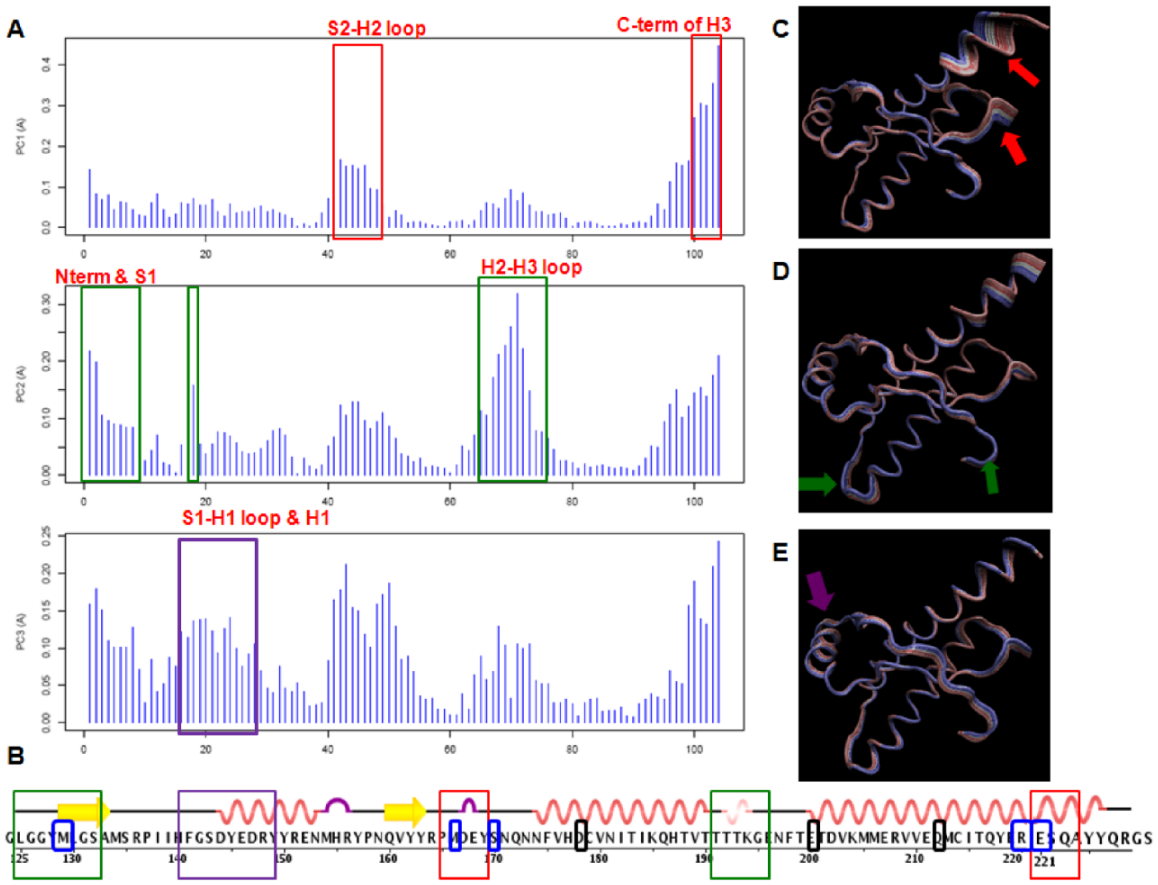


Figure 3.2: PCA analysis results of 11 hPrP structures.

(A) Contribution of each residue of hPrP to the first three principal components. Subdomains of concerted displacement in each PC are indicated by colored boxes and labeled.

(B) Subdomains of concerted displacement in each of the PCs are highlighted against the reference structure 1QLZ (WT hPrP), and color-coded by their first appearance in a PC. From our dataset, pathogenic mutations causing familial disease (D178N, E200K, Q212P, causing FFI, CJD, and GSS, respectively) are indicated (black boxes), as well as nonpathogenic variants (M129V, M166V or M166C, S170N, R220K, E221C)(blue boxes).

(C-E) Structural interpolation of atomic displacements from the mean structure for PC1, PC2, and PC3, respectively (reference structure 1QLZ). Subdomains exhibiting displacement in each PC are indicated by arrows, and the arrows are color-coded to match the boxed subdomains in (A).

For comparison, we also performed a PCA analysis just on the structural variation observed in the WT PrPs (totaling 4 NMR ensembles), while excluding NMR ensembles of mutant and variant PrP structures. The resultant residue contribution plot indicates that all five subdomains of concerted variation contribute to PC1 of the WT dataset **(Figure 3.3, section A)**, implying that they share equal degrees of importance in representing variance between the structures (PC1 captured 30% of the variance of the dataset) **(Figure 3.3, section B)**. Intriguingly, displacement of the H2-H3 loop and the C terminus of H3 are not readily observed in PC2, but are observed in PC3. The lack of additional clustering between the NMR ensembles in PC3, except for the dispersion of models within each NMR ensemble, suggests that these subdomains might play a greater role in discerning conformational changes between models of the NMR ensembles **(not shown)**. Conversely, the N terminus and the S1-H1 loop are readily observed in PC1 and PC2, but not in PC3, showing that these regions play a greater role in separation of the NMR ensembles, instead of inter-model variation.

To check which subdomains vary in a mutant-specific way, we performed three separate analyses, each analysis consisting of the set of WT and variant hPrP structures (encircled by the black oval in **Figure 3.4, section A**) and an NMR ensemble from each of the CJD, FFI, and GSS mutant structures **(Figure 3.4, sections B-D)**. The resultant conformer plots indicate that the pathogenic mutant structures are successfully separated from the WT and non-pathogenic hPrP structures **(Figure 3.4, sections B-D)**. Comparison of residue contribution to each PC indicates that the C-terminus of H3, as well as the S2-H2 loop, differentiate the mutant structures for all analyses, as both subdomains appear in PC1 **(Figure 3.4, sections B-D)**. This observation is reinforced by comparison to the residue contribution plot of the WT, variant, and mutant hPrP structures **(Figure 3.4, section A)**. The remaining subdomains representing the N terminus, S1-H1 loop and H1, and the H2-H3 loop display different levels of importance that are reflective on each of the mutant structures. For example, the H2-H3 loop is strong contributor to conformational separation of the CJD mutant structure, as it appears in PC1 in the residue contribution plot **(Figure 3.4, section B)**, compared to the FFI mutant where it appears in PC3 **(Figure 3.4, section C)**. Similarly, the S1-H1 loop and N terminus of H1 exhibit greater importance in differentiating the GSS mutant structure

(Figure 3.4, section D), as they appear in a later PC for the FFI and CJD structures **(Figure 3.4, sections B-C)**. To ascertain our observations, we calculated the residue difference profile between each of the datasets in **(Figure 3.4, sections B-D)** with hPrP WT and variant dataset (black oval in **Figure 3.4, section A**) for PC1 **(Supplementary Figure S3.1)**. The resultant plots **(Supplementary Figure S3.1)** indicates the residue contribution that is specific to each of the hPrP mutant structures, from which we confirm our observations that the S2-H2 loop exhibits the greatest conformational perturbation for all three mutant structures, and that the H2-H3 loop is clearly important for structural differentiation of the CJD mutant **(Supplementary Figure S3.1)**.

In aggregate, these PCA analyses succeed in delineating and ranking structural subdomains in terms of their relative importance for different pathogenic mutants.

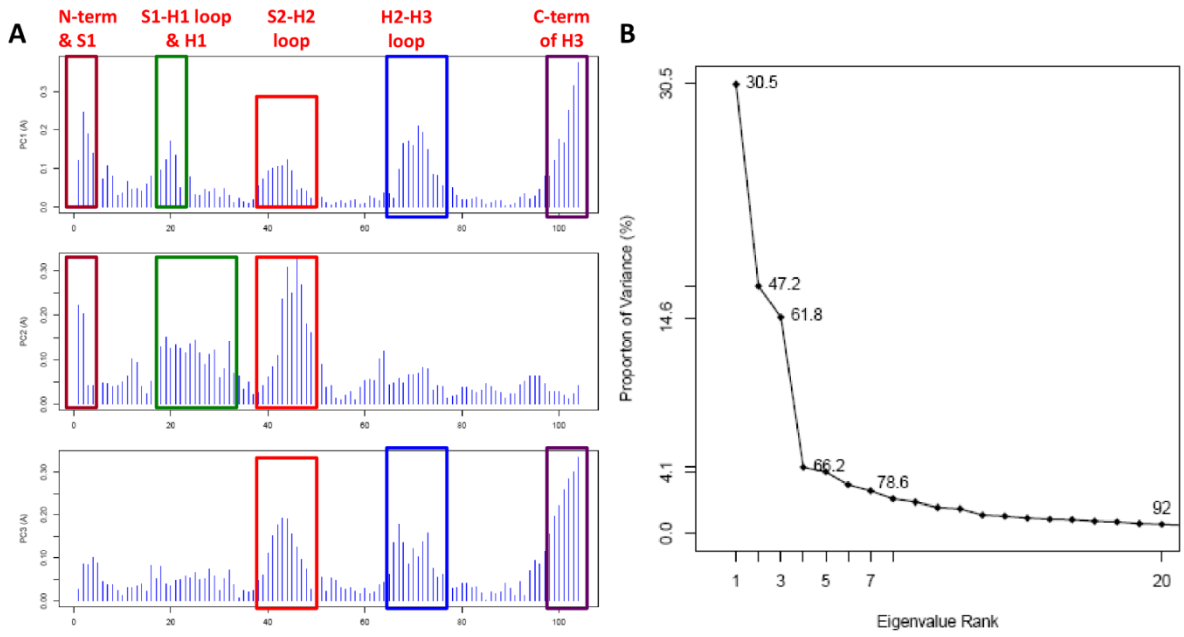


Figure 3.3: PCA analysis of the WT hPrP subset.

(A) Contribution of each residue to the first three principal components (reference structure 1QLZ). Each subdomain of concerted displacement is indicated by a box that is color-coded across all 3 PCs.

(B) Eigenvalue contribution of PCs to variance of the dataset.

Figure 3.4 (next page): Comparative analysis of conformer plots, residue contribution, and structural interpolation of hPrP mutant NMR ensembles structures versus WT and variant hPrP.

Each row of the figure represents one PCA analysis and contains, from left to right, a conformer plot, residue contribution plot, and structural interpolation diagram. An explanation of the conformer plots is provided in **Figure 3.1**. Residue contribution to each PC is color-coded by PC (red=PC1, green=PC2, purple=PC3) in all residue contribution plots. For structural interpolation diagrams, PC1 is represented as equidistant atomic displacements from the mean structure (reference 1QLZ), and corresponding subdomains are indicated (red arrows).

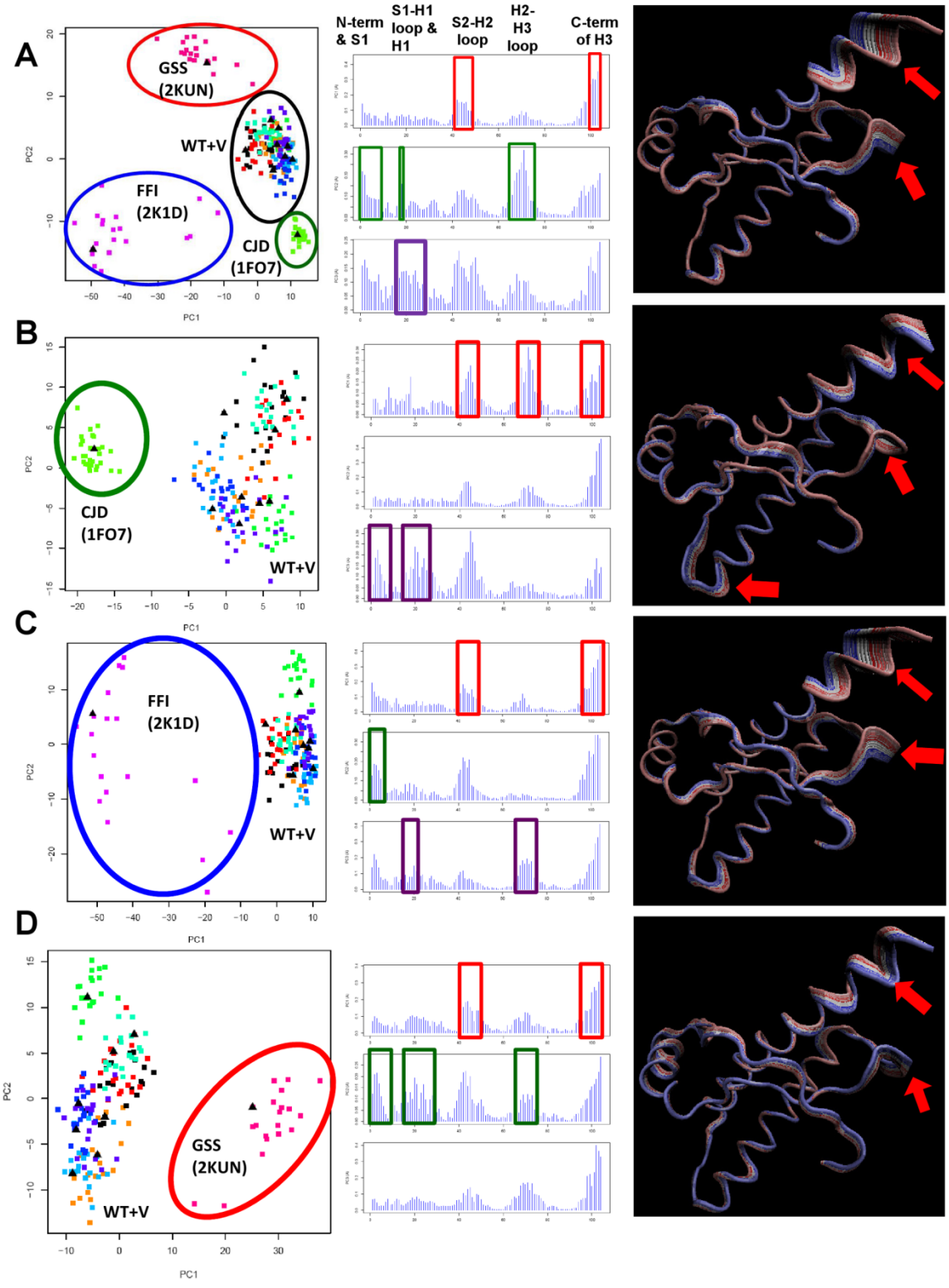
(A) Combined set of WT, variant and mutant hPrP NMR ensembles plotted onto PC1 and PC2. The conformer plot is identical to **Figure 3.1, section A**. In the conformer plot, NMR ensembles of mutant structures are encircled in green, red, and blue ovals and labeled by their corresponding human disease, as well as the PDB code corresponding to the NMR ensemble (in brackets). The set of WT and variant hPrP structures (encircled by the black oval) have been labeled as WT+V.

For rows **(B-D)**, each analysis consists of the set of WT+V and an NMR ensemble from each of the CJD, FFI, and GSS mutant structures, respectively. The NMR ensemble of the mutant structure is encircled by an oval (color-coded to (A)), and labeled by the human disease it represents, and the PDB code corresponding to the NMR ensemble (in brackets).

(B) CJD mutant (PDB 1FO7) and WT+V,

(C) FFI mutant (PDB 2K1D) and WT+V,

(D) GSS mutant (PDB 2KUN) and WT+V.


A
GSS
(2KUN)
WT+V
FFI
(2K1D)
CJD
(1FO7)
PC1
PC2
N-term & S1
S1-H1 loop & H1
S2-H2 loop
H2-H3 loop
C-term of H3
B
CJD
(1FO7)
WT+V
C
FFI
(2K1D)
WT+V
D
GSS
(2KUN)
WT+V

3.4.2 Analysis of Mouse PrP (mPrP) Proteins

We conducted PCA analysis on a set of 14 wildtype, variant, and mutant mouse PrPs NMR ensembles (280 models in total) to examine structural differences between mPrP structures and compare these changes to hPrP. Aside from WT mPrP [PDBs 1XYX, 2L1H, 2L39], 9 PrPs contain mutations in the S2-H2 region (between residues 166-175), and 2 PrP structures [PDBs 2KFM, 2L1K] contain mutations at the C-terminus of H3 (Y255A and Y226A). PCA analysis of mPrP including 2KFM and 2L1K reveals a prominent concerted variation of the C-terminus of H3 that far exceeds any other atomic displacement in the protein, for all three PCs **(Supplementary Figure S3.2)**. One might argue that 2KFM and 2L1K, as the only two structures with conformational differences in H3, are "conformational outliers" that contribute to the displacement of the H3 region in all PCs and overshadow structural differences of the H2-H3 loop. To test this hypothesis, we re-ran the analysis without 2KFM and 2L1K, such that the mPrP dataset consisted only of the WT and variant structures and those with mutations in the S2-H2 loop. Contrary to our expectations, the observed pattern of atomic displacements indicates that the H3 subdomain, in addition to the N terminus of the proteins, remains responsible for conformational variation.

3.4.3 Analysis of Wildtype PrP proteins

PCA was conducted on NMR ensembles of 16 species of WT PrP (21 PDB ensembles corresponding to 420 models in total) **(Figure 3.5)**. Among the species studied, 8 species (mouse, bovine, human, hamster, cat, pig, elk, bank vole) are known to develop TSEs, and 7 species (dog, horse, rabbit, chick, turtle, frog, and wallaby) are "TSE-non-susceptible", taken collectively here to refer to PrP species that are experimentally proven to be resistant to TSEs or for which TSEs remain undetected. In our analysis, sheep is the only species which has been considered in both categories, as sheep with the H168 polymorphism [PDB 1XYU] are TSE-susceptible, but those with the R168 variant [PDB 1Y2S] are highly resistant to disease **[28]**. PCA successfully clusters many of the TSE-non-susceptible species from TSE-susceptible ones, as indicated by the conformer plots **(Figure 3.5, sections A-C)**. PC1 separates chicken (chPrP) and turtle (tPrP) from the rest of the species, such that they form their own

subgroup **(Figure 3.5, section A)**. This is to be expected since they are divergent species evolutionarily. Detailed analysis of residue contribution in this PC indicates that the H2-H3 loop undergoes a significant displacement relative to the rest of the protein **(Figure 3.5, sections G-H)**. However, unexpectedly from an evolutionary point of view, PC2 also contributes to the clustering of the two TSE-non-susceptible species, frog and rabbit **(Figure 3.5, sections A-B)** (when n=3 in hierarchical clustering). Residue contribution to PC2 characterizes the concerted maximum displacement of the S2-H2 loop and the H1 helix **(Figure 3.5, sections G-H)**. With the exception of an additional clustering for pig that is introduced in PC3 **(Figure 3.5, sections B-C)**, analysis of the residue contribution to PC3 does not introduce any newer subdomains than those identified in PC1 or PC2. Thus, the first two PCs are sufficient in describing the range of structural differences between PrP species.

As the H2-H3 loop is longest in chPrP compared to other PrP species **[5,15]**, we wished to assess whether the concerted displacement of the H2-H3 loop in PC1 is the biased result of major conformational differences in chPrP. To this end, we performed a PCA analysis on all the WT PrPs without chPrP **(Figure 3.5, sections D-F)**. Despite the removal of chPrP, the dominant feature described by PC1 remains the displacement of the H2-H3 loop, followed by the displacement of the S2-H2 loop and H1 in PC2 **(not shown)**. Similarly, no additional regions of displacement are witnessed in PC3. With respect to conformational clustering, removal of chPrP has decreased the amount of variation observed in the first 3 PCs (46% compared to 51% with chPrP). Conformational clusters of the dataset without chPrP indicate that the turtle, frog, rabbit, and cat species cluster further away from the TSE-susceptible species **(Figure 3.5, sections D-F)**, and the clustering of the NMR ensemble for pig PrP is also observed in PC3 **(Figure 3.5, sections E-F)**. However, an additional clustering of the sheep resistant R168 polymorphism (PDB 1Y2S) is observed at PC3, while the TSE-susceptible sheep polymorphism H168 (PDB 1XYU) remains closely clustered with the TSE-susceptible PrPs **(Figure 3.5, sections E-F)**. In summary, we demonstrate that our PCA analysis detects major "structural signatures" for PrPs of different evolutionary groups, and highlight PrP subdomains that are worthwhile to explore in TSE-transmissibility studies.

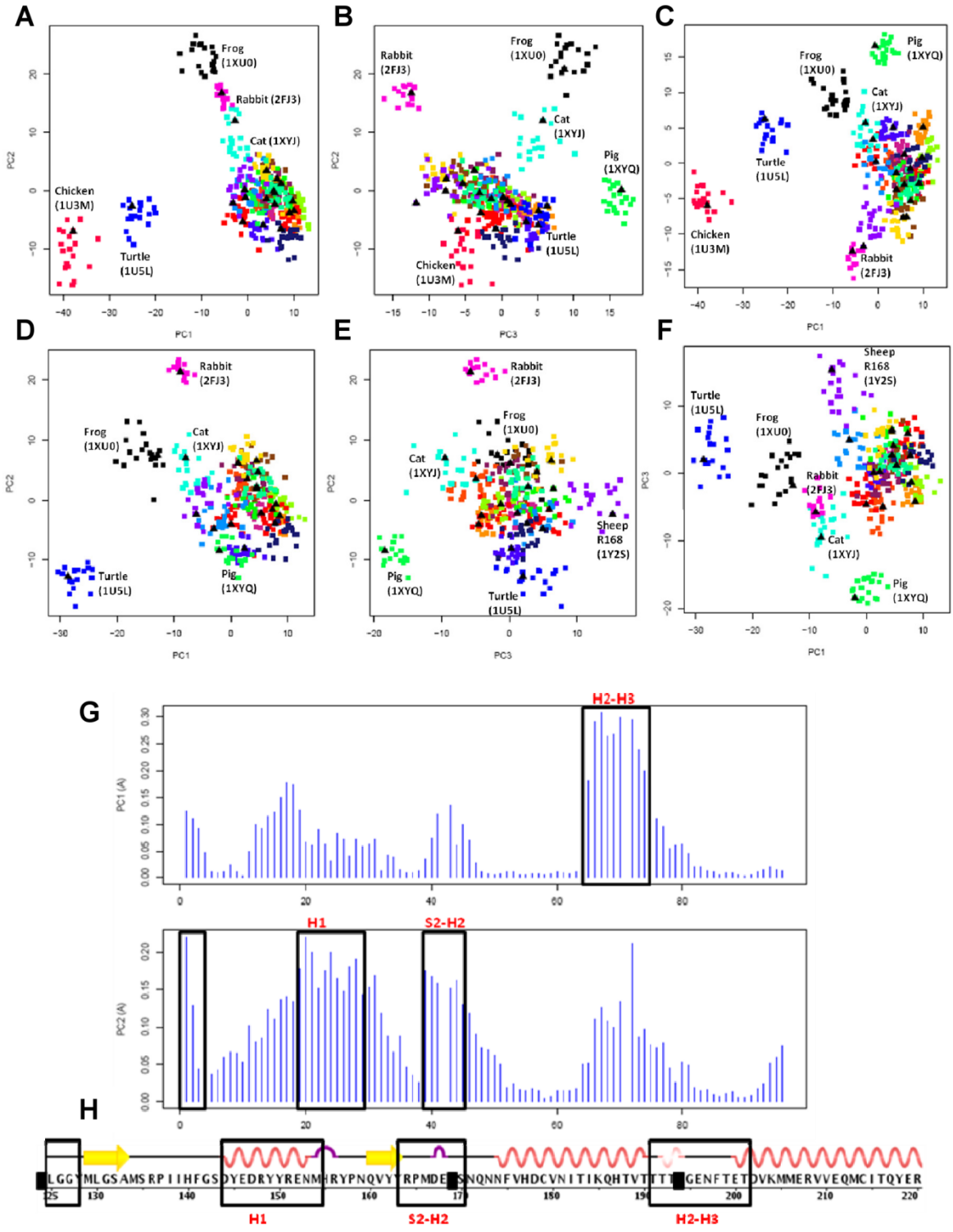

A
Frog (1XU0)
Rabbit (2FJ3)
Cat (1XYJ)
Chicken (1U3M)
Turtle (1U5L)
PC1
PC2
B
Rabbit (2FJ3)
Frog (1XU0)
Cat (1XYJ)
Pig (1XYQ)
Turtle (1U5L)
Chicken (1U3M)
PC3
PC2
C
Pig (1XYQ)
Frog (1XU0)
Cat (1XYJ)
Turtle (1U5L)
Chicken (1U3M)
Rabbit (2FJ3)
PC1
D
Rabbit (2FJ3)
Frog (1XU0)
Cat (1XYJ)
Turtle (1U5L)
Pig (1XYQ)
PC1
PC2
E
Rabbit (2FJ3)
Frog (1XU0)
Cat (1XYJ)
Sheep R168 (1Y2S)
Pig (1XYQ)
Turtle (1U5L)
PC3
PC2
F
Sheep R168 (1Y2S)
Turtle (1U5L)
Frog (1XU0)
Rabbit (2FJ3)
Cat (1XYJ)
Pig (1XYQ)
PC1
PC3
G
H2-H3
PC1 (A)
H1
S2-H2
PC2 (A)
H
H1
S2-H2
H2-H3

Figure 3.5 (previous page): PCA analysis of the 21 NMR ensembles of WT PrP structures.

(A-C) Projection of the structures, including chPrP, onto PCs 1-3

(D-F) Projection of structures, excluding chPrP, onto PCs 1-3.

For **(A-F)**, each point on the conformer plot represents an NMR model, and the models are colored to reflect NMR ensembles. For each NMR ensemble, the NMR representative model that has been selected by OLDERADO **[42]** is indicated by a black triangle. Identifiable clusters of NMR ensembles have been labeled by the species they represent, with the corresponding PDB code of the ensemble in brackets.

(G) Regions of concerted displacement in PC1 and PC2 of the residue contribution plot.

(H) Regions of concerted displacement are labeled (black boxes) onto the primary structure (reference structure 1QLZ (hPrP)). Residues that do not contribute to the core alignment are shaded in black.

3.4.4 Analysis of Mammalian WT PrPs

PCA analyses of the entire WT dataset **(Figure 3.5, sections A-C)** raises the following question: does the structural variation in these analyses reflect upon species evolutionary relationships, and is there discernible clustering that reflects TSE susceptibility and non-susceptibility/resistance? Analysis of WT PrP reveals that distantly-related, non-mammalian species (frog, chicken, and turtle) form separate clusters from the mammalian cluster in the conformer plot **(Figure 3.5, section A)**. To discern the behavior of PrP subdomains in the evolutionary and structural separation of a large subset of closely-related species, we ran PCA on a set of 13 mammalian TSE-non-susceptible and TSE-susceptible PrP NMR ensembles. The resultant conformer plots **(Figure 3.6, section A-C)** show that rabbit and pig PrP structures quickly separate from the remaining PrPs. Analysis of residue contribution to the PCs indicates a different pattern of "subdomain importance" that differentiates between the mammalian PrPs **(Figure 3.7, section A)**, compared to the complete WT species set that includes non-mammalian PrPs **(Figure 3.5, section G)**. The residue contribution plot of the mammalian PrPs **(Figure 3.7, section A)** indicates that the C-terminus of the H3, as opposed to the H2-H3 loop, exhibits the largest atomic displacement in PC1, while the remaining four subdomains appear in PC2 and PC3.

We compared subdomain displacement of the mammalian dataset (n=13 total species) **(Figure 3.7, section A)** to subsets of TSE-non-susceptible mammals (n=5 species, including Sheep R168 variant) **(Figure 3.7, section B)** and TSE-susceptible mammals (n=9 species, including Sheep H168 variant) **(Figure 3.7, section C)**. With respect to the combined set of mammalian and non-mammalian TSE-non-susceptible PrP structures **(presented in Supplementary Figure S3.3, part B)**, removal of the non-mammalian PrPs from that set shifts subdomain importance from the H2-H3 loop **(Supplementary Figure S3.3, part B)** to the C-terminus of H3 in the TSE-non-susceptible mammalian dataset **(Figure 3.7, section B)**, such that the pattern of conformational variation and subdomain importance is similar to the total WT mammalian dataset **(Figure 3.7, section A)**. Notably however, H1 and its flanking loops still exhibit strong displacement at PC2 in both TSE-non-susceptible residue contribution plots **(Figure 3.7, section B, and Supplementary Figure S3.3)**, which suggests that for

all TSE-non-susceptible species, including or excluding non-mammals **(Figure 3.7, section B, and Supplementary Figure S3.3)**, H1 represents a large percentage of conformational variation within that dataset.

It is interesting to note that PCA analysis of mammalian PrPs (n=13), and TSE-non-susceptible mammals (n=5), indicates that TSE-non-susceptible mammals (ex: horse, wallaby, rabbit) exhibit a "structural differentiation", such that they cluster at the periphery of the conformational space away from TSE-susceptible mammals **(Figure 3.6, sections A, D-F)**. This indicates different structural solutions towards resistance that don't necessarily coincide with evolutionary divergence. This is clearly demonstrated by examination of a PC-based cluster dendrogram of all of the 16 PrP NMR ensembles (420 models) under study and of a neighbor-joining tree for the PrP sequences of the 16 species **(Supplementary Figure S3.4)**; horse and wallaby, for example, are closely clustered together in the PC-based dendrogram, even though they are evolutionarily divergent species.

