The Dissertation Committee for Tzintzuni Inde Garcia Certifies that this is the approved version of the following dissertation:

Toward the Rational Design of Mechanical Proteins

Committee:

Andres F. Oberhauser, Supervisor

Werner Braun, Co-Supervisor

Krishna Rajarathnam, Committee Chair

Wlodzimierz Bujalowski

Wolfgang Obermann

R. Bryan Sutton

B. Montgomery Pettitt

Dean, Graduate School