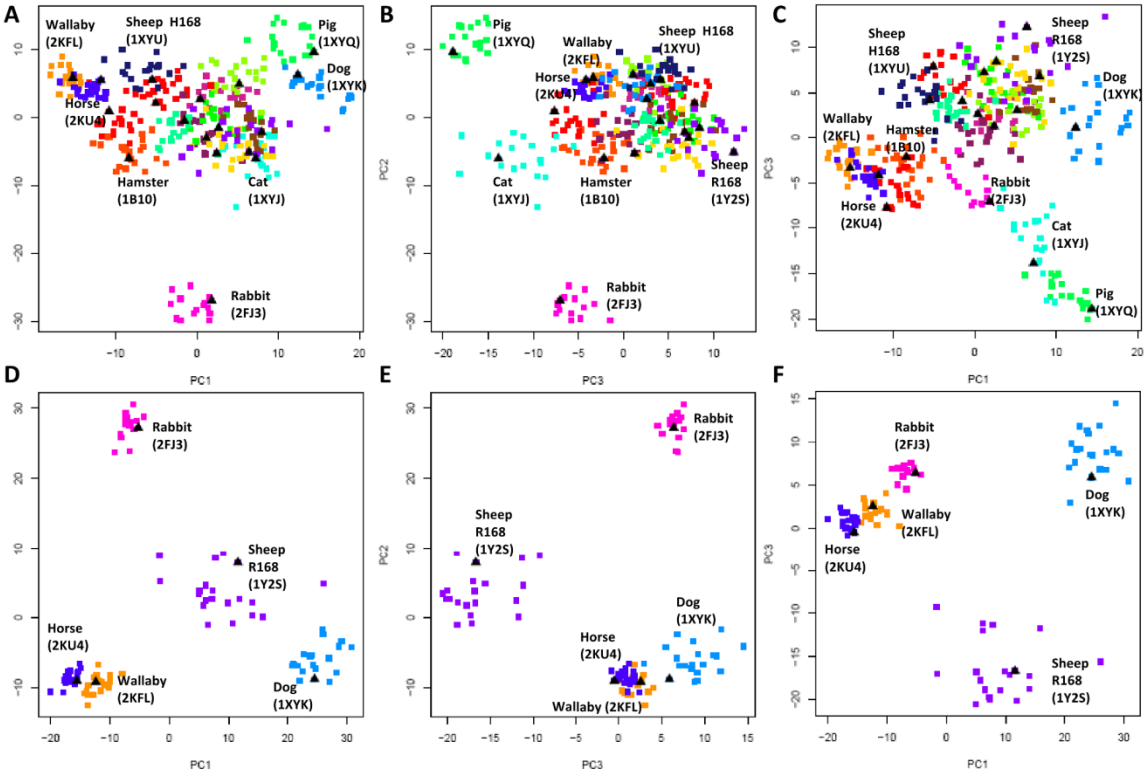


Figure 3.6: Projection of mammalian PrP NMR ensembles onto PCs 1-3.

(A-C) Mammalian PrP structures (n=13 species)

(D-F) TSE-non-susceptible mammals (n=5 species, including Sheep R168 variant).

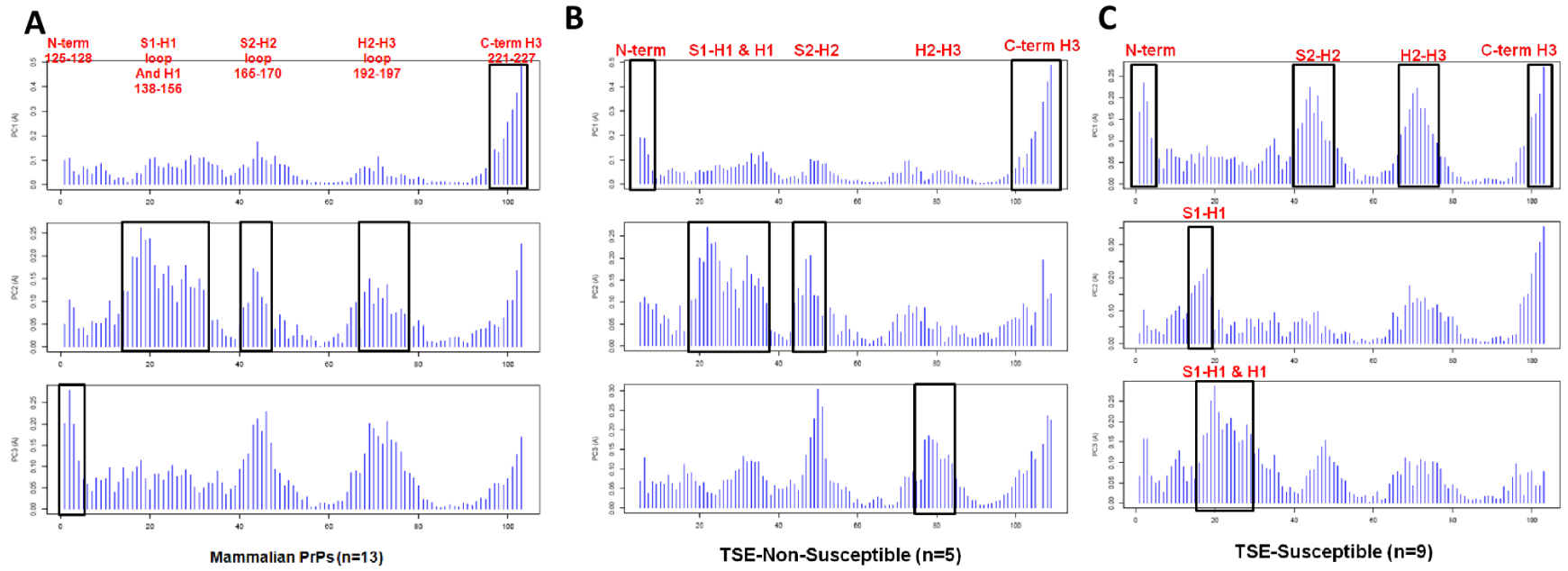


Figure 3.7: Residue contribution to PCs of TSE-non-susceptible, TSE-susceptible, and combined dataset of mammalian PrP. **(A)** The combined mammalian dataset (n=13 species). **(B)** TSE-non-susceptible mammals (n=5 species, including Sheep R168 variant). **(C)** TSE-susceptible mammals (n=9 species, including Sheep H168 variant). Notably, sheep has been included in both species counts, as the sheep polymorphism R168 is non-susceptible, while H168 is susceptible. For all conformer plots, structures are colored by PDB name to reflect NMR ensembles, and identifiable clusters of NMR ensembles have been labeled by the species they represent.

3.4.5 Summary of PCA analyses on PrP datasets

Five subdomains displaying structural plasticity in PrP have been identified in NMR ensembles of hPrP, mPrP, and WT datasets **(Figure 3.8)**. The pattern of concerted displacement of these subdomains for all three PCs, for each of the datasets, is summarized **(Table 3.1)**.

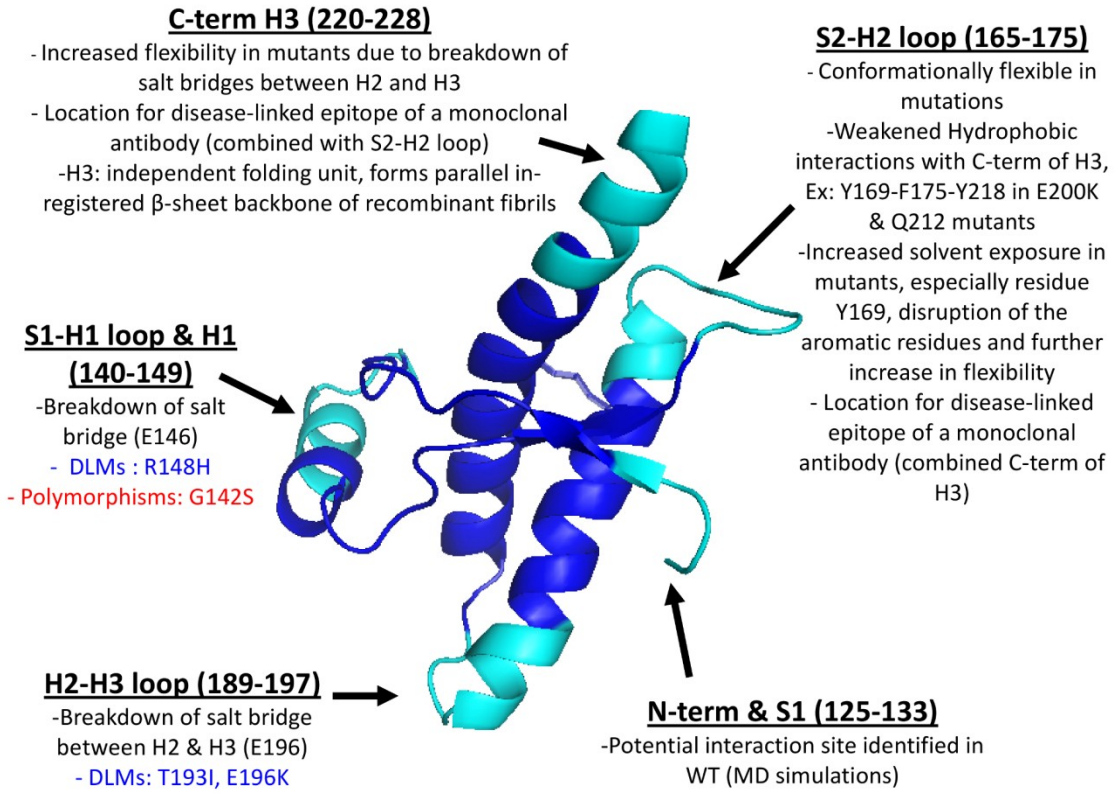


Figure 3.8: Conformationally variable subdomains in hPrP.
Subdomains are colored in cyan, and labeled by region. Important polymorphisms and disease-linked (DLMs) mutations in each section are also depicted.

TABLE 3.1: Summary of PCA analyses on PrP datasets.

DATASET	DESCRIPTION	# PDBs	# Models	Subdomain Contribution to each PC		
				PC1	PC2	PC3
Human (hPrP)	NMR ensembles of wildtype, variant, and mutant PrPs	11	230	S2-H2 loop, C-terminus H3	N-terminus & S1, H2-H3 loop	S1-H1 loop & H1
Wildtype hPrP	4 non-pathogenic hPrP	4	80	All subdomains	H2-H3 loop, C-terminus H3	N-terminus, S1-H1 loop
Mutant hPrP	7 Variant and mutant PrP structures	7	150	S2-H2 loop, C-terminus H3	N-terminus, H2-H3 loop	S1-H1 loop
Wildtype hPrP + CJD Mutant	8 WT & Variant hPrP structures + the CJD Mutant structure (E200K), (PDB 1FO7)	9	190	S2-H2 loop, H2-H3 loop, C-terminus H3	-	N-terminus, S1-H1 loop & H1
Wildtype hPrP + FFI Mutant	8 WT & Variant hPrP structures + the FFI Mutant structure (D178N), (PDB 2K1D)	9	180	S2-H2 loop, C-terminus H3	N-terminus	S1-H1 loop & H1, H2-H3 loop
Wildtype hPrP + GSS Mutant	8 WT & Variant hPrP structures + the GSS Mutant structure (Q212P), (PDB 2KUN)	9	180	S2-H2 loop, C-terminus H3	N-terminus, S1-H1 loop & H1, H2-H3 loop	-
WT PrP with chicken	16 PrP species, including chicken (chPrP), both TSE-susceptible & TSE-non-susceptible	21	420	H2-H3 loop	H1, S2-H2 loop	-
WT PrP without chicken	15 PrP species, excluding the conformational outlier chPrP	20	400	H2-H3 loop	H1, S2-H2 loop	-
Mammalian PrPs	13 WT PrP Species, excluding chicken, turtle, frog	18	360	C-terminus H3	S1-H1 loop & H1, S2-H2 loop, H2-H3 loop	N-terminus
TSE-non-susceptible Mammals	Subgroup of mammalian PrPs (n=5 species, including Sheep R168), excluding chicken, turtle, frog	5	95	C-terminus H3, N-terminus	S1-H1 loop & H1, S2-H2 loop	H2-H3 loop
TSE-susceptible Mammals	Subgroup of mammalian PrPs (n=9, including Sheep H168)	9	265	N-terminus, S2-H2 loop, H2-H3 loop, C-terminus H3	S1-H1 loop	H1
Mouse (mPrP)	3 non-pathogenic and 11 mutant PrP structures	14	280	C-terminus H3	-	-
Mouse (mPrP) without 2KFM & 2L1K	Removal of conformational outliers 2KFM and 2L1K to determine influence on variation in S2-H2	12	240	C-terminus H3	-	-

3.5 DISCUSSION

3.5.1 Delineating and ranking important PrP conformational subdomains

We have conducted exhaustive PCA analyses on a large set of PrP globular structures, as well as several subsets representing particular species of interest (human and mouse), or groupings which hold biological significance (TSE susceptibility or non-susceptibility); from these analyses we identified five conformationally variable subdomains in PrP undergoing varying levels of correlated movements in all datasets, and which are thought to be significant for the PrP conformational conversion process that underlies prion disease. We have demonstrated the benefits of exploring prion protein conformational variation using PCA, and the importance of the identified subdomains towards understanding the PrP conformational conversion process.

One obvious concern with the PCA analysis is that increased structural plasticity in the loop regions and protein termini would bias selection toward these regions, and outweigh identification of other regions in ordered, structured subdomains of the protein. However, for several of our PCA runs, structural variation within the protein datasets does not directly result from increased displacement in protein termini (the WT PrP set is an obvious example). In datasets where termini play a significant role in conformational differentiation of the structures, this variation is supported by weakened NMR definition in the protein (for example, hPrP and its variants vary in length and definition of residues 220-228 of H3 **[26]**). Additionally, our analysis identified structural variation within regions with repetitive secondary structures (ex: S1 and H1). Finally, for all PrP datasets we considered, structural plasticity of the loop regions has only been identified for selected portions of the loops, not the entire loop. For example, we only identify the latter half of the S1-H1 loop as conformationally variable in the hPrP and WT datasets, but the first half of the loop (residues 134-138, hPrP numbering) is relatively invariable.

To our knowledge, the presented work is the first study to perform a multivariate PCA on the native globular structures of PrP. Generally, few publications on prion structural biology have utilized multivariate analysis to comprehend the structural complexity of this protein and model protein flexibility computationally, with the

exception of a couple that have conducted PCA of MD simulations to determine protein flexibility **[10,15]**. Strikingly, some of the structurally variable subdomains we have identified (e.g., the S2-H2 loop) are "complementary" to the 'domains of collective movement' (rigid domains) identified by these studies **[10,15]**. Much of the computational analysis on PrP structures, however, involves the use of molecular dynamic simulations **[13,29,33,34]**, or longer dynamic simulations such as normal mode analysis (NMA) **[35]**. Such methods, as in the case with molecular dynamics, are continuously challenged by their computational expense, involvement of complex force fields, size of the query protein, and long time spans required to run the simulations **[23,36]**. Comparatively, our PCA analysis on native PrP, without the reliance on any structures generated by long- or short-term dynamics studies, succeeds in identifying key regions that may be involved in the conversion process and which have been previously highlighted in MD and NMA studies **[10,14,15,29,34,35]**. Accordingly, PCA is advantageous in rapid identification of important subdomains in PrP while saving computational time and effort, and may be used as starting point to identify key subdomains that can be further analyzed over longer time scales using protein dynamics.

This study is the first large-scale analysis of multiple NMR ensembles for a specific protein, and it poses unique challenges for principal component data analysis and interpretation. While static X-ray structures only provide a snapshot of potential motions of proteins, ensemble analysis of multiple X-ray structures may provide insight into the conformational changes of proteins and elucidate structural mechanisms of biological activity. The abundance of X-ray models for several protein families in the PDB facilitated PCA analysis of these proteins **[37,38,39]**, and development of computer tools for systematic multivariate analysis of X-ray ensembles is gaining increasing importance **[40,41]**. In the case of the PrP family, however, few X-ray structures of PrP exist in the PDB (<40% of all deposited PrP structures in the PDB), and even fewer structures represent globular PrP (as opposed to peptide segments, for example). For this analysis, we could only identify 11 relevant crystal structures, as opposed to the 41 NMR structures we have selected. Use of a reduced sample size based on X-ray structures severely limits the number of PCA analyses that could be performed on PrP subgroups and produces inaccurate estimates of collective motions in PrP. Structural analyses with

multiple NMR ensembles, while increasing the sample size multi-fold, poses a considerable analytical challenge as two sources of structural variation need to be considered: variation of models within an ensemble, and variation between ensembles. As variation between ensembles is expected, and sought for by PCA, eliminating variation within the ensemble remains an issue. To reduce the effect of inter-model variation, we have opted to use entire NMR ensembles, as random selection of any model may inadvertently introduce biases if the selected model is a structural outlier within the ensemble. Additionally, where selection of ensemble representatives was warranted, we used OLDERADO **[42]** to select for models representing the largest central core of the NMR ensemble, i.e., the "average" of the ensemble. Accordingly, PCA on the NMR ensembles allowed for identification of structural differences between NMR ensembles, but also successfully outlined inter-model differences within the ensembles.

3.5.2 The structural response to pathogenic mutation

Our PCA analysis has indicated that different subdomains are variable in different pathogenic mutants of PrP structures. Our PCA analysis has succeeded in providing a ranking for these subdomains that correlates with pathogenicity. In hPrP for example, by comparing displacements in residue contribution plots of the combined hPrP dataset and the mutant hPrP subset, we have demonstrated that the S2-H2 loop (residues 165-175) and the C-terminus of H3 (residues 220-228) are the first subdomains to differentiate pathogenic and nonpathogenic PrP structures. The S2-H2 loop is one of the most affected regions of PrP in terms of structure and flexibility, and may influence stability of PrP during $PrP^{C} \rightarrow PrP^{SC}$ conversion **[29]**. Mutant hPrPs exhibit weakened hydrophobic intramolecular interactions between this loop and the H3 helix, compared to native hPrP **[29]**. Weakened interactions between Y169-F175-Y218 have been reported for the E200K and Q212P mutants, as well as M166-Y225 π-stacking interactions **[29]**. The mutual orientation of aromatic residues in S2-H2 loop is affected by increased solvent exposure of Y169 in mutant PrP, yielding higher flexibility and greater solvent exposure of these hydrophobic residues compared to the observed stabilized aromatic interactions of Y163-Y169-F175 in the native hPrP **[24,29,43]**. As weakened hydrophobic interactions of the S2-H2 loop also weaken the interactions with H3 helix, it is not a

surprise that the C-terminus of H3 (residues 220-228) is equally important in differentiating wildtype from mutant PrPs. The C-terminus of H3 is observed to gain flexibility as a result of a breakdown in salt bridges between the H2 and H3 helices **[13,14,29]**. Interestingly, our conformer plot of all hPrP structures succeeds in separating the E200K, Q212P, and other pathogenic mutants displaying similar behavior (ex: D178N) from the remaining hPrP, reflecting on the specificity of the PCA in differentiating the structures by the plasticity of S2-H2 loop. This is particularly intriguing, as the S2-H2 loop (residues 166-170, hPrP numbering) and the C-terminal of H3 (residues 215-230) form a solvent-accessible disease-linked epitope for monoclonal antibody, and may serve as a recognition area for "protein X" involved in the conversion process **[44]**. Additionally, the S2-H2 loop has been observed to exhibit varying levels of flexibility within TSE-susceptible species, and is rigid in TSE-non-susceptible species, making it a prime candidate for PrP transmissibility studies **[10,15]**.

3.5.3 PrP structural evolution and TSE susceptibility

PCA of WT PrP structures has summarized areas that change concertedly over evolution, *e.g.* the H2-H3 loop. This was a particularly interesting result, as the H2-H3 loop is longer for chicken (the most outlying protein structure) than in any other species, and compared to other TSE-non-susceptible PrPs, is a flexible subdomain within that protein **[5]**. Generally, the structural variation observed does not correlate phylogenetically with organismal speciation. Intriguingly, the two most 'non-phylogenetic' clusterings are for TSE-non-susceptible species, rabbit (a placental mammal) clustering with frog, and horse (a placental mammal) clustering with wallaby (a marsupial). This is evidence for evolutionary 're-visiting' of different structural solutions to TSE resistance, in different evolutionary lineages. PCA profiles clearly show that different PrP subdomains vary amongst the TSE-susceptible and TSE-non-susceptible mammalian subsets. Also, the NMR ensembles for TSE-non-susceptible mammalian PrP structures tend to be peripheral on the PCA conformer plots, and overall, show a greater structural diversity, suggesting that TSE susceptibility may be linked to a greater degree of PrP structural similarity between infecting and receiving species/organisms.

To conclude, we performed an exhaustive analysis of PrP globular structures to identify subdomains of conformational change, as these subdomains of structural plasticity may contribute to PrP conversion and misfolding, and ultimately, to TSEs. Our PCA analysis succeeds in ranking these subdomains of as a function of species variation and disease-susceptibility. This is the first study to perform a multivariate PCA analysis on the native structures of the globular PrP, and one of very few studies to conduct PCA on NMR ensembles to detect biologically significant conformational variability in proteins and protein families. Our identified subdomains within PrP for all datasets studied compare favorably against those identified in computationally-intensive dynamic simulations and experimental data, suggesting that PCA analysis of the native structures can be used as a fast, reliable starting point to identify regions of interest that may warrant further analysis by computational and experimental methods.

3.6 MATERIALS AND METHODS:

3.6.1 PDB Structures

We collated all known PrP structures in the RCSB Protein Data Bank **[45]**, by searching for all proteins within the 'Prion-like' family and superfamily of SCOP **[46]**, proteins which match the architecture of the Major Prion Protein as specified in CATH **[47]** (Mainly alpha, orthogonal bundle, 1.10.790), as well as searches based on PFAM **[48]** Hidden Markov Models (HMMs) representing the Prion-like protein Doppel [PF11466], Prion/Doppel alpha-helical domain [PF00377], and the major prion protein bPrP-N terminal [PF11587]. These searches yielded a total of 112 prion PDB structures, from which only PrP globular domains were selected. The list of PrP globular domains was further refined to exclude dimers (ex: [PDB 3O79]), domain-swapped structures (ex: [PDB 1I4M]), and pdb models representing the average minimized structure of an NMR ensemble (ex: [1E1J, 1E1S, 1E1W, 1FKC, 1HJM, 1QLX, 1QM0, 1QM2] in human PrP, [1AG2] in mouse PrP, and [1DWY], [1DX0] in bovine PrP). A total of 41PDB structures, all of which are NMR-derived, were selected for analysis.

The analysis was performed on three separate cohorts of PrP globular proteins: (i) all human PrP (hPrP), (ii) all mouse PrP (mPrP), and (iii) all wildtype (WT) PrP, representing 16 species of PrP.

(i) The 11 PDB files of hPrP include: [**1E1G**, **1E1P**, **1E1U**, **1FO7**, **1H0L**, **1HJN**, **1QLZ**, **1QM1**, **1QM3**, **2K1D**, **2KUN]**

(ii) The 14 PDB files of mPrP include: [**1XYX**, **2K5O**, **2KFM**, **2KFO**, **2KU5**, **2KU6**, **1Y15**, **1Y16, 2L1D, 2L1E, 2L1H, 2L1K, 2L39, 2L40]**

(iii) The 21 PDB files of WT PrP include (species in parenthesis):[**1XYX]** (mouse); [**1DWZ**, **1DX1]** (bovine); [**1HJN**, **1QLZ**, **1QM1**, **1QM3]** (human); [**1Y2S** , **1XYU]** (sheep); [**1B10]** (hamster); [**1XYJ]** (cat); [**1XYQ]** (pig); [**1XYW]** (elk); [**2K56]** (bank vole); [**1XYK]** (dog); [**2KU4]** (horse); [**2FJ3]** (rabbit); [**1U3M]** (chicken); [**1U5L]** (turtle); [**1XU0]** (frog); [**2KFL]** (wallaby)

3.6.2 NMR Ensembles

For each of the datasets studied, an analysis was performed all models of the PDB NMR Ensembles, as well as the subset of representative models for each ensemble, identified using EBI OLDERADO **[42]**.

3.6.3 Structural Superposition & Principal Component Analysis (PCA) of PrP structures

For each dataset being studied, a multiple sequence alignment of all structures, based on ATOM residues, was generated using EBI MUSCLE **[49]**. This alignment and the corresponding structures were used as input in the Bio3D **[41]** package within the R statistical program **[50]**. Iterated rounds of structural superposition of PrP structures by Cα atoms, ignoring gap/insertion regions and missing residues, was performed to identify invariant core residues of PrP with a 1°A core cutoff. The structurally invariant core was used as a reference frame for structural alignment of the PrP NMR models, and Cartesian coordinates of the aligned Cα atoms were used as input for principal component analysis (PCA).

PCA maps high-dimensional data into fewer dimensions by a linear transformation **[16]**, and has been employed in several studies to provide insight into the

nature of conformational changes within proteins and protein families. In this study, PCA finds axes along which the high-dimensional ensemble of PrP protein structures can be best separated. The input is a coordinate matrix, X, composed of N by P dimensions, where N represents the number of structures and P represents three times the number of residues **[23,39]**, and each row of the matrix corresponds to the Cα coordinates of each structure. PCA is based on diagonalization of the covariance matrix, C, with elements C_{ij} built from X as follows:

$$C_{ij} = < (X_i - <X_i>)(X_j - <X_j>) >$$

where
i,j = all pairs of 3N Cartesian coordinates
< > = average over N atoms under consideration

Principal components (orthogonal eigenvectors) describe axes of maximal variance of the distribution of structures, and eigenvalues provide the percentage of variance (total mean square displacement) of atom positional fluctuations captured along each PC. Projecting PrP structures onto the conformational subspace defined by the largest PCs produces a low-dimension "conformer plot" which allows for the identification of dominant conformational changes and the characterization of inter-conformer relationships **[41]**. Additionally, the relative displacement of each residue described by a given PC can be represented in a "residue contribution" plot. Collectively, both plots allow for the identification of "conformationally variable subdomains" that are responsible for conformational clustering of the PrP structures, and which contribute to the structural variation observed in the datasets. These subdomains represent the largest segments of structural plasticity within the prion protein, making them candidate sites in the PrP conversion process.

Variation within models of an NMR ensemble poses a challenge for PCA analysis: how does the selection of a particular model influence the structural variation of a dataset? To test the extent to which inter-model variation within an NMR ensemble influences identification of variable PrP subdomains, we conducted PCA analyses on randomly selected NMR models within the hPrP and mPrP datasets. Using the total hPrP (11 PDBs) and mPrP (14 PDBs) datasets listed above, an NMR model was selected at

random from each of the NMR ensembles within that set, creating a subset of 'representative' NMR models for all the structures. The process was repeated 50 times and PCA was performed on each of the selected subsets. These random PCA runs on NMR models **(Supplementary Figures S3.5, S3.6)** succeed in identifying the same variable subdomains as those identified using ensembles, for hPrP **(Supplementary Figure S3.5)**, and for mPrP **(Supplementary Figure S3.6)**.

3.6.4 Molecular Graphics

Molecular figures have been rendered using PyMOL **[51]** and VMD **[52]**.

3.7 REFERENCES

1. Aguzzi A, Sigurdson C, Heikenwaelder M (2008) Molecular mechanisms of prion pathogenesis. Annual Review of Pathology-Mechanisms of Disease. Palo Alto: Annual Reviews. pp. 11-40.
2. Prusiner SB (1998) Prions. Proceedings of the National Academy of Sciences 95: 13363-13383.
3. van der Kamp MW, Daggett V (2009) The consequences of pathogenic mutations to the human prion protein. Protein Engineering Design & Selection 22: 461-468.
4. Zahn R, Liu AZ, Luhrs T, Riek R, von Schroetter C, et al. (2000) NMR solution structure of the human prion protein. Proceedings of the National Academy of Sciences of the United States of America 97: 145-150.
5. Calzolai L, Lysek DA, Perez DR, Guntert P, Wuthrich K (2005) Prion protein NMR structures of chickens, turtles, and frogs. Proceedings of the National Academy of Sciences of the United States of America 102: 651-655.
6. Garcia FL, Zahn R, Riek R, Wuthrich K (2000) NMR structure of the bovine prion protein. Proceedings of the National Academy of Sciences of the United States of America 97: 8334-8339.
7. Gossert AD, Bonjour S, Lysek DA, Fiorito F, Wuthrich K (2005) Prion protein NMR structures of elk and of mouse/elk hybrids. Proceedings of the National Academy of Sciences of the United States of America 102: 646-650.
8. Lysek DA, Schorn C, Nivon LG, Esteve-Moya V, Christen B, et al. (2005) Prion protein NMR structures of cats, dogs, pigs, and sheep. Proceedings of the National Academy of Sciences of the United States of America 102: 640-645.
9. Riek R, Hornemann S, Wider G, Billeter M, Glockshuber R, et al. (1996) NMR structure of the mouse prion protein domain PrP(121-231). Nature 382: 180-182.
10. Blinov N, Berjanskii M, Wishart DS, Stepanova M (2009) Structural Domains and Main-Chain Flexibility in Prion Proteins. Biochemistry 48: 1488-1497.
11. Christen B, Hornemann S, Damberger FF, Wuthrich K (2009) Prion Protein NMR Structure from Tammar Wallaby (Macropus eugenii) Shows that the beta 2-alpha 2 Loop Is Modulated by Long-Range Sequence Effects. Journal of Molecular Biology 389: 833-845.
12. Lee S, Antony L, Hartmann R, Knaus KJ, Surewicz K, et al. Conformational diversity in prion protein variants influences intermolecular beta-sheet formation. Embo Journal 29: 251-262.
13. Meli M, Gasset M, Colombo G (2011) Dynamic Diagnosis of Familial Prion Diseases Supports the beta 2-alpha 2 Loop as a Universal Interference Target. PLoS ONE 6: 10.
14. Rossetti G, Giachin G, Legname G, Carloni P (2010) Structural facets of disease-linked human prion protein mutants: A molecular dynamic study. Proteins-Structure Function and Bioinformatics 78: 3270-3280.
15. Santo KP, Berjanskii M, Wishart DS, Stepanova M (2011) Comparative analysis of essential collective dynamics and NMR-derived flexibility profiles in evolutionarily diverse prion proteins. Prion 5.
16. Jolliffe IT (2002) Principal Component Analysis: Springer New York.

17. Andersson CD, Chen BY, Linusson A Mapping of ligand-binding cavities in proteins. Proteins: Structure, Function, and Bioinformatics 78: 1408-1422.
18. Naumann T, Matter H (2002) Structural Classification of Protein Kinases Using 3D Molecular Interaction Field Analysis of Their Ligand Binding Sites: Target Family Landscapes. Journal of Medicinal Chemistry 45: 2366-2378.
19. Berglund A, Rosa MCD, Wold S (1997) Alignment of flexible molecules at their receptor site using 3D descriptors and Hi-PCA. Journal of Computer-Aided Molecular Design 11: 601-612.
20. Okazaki K-i, Takada S (2011) Structural Comparison of F1-ATPase: Interplay among Enzyme Structures, Catalysis, and Rotations. Structure 19: 588-598.
21. Gunnarsson I, Andersson P, Wikberg J, Lundstedt T (2003) Multivariate analysis of G protein-coupled receptors. Journal of Chemometrics 17: 82-92.
22. Miguel LT, George N. Phillips, Jr., Lydia EK (2002) A dimensionality reduction approach to modeling protein flexibility. Proceedings of the sixth annual international conference on Computational biology. Washington, DC, USA: ACM.
23. Teodoro ML, Phillips GN, Kavraki LE (2003) Understanding Protein Flexibility through Dimensionality Reduction. Journal of Computational Biology 10: 617-634.
24. Ilc G, Giachin G, Jaremko M, Jaremko L, Benetti F, et al. (2010) NMR Structure of the Human Prion Protein with the Pathological Q212P Mutation Reveals Unique Structural Features. PLoS ONE 5: 11.
25. Zhang Y, Swietnicki W, Zagorski MG, Surewicz WK, Sönnichsen FD (2000) Solution Structure of the E200K Variant of Human Prion Protein. Journal of Biological Chemistry 275: 33650-33654.
26. Calzolai L, Lysek DA, Güntert P, von Schroetter C, Riek R, et al. (2000) NMR structures of three single-residue variants of the human prion protein. Proceedings of the National Academy of Sciences 97: 8340-8345.
27. Zahn R, Güntert P, von Schroetter C, Wüthrich K (2003) NMR Structure of a Variant Human Prion Protein with Two Disulfide Bridges. Journal of Molecular Biology 326: 225-234.
28. Belt P, Muileman IH, Schreuder BEC, Bosderuijter J, Gielkens ALJ, et al. (1995) IDENTIFICATION OF 5 ALLELIC VARIANTS OF THE SHEEP PRP GENE AND THEIR ASSOCIATION WITH NATURAL SCRAPIE. Journal of General Virology 76: 509-517.
29. Rossetti G, Cong X, Caliandro R, Legname G, Carloni P (2011) Common Structural Traits across Pathogenic Mutants of the Human Prion Protein and Their Implications for Familial Prion Diseases. Journal of Molecular Biology 411: 700-712.
30. McColl IH, Blanch EW, Gill AC, Rhie AGO, Ritchie MA, et al. (2003) A New Perspective on Beta-Sheet Structures Using Vibrational Raman Optical Activity: From Poly(l-lysine) to the Prion Protein. Journal of the American Chemical Society 125: 10019-10026.
31. Martin TC, Moecks J, Belooussov A, Cawthraw S, Dolenko B, et al. (2004) Classification of signatures of Bovine Spongiform Encephalopathy in serum using infrared spectroscopy. Analyst 129: 897-901.

32. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2009) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 37: D5-D15.
33. van der Kamp M, Daggett V (2011) Molecular Dynamics as an Approach to Study Prion Protein Misfolding and the Effect of Pathogenic Mutations
Prion Proteins. Springer Berlin / Heidelberg. pp. 169-197.
34. van der Kamp MW, Daggett V (2010) Pathogenic Mutations in the Hydrophobic Core of the Human Prion Protein Can Promote Structural Instability and Misfolding. Journal of Molecular Biology 404: 732-748.
35. Samson AO, Levitt M (2011) Normal Modes of Prion Proteins: From Native to Infectious Particle. Biochemistry 50: 2243-2248.
36. Freddolino PL, Harrison CB, Liu Y, Schulten K Challenges in protein-folding simulations. Nat Phys 6: 751-758.
37. Bakan A, Bahar I (2009) The intrinsic dynamics of enzymes plays a dominant role in determining the structural changes induced upon inhibitor binding. Proceedings of the National Academy of Sciences 106: 14349-14354.
38. Grant BJ, McCammon JA, Caves LSD, Cross RA (2007) Multivariate Analysis of Conserved Sequence-Structure Relationships in Kinesins: Coupling of the Active Site and a Tubulin-binding Sub-domain. Journal of Molecular Biology 368: 1231-1248.
39. Yang L, Song G, Carriquiry A, Jernigan RL (2008) Close Correspondence between the Motions from Principal Component Analysis of Multiple HIV-1 Protease Structures and Elastic Network Modes. Structure 16: 321-330.
40. Bakan A, Meireles LM, Bahar I ProDy: Protein Dynamics Inferred from Theory and Experiments. Bioinformatics 27: 1575-1577.
41. Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD (2006) Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22: 2695-2696.
42. Kelley LA, Sutcliffe MJ (1997) OLDERADO: On-line database of ensemble representatives and domains. Protein Science 6: 2628-2630.
43. Corsaro A, Thellung S, Bucciarelli T, Scotti L, Chiovitti K, et al. (2010) High hydrophobic amino acid exposure is responsible of the neurotoxic effects induced by E200K or D202N disease-related mutations of the human prion protein. The International Journal of Biochemistry & Cell Biology 43: 372-382.
44. Telling GC, Scott M, Mastrianni J, Gabizon R, Torchia M, et al. (1995) Prion propagation in mice expressing human and chimeric PrP transgenes implicates the interaction of cellular PrP with another protein. Cell 83: 79-90.
45. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Research 28: 235-242.
46. Chandonia J-M, Hon G, Walker NS, Lo Conte L, Koehl P, et al. (2004) The ASTRAL Compendium in 2004. Nucl Acids Res 32: D189-192.
47. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, et al. (1997) CATH - a hierarchic classification of protein domain structures. Structure 5: 1093-1108.
48. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. The Pfam protein families database. Nucleic Acids Research 38: D211-D222.

49. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792-1797.
50. Team RDC (2011) R: A Language and Environment for Statistical Computing.
51. Schrodinger, LLC (2010) The PyMOL Molecular Graphics System, Version 1.3r1.
52. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. Journal of Molecular Graphics 14: 33-38.

3.8 SUPPLEMENTARY MATERIAL

Supplemental data includes 6 figures that can be found in **Appendix B**.

CHAPTER IV

Origins and Evolution of the HET-s Prion-Forming Protein: Searching for Other Amyloid-Forming Solenoids

Chapters II and III investigated conformational changes in the secondary and tertiary structures of native prion and amyloid-forming proteins to identify protein stretches with the potential for fibril formation. In Chapter IV, a functional, structural, and evolutionary analysis of homologs is conducted against the HET-s prion-forming domain (PFD), the first atomic structure of a functional amyloid fibril to date. Searching for homologs to the HET-s PFD aims to identify potential fibril-forming proteins that are amenable to adopting this specific, highly-structured β-aggregate. The results of this study shed light on the distribution of the HET-s β-solenoid fold in proteomes, and the overall implications towards identification of newer amyloid-forming proteins that can adopt a similar amyloid fold.

A version of this chapter is originally published as:
Gendoo DMA, Harrison PM (2011) Origins and Evolution of the HET-s Prion-Forming Protein: Searching for Other Amyloid-Forming Solenoids. PLoS ONE 6(11): e27342. doi:10.1371/journal.pone.0027342

4.1 ABSTRACT

The HET-s prion-forming domain from the filamentous fungus *Podospora anserina* is gaining considerable interest since it yielded the first well-defined atomic structure of a functional amyloid fibril. This structure has been identified as a left-handed beta solenoid with a triangular hydrophobic core. To delineate the origins of the HET-s prion-forming protein and to discover other amyloid-forming proteins, we searched for all homologs of the HET-s protein in a database of protein domains and fungal genomes, using a combined application of HMM, psi-blast and pGenThreader techniques, and performed a comparative evolutionary analysis of the N-terminal alpha-helical domain and the C-terminal prion-forming domain of HET-s. By assessing the tandem evolution of both domains, we observed that the prion-forming domain is restricted to S*ordariomycetes*, with a marginal additional sequence homolog in *Arthroderma otae* as a likely case of horizontal transfer. This suggests innovation and rapid evolution of the solenoid fold in the Sordariomycetes clade. In contrast, the N-terminal domain evolves at a slower rate (in Sordariomycetes) and spans many diverse clades of fungi. We performed a full three-dimensional protein threading analysis on all identified HET-s homologs against the HET-s solenoid fold, and present detailed structural annotations for identified structural homologs to the prion-forming domain. An analysis of the physicochemical characteristics in our set of structural models indicates that the HET-s solenoid shape can be readily adopted in these homologs, but that they are all less optimized for fibril formation than the *P. anserina* HET-s sequence itself, due chiefly to the presence of fewer asparagine ladders and salt bridges. Our combined structural and evolutionary analysis suggests that the HET-s shape has "limited scope" for amyloidosis across the wider protein universe, compared to the 'generic' left-handed beta helix. We discuss the implications of our findings on future identification of amyloid-forming proteins sharing the solenoid fold.

4.2 INTRODUCTION

The exact atomic structure adopted by amyloid fibrils is a topic of intense debate, as high molecular weights and the polymeric character and insolubility of amyloid fibrils

remain obstacles for high resolution structure determination methods such as nuclear magnetic resonance (NMR) spectroscopy **[1,2,3]**. Several structural studies of peptide amyloid fibrils have shown that the fibrils are arranged in a "cross-beta" sheet, a pattern characterized by repetitive arrays of beta-sheets that are parallel to the fibril axis, with their strands perpendicular to the axis **[1,2,3,4,5]**. While atomic-resolution structures of the infectious fibrils for many prions and amyloid-forming proteins are still lacking, recent studies have presented the first well-defined atomic structure of a functional amyloid, based on amyloid fibrils of the HET-s yeast prion **[6,7]**.

The *het-s* gene locus has two antagonistic alleles, *het-s* and *het-S*, which encode for HET-s and HET-S, respectively, and which give rise to the compatibility phenotypes [Het-s] and [Het-S] **[8,9,10]**. In comparison to its polymorphic variant, HET-S, only HET-s undergoes a transition to an infectious prion state. The HET-s prion of the filamentous fungus *Podospora anserina* is involved in heterokaryon incompatibility, a programmed cell death reaction that regulates the fusion between genetically distinct individuals **[8,9,10,11]**. HET-s is a 289 residue protein with an N-terminal domain (residues 1-227) and a prion-forming C-terminal domain (residues 218-289). The crystal structure of the HET-s N-terminal domain comprises an alpha-helical fold of 8-9 helices and a short two-stranded beta sheet **[8]**. The HET-s prion forming domain (PFD) is necessary and sufficient for amyloid formation *in vitro*, as well as prion propagation *in vivo* **[8,11,12]**. Fibrils formed from this PFD are described as a left-handed β-solenoid composed of four parallel, stacked pseudo-repeated β-helices; the pseudo-repeats are a result of one molecule forming two turns of the solenoid **[6,7]**. The first three β-strands of each pseudo-repeat enclose a dense triangular hydrophobic core **[6,7]**. In addition to intra- and inter-molecular hydrogen bonds between the pseudo-repeats, the solenoid structure is also stabilized by favourable side-chain contacts, such as salt bridges, between oppositely charged residues facing outside of the triangular core **[6,7]**.

Since its discovery, the HET-s solenoid, both in its native and fibrillar forms, has been well characterized **[6,7,10,11]**. However, studies on the evolutionary analysis of this fold, and identification of possible homologs to HET-s, remain largely lacking, despite the observation that a structural homolog of HET-s contributes to efficient cross-seeding of the amyloid form **[10]**. Accordingly, analysis of the evolution of the complete HET-s

protein may allow for the identification of newer, potential amyloid-forming proteins that can adopt the HET-s solenoid shape. To this end, we perform an exhaustive search for all homologs of the prion-forming solenoid, as well as the homologs to the HET-s N-terminal domain. Based on our findings, we perform an evolutionary analysis of both domains to determine when the solenoid fold arose in evolution, and its point of attachment to the HET-s N-terminal domain. Additionally, we identify and model structural homologs to the C-terminal solenoid fold, and we present an analysis of the conserved physicochemical properties we have observed in these generated solenoids, and how they compare to the current understanding of the β-solenoid structure. Our data sheds light on the relationship between the HET-s solenoid fold and understanding the amyloid disease state.

4.3 METHODS

4.3.1 Datasets

We downloaded the NCBI NR (non-redundant database: 14,261,927 protein sequences, database assembly dated 5/31/2011) from ftp://ftp.ncbi.nih.gov/blast/db/FASTA/. The Podospora anserina proteome (21,408 sequences) was downloaded from the NCBI Taxonomy Browser **[13]** (Taxonomy ID 5145). An additional 99 fungal proteomes (including mitochondrial proteomes, where available) from finished and ongoing projects were downloaded from the Broad Fungal Genome Initiative **[14]**. The 100 proteomes **(Supplementary List S4.1)** were grouped together into one in-house database (total of 715,255 protein sequences), and will be collectively referred to as BROAD throughout the manuscript.

4.3.2 Identification of HET-s homologs using sequence analysis

Using the genomes from NR and BROAD, we searched for homologs to the HET-s protein using (i) the N-terminal domain (residues 1-227), and (ii) the C-terminal prion-forming domain (PFD) (residues 218-289). For each query, sequence similarity searches were performed using Psi-blast [version 2.2.23] **[15]** with default parameters and masking for low complexity regions. Searches were performed until convergence was

reached or up to a maximum of 20 iterations, whichever was earlier. Significant hits were considered with E value<0.0001.

HMMs (Hidden Markov Models) for each of the queried regions were generated using HMMER [version 3.0, March 2010] **[16]**, based on blastp [version 2.2.23] **[15,17]** hits of each query against the NR database. For the N-terminal domain, 86 hits were identified from which only significant hits (E< 0.0001) were used to create the HMM (n=52). For the PFD, separate HMMs were generated for significant hits (E< 0.0001) to the PFD from blastp (n=7) as well as psiblast (n=12). HMMs were also generated using the entire sequences of all members that shared a conserved prion domain (n=12), as indicated by CDART (Conserved Domain Architecture Retrieval Tool) **[18]**. The CDART sequences were also refined and an HMM was generated only from the subsequences that match the prion-forming domain itself. A final HMM for the prion domain was generated based on sequences of the HET-s_218-289 family from Pfam (PF11558) (n=2) **[19]**. While such small number of sequences may raise concern about the quality of the resulting PFD HMMs, for HMMs generated from blastp, psiblast, or pfam multiple sequence alignments, we opted to generate these domain-specific HMMs to reduce the number false positive homologs to the solenoid fold when querying the HMM against NR, as opposed to relying on an HMM based on a multidomain (Nterm and Cterm PFD) sequence alignment. The pfam-based HMM is an extreme case of a "restricted" HMM, but which reflects on the highly restricted nature of the HET-s solenoid. Conserved protein domains were identified by querying the HMMs against the NR database to increase chances of detecting remote homologs to the Nterm and C-term PFD.

4.3.3 Identification of structural homologs based on protein fold recognition

All significant hits from Psiblast runs against NR and BROAD, as well as significant hits from HMMER searches were threaded against the HET-s solenoid [PDB: 2RNM] chains A-E, using pGenThreader **[20]**. Corresponding alignments of the significant hits were used to generate 3D models with MODELLER **[21]**. If needed, these alignments were modified based on sequence-alignments of the C-terminal region of

HET-s and its homologs **[10]**. 500 models for each protein were generated and the best model was selected with the lowest Discrete Optimized Protein Energy (DOPE) score. Stereochemistry of the models was assessed using the PROCHECK summary **[22]** of EBI PDBsum **[23]**. Selected models were viewed and rendered in PyMOL **[24]**. The RMSD calculation between the generated model and 2RNM template was calculated based on a structural alignment using the 'super' function in PyMOL **[24]**. Where applicable, the presence of salt bridges at specific positions within the models was determined using the ESBRI Server **[25]**.

4.3.4 Functional analysis of homologs

We downloaded a non-redundant set of 'genetic' single-chain domain protein sequences (n=10,569) from ASTRALSCOP, based on PDB SEQRES records (release 1.75). This was the non-redundant set made such that all sequences in it have pairwise similarity ≤ 40%. Entire protein sequences of all the identified homologs to the prion-forming and N-terminal domains were searched against this dataset using Blastp [version 2.2.23] **[15,17]**. Significant hits from ASTRALSCOP ($E \leq 0.0001$) were submitted to the SUPERFAMILY HMM search engine for further classification of protein domains and protein domain families **[26,27]**. To search for HET-s/LopB (HeLo) domains specifically, an HMM was constructed based on a previously identified loss-of-pathogenicity (LopB) protein and HeLo domains (n=24 sequences) **[8,28]**, and queried against the entire sequence of the N-terminal homologs identified from this study. Significant hits were selected based on a cutoff $E \leq 0.0001$. Protein sequences of identified structural homologs to the HET-s PFD were also searched against the Conserved Domain Database (38,392 PSSMs) using the NCBI CD-Search and Batch Web CD-Search Tools **[29,30,31]**.

4.3.5 Phylogenetic analysis

The NCBI taxonomy browser **[13]** and the taxonomy common tree generation tool (http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree/wwwcmt.cgi) were used to determine the taxonomic lineage for identified homologs. Additional taxonomic trees were generated using the Interactive Tree of Life (iTOL) server **[32]**. PHYLIP v3.69 **[33]**

was used to make neighbor-joining majority-rule consensus trees based on MUSCLE **[34]** multiple alignments. These trees were produced based on 100 replicates using the PHYLIP *seqboot*, *protdist*, *neighbor*, and *consense* programs. Briefly, 100 bootstrapped datasets were generated using *seqboot*. Bootstrapped datasets were then used as input into *protdist*, and distance matrices were generated for all sets using the Janet-Taylor-Thornton (JTT) matrix, with default parameters. Neighbor joining trees were generated based on these distance matrices using *neighbor*. Lastly, the *consense* tool was used to pick the final neighbor-joining bootstrapped tree. Selected trees were viewed using TreeDyn **[35]** within the Phylogeny.fr server **[36]**. Similarity matrices for N- and C-terminal domains of PFD homologs were generated based on the BLOSUM matrix using the EBI ClustalW **[37]** program, at default settings.

To make the neighbor-joining tree for phylogenetic analysis of horizontal transfer, we used the CLUSTALW **[37]** phylogenetic option, with 1000 bootstrap iterations. The tree was visualized using ProWeb tree server (*www.proweb.org/treeviewer/*).

4.4 RESULTS

4.4.1 Identification of homologs to the HET-s domains

Homologs of the HET-s N-terminal and prion-forming domain (PFD) have been searched against the non-redundant database (NR) and genomes from the Broad Fungal Genome Initiative (here, termed 'BROAD'), using Psiblast and HMMER as described in *Methods*. A total of 408 significant hits against both domains were observed, 217 hits were from NR and an additional 191 hits were from BROAD. In the initial comprehensive homology search, 29 hits were observed to match the prion-forming domain (PFD), and 400 hits matched against the HET-s N-terminal domain. Using Blastclust to remove identical sequences (100% identity cutoff), 16 hits to the PFD and 338 hits to the N-terminal domain are observed.

4.4.2 Evolution of the Prion-Forming Domain

Despite the inclusion of the NR database, which represents all kingdoms of life, all the identified homologs of the prion-forming domain are restricted to the fungal

kingdom, and they all belong to Saccharomyceta, more specifically, the Sordariomyceta **(Figure 4.1)**. Twenty-nine homologs to the PFD were identified using Psiblast and HMMer, in the initial comprehensive homology search. Manual curation to remove different genbank entries for the same gene (including provisional genbank entries), as well as removal of allelic variants with very high sequence similarity (>80% sequence identity) yielded 10 homologs to the PFD that were used in further evolutionary study **(Supplementary List S4.2)**. In addition to *Podospora anserina*, these 10 homologs were from 4 other fungal species, including *Nectria haematococca mpVI* 17-13-4, *Fusarium oxysporum*, *Fusarium graminearum (Gibberella zeae)*, and *Fusarium verticilliodes* **(Figure 4.1)**. Almost all of these hits from our initial homology search have been previously identified as homologs to HET-s **[37]**, with the exception of a newly identified homolog, EEU39630.1 [GI: 256726268] from *Nectria haematococca mpVI* 17-13-4.

Interestingly, searching through non-significant hits to the HET-s PFD revealed the presence of newly-identified remote HET-s homologs that lend a more complete picture about the evolution of the HET-s PFD within fungi. We identified a HET-s homolog with a PFD domain in *Grosmannia clavigera* kw1407 [Genbank: EFX05012.1, GI: 320592582], which is a species that also belongs to the Sordariomyceta **(Figure 4.1)**. This protein was identified in the NR database with marginal significance levels (E<=0.010 in psiblast iterations). Performing a reverse PSI-BLAST of this homologous PFD domain in the NR database yields a significant match to *Podospora anserina* HET-s residues 218-282 (E-value<0.005). We have also observed the presence of another small s protein annotation in *Arthroderma otae* CBS 113480 (anamorph: *Microsporum canis* CBS 113480), which is a more divergent Saccharomyceta species **(Figure 4.1)**. This protein was identified in both the NR [Genbank: XP_002843091, GI: 296804478] and BROAD (MCYG_08174) datasets with marginal significance levels in BROAD (E<=0.030 in psiblast iterations). Unlike the PFD homolog identified in *G. clavigera*, which spans almost the entire length of the PFD (68 residues in *G. clavigera* compared to 72 residues in HET-s), the subsequence of *A.otae* matching against the PFD is much shorter (49 residues). By taking the segment in *A.otae* that matches only the PFD of HET-s, and performing a reverse PSI-BLAST with default parameters for short sequences, we find a significant match to *Podospora anserina* HET-s residues 271-289

(E-value<0.005). Interestingly, the N-term of the *A.otae* small s protein exhibits significant homology to the N-term of HET-s (E-value 2e-35 in a web-based search). Given that the remote homology of the *A. otae* segment to HET-s PFD is unlikely to occur beside a homology to the N-terminal HET-s domain, simply by chance, this marginally detectable homology likely indicates a horizontal transfer from the *Sordariomycetes* to *Arthroderma otae* (a *Eurotiomycetes* species). Indeed, the most similar sequences to the N-terminal domain of the *A. otae* protein come from the *Sordariomycetes* species *P. anserina* and *Fusarium oxysporum* (43% and 42% respectively, over 215 residues). Also, 6/10 of the most similar N-terminal domain sequences come from *Sordariomycetes* species, and not *Eurotiomycetes*). To investigate further this likely horizontal transfer, neighbor-joining phylogenetic analysis was performed on the N-terminal domains of HET-s orthologs that significantly align to the *A. otae* N-terminal domain protein sequence **(Supplementary Figure S4.1)**. Regardless of the parameters used, the *A. otae* sequence always clusters with high bootstrap support (>80%) with the sequence from *Fusarium oxysporum*, within a larger grouping of *Sordariomycetes* sequences (green box in **Supplementary Figure S4.1**). Indeed, this is the only well-supported clustering between sequences from different phylogenetic fungal classes.

To compare the evolution of the N-terminal and C-terminal (prion-forming) domains that occur in the HET-s protein, we generated a similarity matrix for all proteins containing significant homologs of both HET-s domains (n=11) **(Figure 4.2, Supplementary Table S4.1)**. We compared all pairwise similarities for the N-terminal domains to the corresponding pairwise similarities for the C-terminal PFD **(Figure 4.2, Supplementary Table S4.1)**. The plot clearly shows that the C-terminal PFD is evolving more rapidly that the N-terminal domain, with higher percentages of sequence identity between the N-terminal domains as opposed to the C-terminal domains, and only one pairwise comparison in disagreement amongst HET-s sequences from species other than *Podospora anserina*. Despite this, the majority-rule consensus neighbor-joining trees have similar clusterings of sequences (ignoring the tree branchings with <60% support) **(Figure 4.3)**. Taken collectively, the rapid evolution of the HET-s PFD we have demonstrated, coupled with the limited phyletic distribution of PFD homologs we have

observed, suggests innovation of the PFD in Sordariomyceta, followed by rapid evolution in this domain, relative to the N-terminal domain. The additional marginal homolog in *A. otae* most likely arose by horizontal transfer, after innovation of the domain in *Sordariomycetes*.

Figure 4.1 Taxonomic lineage of homologs to the HET-s PFD.

The expanded taxonomic lineage of all species is presented. Actual species of HET-s PFD homologs are highlighted in bold and underlined. The non-significant PFD homologs from *Arthroderma otae* and *Grosmannia clavigera* (red boxes) are also included for comparison.

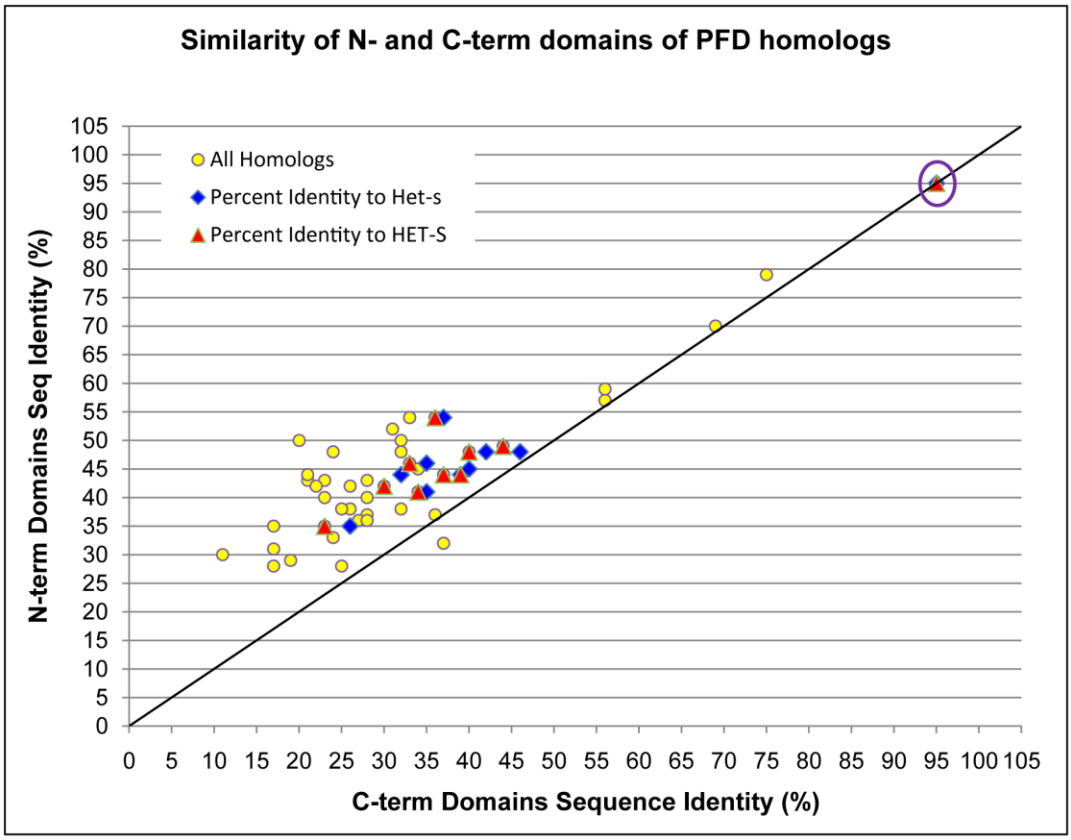


Figure 4.2 Graphical representation of the similarity matrix between N- and C-terminal homologs of the PFD.

Each point on the graph represents the percent similarity of the C-terminal domains and the N-terminal domains for a pair of PFD homologs. In addition to HET-s, ten PFD homologs are represented. Pairs of homologs that include the HET-s or HET-S proteins have been colored differently for comparison. Comparison of *Podospora anserina* sequences to each other are circled (purple).

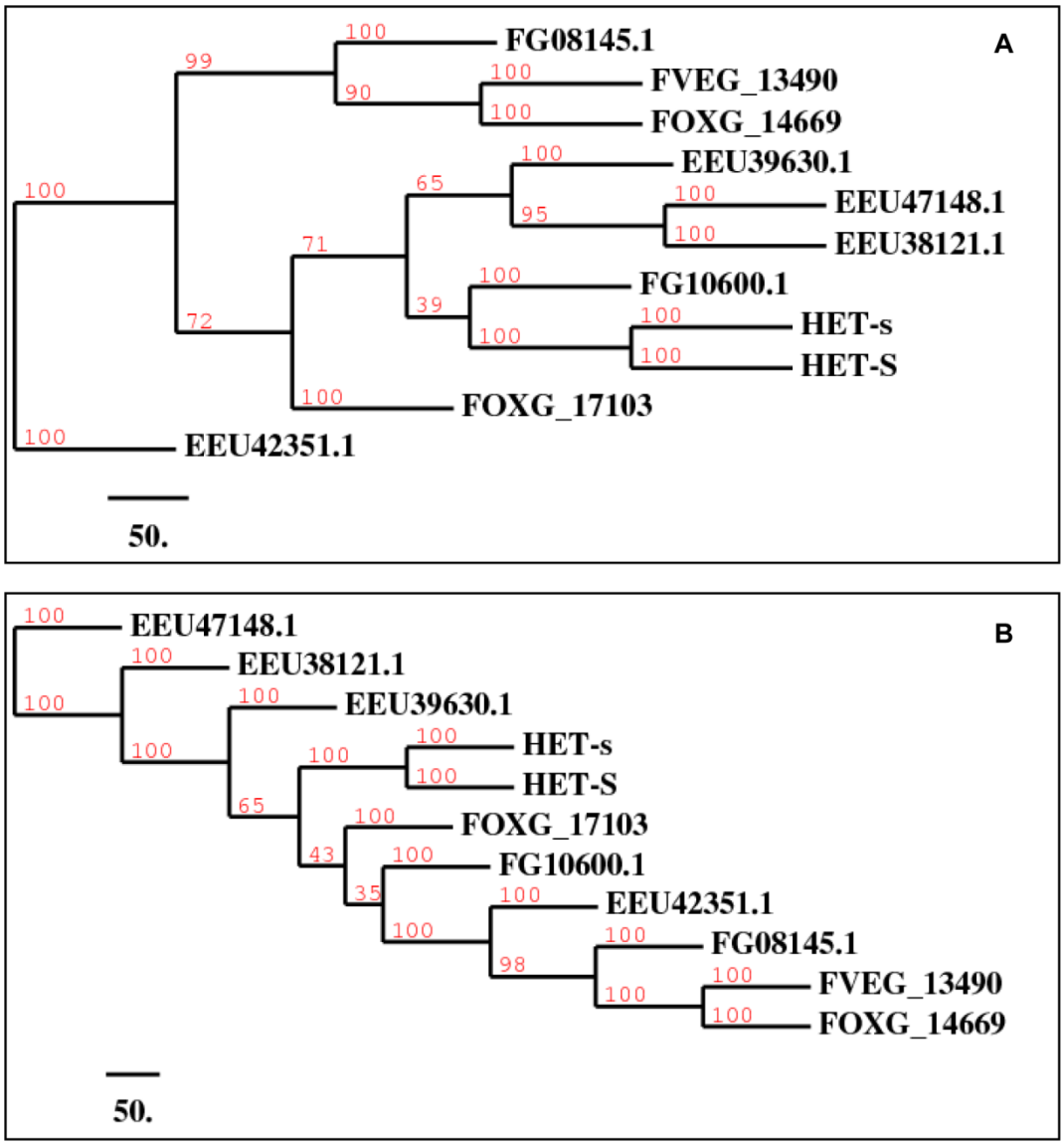


Figure 4.3 Phylogenetic trees of homologs to the HET-s prion-forming and N-terminal domains. (A) Phylogenetic tree of homologs to the HET-s prion-forming domain. (B) Phylogenetic tree of homologs to the HET-s N-terminal domain. The generated trees are neighbor-joining majority-rule consensus trees composed of 11 sequences. Sequences starting with EEU represent *Nectria haematococca* mpVI 77-13-4, FOXG represent *Fusarium oxysporum* f. sp. lycopersici, FVEG represent *Fusarium verticillioides* (*Gibberella moniliformis*), and FG represent *Fusarium graminearum* (*Gibberella zeae*). Branch numbers indicate the number of times the partition of the species into two sets which are separated by that branch occurs among the trees, out of 100 trees, as described by Phylip consense program **[33]**.

4.4.3 Distribution of the HET-s solenoid fold in HET-s homologs

Threading of all identified homologs to the HET-s N-terminal and PFD against the prion-forming solenoid [PDB: 2RNM] using pGenThreader **[20]** , identified 11 structural homologs from 5 species, almost all of which had already been previously identified in the sequential analysis **(Table 4.1)**. One of these homologs (FG10600.1) has been addressed in a previous publication and a model similar to HET-s has been proposed based on experimental analysis **[10]**. Two of the identified homologs (FOXG17103 and FOXG17314) are 100% identical and were considered henceforth as one model **(Table 4.1)**. Interestingly, in addition to these homologs that have been identified both by sequential and structural analysis, we also identified one further potential structural homolog through threading alone, *i.e.*, TSTA_087480, in *Talaromyces stipitatus* **(Table 4.1)**. However, for this case, absence of other known homologs to TSTA_087480 precludes further bioinformatic analysis.

We were able to successfully generate solenoid structural models for all identified structural threadings of the C-terminal PFD using the MODELLER tool **[21]** and pGenThreader-generated sequence alignments **(Figure 4.4)**. The RMSD and PROCHECK **[22]** calculations of our generated models compare favorably against the template solenoid fold [PDB: 2RNM] **(Table 4.1)**. Similar to the HET-s PFD, the modeled proteins adopt a pseudorepetitive structure, where one chain is composed of two turns of the solenoid, in addition to a conserved triangular hydrophobic core with similar compositions of alanine (A) and the bulky hydrophobic residues of valine (V), isoleucine (I), and phenyalanine (F) **(Figure 4.4, Figure 4.5)**. The asparagine ladder, as previously noted by Wasmer *et al.* **[10]** also remains largely conserved throughout the homologs **(Figure 4.5)**, although in some sequences, asparagines ladder residues are missing at the appropriate positions. Few of the models retain the ability for formation of a salt bridge pair at positions comparable to that of the 3 salt bridges of the PFD structure. Additionally, we have observed changes in the length of the pseudorepeats which may hinder the formation of a stable, repetitive fibril. For example, we have observed that the first pseudorepeat “rung” is shorter by 2 residues than the second rung in the homologs FVEG13490, FG08145, and FOXG14669. This length difference would yield an irregular fibrillar stacking of the solenoid.

We attempted to model structurally the small s proteins of the more divergent PFD sequence homologs from *Grosmannia clavigera* and *Arthroderma otae,* to determine if the conserved physicochemical properties of the HET-s structure could be observed in these marginal remote homologs. The small s protein from *G.clavigera* could easily be modeled against the solenoid structure, and similar to the other homologs, retains pseudorepeats, a conserved hydrophobic core, and asparagines ladders. Contrastingly, for the *A.otae* small s protein, all threading attempts using the entire sequence were ranked as "GUESS" in pGenThreader **[20]**, with the exception of chain A of the solenoid structure [PDB: 2RNM], which ranked as "LOW" at 19% sequence identity. Interestingly, an unambiguous sequence alignment in the *A. otae* sequence could be generated for only one rung of the PFD solenoid (not shown), indicating perhaps that it comprises an obligate oligomer with a single solenoid rung.

Table 4.1: HET-s homologs showing significant structural homology to the 2RNM solenoid.

Threading Score	Accession Number[a]	Protein	DB[b]	Structural Model			
				Template Chain	% Identity[c]	RMSD[d]	Procheck[e]
LOW	[GI: 242774612]	Hypothetical protein, *Talaromyces stipitatus*, TSTA_087480	NR	C	17.7	0.816	88.1
LOW	EEU47148.1	Hypothetical protein, *Nectria haematococca mpVI* 77-13-4	BROAD	A	36.7	0.616	84.7
LOW	EEU42351.1	Hypothetical protein, *Nectria haematococca mpVI* 77-13-4	BROAD	C	31.6	0.736	82.8
LOW	EEU39630	Hypothetical protein, *Nectria haematococca mpVI* 77-13-4	BROAD	C	24.1	1.048	82.3
MEDIUM	EEU38121.1	Hypothetical protein, *Nectria haematococca mpVI* 77-13-4	BROAD	A	35.4	0.487	77.6
LOW	FOXG14669	Conserved hypothetical protein, *Fusarium oxysporum*	BROAD	C	34.2	1.460	81.8
LOW	FOXG17103 or FOXG17314	Conserved hypothetical protein, *Fusarium oxysporum*	BROAD	C	29.1	1.073	80
LOW	FVEG13490	*Fusarium verticilliodes,* hypothetical protein	BROAD	C	26.6	1.172	80.7
LOW	FG 08145.1 [GI: 46127535]	Hypothetical protein, *Fusarium graminearum*	NR	D	31.6	0.667	75
MEDIUM	FG 10600.1 [GI: 46138171]	Hypothetical protein, *Fusarium graminearum*	NR	A structure based on experimental analysis is proposed by Wasmer *et al*, 2010 **[10]**			
LOW	[GI: 320592582]	Small s protein, *Grosmannia clavigera* kw1407	NR	A	26.6	0.492	83.3

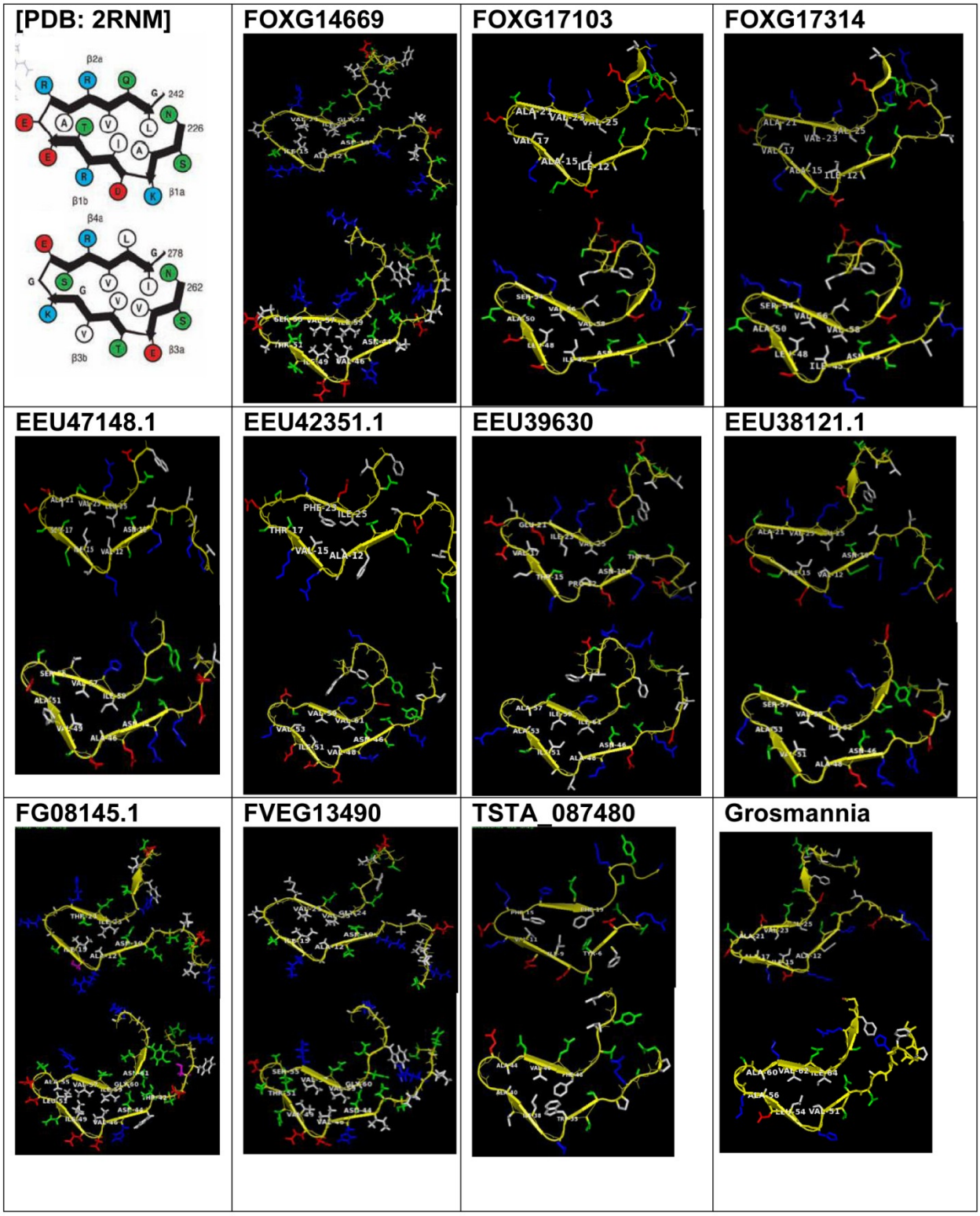
a: The Genbank (GI) identification number from NR and BROAD accession numbers are provided, where available.
b: NR: non-redundant database, BROAD: Broad Fungal Genomes Initiative
c: Percentage identity based on comparison with template in pGenThreader.
d: RMSD calculations are performed against the NMR model 9 of the [PDB: 2RNM] template.
e: Represents percentage of residues in the most favored region

---

Figure 4.4 (next page)

Models of HET-s homologs with structural homology to the HET-s PFD.

The original solenoid [PDB: 2RNM] **[6,7]** is shown in the top left corner. Ten structural homologs are represented, including the small s protein homolog from *G. calivigera*. The structure of FG10600.1 has been shown by Wasmer *et al* **[10]** and is not included here. For each structure, two rungs for each solenoid are represented, with the first rung at the top. Amino acids are color-coded as follows: acidic (Asp, Glu) in red, basic (Arg, Lys, His) in blue, nonpolor (Met, Phe, Pro, Trp, Val, Leu, Ile, Ala) in white, polar (Ser, Thr, Asn, Gln, Tyr) in green, and the protein backbone in yellow.


[PDB: 2RNM]
FOXG14669
FOXG17103
FOXG17314
EEU47148.1
EEU42351.1
EEU39630
EEU38121.1
FG08145.1
FVEG13490
TSTA_087480
Grosmannia

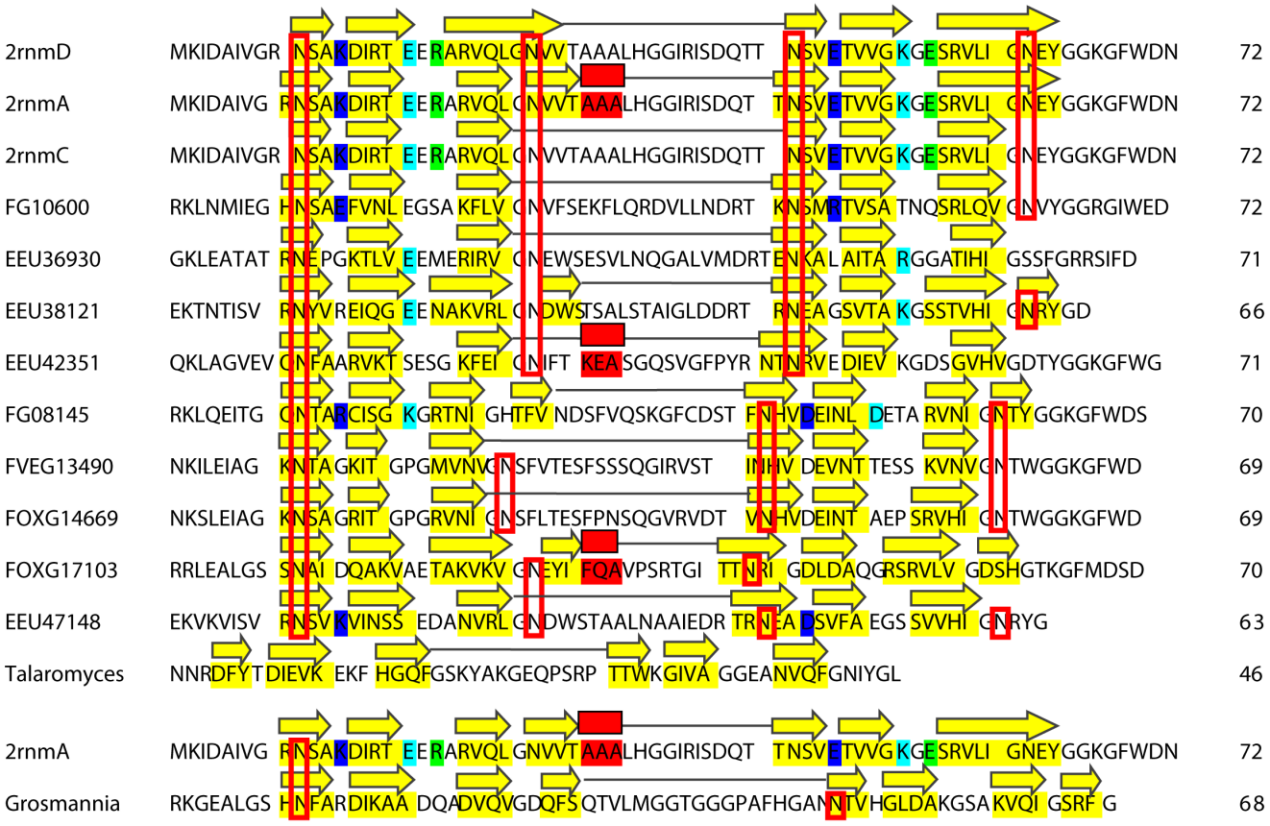


Figure 4.5 Conserved physicochemical properties of the HET-s structure in homologous solenoid models. The three chains of 2RNM (A, C, D), used as templates in MODELLER, are represented (top). Beta sheet positions (as rendered in PyMOL) are highlighted in yellow (unless colored to represent other physicochemical properties), and helices are highlighted in red. The three salt bridge pairs (K229-E265, E234-K270, R236-E272) of HET-s are highlighted in dark blue, light blue, and light green, respectively. Asparagine ladders are represented in red boxes. The same coloring scheme has been adapted to the 11 generated models against the HET-s solenoid structure. The small s protein from *G. clavigera* (here, represented as Grosmannia), is also included for comparison against its 2rnm:A template. For all models, gapped positions have been removed for clarity, and the number of amino acids spanning the HET-s PFD are indicated (out of 72 residues). The secondary structure of each model has been placed above each sequence. Beta-strands are represented by yellow arrows and alpha-helices by red boxes. Blank spaces between the yellow arrows represent β arcs within each solenoid rung, and the long connecting loop between the two solenoid rungs in each model has been represented by a grey flat line.

4.4.4 Evolution of the HET-s N-terminal Domain across fungal clades

As opposed to the prion domain, which was likely innovated in Sordariomycetes, homologs to the HET-s N-terminal domain are more widespread within fungi **(Figure 4.6)**; however, the domain was not discovered outside of the fungal kingdom. As noted above, analysis of the N-terminal domains of the PFD homologs indicates that, while almost all of the domains share <50% identity with the HET-s or HET-S N-terminal domains, the sequence similarity between these domains still exceeds that of the PFDs **(Figure 4.2)**. Comparing the N-terminal domains of the homologs to one another also indicated that 8 pairs of homologous sequences (aside from those involving HET-s or HET-S) share >50% sequence identity, twice the number observed for the C-terminal PFDs **(Supplementary Table S4.1)**.

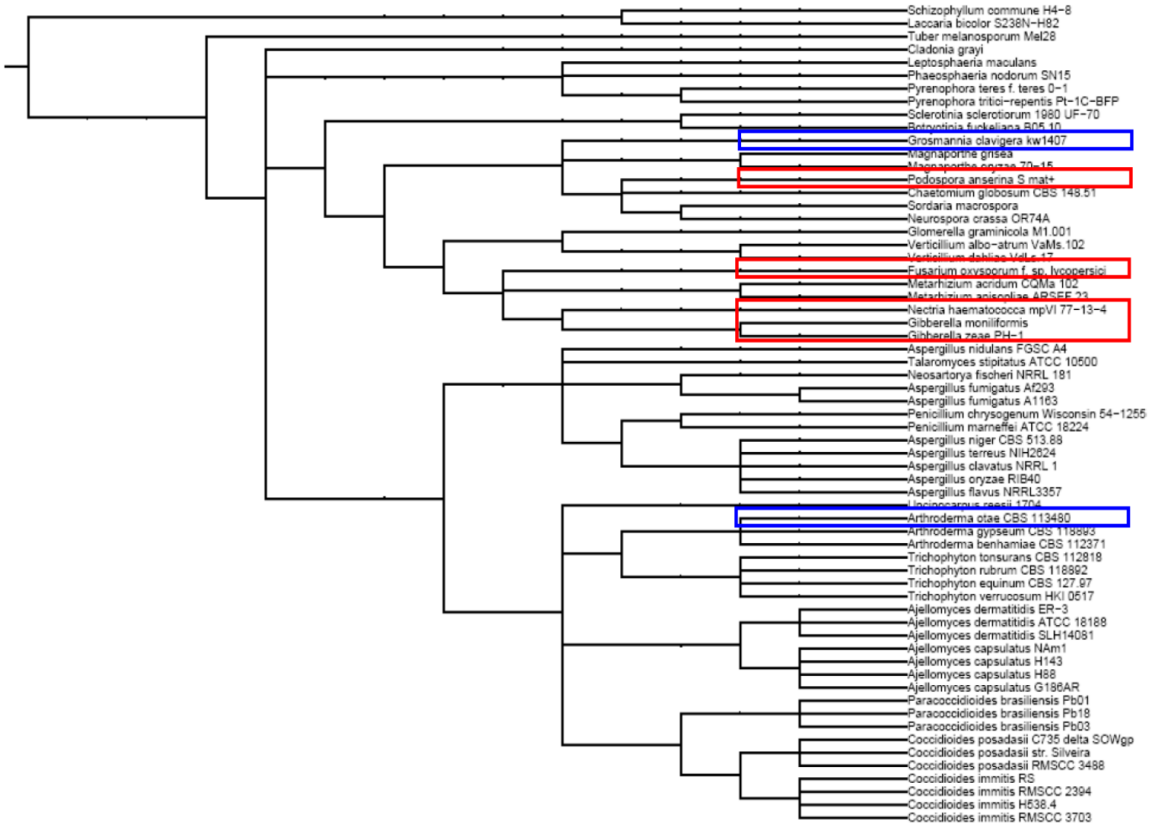


Figure 4.6 Taxonomic lineage of homologs to the N-Term domain. Species with proteins homologous to the prion domain are highlighted in the red box. The marginal additional homologs observed in *Grosmannia clavigera* and *Arthroderma otae* are highlighted in the navy boxes.

While an initial screen of the homologous sequences that contain the N-terminal HET-s domain indicates that many are labeled as hypothetical or predicted proteins, protein domain assignments reveal a wide diversity of domain architectures in HET-s homologs **(Figures 4.7 & 4.8)**. Forty HET-s homologs were mapped to 65 SCOP domains **(Supplementary Tables S4.2 and S4.3)**. Using the SUPERFAMILY HMM search engine **[26,27]**, these domains could be categorized into 10 superfamilies, with ankyrin being the most prevalent, followed by the WD40 repeat-like and the UBC-like domains **(Figure 4.7)**. A phylogenetic analysis of these 40 homologs indicates that the ankyrin repeat is largely predominant in Sordariomycetes **(Figure 4.8)**. Using HMMs, we also checked for the presence of HeLo (HET-s/LopB) domains in the entire sequences of identified homologs to the HET-s N-terminal domain, and we identified 212 HeLo domains in that set **(Supplementary Table S4.4)**. The HeLo domain had been previously identified based on >30% sequence similarity between the HET-s N-terminal domain and a fungal loss-of-pathogenicity (LopB) protein from *Leptosphaeria maculans* **[8,28]**. In this study, we identified a second LopB protein [GI: 189205459] from *Pyrenophora tritici-repentis* Pt-1C-BFP with 30% similarity and 14% identity to the N-terminal domain. Searching for the conserved HeLo domains using the HMM also yielded a significant match to a HET-s/LopB domain from *Metarhizium anisopliae* ARSEF 23 [GI: 322703231, E-value 1.6e-10], as well as marginally significant matches [GI: 310797955, GI: 317157340, GI: 317033349] in several proteins from *Glomerella graminicola*, *Aspergillus oryzae* RIB40, and *Aspergillus niger* CBS 513.88, respectively [corresponding E-values 0.0042, 0.00082, 0.00083]. We visually inspected the remaining homologs of the N-terminal for any other HeLo domain-containing proteins and identified 3 more hits that are classified as containing a HeLo domain but which are not detected using the HMM ([GI:212532807] from *Penicillium marneffei* ATCC 18224, [GI:242776556] from *Talaromyces stipitatus* ATCC 10500, and [GI: 327353076] from Ajellomyces dermatitidis ATCC 1818).

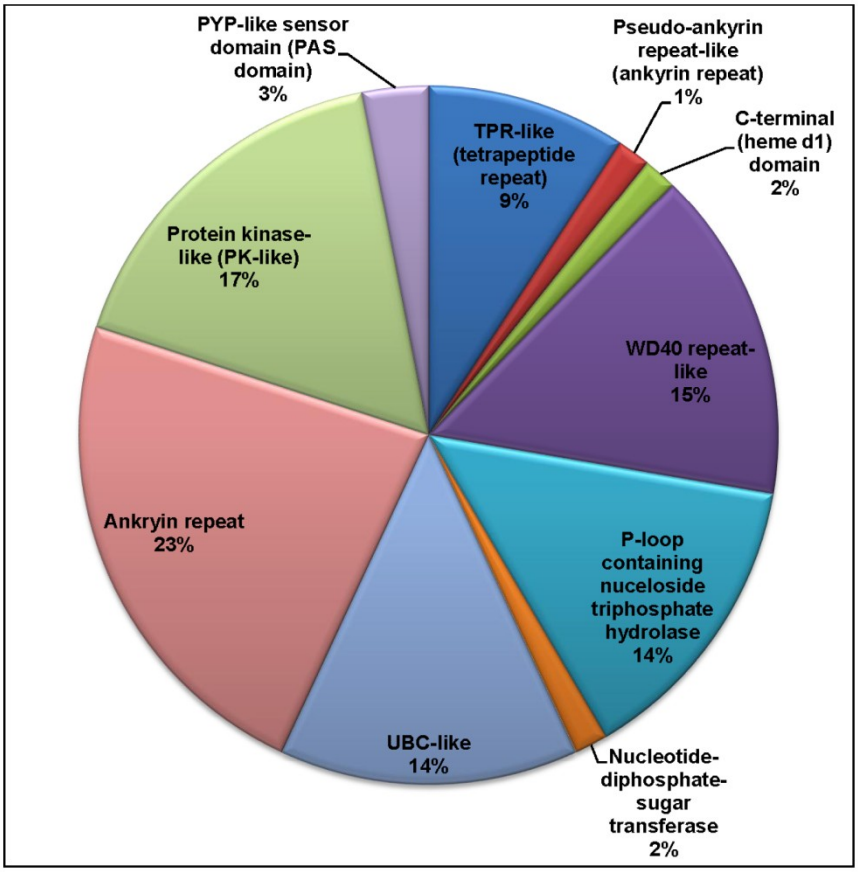


Figure 4.7 Classification of 65 SCOP domains into superfamilies. These are the SCOP domain superfamilies that co-occur with the HET-s protein domains. Ten superfamilies are represented.

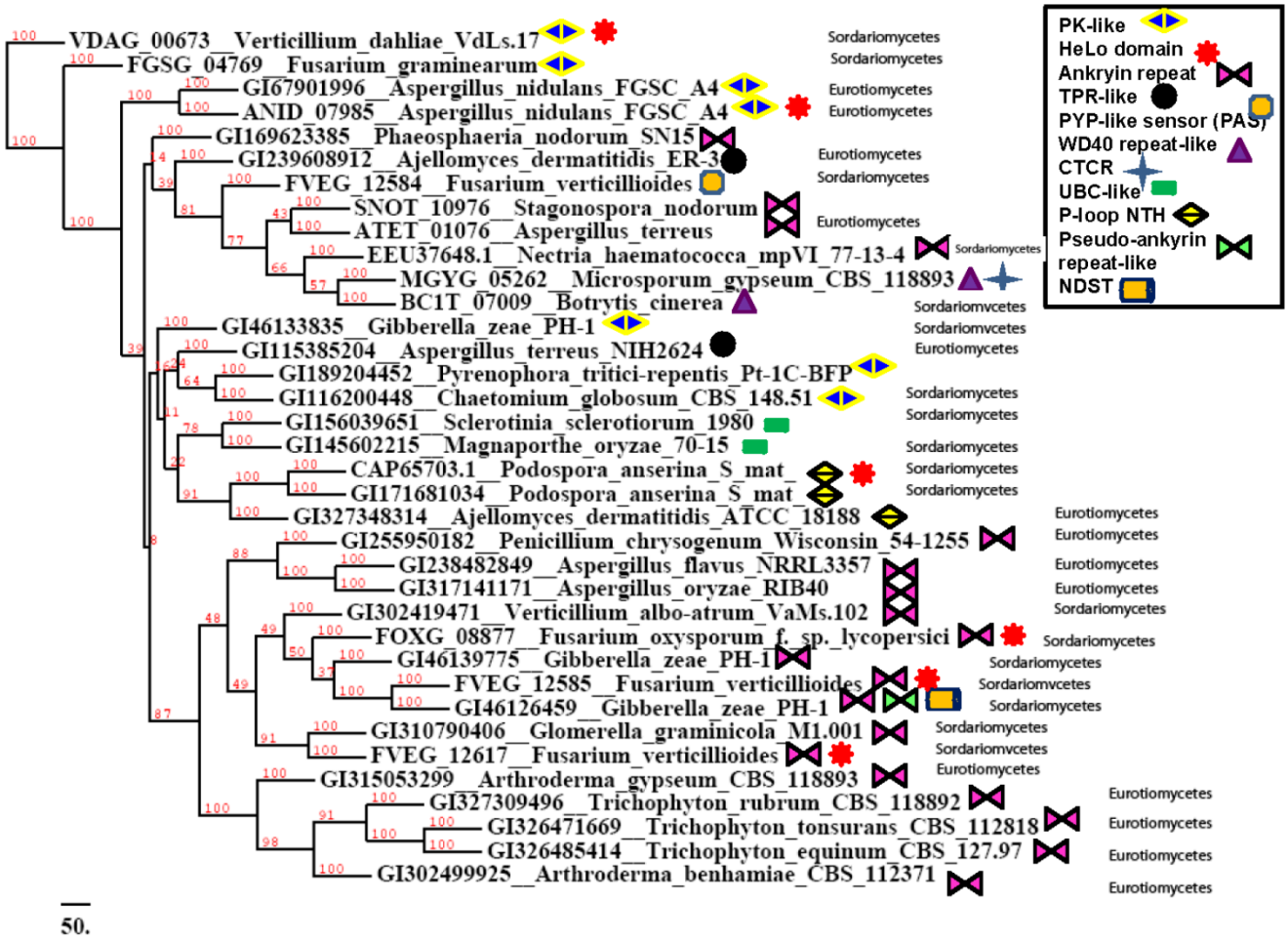


Figure 4.8 SUPERFAMILY associations with the N-terminal homologs (n=36). A majority-rule consensus tree is generated for N-terminal homologs which are significantly associated with families identified using SUPERFAMILY. The clades of the different proteins are also annotated, especially for proteins which belong to 'Sordariomycetes' or 'Eurotiomycetes'. Superfamilies associated with each protein are indicated, and are abbreviated as follows:

CTCR: C-term (heme d1) of cytochrome reductase

P-loop NTH: P-loop containing nucleoside triphosphate hydrolase

NDST: Nucleoside-diphospho-sugar transferase

4.5 DISCUSSION

The HET-s solenoid remains the only atomic resolution of a fibril known to date, which raises an intriguing question of whether other amyloid-forming proteins that adopt the HET-s solenoid shape exist, and whether they can be identified. To probe this question, we have performed an exhaustive study for homologs of the HET-s prion-forming solenoid domain to identify potential amyloid-forming proteins that adopt such a shape in their native form or fibril states. Additionally, we investigated the evolutionary relationship between the prion-forming solenoid, and the HET-s N-terminal domain.

Our evolutionary analysis of the prion-forming domain reveals that the PFD, compared to the N-terminal domain, has limited phyletic distribution and has evolved rapidly. Despite the use of the NR database and multiple queries based on psi-blast and HMMs of the PFD, all results converge to the same set of homolog hits (n=11). This indicated that a "restricted" profile HMM based on a small number of blast sequences has not influenced the results. Remote homologs to the *P. anserina* PFD were identified (in *G. clavigera* and *A. otae*), but with the exception of the remote homolog from *A.otae*, all the PFD homologs remain restricted to one fungal clade, Sordariomycetes. In several species, the HET-s homologs exist as paralogous gene families, as we observed a single HET-s protein in *Podospora anserina*, two in F. *graminearum a*nd four in *N. haematococca*. A comparison of the sequence similarities for the PFD and N-terminal domain of these homologs indicates a rapid divergence of the PFD compared to their companion N-terminal alpha-helical domains, as indicated by their sequence similarity matrix **(Figure 4.2, Supplementary Table S4.1)**. In stark contrast to the limited phyletic distribution of the PFD, we have identified a set of N-terminal homologs almost 14 times larger than the PFD homolog set, and not surprisingly, with a larger evolutionary spread within fungi **(Figure 4.6)**. Based on the phyletic distribution of these domains, the evolutionary point of attachment of the HET-s N-terminal domain and prion-forming domain can be attributed to Sordariomyceta, with a marginal homolog in *A.otae* that probably arose by horizontal transfer. Parsimoniously, horizontal transfer is a more likely event compared to multiple parallel gene loss events of the PFD in several fungal clades associated with the N-terminal domain.

The striking abundance and widespread phyletic distribution of homologs to the N-terminal domain implies that it may serve several functions beyond heterokaryon incompatibility and amyloidogenicity in many fungal species. Our protein domain assignment analysis of the homologous sequences that contain the N-terminal domain identified a wide diversity of protein domain partners. While many of the homologs to the N-terminal domain are hypothetical proteins, we have successfully identified 10 proteins superfamilies, based on SCOP and SUPERFAMILY, in 10% of our homolog dataset **(Figure 4.7)**. The most common superfamily is the ankyrin repeat, followed by the protein kinase-like (PK-like) domain, WD40 repeat-like, and UBC-like domains, among others. Interestingly, all of the above-mentioned families are involved in protein-protein interactions. The ankyrin repeat is of particular interest, as this repeat is predominant in the HET-s homologs in Sordariomycetes **(Figure 4.8)**. This repeat is a common protein-protein interaction motif found in a variety of functionally diverse proteins such as enzymes, toxins, and transcription factors **[38]**. Similarly, proteins containing WD40 or tetratricopeptide (TPR) repeats serve as platforms for protein complexes **[39,40,41]**; WD40 repeats are found in G proteins that participate in transmembrane signaling machinery, as well as proteins involved in RNA-processing complexes **[39,40]**.

In addition to protein-protein interactions, another underlying functionality we have observed, both in the HET-s N-terminal and prion-forming domains, is that of 'pathogenicity'. While previous studies of the N-terminal homologs did not identify any homologs with a known function, a new HET-s/LopB (HeLo) domain had been identified based on a 31% similarity of the HET-s N-terminal domain to the loss-of-pathogenicity (LopB) protein from the Dothideomycete fungus *Leptosphaeria maculans,* a fungus that causes blackleg disease of *Brassica napus* **[8,28]**. In current literature, 23 representative HeLo domains have been identified to date **[8,28]**. We searched for these proteins in our list of homologs, and in addition to these representative proteins, we identified a second loss-of-pathogenicity protein (LopB) in the Dothideomycete fungus *Pyrenophora tritici-repentis*, and 212 HeLo domains in more than 40 species **(Supplementary Table S4.4)**. Notably, we observed that the species of many of the PFD structural homologs we have identified, such as *Nectria haematococca mpVI* 17-13-4, *Fusarium oxsyporum*, and

Fusarium graminearum, are all plant pathogens, causing diseases such as wheat headblight disease and *Fusarium* wilt disease **[42,43]**.

Our evolutionary search for sequential homologs to the HET-s PFD, and subsequent analysis on structural homologs to the HET-s solenoid structure, sheds light on the contribution of the HET-s solenoid fold to fibril formation and stability in amyloid-forming proteins. As the HET-s solenoid shape remains the only atomic structure for a fibril to date, to what extent do other proteins share this fold? From an evolutionary perspective, our analysis of the PFD solenoid, and the limited phyletic distribution of PFD structural homologs we have observed, suggest that the HET-s solenoid shape has 'limited scope' for amyloidosis. The restriction of this particular left-handed β-solenoid to filamentous ascomycotes strikingly contrasts against that of a 'generic' left-handed beta-helix found in almost all phyla **[44]**, and which is the current proposed model for fibrils of prions and other amyloid-forming proteins that are not necessarily fungal **[45,46,47,48,49,50,51]**. Interestingly, at face value, the HET-s solenoid is an attractive candidate for the formation of stable fibrils in the structural homologs we have identified: this shape is easily modelled in the homologs we have identified (despite poor sequence identity), and could even be modelled in remote homologs to the PFD, such as the small s protein of *G. clavigera* **(Figure 4.4 and Figure 4.5),** and even in *A.otae***.** Several characteristic physicochemical properties of HET-s remained conserved within these models, such as a conserved triangular hydrophobic core with enrichment for hydrophobic bulky residues, and conserved asparagine ladders at comparable positions to the HET-s PFD **(Figure 4.5)**. Such characteristics are amenable for fibril formation in some structural homologs such as FG10600.1, whereby the structural conservation in this solenoid allowed for HET-s and FG10600.1 amyloid cross-seeding experiments **[10]**. However, a closer inspection of structural homologs to the PFD indicates that the potential for salt-bridge formation is largely lacking, with several homologs only partaking in one possible salt-bridge pair compared to the 3 salt bridges in HET-s **(Figure 4.5)**. Additionally, in at least three of the structural homologs we have analyzed, we observe a discrepancy in the length of the rungs composing the pseudorepetitive solenoid, such that the first rung is shorter than the second rung in the solenoid monomer. If these homologs do indeed form fibrils, they would be built on the

stacking of structurally different units, and as such, there would a noticeable "shift" in the hydrophobic core, asparagine ladders, and salt bridges between different units of the solenoid. These shifts in the inter- and intra-molecular bonds of the solenoid monomers may hinder stability of the resultant fibril; this remains to be determined by experimental analysis. Based on our analysis however, the contribution of the HET-s shape to future amyloid forming proteins is quite limited, and for many of the structural homologs that can adopt that shape, structural and energetic hindrances would need to be overcome before formation of a stable fibril.

We have performed an evolutionary, functional, and structural bioinformatics analysis of homologs to the HET-s prion-forming domain, and we compare our findings against the identified homologs of the HET-s N-terminal domain. Based on phylogenetic analysis, we conclude that the HET-s PFD has a limited phyletic distribution in the kingdom of life, especially within fungi, but is also highly evolving compared to the N-terminal domain. Using fold recognition techniques, we have predicted a set of PFD homologous structures which are amenable to adopting a β-solenoid fold, but which lack many of the characteristics of the HET-s solenoid that promote the formation of stable fibrils. Accordingly, we conclude that the HET-s shape has 'limited scope' for amyloidosis across the wider protein universe. Additionally, we assessed the tandem evolution of the HET-s N-terminal and prion-forming domains and identified functional linkages of the N-terminal homologs. Our research suggests that the HET-s N-terminal domain has a widespread phyletic distribution and may contribute to several protein-protein interactions besides heterokaryon incompatability.

4.6 REFERENCES

1. Kajava AV, Baxa U, Wickner RB, Steven AC (2004) A model for Ure2p prion filaments and other amyloids: The parallel superpleated Î²-structure. Proceedings of the National Academy of Sciences of the United States of America 101: 7885-7890.
2. Kajava AV, Squire JM, Parry DAD, Andrey Kajava JMS, David ADP (2006) [beta][hyphen (true graphic)]Structures in Fibrous Proteins. Advances in Protein Chemistry: Academic Press. pp. 1-15.
3. Kajava AV, Steven AC, Andrey Kajava JMS, David ADP (2006) [beta][hyphen (true graphic)]Rolls, [beta][hyphen (true graphic)]Helices, and Other [beta][hyphen (true graphic)]Solenoid Proteins. Advances in Protein Chemistry: Academic Press. pp. 55-96.
4. Dobson CM (2005) Structural biology: Prying into prions. Nature 435: 747-749.
5. Nelson R, Sawaya MR, Balbirnie M, Madsen AO, Riekel C, et al. (2005) Structure of the cross-[beta] spine of amyloid-like fibrils. Nature 435: 773-778.
6. Van Melckebeke Hln, Wasmer C, Lange A, Ab E, Loquet A, et al. (2010) Atomic-Resolution Three-Dimensional Structure of HET-s(218-289) Amyloid Fibrils by Solid-State NMR Spectroscopy. Journal of the American Chemical Society 132: 13765-13775.
7. Wasmer C, Lange A, Van Melckebeke H, Siemer AB, Riek R, et al. (2008) Amyloid Fibrils of the HET-s(218-289) Prion Form a {beta} Solenoid with a Triangular Hydrophobic Core. Science 319: 1523-1526.
8. Greenwald J, Buhtz C, Ritter C, Kwiatkowski W, Choe S, et al. (2010) The Mechanism of Prion Inhibition by HET-S. Molecular Cell 38: 889-899.
9. Saupe SJ (2011) The [Het-s] prion of Podospora anserina and its role in heterokaryon incompatibility. Seminars in Cell & Developmental Biology In Press, Corrected Proof.
10. Wasmer C, Zimmer A, Sabaté R, Soragni A, Saupe SJ, et al. (2010) Structural Similarity between the Prion Domain of HET-s and a Homologue Can Explain Amyloid Cross-Seeding in Spite of Limited Sequence Identity. Journal of Molecular Biology 402: 311-325.
11. Balguerie A, Reis SD, Ritter C, Chaignepain S, Coulary-Salin B, et al. (2003) Domain organization and structure-function relationship of the HET-s prion protein of Podospora anserina. EMBO J 22: 2071-2081.
12. Coustou V, Deleu C, Saupe S, Begueret J (1997) The protein product of the het-s heterokaryon incompatibility gene of the fungus Podospora anserina behaves as a prion analog. Proceedings of the National Academy of Sciences 94: 9773-9778.
13. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, et al. (2009) Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 37: D5-D15.
14. Cuomo CA, Birren BW, Jonathan W, Christine G, Gerald RF (2010) The Fungal Genome Initiative and Lessons Learned from Genome Sequencing. Methods in Enzymology: Academic Press. pp. 833-855.

15. Altschul SF, Madden TL, SchÃ¤ffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389-3402.
16. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763.
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of Molecular Biology 215: 403-410.
18. Geer LY DM, Lipman DJ, Bryant SH (2002) CDART: protein homology by domain architecture. Genome Res 12: 1619-1623.
19. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Research 38: D211-D222.
20. Lobley A, Sadowski MI, Jones DT (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics 25: 1761-1767.
21. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al. (2002) Comparative Protein Structure Modeling Using Modeller. Current Protocols in Bioinformatics: John Wiley & Sons, Inc.
22. Laskowski R A MMW, Moss D S and Thornton J M (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26: 283-291.
23. Laskowski RA, Chistyakov VV, Thornton JM (2005) PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids. Nucleic Acids Research 33: D266-D268.
24. Schrodinger, LLC (2010) The PyMOL Molecular Graphics System, Version 1.3r1.
25. Costantini S, Colonna G, Facchiano AM (2008) ESBRI: a web server for evaluating salt bridges in proteins. Bioinformation 3: 137-138.
26. Gough J, Chothia C (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucl Acids Res 30: 268-272.
27. Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. Journal of Molecular Biology 313: 903-919.
28. Fedorova N, Badger J, Robson G, Wortman J, Nierman W (2005) Comparative analysis of programmed cell death pathways in filamentous fungi. BMC Genomics 6: 177.
29. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. (2009) CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Research 37: D205-D210.
30. Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. Nucleic Acids Research 32: W327-W331.
31. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, et al. (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research 39: D225-D229.
32. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23: 127-128.
33. Felsenstein J (1989) PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166.

34. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792-1797.
35. Chevenet F, Brun C, Banuls A-L, Jacq B, Christen R (2006) TreeDyn: towards dynamic graphics and annotations for analyses of trees. BMC Bioinformatics 7: 439.
36. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, et al. (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Research 36: W465-W469.
37. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, et al. (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Research 31: 3497-3500.
38. Bork P (1993) Hundreds of ankyrin-like repeats in functionally diverse proteins: Mobile modules that cross phyla horizontally? Proteins: Structure, Function, and Bioinformatics 17: 363-374.
39. Li D, Roberts R (2001) Human Genome and Diseases:¶WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. Cellular and Molecular Life Sciences 58: 2085-2097.
40. Smith TF, Gaitatzes C, Saxena K, Neer EJ (1999) The WD repeat: a common architecture for diverse functions. Trends in Biochemical Sciences 24: 181-185.
41. Blatch GL, Lässle M (1999) The tetratricopeptide repeat: a structural motif mediating protein-protein interactions. BioEssays 21: 932-939.
42. Bai G, Shaner G (2004) Management and resistance in wheat and barley to Fusarium head blight. Annual Review of Phytopathology 42: 135-161.
43. Takken F, Rep M (2010) The arms race between tomato and Fusarium oxysporum. Molecular Plant Pathology 11: 309-314.
44. Choi JH, Govaerts C, May BCH, Cohen FE (2008) Analysis of the sequence and structural features of the left-handed β-helical fold. Proteins: Structure, Function, and Bioinformatics 73: 150-160.
45. Govaerts C, Wille H, Prusiner SB, Cohen FE (2004) Evidence for assembly of prions with left-handed beta-helices into trimers. Proc Natl Acad Sci USA 101: 8342 - 8347.
46. Choi JH, May BCH, Wille H, Cohen FE (2009) Molecular Modeling of the Misfolded Insulin Subunit and Amyloid Fibril. Biophysical Journal 97: 3187-3195.
47. Guo J-t, Wetzel R, Xu Y (2004) Molecular modeling of the core of Aβ amyloid fibrils. Proteins: Structure, Function, and Bioinformatics 57: 357-364.
48. Langedijk JPM, Fuentes G, Boshuizen R, Bonvin AMJJ (2006) Two-rung Model of a Left-handed [beta]-Helix for Prions Explains Species Barrier and Strain Variation in Transmissible Spongiform Encephalopathies. Journal of Molecular Biology 360: 907-920.
49. Stork M, Giese A, Kretzschmar HA, Tavan P (2005) Molecular Dynamics Simulations Indicate a Possible Role of Parallel [beta]-Helices in Seeded Aggregation of Poly-Gln. Biophysical Journal 88: 2442-2451.
50. Zanuy D, Gunasekaran K, Lesk AM, Nussinov R (2006) Computational Study of the Fibril Organization of Polyglutamine Repeats Reveals a Common Motif Identified in [beta]-Helices. Journal of Molecular Biology 358: 330-345.

51. Iconomidou VA, Vriend G, Hamodrakas SJ (2000) Amyloids protect the silkmoth oocyte and embryo. FEBS Letters 479: 141-145.

4.7 SUPPLEMENTAL DATA

Supplemental data includes Supplementary Methods (2 lists) and Supplementary Results in the form of 1 figure and 4 tables. Supplemental data can be found in **Appendix C**, as well as online with this article at:

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0027342

CHAPTER V

Conclusions

Identification of ambiguous encoding in protein secondary sequences and tertiary structure is paramount to develop an understanding of key protein segments underlying many conformational diseases. Critical assessment of these structurally-ambivalent peptides is especially important for virulent proteins whose amyloid formation leads to a vast array of neurological disorders. Much of the research in this thesis centers specifically on the prion protein, for which a self-templating mechanism towards fibril and amyloid formation has been proposed, and which has become a precedent for a growing paradigm witnessed in many amyloid-forming proteins. The analysis on PrP serves as a template for which similar studies can be conducted on amyloid-forming proteins, given the parallels observed in amyloid formation, structures and morphologies across these proteins. The presented thesis exploits on a common theme underlying amyloid formation: the necessity for a protein to adopt a conformational change into a β-sheet rich isoform that facilitates its packing into pathogenic assemblies. To this end, Chapters II and III analyze for conformational change in PrP and amyloid-forming proteins via bioinformatic analyses of secondary sequences and tertiary protein structures, respectively. Chapter IV analyzes for conformational change on a grander scope, by attempting to identify protein sequences that are able to adopt an alternative fold that resembles that of an amyloid fibril.

Prediction and identification of amyloidogenic segments in proteins remains a compelling and challenging endeavor, despite experimental and computational advancements in this area. The premise that conformationally-variable segments in proteins are potentially amyloidogenic is a straightforward one, but as I have demonstrated in Chapter II, this relationship is not clear-cut across the protein universe. Understandably, secondary structural conformational change is an intrinsic factor in algorithms that predict amyloidogenic segments in proteins, and its exact association with amyloidogenicity warrants a study in its own right. The novelty of the research presented in Chapter II is the elucidation of the statistical relationship between amyloidogenicity and each of the structurally ambivalent discordant and chameleon peptide classes. Despite that discordance was hypothesized to be associated with amyloidogenicity more than a decade ago, the work presented here demonstrates, for the first time, that discordance is weakly tied to amyloidogenicity and may serve instead other protein

functions that rely on conformational variability, such as protein-ligand interactions and viral replication. Compared to discordance, the natural phenomenon of chameleonism observed in proteins is no more advantageous in amyloidogenesis, as enrichment of chameleon peptides in amyloid-forming proteins is severely lacking. From a biological perspective, this study categorically refutes the erroneous assumption that 'every protein segment demonstrating conformational change from α→β secondary structure is equally prone to amyloidogenicity'. From a computational perspective, this study stresses the shortcomings of secondary structure prediction algorithms to predict amyloidogenic segments. Indeed, identified enrichment levels of discordant and chameleon peptides in amyloid-forming proteins, however slight, are heavily dependent on the predictive algorithm being used to identify them. Such inconsistencies are problematic for sequence-based amyloid prediction algorithms that rely on 'secondary structure propensity' as a main factor in their decision making.

Assessment of protein flexibility by analysis of tertiary protein structures is a logical step that overcomes many limitations posed by secondary-structure algorithms in proper prediction of amyloidogenic segments. Computational modeling of protein folding facilitates the understanding of dominant protein motions underlying key biological functions, knowledge that can be harnessed to improve prediction of amyloidogenic segments in proteins involved in conformational disease. While this seems to be a forgone conclusion, the greatest caveat of such an approach is extracting true biological information from hordes of background noise. Experimental methods towards elucidation of structures, such as X-ray and NMR, are designed mainly to provide a static 3D representation of biomolecules, and accordingly, fail to provide a full description of structural changes undergone by proteins as function of time **[1]**. Overcoming this constraint using Molecular Dynamic simulations has introduced its own problems however, mainly the requirement for long time-scales to observe a single folding event, and a large number of trajectories to obtain reasonable statistics and provide a complete picture of the folding process **[2]**. Chapter III adopts a different approach that maximizes on the structural information provided by X-ray and NMR structures while refraining from the multidimensional complexities of MD simulations.

The presented PCA analysis in Chapter III adopts a novel and interesting approach to analyzing protein motions within prion proteins. While the use of PCA in multidimensionality reduction and elucidation of protein motion is not a novel concept in itself, the novelty of the presented work is two-fold, both biological and computational. Biologically, this novelty stems from the application of the technique onto native globular structures of the Prion family, the first study of its kind for PrP, and indeed any amyloid-forming protein. Accordingly, this analysis exemplifies futuristic studies that can be conducted on amyloid-forming proteins and the benefits that can be reaped from such an analysis. Computationally, the application of the technique onto multiple NMR structures represents a new target base for PCA-based analysis, and the use of a PC-based domain ranking method represents a new overture towards utilizing knowledge reaped from PCA in understanding biological phenomena.

The proposed PC-based domain rankings presented in Chapter III succeed in reflecting species variation and TSE susceptibility of PrPs, and have applications in future studies of PrP mutant structures in specific species and across the wider spectrum of TSE-non-susceptible or TSE-susceptible prion structures. In the analysis of human PrP mutant structures for example, I have demonstrated domain perturbations arising from each mutation that reflect upon the human TSE diseases of GSS, FFI, and CJD. The highest-ranking domains in terms of their importance with each of the diseases indicates 'localized hot spots' within these mutant structures that can be further assessed experimentally, and which can ultimately serve as potential target sites for inhibitors against that particular disease. An analysis of PrP structures across multiple species has also succeeded in identifying "structural signatures" for PrPs of different evolutionary groups. This is particularly intriguing for TSE transmissibility studies, given the added observation that the structural differentiation observed in the PrP structures does not necessarily coincide with evolutionary divergence. These results raise new questions about the possible interplay between different structural solutions to TSE resistance.

The results presented in Chapter III are strengthened by experimental and computational observations that reflect the importance of the identified domains in the PrP conversion process, as well as the strength of the employed technique. As we have demonstrated in the discussion Chapter III, the results are supported by experimental and

computational analyses that have highlighted variable domains of PrP during the PrP conversion process. Arguably, another way to confirm validity of the conformationally flexible domains identified here is to compare them against studies that have identified rigid domains of PrP, which would represent the 'reverse complement' of this analysis. PCA had been applied on MD simulations of PrP **[3,4]** to characterize local flexibility in 8 different species, including human, cow, elk, cat, hamster, chicken, turtle and frog. In these studies, the authors identified 'dynamic domains' that represent correlated motion of protein movement (same direction, speed, and time) throughout a simulation; these dynamic domains represent rigid domains of the protein. Based on this analysis, the H2 & H3 helices represent the largest dynamic domain, followed by H1, while the S2-H2 loop exhibits varying degrees of flexibility in TSE-resistant species **[3,4]**. These results complement the presented work in Chapter IV and indicate that the method presented in this thesis is can produce competent results in a faster and reliable manner.

Collectively, Chapters II and III test for conformational changes in secondary sequences and tertiary structures of prions and amyloid-forming proteins. The latter analysis in Chapter III suggests key areas of PrP that are prone to conversion between the PrP^{C} and PrP^{SC} isoforms. While this conversion process remains disputable, an equally interesting puzzle, whether in PrP or other amyloidogenic proteins, is the final adopted shape of an amyloid fibril. As previously discussed, this shape has been alluded to experimentally in the form a common cross-β spine core and parallel in-register arrangements on protein peptides, as well as speculative LβH-based structures derived from molecular modeling and simulation exercises. The recently solved amyloid fibril structure of HET-s raises the question of whether the HET-s solenoid fold serves as an archetype for the complete atomic structure of all amyloid fibrils. At the onset, the HET-s solenoid fold, being a left-handed beta-helix with the common characteristics of a cross-β spine arrangement, seems to be an ideal prototype. Arguably however, LβH-structures have not been documented in humans or mammals, which encompass the larger number of pathogenic, amyloid-forming proteins. The analysis presented in Chapter IV answers these contradictions by assessing the prevalence of the HET-s β-solenoid across the wider protein universe. Such an assessment also serves the added purpose of identifying new, potential amyloid forming proteins that may adopt the solenoid fold in their native form

or fibril states. The limited scope of the HET-s solenoid fold in amyloidosis, as demonstrated using structural and evolutionary analyses, indicates that the HET-s solenoid fold is not the typical atomic structure of an amyloid fibril. This still however leaves the possibility of 'generic' left-handed beta-helix amyloid fold, but that remains to be discovered by experimental methods.

The research presented within this thesis lays the groundwork for similar and future studies that can be conducted on known amyloid-forming proteins or potential amyloidogenic candidates. Detection of amyloidogenic segments in proteins remains of paramount importance towards understanding the role these peptides and domains play in amyloidosis in conformational diseases.

REFERENCES:

1. Teodoro ML, Phillips GN, Kavraki LE: Understanding Protein Flexibility through Dimensionality Reduction. Journal of Computational Biology 2003, 10(3-4):617-634.

2. Freddolino PL, Harrison CB, Liu Y, Schulten K: Challenges in protein-folding simulations. Nat Phys 2010, 6(10):751-758.

3. Blinov N, Berjanskii M, Wishart DS, Stepanova M: Structural Domains and Main-Chain Flexibility in Prion Proteins. Biochemistry 2009, 48(7):1488-1497.

4. Santo KP, Berjanskii M, Wishart DS, Stepanova M: Comparative analysis of essential collective dynamics and NMR-derived flexibility profiles in evolutionarily diverse prion proteins. Prion 2011, 5(3).

APPENDIX

APPENDIX A

Supplemental Data for Chapter II

Discordance and chameleon sequences: Their distribution and implications for amyloidogenicity

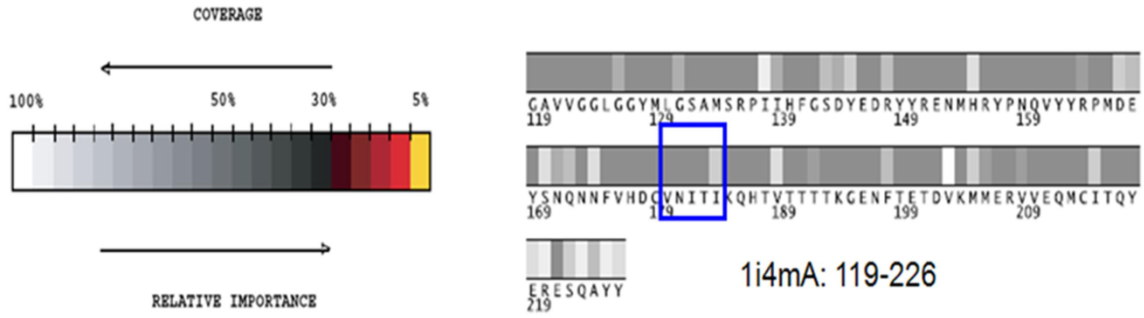
A.1: Figures

FIGURE S2.1

(A)

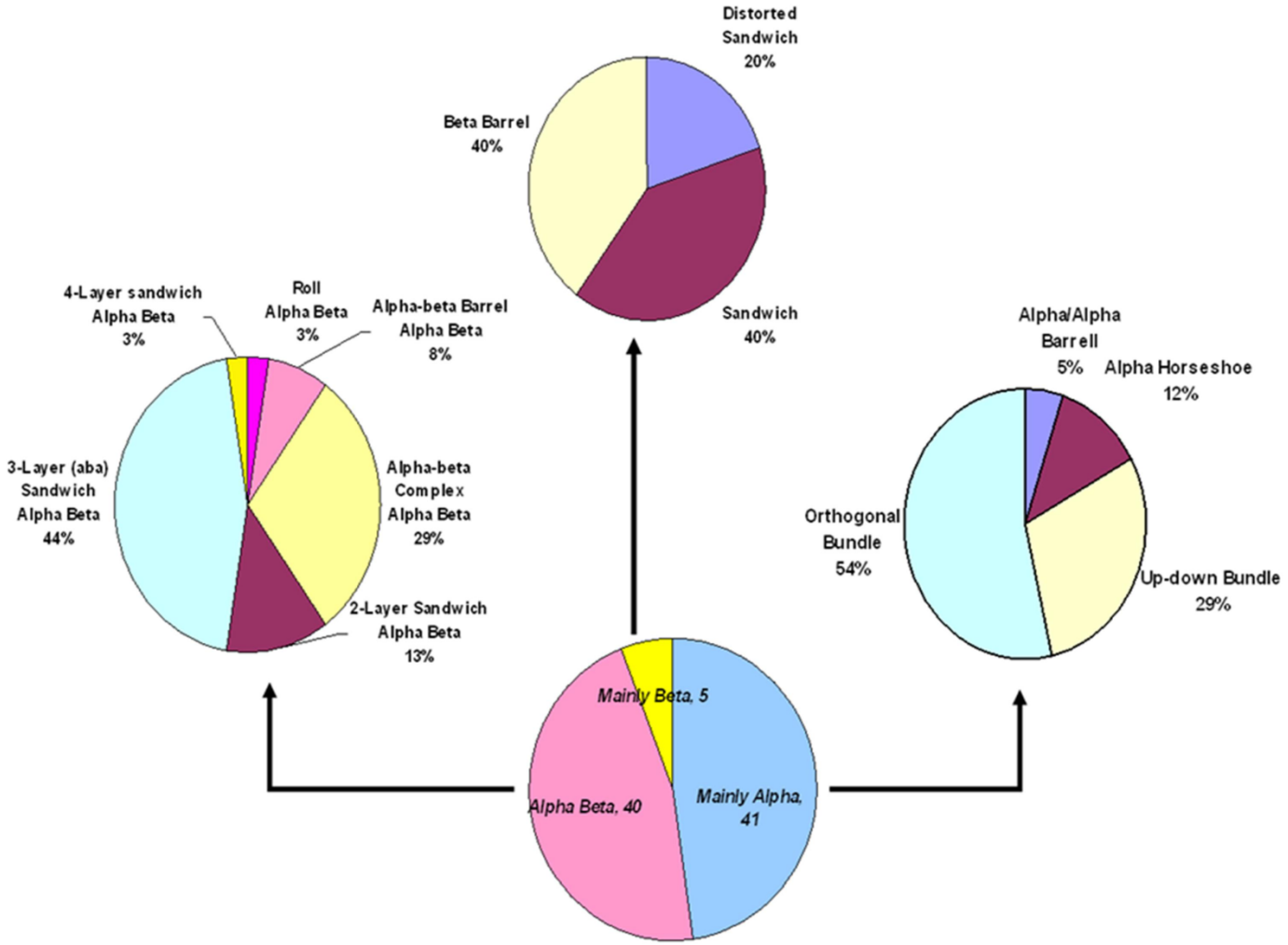
1I4M:A							
Residue Positions	Amino Acid	Normalized Score	Cons.	MSA Data Residue Variety Per Position		Residues Variety Per Position in MSA (%)	
				MSA	Residues	Highest	Lowest
62	V	-0.511	9	16/16	V	V: 100	0
63	N	-0.513	9	16/16	N	N: 100	0
64	I	-0.512	9	16/16	I	I: 100	0
65	T	-0.516	9	16/16	T	T: 100	0
66	I	0.023	5*	16/16	I,V	V: 62.5	I: 37.5

(B)



Conservation analysis for the PrP Protein family

FIGURE S2.2



Distribution of CATH architectures in 86 discordant proteins

A.2: Tables

TABLE S2.1: Complete List of 119 discordant stretches identified from the SCOP domain dataset. The amino acids positions of the stretches (numbered from the start of each protein chain) are recorded. Details of the discordant protein (PDB, chain, CATH, Protein representations, and species) are shown.

A Snapshot of the table is shown here. For the entire table, please consult the online supplementary material of this article (Supporting Information Table I) at: http://onlinelibrary.wiley.com/doi/10.1002/pro.590/full

Uniprot Accession	SCOP domain	PDB & Chain	Discordant Stretch	Length	Region Area	CATH	Protein	Organism
P03485	d1aa7a_	1aa7:A	fvftlt	6	62-67	Mainly alpha, up-down bundle	Influenza virus matrix protein M1	Influenza A virus [TaxId: 11320]}
P03697	d1avqa_	1avq:A	smwvt	5	163-167	Alpha Beta, alpha-beta complex	lambda exonuclease	Bacteriophage lambda [TaxId: 10710]}
P03697	d1avqa_	1avq:A	rtlfef	6	91-96	Alpha Beta, alpha-beta complex	lambda exonuclease	Bacteriophage lambda [TaxId: 10710]}
P00709	d1b9oa_	1b9o:A	lictmf	6	26-31	Mainly alpha, orthogonal bundle	alpha-Lactalbumin	Human (Homo sapiens) [TaxId: 9606]}
P22915	d1bjaa_	1bja:A	tyiik	5	4-8	Mainly alpha, orthogonal bundle	Transcription factor MotA, activation domain	Bacteriophage T4 [TaxId: 10665]}

TABLE S2.2:

Discordant stretches from SCOP which exhibit chameleon conformational properties. 28 discordant stretches with chameleonic properties are identified. Corresponding chameleon proteins are represented in the format (PDB:chain).DS=Discordant Stretch

PDB	DS Region	DS Sequence	Protein	Chameleon Sequences
1kiy:A	94-99	TYTLVL	Trichodiene synthase	1v5v:A, 1v5v:B
1i4m:A	62-66	VNITI	Prion protein domain	1e4k:C, 1e4j:A, 1fnl:A, 1hf1, 1op8:F, 1op8:C, 1op8:D, 2vov:A, 2vow:A, 2vox:A
1tca:A	232-236	FSYVV	Triacylglycerol lipase	1iic:A, 1iic:B, 1iid:A, 2nmt:A, 2p6e:F, 2p6e:E, 2p6e:C, 2p6e:A, 2p6e:B, 2p6e:D, 2p6f:F, 2p6f:E, 2p6f:C, 2p6f:A, 2p6f:B, 2p6f:D, 2p6g:F, 2p6g:E, 2p6g:C, 2p6g:A, 2p6g:B, 2p6g:D
1jma:A	233-237	VYSLK	HSV glycoprotein D	1a22:B, 1axi:B, 1hwg:B, 1hwg:C, 1hwh:B, 3hhr:B, 3hhr:C, 1kf9:F, 1kf9:E, 1kf9:C, 1kf9:B, 2aew:A, 2aew:B
1gux:B	108-112	IIVFY	Retinoblastoma tumor suppressor domains	1amo:A, 1amo:B, 1b1c:A, 1j9z:A, 1j9z:B, 1ja0:A, 1ja1:A, 1ja1:B
1u5u:A	346-350	LRVAV	Allene oxide synthase-lipoxygenase protein, N-terminal domain	1bl7:A, 1di9:A, 1kv1:A, 1wfc:A, 1nh7:A, 1nh8:A, 1ouk:A, 1ouy:A, 1ove:A, 1oz1:A, 1r3c:A, 1w7h:A, 1w82:A, 1w83:A, 1w84:A, 1wbn:A, 1wbo:A, 1wbs:A, 1wbt:A, 1wbv:A, 1wbw:A, 1yqj:A, 1zyj:A, 1zz2:A, 2bak:A, 2bal:A, 2baq:A, 2cvh:A, 2cvh:B, 2fju:B, 2fsl:X, 2fsm:X 2fso:X, 2fst:X, 2gfs:A, 2i0h:A, 2npq:A, 2okr:A, 2okr:D, 2onl:B, 2onl:A, 2p5a:A, 2pkj:A, 2ptj:A, 2pv5:A, 2pv8:A, 2rg6:A, 3bv2:A, 2bv3:A, 3bx5:A, 3c5u:A, 3d7z:A, 3d83:A, 3dt1:A, 3e92:A, 3e93:A
1i17:A	27-31	RYYAA	Prion-like protein Doppel	1cfr:A
1fjh:A	82-86	VVSVN	3-alpha-	1dfa:A, 1ef0:A, 1vde:A, 1vde:B,

			hydroxysteroid dehydrogenase	2thi:A, 2thi:B, 3thi:A, 1jva:A, 1jva:B, 1lws:A, 1um2:A, 1um2:B, 2r4i:C, 2r4i:A, 2r4i:B, 2r4i:D
1ccw:B	306-310	GVIVT	Glutamate mutase, large subunit	1hzp:A, 1hzp:B, 1m1m:A, 1m1m:B, 1okk:D, 1rj9:A, 1u6e:A, 1u6e:B, 1u6s:A, 1u6s:B, 2ahb:A, 2ahb:B, 2aj9:A, 2aj9:B, 2cnw:F, 2cnw:E, 2cnw:D, 2iyl:D, 2j7p:E, 2j7p:D, 2q9a:A, 2q9a:B, 2q9b:A, 2q9b:B, 2q9c:A, 2q9c:B, 2qnx:A, 2qnx:B, 2qny:A, 2qny:B, 2qnz:A, 2qnz:B, 2qO0:A, 2qO0:B, 2qO1:A, 2qO1:B, 2qx1:A, 2qx1:B, 3dii:A, 3dii:B, 3dij:A, 3dij:B
1f9a:A	80-84	IWVSY	Nicotinamide mononucleotide (NMN) adenylyltransferase	1jnd:A, 1jne:A
1gxy:A	70-74	TALVA	Eukaryotic mono-ADP-ribosyltransferase ART2.2	1jxh:A, 1jxi:A, 1jxi:B, 2eay:A, 2eay:B, 2uzh:C, 2uzh:A, 2uzh:B, 3ddy:A, 1llj:A
1mgp:A	259-263	VITTH	Hypothetical protein TM841	1r1a:1, 2hwd:1, 2hwe:1, 2hwf:1, 1r1a:4
1v74:A	99-103	RIYLE	Colicin D nuclease domain	1s3o:A, 1s3o:B, 2dud:A, 2dud:B,3ull:A, 3ull:B
1iu4:A	66-70	GVTWV	Microbial transglutaminase	1orv:C, 1orv:B, 1orv:A, 1orv:D, 1orw:C, 1orw:A, 1orw:B, 1orw:D, 2aj8:A, 2aj8:B, 2aj8:D, 2ajb:C, 2ajb:A, 2ajb:B, 2ajb:D, 2ajc:C, 2ajc:A, 2ajc:B, 2ajc:D, 2ajd:C, 2ajd:B, 2ajd:D, 2bua:C, 2bua:A, 2bua:B, 2bua:D, 2buc:C, 2buc:A, 2buc:B, 2buc:D
1nth:A	273-277	TTIVD	Monomethylamine methyltransferase MtmB	1nfg:C, 1nfg:A, 1nfg:B, 1nfg:D, 1nu5:A
1ors:C	21-25	IVVVV	Potassium channel KVAP	2iut:A, 2iut:B, 2iuu:F, 2iuu:A, 2iuu:E, 2iuu:B, 2iuu:D, 2iuu:C
1v32:A	67-71	RLVLL	Hypothetical protein AT5G08430 (rafl09-47-k03)	1zsw:A, 2dbs:A, 2dbs:B, 2i6t:A, 2i6t:B, 2k6v:A, 3dl2:A, 3dl2:B

1k8k:G	129-133	IVRVL	Arp2/3 complex 16 kDa subunit ARPC5	2dgy:A
1w7b:A	267-271	VLIRI	Annexin II	2h6e:A
3pvi:A	88-92	VIIAK	Restriction endonuclease PvuII	2g8l:A, 2g8l:B, 3bgw:A, 3bh0:A,
1jb0:A	399-403	IFMVR	Apoprotein a1, PsaA	2vvg:A, 2vvg:B
1kf6:D	14-18	FWGLF	Fumarate reductase subunit FrdD	3cmg:A
2mev:1	33-37	VAFFY	Mengo encephalomyo carditis virus coat proteins	1tjy:A, 1tm2:A
1xg7:A	149-153	IVFTV	Hypothetical protein PF0904	2hew:F, 2hey:F, 2hey;G
1muk:A	505-509	SVAIL	Reovirus polymerase lambda3	1knx:C, 1knx:B, 1knx:D
1otk:A	6-10	TAYTL	Phenylacetic acid degradation protein PaaC	1lpd:A, 1rl0:A
1r1g:A	10-14	VKCVA	Neurotoxin bmk37	1bdg:A
1gu2:A	32-36	KIFFN	Cytochrome c"	1d9k:A, 1kb5:H

TABLE S2.3:

Complete list of discordant and chameleon segments in the prion-like superfamily. For segments which are also chameleonic, corresponding PDBs **(PDB:Chain)** are shown. Chameleon sequences recorded are only those which are observed in known helices. Predicted fibril-forming hexapeptides are provided for the entire protein sequence.

A Snapshot of the table is shown here. For the entire table, please consult the online supplementary material of this article (Supporting Information Table III) at: http://onlinelibrary.wiley.com/doi/10.1002/pro.590/full

Prion Database Entry	Species	Discordant Segment & Sequence	Chameleon Segment & Sequence	Predicted Fibril Hexapeptide
1xyx:A	PrP Mouse [Mus musculus]	60-64: VNITI	82-86 : DVKMM	1-6: VVGGLG
			87-91 : ERVVE	16-21: RPMIHF
			80-84 : ETDVK	18-23: MIHFGN
			88-92 : RVVEQ	19-24: IHFGND
			60-64 : VNITI	26-31: EDRYYR
			95-99 : VTQYQ	2-7: VGGLGG
			89-93 : VVEQM	27-32: DRYYRE
				28-33: RYYREN
				29-34: YYRENM
				30-35: YRENMY
				32-37: ENMYRY
				56-61: VHDCVN
				57-62: HDCVNI
				58-63: DCVNIT
				59-64: CVNITI
				60-65: VNITIK
				61-66: NITIKQ
				62-67: ITIKQH
				63-68: TIKQHT
				64-69: IKQHTV
				65-70: KQHTVT
				66-71: QHTVTT
				67-72: HTVTTT
				68-73: TVTTTT
				86-91: MERVVE
				87-92: ERVVEQ
				88-93: RVVEQM
				89-94: VVEQMC
				90-95: VEQMCV
				91-96: EQMCVT
1b10:A	PrP Golden hamster [Mesocricetus auratus]	56-60: VNITI	83-87 : ERVVE	22-27: EDRYYR
			5-9 : MLGSA	23-28: DRYYRE
			62-66 : QHTVT	24-29: RYYREN
			84-88 : RVVEQ	25-30: YYRENM
			77-81 : TDIKI	44-49: QYNNQN
			56-60 : VNITI	52-57: VHDCVN
			85-89 : VVEQM	53-58: HDCVNI
				54-59: DCVNIT
				55-60: CVNITI
				56-61: VNITIK
				57-62: NITIKQ

TABLE S2.4: Discordant Proteins with metal-ion binding properties. All discordant proteins that are metal-ion binding are listed (first row). Subsequent rows indicate discordant proteins associated with GO children terms (numbered according to GO tree). Proteins which exhibit hydrolase activity are in bold & italics.

Molecular Function	Number of Discordant Proteins	Discordant Proteins
Metal ion binding GO:0046872	15	Single-stranded DNA-binding protein; Polyphenol oxidase I, chloroplastic; Vanadium chloroperoxidase; Cytochrome c"; ***Lambda Exonuclease;*** ***Peptide deformylase;*** Hemophore HasA; ***Peptidyl-Lys metalloendopeptidase; Type-2 restriction enzyme NgoMIV;*** Allene oxide synthase-lipoxygenase protein; Cytochrome c oxidase subunit 1; Cytochrome b6; Photosystem I P700 chlorophyll a apoprotein A1; ***Type-2 restriction enzyme PvuII; Extracellular small neutral protease***
1. Calcium Ion Binding GO:0005509	6	Annexin; Annexin A2; SPARC; Alpha-lactalbumin; Allene oxide synthase-lipoxygenase protein; ***Extracellular small neutral protease***
2. Magnesium Ion Binding GO:0000287	6	BstYI; ***Lambda Exonuclease***; Intron-associated endonuclease 1; Photosystem I P700 chlorophyll a apoprotein A1; ***Type-2 restriction enzyme NgoMIV;*** ***Type-2 restriction enzyme PvuII***
3.Transition Metal Ion Binding GO:0046914	------	----------
3A.Copper Ion Binding GO:0005507	4	Major prion protein; SPARC; Cytochrome c oxidase subunit 1; Polyphenol oxidase I, chloroplastic
3B. Iron Ion Binding GO:0005506	6	Cytochrome c"; Cytochrome b6; ***Peptide deformylase***; Hemophore HasA; Allene oxide synthase-lipoxygenase protein; Photosystem I P700 chlorophyll a apoprotein A1;
3C. Zinc Ion Binding GO:0008270	3	Single-stranded DNA-binding protein; ***Peptidyl-Lys metalloendopeptidase***; ***Extracellular small neutral protease***;
3D. Cobalt Ion Binding GO:0050897	2	Methylaspartate mutase E chain; Adenosylcobalamin-dependent ribonucleoside-triphosphate reductase

TABLE S2.5:

List of Pathogenic and Non-pathogenic Amyloid-forming Proteins (n=50) used in this study for chameleon analysis.

Uniprot Identifier	PDB	Protein	Pathogenic (PA) or Non-pathogenic (NA)
ACYP2_HORSE	1APS	Acylphosphatase	PA
ADH1_YEAST	2HCY	Alcohol dehydrogenase 1, Saccharomyces cerevisiae	NA
HMGB1_RAT	1CKT	Amphoterin, rat	PA
HMGB1_RAT	2GZK	Amphoterin, rat	PA
A4_HUMAN	1AAP	Amyloid beta A4 protein	PA
A4_HUMAN	1MWP	Amyloid beta A4 protein	PA
A4_HUMAN	1RW6	Amyloid beta A4 protein	PA
A4_HUMAN	2FMA	Amyloid beta A4 protein	PA
A4_HUMAN	2G47	Amyloid beta A4 protein	PA
A4_HUMAN	3DXE	Amyloid beta A4 protein	PA
ANDR_HUMAN	2AX6	androgen receptor protein	PA
ANDR_HUMAN	3BTR	androgen receptor protein	PA
CY552_HYDTT	1YNR	Apo-cytochrome C552, H. thermophilus	NA
APOA1_HUMAN	2A01	Apolipoprotein A1	PA
APOA2_HUMAN	2OU1	Apolipoprotein A2	PA
APOC2_HUMAN	1I5J	Apolipoprotein C-II	NA
SYUA_HUMAN	1XQ8	a-synuclein	PA
ATX1_HUMAN	1OA8	Ataxin-1	PA
ATX2_HUMAN	3KTR	Ataxin-2	PA
ATX3_HUMAN	2JRI	Ataxin-3	PA
ANF_HUMAN	1YK0	Atrial Naturetic factor	PA
B2MG_HUMA	1K5N	B2 Microglobulin	PA

N			
LACB_BOVIN	1BEB	B-lactoglobulin	NA
F13A_HUMAN	1EX0	Coagulation factor XIII, H. sapiens	NA
CSPB_BACSU	2ES2	Cold Shock Protein, cspB, Bacilus subtilis	NA
CYTC_HUMAN	3GAX	Cystatin C	PA
CY1_BOVIN	1PPJ	Cytochrome c, B.taurus	NA
COIA1_HUMAN	1BNL	Endostatin	NA
COIA1_HUMAN	3HSH	Endostatin	NA
FIBA_HUMAN	1FZD	Fibrinogen a-chain	PA
FINC_MOUSE	1MFN	Fibronectin, Mus musculus	NA
GAGA_DROME	1YUI	GAGA factor, Drosophila	NA
GELS_HUMAN	1KCQ	Gelsolin	PA
GELS_HUMAN	1T44	Gelsolin	PA
GELS_HUMAN	2FH1	Gelsolin	PA
Q03689	2RNM	Het-S	NA
GB_HHV1K	2GUM	HSV glycoprotein B	NA
CYTB_HUMAN	2OCT	Human Stefin B/Cystatin B	PA
INS_HUMAN	1MSO	Insulin	PA
IAPP_HUMAN	3G7W	Islet Amyloid Polypeptide	PA
BGH3_HUMAN	2VXP	Kerato-epithelin/Transforming growth factor-beta-induced protein ig-h3	PA
TTHY_HUMAN	1F86	Lactoferrin/lactotransferrin	PA
TRFL_HUMAN	1H45	Lactoferrin/lactotransferrin	PA
LAMA1_MOUSE	2JD4	Laminin alpha-1 chain, G-like domain, mouse	PA
LYSC_HUMAN	1JSF	Lysozyme	PA
AMPM_PYRFU	1XGS	Methionine aminopeptidase, P. furiosus	NA
MONB_DIOCU	2O9U	Monellin	NA
MYG_HORSE	2V1F	Myoglobin, horse heart	NA

PABP2_HUMAN	3B4D	Nuclear poly(A) binding protein	PA
P53_HUMAN	1DT7	p53 protein	NA
P53_HUMAN	1OLG	p53 protein	NA
P53_HUMAN	2B3G	p53 protein	NA
P53_HUMAN	3D06	p53 protein	NA
P53_HUMAN	3DAC	p53 protein	NA
P85A_BOVIN	1OO3	p85 phosphatidyl inositol-3-kinase (SH3 domain)	NA
P85A_BOVIN	1PNJ	p85 phosphatidyl inositol-3-kinase (SH3 domain)	NA
P85A_BOVIN	1QAD	p85 phosphatidyl inositol-3-kinase (SH3 domain)	NA
P85B_BOVIN	3L4Q	Phosphatidylinositol 3-kinase (PI3-SH3)	NA
PGK_YEAST	1FW8	phosphoglycerate kinase, yeast	NA
PRIO_HUMAN	3HAK	Prion	PA
CBPA2_HUMAN	1DTD	Procarboxypeptidase A2 activation domain, H. sapiens	NA
CBPA2_HUMAN	1O6X	Procarboxypeptidase A2 activation domain, H. sapiens	NA
PRL_HUMAN	3D48	Prolactin	PA
RNSA_STRAU	1LNI	RNase Sa, S. aureofaciens	NA
SODC_HUMAN	1MFM	Superoxide dismutase	PA
TBP_HUMAN	1CDW	TATA-box-binding protein	PA
LIPB_CANAR	3ICV	Triacylglycerol lipase, C. antartica	NA
URE2_YEAST	1K0D	Ure2p yeast	NA

TABLE S2.6:

List of Amyloidogenic Determinants (n=45) for chameleon analysis. For determinants for which a 3D structure is known, the corresponding PDB:Chain and start and end positions are indicated. Of the determinants with a 3D structure representation, 17 determinants had helices >5 residues long (bold).

PROTEIN	AMYLOIDOGENIC DETERMINANT	PDB	Start Pos.	End Pos.
a-Synuclein	**GVATVA** **VGGAVVTGVTAVAQKTV** **GSIAAAT**	1XQ8:A	51 66 86	56 82 92
b-Lactoglobulin	DIQKVAGTWY KYLLFCMENS SLVCQCLVRTP HIRLSFN	1BEB:A	11 101 116 146	20 110 126 152
B2-Microglobulin	SNFLNCYVSGFHPSDIEVDLLK KDWSFYLLYYTEFTPTEKDEYACRVNHVTLSQPKIVKWDRDM	1B0G:B	20 58	41 99
Acylphosphatase, human muscle	**RVQGVCFRMYTEDEAR** SKLEYSNFSIRY	1APS:A	16 87	31 98
Amphoterin, rat	**MSSYAFFVQTCREEHK**	1CKT:A	12	27
Apolipoprotein C-II	MSTYTGIFTDQ	1SOH:A	60	70
Cold shock protein, cspB, *Bacillus subtilis*	MLEGKVKWFNSEKGFGFIEVEGQDDVFVHFSAIQGEGFKTLEEGQA VSFEIVEGNRGPQAANVTKEA	2ES2:A	1	67
Gelsolin	**SFNNGDCFILDLGNNIHQWCGSNSNRYER**	1KCQ:A	182	210
Human complement receptor type 1	STNRENFHYGSVVTYRS	1GKG:A	1038	1054
Insulin	**LYQLEN** **LVEALYL**	1XDA:A 1XDA:B	13 11	18 17
Kerato-epithelin	**FSMLVAAIQSAGLTETLN**	1X3B:A	515	532
Lactoferrin	NAGDVAFV	1LFH:A	538	545

Laminin alpha-1 chain, Glike domain, mouse	SAKVDAIGLEIV	2JD4:A	2919	2930
Lysozyme, human	IFQINS	1REX:A	56	61
Myoglobin, horse heart	**GLSDGEWQQVLNVWGKVEADIAGHGQEVL** **IKYLEFISDAIIHVLHSK**	1WLA:A	1 101	29 118
Prion protein, human, hPrP	AGAAAAGAVVGGLGG **SAMSRPIIHFGSDYEDRYYRENMHRYPNQ** SNQNNF **DCVNITIKQHTVTTTT**	1QLX:A	113 132 170 178	127 160 175 193
Prolactin	**GAARCQVTLRDLFDRAVVLSHYIHNLSS** RYTHGRGFITKAINS	1RW5:A	7 43	34 57
RepA of Pseudomonas pPS10 plasmid	**LVLCAASLI**	1HKQ:A	26	34
Transthyretin	YTIAALLSPYS	1TTA:A	105	115
AB protein precursor	**EVHHQKLVFFAEDVG**	1AMB:A	11	25
Islet amyloid polyprotein, (Amylin) (AIAPP)	SNNFGAILSS TNVGSNTY	1KUW:A 2KB8:A	20 30	29 37
Lung surfactant protein C	**VVVVVVVLVVVVIV**	1SPF:A	9	22
Tau (neurofibrillary tangles)	VQIVYK	3FQP:A	1	6
Polyadenine-binding protein 2	AAAAAAAAAA	---	---	---
Tau (neurofibrillary tangles)	VQIVYK	---	---	---
(Pro)calcitonin (ACal)	DFNKF	---	---	---
ABri	RTVKKNIIEEN	---	---	---
ADan	LFLNSQEKHY	---	---	---
APin	MPYVFSFKMPQEQGQMFQYYPVYMVLPWEQPQQTVRRSPQQTRQQQ	---	---	---
a-S2C Casein (ACas)	ALNEINQFYQKFPQYLQYLYQGPIVLNPWDQVKRNAVPIPTPTLNR	---	---	---
Serum amyloid A protein (AA) or its fragments	SFFSFLGEAFD	---	---	---

APPENDIX B

Supplemental Data for Chapter III

The Landscape of the Prion Protein's Structural Response to Mutation Revealed by Principal Component Analysis of Multiple NMR Ensembles

B.1: FIGURES

Figure S3.1: Difference profile demonstrating residue contribution towards PC1 for the CJD, FFI, and GSS mutant structures. Each row of the plot represents the residue difference profile between each of the datasets in **(Figure 3.4, sections B-D)** with the hPrP WT and variant dataset (black oval in **Figure 3.4, section A**) for PC1. Negative values indicate residues that differentiate between WT structures, positive values indicate residues that differentiate the mutant structure from the remaining WT and variant dataset.

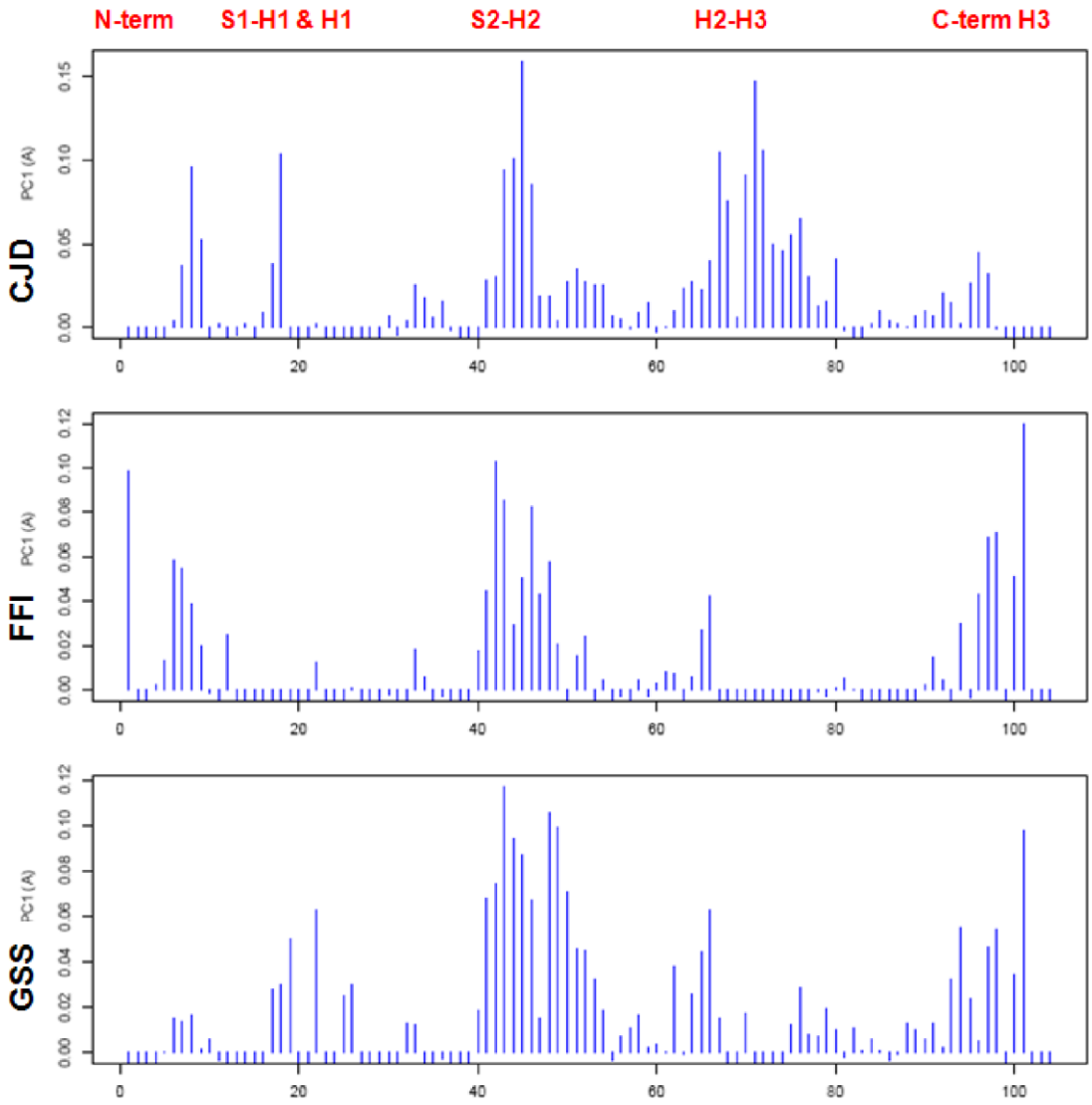


Figure S3.2: PCA analysis of mPrP structures. Contribution of each residue to the first three principal components is indicated, and subdomains displaying concerted atomic displacement in each PC are labeled (black box) and numbered (reference structure 1XYX).

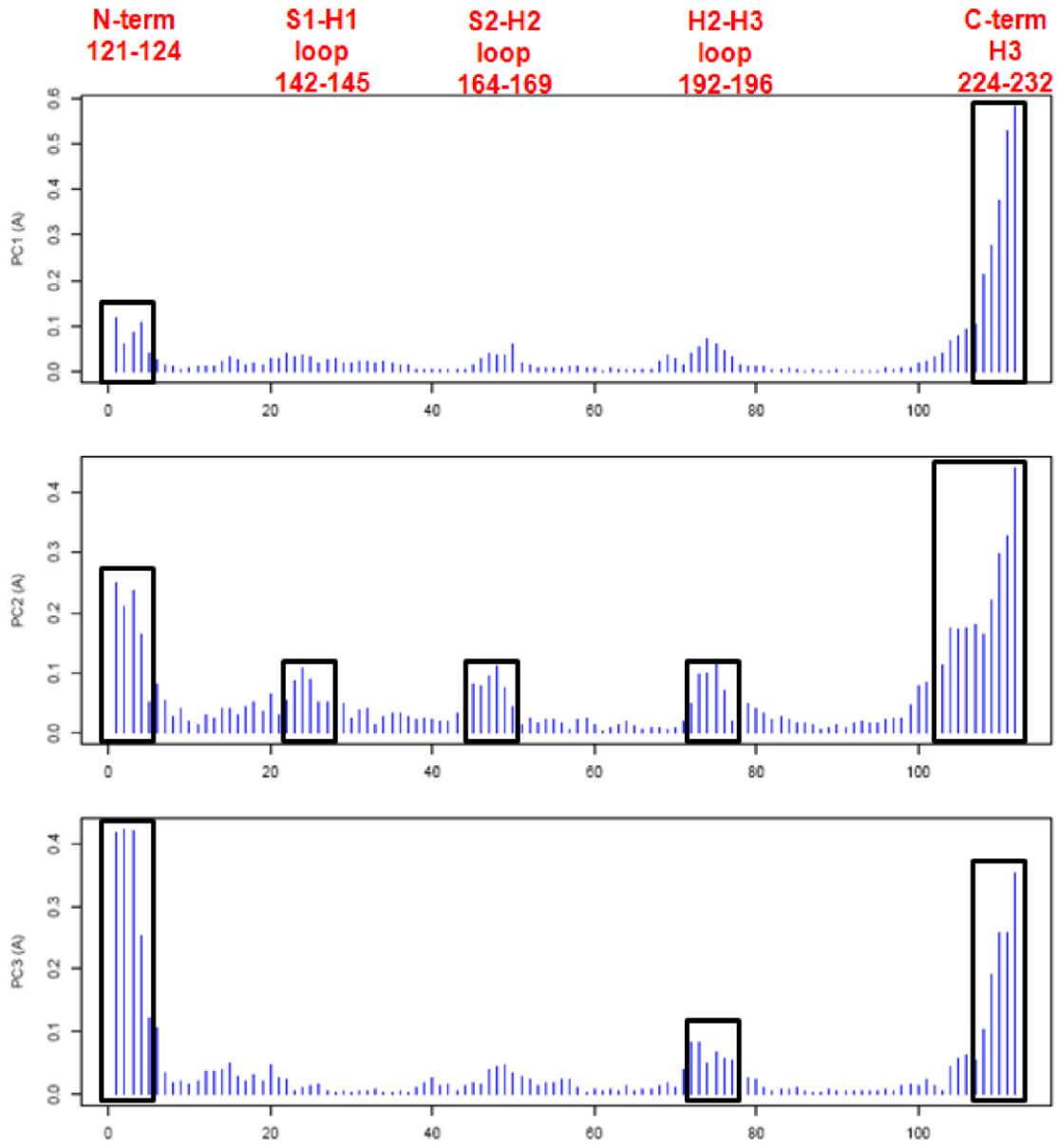


Figure S3.3: Results of PCA on TSE-susceptible and TSE-Non-Susceptible PrP subsets.

(A) Residue contribution to the first three PCs in the TSE-susceptible subset, based on reference structure 1QLZ. Coincidentally, this set consists entirely of mammalian species, and is thus identical to **Figure 3.7, section C**, but has been placed here for comparison with (B).

(B) Residue contribution to the first three PCs in the TSE-non-susceptible subset, based on reference structure 1XYK. This set consisted of both mammalian and non-mammalian species.

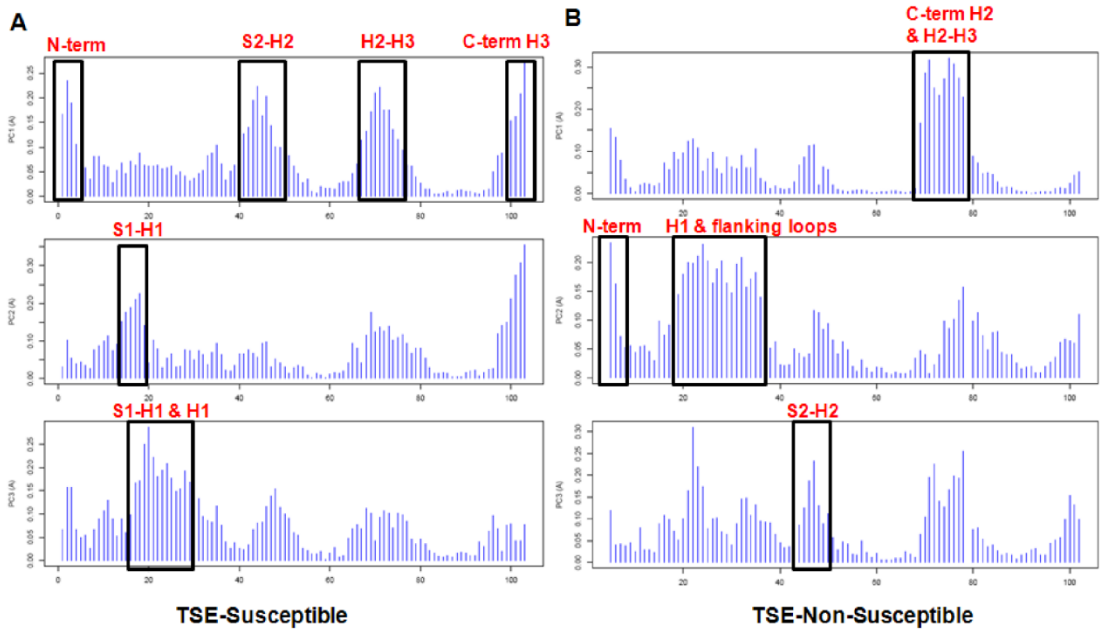
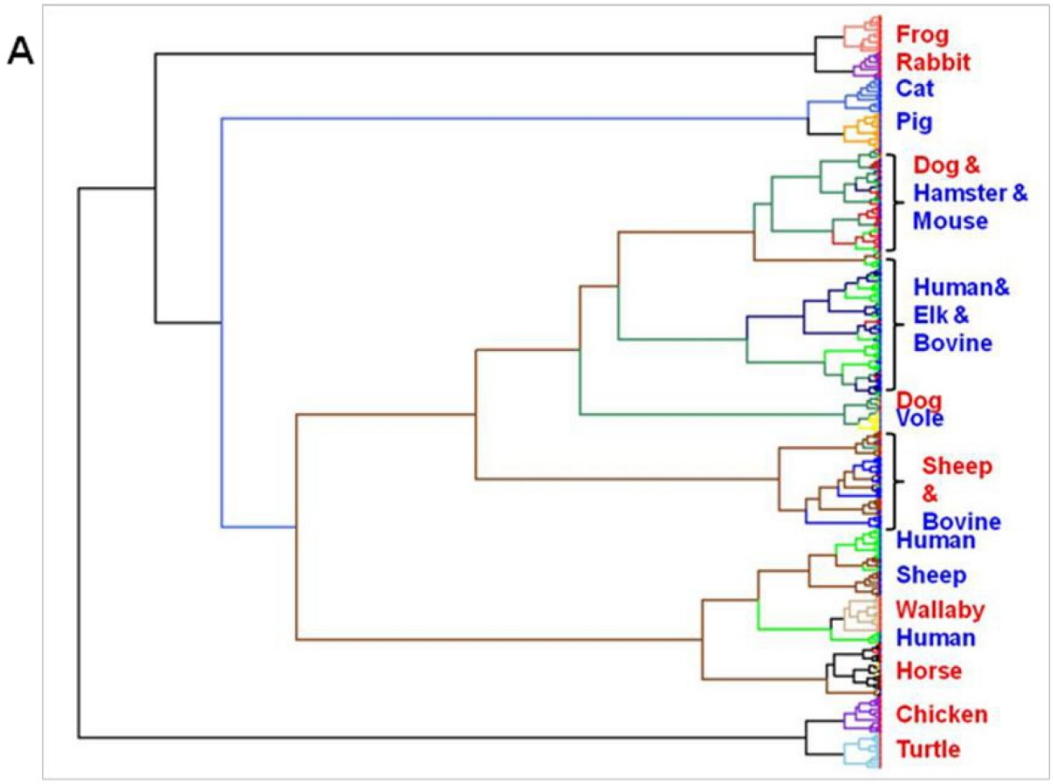


Figure S3.4 (next page): Comparison of Neighbor-joining tree and PC-based dendrogram of 16 WT PrP species (n=420 models).

(A) PC-based dendrogram of 420 models. Edges of the tree are colored to reflect different species. Species have been labeled and colored blue or red to reflect TSE-susceptibility or resistance, respectively.

(B) Neighbor joining tree of 16 PrP species representatives generated by ClustalW, using the Blosum algorithm. This is a bootstrapped tree (100 bootstraps). Bootstrap values are indicated.


A
Frog
Rabbit
Cat
Pig
Dog &
Hamster &
Mouse
Human&
Elk &
Bovine
Dog
Vole
Sheep
&
Bovine
Human
Sheep
Wallaby
Human
Horse
Chicken
Turtle

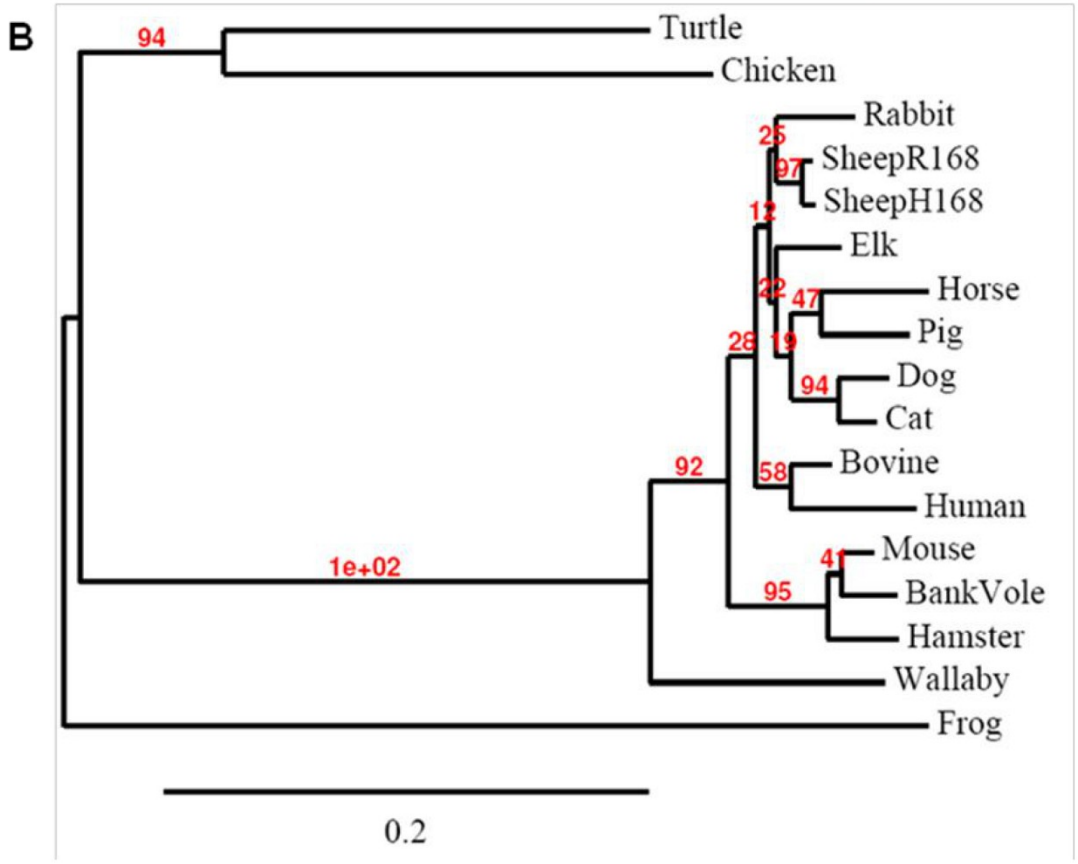

B
94
Turtle
Chicken
Rabbit
25
97
SheepR168
SheepH168
12
Elk
22
47
Horse
Pig
28
19
94
Dog
Cat
92
58
Bovine
Human
1e+02
41
Mouse
95
BankVole
Hamster
Wallaby
Frog
0.2

Figure S3.5: Residue contribution plot for 50 random runs of the hPrP dataset. Using the hPrP dataset of WT, variant, and mutant structures from **Figure 3.1, section A** (11 PDB structures in total), an NMR model was selected at random from each of the NMR ensembles within that set, creating a subset of 11 'representative' NMR models for all the structures. The process was repeated 50 times and PCA was performed on each of the selected subsets. The average of the plots is indicated (black line), and regions of concerted atomic displacement are highlighted and labeled (blue boxes).

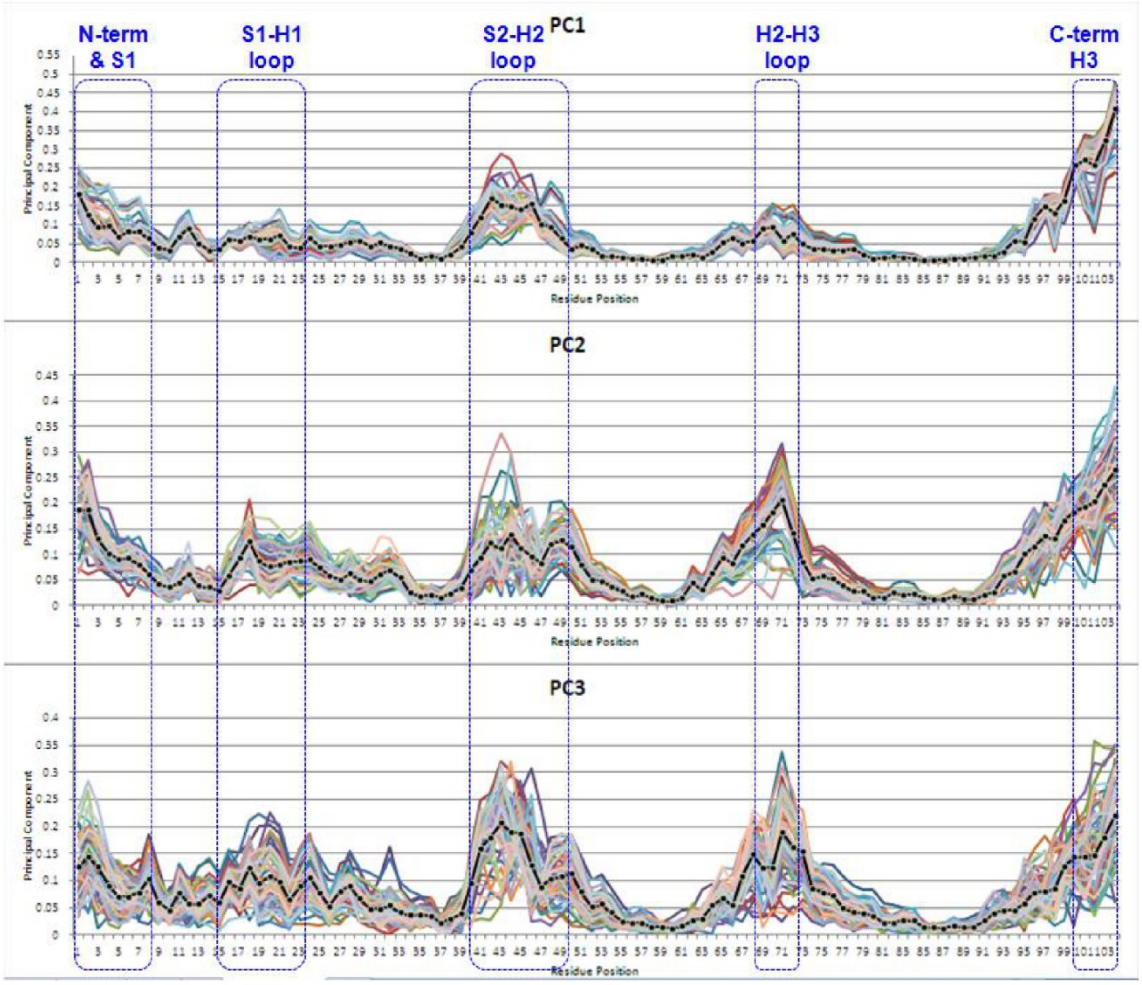
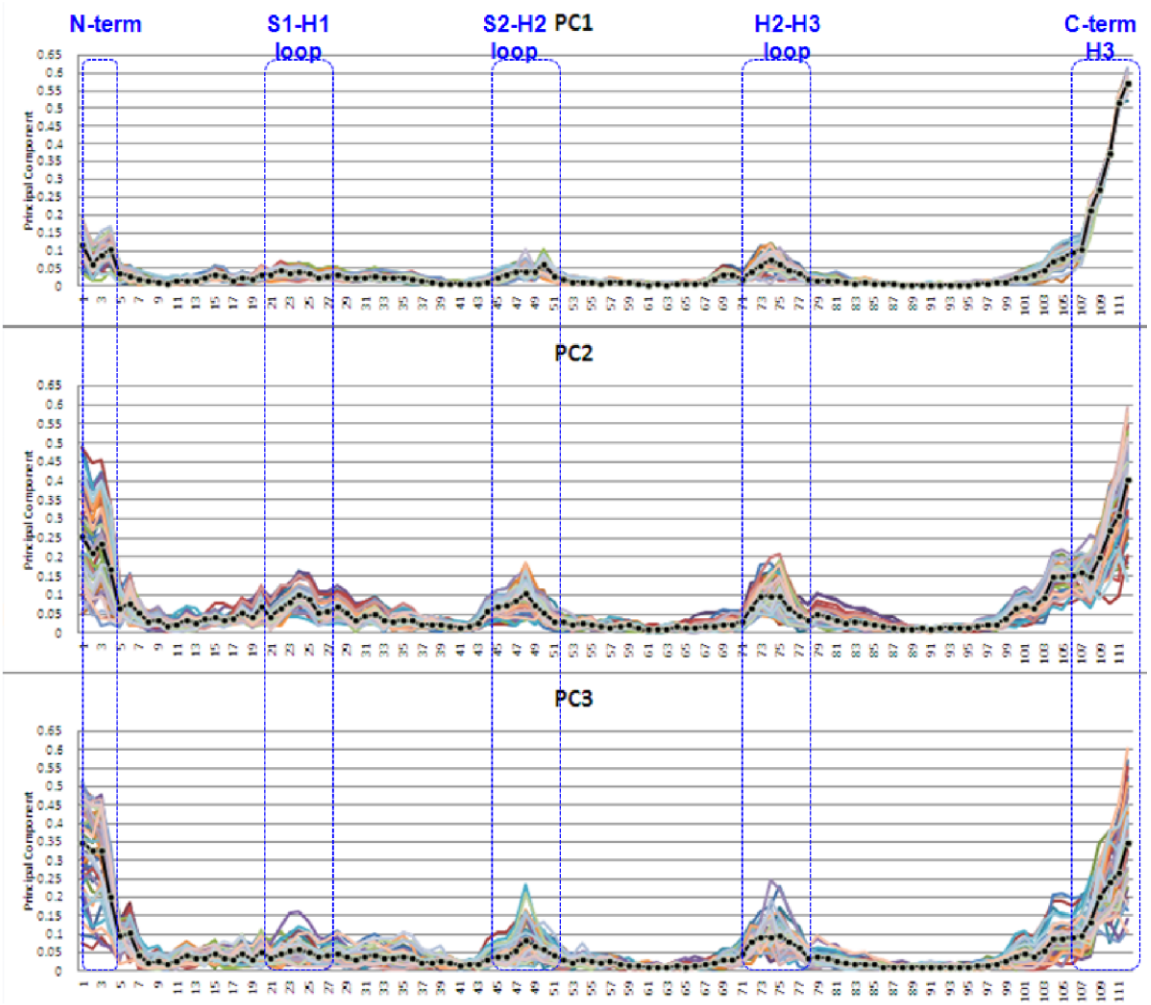


Figure S3.6: Residue contribution plot for 50 random runs of the mPrP dataset. Using an mPrP set of 14 NMR ensembles, an NMR model was selected at random from each of the NMR ensembles within that set, creating a subset of 14 'representative' NMR models for all the structures. The process was repeated 50 times and PCA was performed on each of the selected subsets. The average of the plots is indicated (black line), and regions of concerted atomic displacement are highlighted and labeled (blue boxes).



APPENDIX C

Supplemental Data for Chapter IV

Origins and Evolution of the HET-s Prion-Forming Protein: Searching for Other Amyloid-Forming Solenoids

C.1: DATA LISTS

LIST S4.1 List of Proteomes constituting the BROAD database.

Allomyces macrogynus ATCC 38327
Allomyces macrogynus ATCC 38327 mitochondria
Arthroderma benhamiae CBS 112371
Aspergillus clavatus
Aspergillus flavus
Aspergillus fumigatus
Aspergillus nidulans FGSC A4
Aspergillus niger e_gw1
Aspergillus niger est_GW1_C
Aspergillus niger est_GWPlus_C
Aspergillus niger est_fge1_pg_C
Aspergillus niger est_fge1_pm_C
Aspergillus niger fge1_kg_C
Aspergillus niger fge1_pg_C
Aspergillus niger fge1_pm_C
Aspergillus niger gw1
Aspergillus oryzae
Aspergillus terreus
Batrachochytrium dendrobatidis
Blastomyces dermatitidis ATCC 18188
Blastomyces dermatitidis ER-3
Blastomyces dermatitidis SLH14081
Botrytis cinerea
Candida Albicans sc5314 assembly 21
Candida albicans WO1
Candida guilliermondii
Candida lusitaniae
Candida parapsilosis
Candida tropicalis
Capsaspora owczarzaki ATCC 30864
Chaetomium globosum
Coccidioides immitis H538.4
Coccidioides immitis RMSCC 2394
Coccidioides immitis RMSCC 3703
Coccidioides immitis RS
Coccidioides posadasii RMSCC 3488
Coccidioides posadasii Silveira
Coprinopsis cinerea okayama7#130
Cryptococcus neoformans grubii
Debaryomyces hansenii
Fusarium graminearum
Fusarium oxysporum f. sp. lycopersici
Fusarium solani
Fusarium solani f.pisi
Fusarium solani subsp. pisi
Fusarium verticillioides
Histoplasma capsulatum G186AR

Histoplasma capsulatum G186AR mitochondria
Histoplasma capsulatum H143
Histoplasma capsulatum H88
Histoplasma capsulatum NAm1
Laccaria bicolor
Lodderomyces elongisporus
Magnaporthe grisea (M. oryzae) 70-15
Microsporum canis CBS 113480
Microsporum gypseum CBS 118893
Monosiga brevicollis
Nectria haematococca mpVI 77-13-4
Nectria haematococca mpVI
Neosartorya fischeri
Neurospora crassa OR74A
Neurospora crassa OR74A mitochondria
Paracoccidioides brasiliensis Pb01
Paracoccidioides brasiliensis Pb03
Paracoccidioides brasiliensis Pb03 mitochondria
Paracoccidioides brasiliensis Pb18
Paracoccidioides brasiliensis Pb18 mitochondria
Podospora anserina
Podospora anserina S mat+
Puccinia graminis f. sp. tritici
Puccinia graminis f. sp. tritici mitochondria
Puccinia triticina 1-1 BBBD
Pyrenophora tritici-repentis
Rhizopus oryzae
Rhizopus oryzae RA 99-880 mitochondria
Saccharomyces cerevisiae RM11-1a
Salpingoeca rosetta
Schizosaccharomyces cryophilus OY26
Schizosaccharomyces japonicus yFS275
Schizosaccharomyces japonicus yFS275 mitochondria
Schizosaccharomyces octosporus mitochondria
Schizosaccharomyces octosporus yFS286
Schizosaccharomyces pombe 972h-
Sclerotinia sclerotiorum
Sclerotinia sclerotiorum mitochondria
Spizellomyces punctatus
Stagonospora nodorum
Thecamonas trahens ATCC 50062
Trichophyton equinum CBS127.97
Trichophyton rubrum CBS 118892
Trichophyton rubrum CBS 118892 mitochondria
Trichophyton tonsurans
Trichophyton tonsurans CBS 112818 mitochondria
Trichophyton verrucosum HKI 0517
Uncinocarpus reesii
Ustilago maydis
Verticillium albo-atrum VaMs.102
Verticillium dahliae VdLs.17

LIST S4.2 FASTA sequences of 10 homologs to the HET-s prion-forming domain (PFD).

>gi|2739337|gb|AAB94631.1| small s protein [Podospora anserina]
MSEPFGIVAGALNVAGLFNNCVDCFEYVQLGRPFGRDYERCQLRLDIAKARLSRWGEAVK
INDDPRFHSDAPTDKSVQLAKSIVEEILLLFESAQKTSKRYELVADQQDLVVFEDKDMKP
IGRALHRRLNDLVSRRQKQTSLAKKTAWALYDGKSLEKIVDQVARFVDELEKAFPIEAVC
HKLAEIEIEEVEDEASLTILKDAAGGIDAAMSDAAAQKIDAIVGRNSAKDIRTEERARVQ
LGNVVTAAALHGGIRISDQTTNSVETVVGKGESRVLIGNEYGGKGFWDN
>FOXG_14669T0 | FOXG_14669 | Fusarium oxysporum f. sp. lycopersici conserved hypothetical protein (288 aa)
MAEIFGTVAGAISIAGLFNNCVDCFNYVQIARHFGQDFSRYQLRLDVAKSRLARWGASID
INNNRRFSLIEPADQTVISAQDILQEIVARFETARKISRRYETRTKEQGLRIYTEADLGP
VSHRLHSRFDGITKQRYKSLGLMKKTCWALYDKSYMGRMIDDIIASIEDLEKVFPSTPQL
TSQLVQMEIEEINDEQELELIHDVTEGVDPVLSDASKNKSLEIAGKNSAGRITGPGRVNI
GNSFLTESFPNSQGVRVDTVNHVDEINTAEPSRVHIGNTWGGKGFWD*
>FOXG_17103T0 | FOXG_17103 | Fusarium oxysporum f. sp. lycopersici conserved hypothetical protein (289 aa)
MAEIFGVVASALSVAVLFNNVVDCFEYIQLGRNFGEDYQTCQVKLDIARLRLSRWGDAAK
INNDSRFTEVKPSNNQVRVAKNTLEQLLNLFRNAHTESSNFKLGEGEEELALFDPSTNTN
QAVVALRNTMRDLAHKRQKTTSLSKKISWALYKQKSFMRLIEDIQELLDGLEAIFPQQET
YKRMVEIEIEEVGEGPSLQVLSDAAQETDDLLQEAASRRLEALGSSNAIDQAKVAETAKV
KVGNEYIFQAVPSRTGITTNRIGDLDAQGRSRVLVGDSHGTKGFMDSD*
>FOXG_17314T0 | FOXG_17314 | Fusarium oxysporum f. sp. lycopersici conserved hypothetical protein (289 aa)
MAEIFGVVASALSVAVLFNNVVDCFEYIQLGRNFGEDYQTCQVKLDIARLRLSRWGDAAK
INNDSRFTEVKPSNNQVRVAKNTLEQLLNLFRNAHTESSNFKLGEGEEELALFDPSTNTN
QAVVALRNTMRDLAHKRQKTTSLSKKISWALYKQKSFMRLIEDIQELLDGLEAIFPQQET
YKRMVEIEIEEVGEGPSLQVLSDAAQETDDLLQEAASRRLEALGSSNAIDQAKVAETAKV
KVGNEYIFQAVPSRTGITTNRIGDLDAQGRSRVLVGDSHGTKGFMDSD*
>FVEG_13490T0 | FVEG_13490 | Fusarium verticillioides conserved hypothetical protein (288 aa)
MAEIFGTVASAISMAGLFNNCVDCFSYIQIAKHFGQDFSRYQLRLDVAKCRLARWGESIN
INNDQRFSLAQPTDPMVVLAQGILEEIVARFEAAYKVSRRYTARTEEEGLSICTKADLGA
VSQRVHSRFDVFTKQRYKSLGLMKKTGWALYDKSYMGRMIDDIIASIEDLEKVFPGTPQV
TSQLVEMEIEEVNDEQELEVIQDAAEGLDPLLSDASKNKILEIAGKNTAGKITGPGMVNV
GNSFVTESFSSSQGIRVSTINHVDEVNTTESSKVNVGNTWGGKGFWD*
>gi|256733801|gb|EEU47148.1| hypothetical protein NECHADRAFT_99486 [Nectria haematococca mpVI 77-13-4]
MAEVFGIVTGAIGLAGLFQQCVECFEYVQLGRHFAQDFGMYQLKLDIAKRRLHRWGEAVN
INDNPRFNAPGEDDTLVQEVQAILEEIALLFQTIQKSSKRYTIKAPKEDLECLTEENLQP
VFRRLHAGWTNTTRRPGQKKVNFAKKASWALYDAKNFEKLIEQVSGFLDDLEMLFPAEEL
NRRRLVKLEIEDIADEESLTVLHQTAVEADPLLADVVKEKVKVISVRNSVKVINSSEDAN
VRLGNDWSTAALNAAIEDRTRNEADSVFAEGSSVVHIGNRYG
>gi|256728996|gb|EEU42351.1| hypothetical protein NECHADRAFT_79833 [Nectria haematococca mpVI 77-13-4]
MTEIFGAVSGAISIAALFNDCVDCFEYIQLARHFGKDYSRCQLRLDVAKWRLDRWGAAID
INNDPRFRSGAPANESVRHAQDILREIVGSIEGAYKVSRRYEQSTPDQNRVTLTHADLDP
ASQQLRNEFQTITKKRQDRTSLIRKTGWALYDKKRLGNLIDNIVTSIDELELVFPSVAQA
SVDLARAEIQKVDDQQSLHLIRDAADGLDPVLNDLAKQKLAGVEVQNFAARVKTSESGKF
EIGNIFTKEASGQSVGFPYRNTNRVEDIEVKGDSGVHVGDTYGGKGFWG
>gi|256726268|gb|EEU39630.1| hypothetical protein NECHADRAFT_82003 [Nectria haematococca mpVI 77-13-4]
MGGEAVGVNTEPRFATDNSDDITAQRVCRVLEETRLCFEGVHRLSSRYSPPADSRGLTHS
ELTPVARNLHSRMEDIVHQRQKRGKLLEKASQALYSNKYLDQLIGDIAGLVGNLENLYPV
QMQRRRFVGLEMEAVDDDMSLSTLKNAGSGTDGVLSEVVTNKMQAIADRTEAITGKLEAT

ATRNEPGKTLVEEMERIRVGNEWSESVLNQGALVMDRTENKALAITARGGATIHIGSSFG
RRSIFD
>gi|256724752|gb|EEU38121.1| hypothetical protein NECHADRAFT_48298
[Nectria haematococca mpVI 77-13-4]
MAETFGIVTGAIGLAGLFQQCVECFEYVQLGRHFVQDFGRCRLKLDIAKRRLNRWGEAVN
IHENPQFTDTESEEIQEILEEIANLFDTIQRSSKRYERKAPKEELECLSDENLQPVFRGL
HARWAKISPPKQRDVSLMRKTTWALYDAKYFEKLIGEVTGFVDDLEKVFPAEQAQCHLVQ
IEIEDISDEESLTVLQETANGTDCLLAAAVKEKTNTISVRNYVREIQGEENAKVRLGNDW
STSALSTAIGLDDRTRNEAGSVTAKGSSTVHIGNRYGD
>gi|46127535|ref|XP_388321.1| hypothetical protein FG08145.1 [Gib...
MAEVFGAVAGAIGIAALFNNCIDCFDYIQIARHFGDDFSKYQLRLDVAKCRLSRWGAAIN
VNSDPRFSNNTSKDQTTTLAETLLGEIVARFESAQKSSLLYKTVSRDQEMQVCSEADLGA
VPQRLHSHLRTLTMHRQNRVGLTKKAYWAIYDKNEMGRMIDDIFDLINDLEKVFPATPQA
TSRLAEMEIQEVNDQQGLKMIQDTAQDLDPILADTTKRKLQEITGQNTARCISGKGRTNI
GHTFVNDSFVQSKGFCDSTFNHVDEINLDETARVNIGNTYGGKGFWDS
>gi|46138171|ref|XP_390776.1| hypothetical protein FG10600.1 [Gib...
MAEIFGIVSGALSVAAIFNNCVDTFEYIQLGRRFGEDFQRYQLKLDLAKTRLGRWGEAIS
INNEPRFSSFASADKEVNIAREILEDIASCFEGAQKKSSRYADRADQGELEIFGESDMNP
MLRRLHRHSKDIARQRQKTTSIIKKTKWALYDAKSLERTIDQICSWIDELEKLFPEQSAQ
TQLVEREIEKIDDKPTLEALKDAASGVDPVMEDAVQRKLNMIEGHNSAEFVNLEGSAKFL
VGNVFSEKFLQRDVLLNDRTKNSMRTVSATNQSRLQVGNVYGGRGIWED

C.2: FIGURES

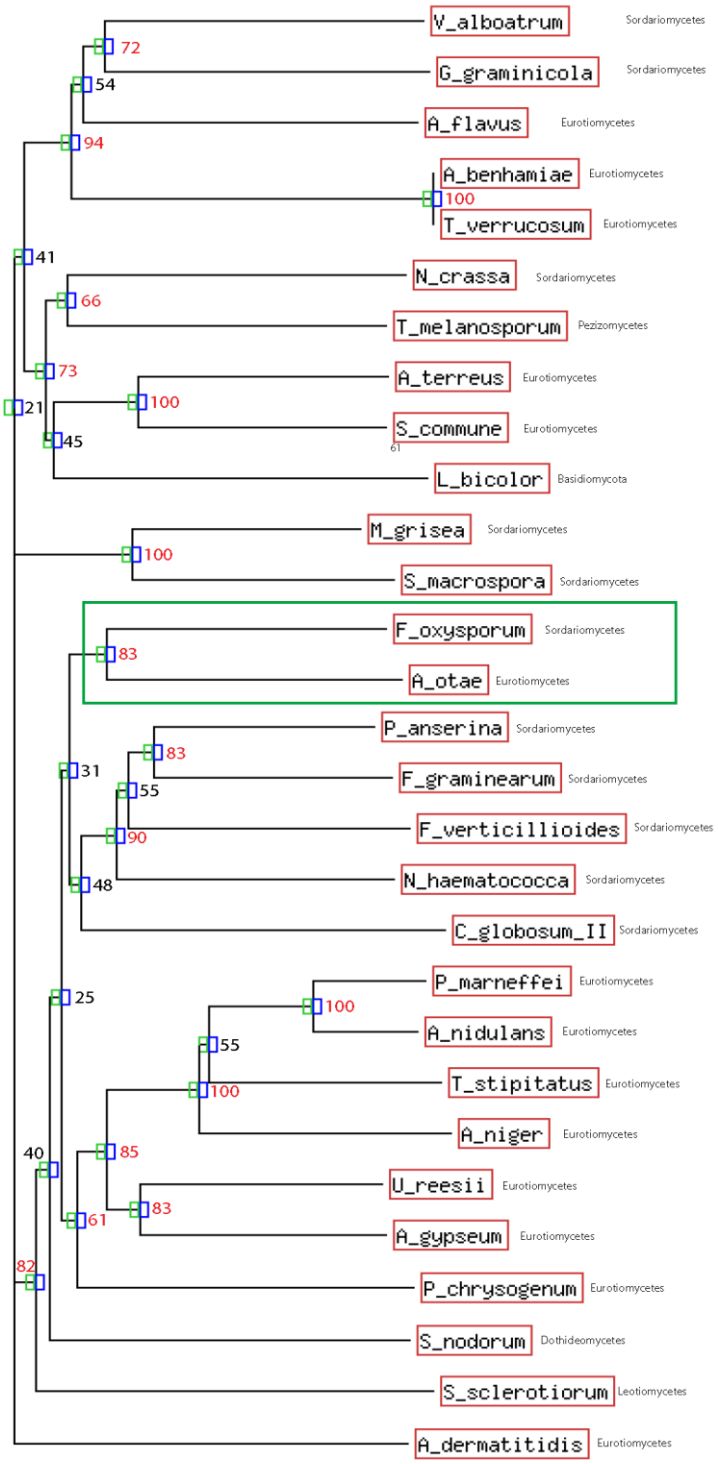


FIGURE S4.1 Neighbor-joining phylogentic tree of the N-terminal domains of Het-s orthologs that significantly align to the *A. otae* N-terminal domain protein sequence. This is a phylogenetic tree made with the neighbor-joining algorithm. The % bootstrap values are labeled at each node. The green box shows the clustering of A. otae with F. oxysporum. The phylogenetic class of each sequence is labeled after the species name (i.e., Sordariomycetes, Eurotiomycetes, etc.).

C.3: TABLES

TABLE S4.1 Blosum similarity matrix for the N-terminal domains and C-terminal domains of the homologs to the PFD. A percent similarity matrix is provided for each of the N-terminal and C-terminal domains based on 10 homologs to the PFD. Naming of the homologs matches the naming scheme of **(Figure 4.3)**.

(A) Similarity matrix for the C-term domain of the homologs to HET-s PFD (n=10)

FOXG 14669	100										
FOXG 17314	24	100									
FVEG 13490	75	28	100								
EEU 47148.1	23	23	26	100							
EEU 42351.1	33	32	31	26	100						
EEU 39630.1	17	25	11	36	19	100					
EEU 38121.1	21	27	22	69	28	37	100				
HET-S	37	30	39	44	33	23	40	100			
HET-s	39	32	40	46	35	26	42	95	100		
FG 08145.1	56	17	56	23	32	17	25	34	35	100	
FG 10600.1	20	21	24	33	32	28	34	36	37	28	100
	FOXG 14669	**FOXG 17314**	**FVEG 13490**	**EEU 47148.1**	**EEU 42351.1**	**EEU 39630.1**	**EEU 38121.1**	**HET-S**	**HET-s**	**FG 08145.1**	**FG 10600.1**

(B) Similarity matrix for the N-term domains of the homologs to the PFD (n=10)

FOXG14669	**100**										
FOXG17314	33	**100**									
FVEG13490	79	37	**100**								
EEU47148.1	43	35	42	**100**							
EEU42351.1	54	38	52	38	**100**						
EEU39630.1	31	28	30	37	29	**100**					
EEU38121.1	43	36	42	70	40	32	**100**				
HET-S	44	42	44	49	46	35	48	**100**			
HET-s	44	44	45	48	46	35	48	95	**100**		
FG08145.1	57	35	59	40	50	28	38	41	41	**100**	
FG10600.1	50	44	48	46	48	36	45	54	54	43	**100**
	FOXG 14669	**FOXG 17314**	**FVEG 13490**	**EEU 47148.1**	**EEU 42351.1**	**EEU 39630.1**	**EEU 38121.1**	**HET-S**	**HET-s**	**FG 08145.1**	**FG 10600.1**

TABLE S4.2

List of N-terminal homologs with significant hits to SCOP domains (n=40). For each homolog, the number of hits against the SCOP domains is also shown.

N-TERM HOMOLOG §	Number of Hits Against SCOP domains
ANID_07985T0	1
ATET_01076 | ATEG_01076 | Aspergillus terreus conserved hypothetical protein (940 aa)	30
BC1T_07009 | BC1G_07009 | Botrytis cinerea vegetative incompatibility protein HET-E-1 (1066 aa)	30
FGSG_04769T0	1
FOXG_08877T0 | FOXG_08877 | Fusarium oxysporum f. sp. lycopersici conserved hypothetical protein (1124 aa)	53
FVEG_12584T0 | FVEG_12584 | Fusarium verticillioides conserved hypothetical protein (492 aa)	2
FVEG_12585T0 | FVEG_12585 | Fusarium verticillioides conserved hypothetical protein (810 aa)	15
FVEG_12617T0 | FVEG_12617 | Fusarium verticillioides conserved hypothetical protein (1154 aa)	71
gi|115385204|ref|XP_001209149.1| predicted protein [Aspergillus ...	2
gi|116200448	1
gi|145602215|ref|XP_359683.2| hypothetical protein [Magnaporthe ...	1
gi|156039651|ref|XP_001586933.1| hypothetical protein SS1G_11962...	9
gi|169623385|ref|XP_001805100.1| hypothetical protein SNOG_14931...	36
gi|170940475|emb|CAP65703.1| unnamed protein product [Podospora anserina S mat+]	8
gi|171681034|ref|XP_001905461.1| hypothetical protein [Podospora...	8
gi|189204452	1
gi|238482849|ref|XP_002372663.1| ankyrin putative [Aspergillus …	85
gi|239608912|gb|EEQ85899.1| tetratricopeptide repeat-containing ...	6
gi|255950182|ref|XP_002565858.1| Pc22g19550 [Penicillium chrysog...	144
gi|256724277|gb|EEU37648.1| hypothetical protein NECHADRAFT_87396 [Nectria haematococca mpVI 77-13-4]	21

gi|261187596|ref|XP_002620217.1| tetratricopeptide repeat-contai...	6
gi|302419471|ref|XP_003007566.1| ankyrin-1 [Verticillium albo-at...	115
gi|302499925|ref|XP_003011957.1| hypothetical protein ARB_01712 ...	60
gi|302661580|ref|XP_003022456.1| hypothetical protein TRV_03406 ...	60
gi|302888950|ref|XP_003043361.1| hypothetical protein NECHADRAFT_87396 [Nectria haematococca mpVI 77-13-4]	21
gi|310790406|gb|EFQ25939.1| hypothetical protein GLRG_01083 [Glo...	83
gi|315053299	58
gi|317141171	76
gi|326471669	53
gi|326485414	54
gi|327309496	61
gi|327348314	1
gi|327354081	6
gi|46126459|ref|XP_387783.1| hypothetical protein FG07607.1 [Gib...	80
gi|46133835|ref|XP_389233.1| hypothetical protein FG09057.1 [Gib...	7
gi|46139775|ref|XP_391578.1| hypothetical protein FG11402.1 [Gib...	84
gi|67901996	1
MGYG_05262T0 | MGYG_05262 | Microsporum gypseum CBS 118893 hypothetical protein (1322 aa)	53
SNOT_10976 | SNOG_10976 | Stagonospora nodorum hypothetical protein (1011 aa)	2
VDAG_00673T0 | VDAG_00673 | Verticillium dahliae VdLs.17 conserved hypothetical protein (612 aa)	1

§ Clusters of 100% Identical Sequences as determined by Blastclust are included below. All sequences are represented in the table, but identical matches for each cluster are highlighted in yellow.

Cluster1: gi|239608912, gi|261187596, gi|327354081

Cluster2: gi|256724277|gb|EEU37648.1|, gi|302888950|ref|XP_003043361.1|

Cluster3: gi|302499925, gi|302661580

TABLE S4.3

SCOP domains that are significant (E<0.0001) to the HET-s N-terminal homolog proteins (n=65). 65 SCOP domains are represented, with a description of each domain, and the number of hits against the homologous proteins of the N-terminal domain.

SCOP DOMAIN	Number of Hits to N-TERM Homologs	SCOP DOMAIN DESCRIPTION
d1a17a_	4	Protein phosphatase 5 {Human (Homo sapiens) [TaxId: 9606]}
d1awcb_	118	GA bindinig protein (GABP) beta 1 {Mouse (Mus musculus) [TaxId: 10090]}
d1bd8a_	69	Cell cycle inhibitor p19ink4D {Human (Homo sapiens) [TaxId: 9606]}
d1bywa_	1	Erg potassium channel, N-terminal domain {Human (Homo sapiens) [TaxId: 9606]}
d1dcqa1	19	Pyk2-associated protein beta {Mouse (Mus musculus) [TaxId: 10090]}
d1elra_	3	Hop {Human (Homo sapiens) [TaxId: 9606]}
d1elwa_	4	Hop {Human (Homo sapiens) [TaxId: 9606]}
d1erja_	11	Tup1, C-terminal domain {Baker's yeast (Saccharomyces cerevisiae) [TaxId: 4932]}
d1gxra_	8	Groucho/tle1, C-terminal domain {Human (Homo sapiens) [TaxId: 9606]}
d1gz8a_	1	Cyclin-dependent PK, CDK2 {Human (Homo sapiens) [TaxId: 9606]}
d1i2ma_	2	Ran {Human (Homo sapiens) [TaxId: 9606]}
d1i7ka_	1	Ubiquitin conjugating enzyme UBC {Human (Homo sapiens), ubch10 [TaxId: 9606]}
d1ihga1	3	Cyclophilin 40 {Cow (Bos taurus) [TaxId: 9913]}
d1iknd_	75	I-kappa-B-alpha {Human (Homo sapiens) [TaxId: 9606]}
d1ixva_	67	26S proteasome non-ATPase regulatory subunit 10, gankyrin {Baker's yeast (Saccharomyces cerevisiae) [TaxId: 4932]}
d1jasa_	1	Ubiquitin conjugating enzyme, UBC {Human (Homo sapiens), ubc2b [TaxId: 9606]}
d1k1aa_	93	bcl-3 {Human (Homo sapiens) [TaxId: 9606]}
d1l0qa2	2	Surface layer protein {Archaeon Methanosarcina mazei [TaxId: 2209]}
d1ll2a_	1	Glycogenin {Rabbit (Oryctolagus cuniculus) [TaxId: 9986]}
d1lv7a_	1	AAA domain of cell division protein FtsH

		{Escherichia coli [TaxId: 562]}
d1mh1a_	2	Rac {Human (Homo sapiens) [TaxId: 9606]}
d1myoa_	110	Myotrophin {Rat (Rattus norvegicus) [TaxId: 10116]}
d1n11a_	103	Ankyrin-R {Human (Homo sapiens) [TaxId: 9606]}
d1n9la_	1	Putative blue light receptor, phot-lov1 domain {Green algae (Chlamydomonas reinhardtii) [TaxId: 3055]}
d1nexb2	10	Cdc4 propeller domain {Baker's yeast (Saccharomyces cerevisiae) [TaxId: 4932]}
d1nr0a1	9	Actin interacting protein 1 {Nematode (Caenorhabditis elegans) [TaxId: 6239]}
d1nr0a2	2	Actin interacting protein 1 {Nematode (Caenorhabditis elegans) [TaxId: 6239]}
d1o6ya_	1	Mycobacterial protein kinase PknB, catalytic domain {Mycobacterium tuberculosis [TaxId: 1773]}
d1ot8a_	86	Neurogenic locus notch receptor domain {Fruit fly (Drosophila melanogaster) [TaxId: 7227]}
d1oy3d_	54	Transcription factor inhibitor I-kappa-B-beta, IKBB {Mouse (Mus musculus) [TaxId: 10090]}
d1p22a2	12	F-box/WD-repeat protein 1 (beta-TRCP1) {Human (Homo sapiens) [TaxId: 9606]}
d1pgua1	1	Actin interacting protein 1 {Baker's yeast (Saccharomyces cerevisiae) [TaxId: 4932]}
d1phka_	1	gamma-subunit of glycogen phosphorylase kinase (Phk) {Rabbit (Oryctolagus cuniculus) [TaxId: 9986]}
d1r2qa_	2	Rab5a {Human (Homo sapiens) [TaxId: 9606]}
d1s70b_	85	Myosin phosphatase targeting subunit 1, MYPT1 {Chicken (Gallus gallus) [TaxId: 9031]}
d1tbga_	10	beta1-subunit of the signal-transducing G protein heterotrimer {Cow (Bos taurus) [TaxId: 9913]}
d1uoha_	151	26S proteasome non-ATPase regulatory subunit 10, gankyrin {Human (Homo sapiens) [TaxId: 9606]}
d1uwha_	1	B-Raf kinase {Human (Homo sapiens) [TaxId: 9606]}
d1vjya_	3	Type I TGF-beta receptor R4 {Human (Homo sapiens) [TaxId: 9606]}
d1vyhc1	8	Platelet-activating factor acetylhydrolase IB subunit alpha {Mouse (Mus musculus) [TaxId: 10090]}
d1w3ba_	3	O-GlcNAc transferase p110 subunit, OGT {Human (Homo sapiens) [TaxId: 9606]}
d1wdya_	83	RNase L, 2-5a-dependent ribonuclease {Human (Homo sapiens) [TaxId: 9606]}
d1wmsa_	2	Rab9a {Human (Homo sapiens) [TaxId: 9606]}
d1wzva1	1	Ubiquitin conjugating enzyme, UBC {Human (Homo sapiens), E2 L6 [TaxId: 9606]}
d1xtqa1	2	GTP-binding protein RheB {Human (Homo sapiens) [TaxId: 9606]}

d1y8xa1	1	Ubiquitin conjugating enzyme, UBC {Human (Homo sapiens), E2 M [TaxId: 9606]}
d1ycsb1	78	53BP2 {Human (Homo sapiens) [TaxId: 9606]}
d1yhwa1	1	pak1 {Human (Homo sapiens) [TaxId: 9606]}
d1yrva1	1	Ubiquitin conjugating enzyme, UBC {Human (Homo sapiens), E2 U [TaxId: 9606]}
d1z06a1	2	Rab-33b {Mouse (Mus musculus) [TaxId: 10090]}
d1z2aa1	2	Rab23 {Mouse (Mus musculus) [TaxId: 10090]}
d1z2ua1	2	Ubiquitin conjugating enzyme, UBC {Caenorhabditis elegans, E2 2 [TaxId: 6239]}
d1zdna1	1	Ubiquitin conjugating enzyme, UBC {Human(Homo sapiens), E2 S [TaxId: 9606]}
d2ajaa1	1	Hypothetical protein LPG2416 {Legionella pneumophila [TaxId: 446]}
d2awfa1	1	Ubiquitin conjugating enzyme, UBC {Human (Homo sapiens), E2 G1 [TaxId: 9606]}
d2c2la1	3	STIP1 homology and U box-containing protein 1, STUB1 {Mouse (Mus musculus) [TaxId: 10090]}
d2fo1e1	69	Lin-12 {Caenorhabditis elegans [TaxId: 6239]}
d2gfsa1	1	MAP kinase p38 {Human (Homo sapiens) [TaxId: 9606]}
d2j4za1	1	Aurora-related kinase 1 (aurora-2) {Human (Homo sapiens) [TaxId: 9606]}
d2java1	1	Serine/threonine-protein kinase Nek2 {Human (Homo sapiens) [TaxId: 9606]}
d2ovrb2	10	F-box/WD repeat-containing protein 7, FBXW7 {Human (Homo sapiens) [TaxId: 9606]}
d2ozaa1	1	MAP kinase activated protein kinase 2, mapkap2 {Human (Homo sapiens) [TaxId: 9606]}
d2uyza1	1	Ubiquitin conjugating enzyme, UBC {Human (Homo sapiens), ubc9 [TaxId: 9606]}
d3blha1	1	Cell division protein kinase 9, CDK9 {Human (Homo sapiens) [TaxId: 9606]}
d3raba_	2	Rab3a {Rat (Rattus norvegicus) [TaxId: 10116]}

TABLE S4.4:

212 HeLo domains identified in N-terminal homologs using HMMER. The HMMER output is shown. **§**

A Snapshot of the table is shown here. For the entire table, please consult the online supplementary material of this article (Table S4.xls) at: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0027342

Full Sequence			Best 1 Domain						
E-value	score	bias	E-value	score	bias	exp	N	Sequence	Description
0	1503	0.2	0	1497.5	0.1	2	1	FGSG_13510T0	| FGSG_13510 | Fusarium graminea
0	1116.2	0	0	1116	0	1	1	gi|115385204|ref|XP_001209149.1|	predicted protein [Aspergillus .
0	1069.2	0.6	0	1068.7	0.4	1.2	1	gi|169623385|ref|XP_001805100.1|	hypothetical protein SNOG_14931.
0	1970	0	0	1969.7	0	1.1	1	gi|46133835|ref|XP_389233.1|	hypothetical protein FG09057.1 [
0	1025.8	0	0	1025.6	0	1	1	gi|9453820|emb|CAB99388.1|	related to small s protein [Neur
4.20E-285	943.4	0	3.50E-284	940.3	0	2	1	gi|256722211|gb|EEU35598.1|	hypothetical protein NECHADRAFT_
4.20E-285	943.4	0	3.50E-284	940.3	0	2	1	gi|302884834|ref|XP_003041311.1|	hypothetical protein NECHADRAFT.
5.20E-282	933.2	0	5.80E-282	933	0	1	1	gi|116205553|ref|XP_001228587.1|	hypothetical protein CHGG_10660.
3.80E-274	907.1	5.1	4.30E-274	906.9	3.6	1	1	gi|296817429|ref|XP_002849051.1|	kinesin light chain [Arthroderm.
2.30E-208	689.2	0.6	3.20E-187	619.2	0	2	2	NCU11978T0	| NCU11978 | Neurospora crassa O
3.00E-186	616	6.7	3.30E-186	615.8	4.6	1	1	gi|212541562|ref|XP_002150936.1|	conserved hypothetical protein .
1.60E-185	613.5	7.6	1.90E-185	613.4	5.3	1	1	gi|67523131|ref|XP_659626.1|	hypothetical protein AN2022.2 [A
6.30E-172	568.6	10.3	7.00E-172	568.4	7.1	1	1	gi|242800108|ref|XP_002483519.1|	conserved hypothetical protein .
1.90E-149	494.1	0	8.10E-127	419.2	0	2	2	gi|164423377|ref|XP_001728051.1|	hypothetical protein NCU11388 [.
1.00E-130	432.1	3.5	1.30E-130	431.7	2.4	1	1	est_fge1_pg_C_40559.1	| est_fge1_pg_C_40559 | Aspergil
1.70E-130	431.3	3.8	2.20E-130	431	2.7	1	1	gi|317032996|ref|XP_003188845.1|	hypothetical protein ANI_1_22360
3.70E-130	430.2	0	4.10E-130	430.1	0	1	1	gi|156039651|ref|XP_001586933.1|	hypothetical protein SS1G_11962.

§ The significance values and bit scores are shown for the entire sequence, as well significance values for the "best domain" if the target sequence only contained the single best-scoring domain.
Exp = expected number of domains in the target sequence, according to HMMER
N = number of domains identified and annotated in the target sequence

Thank you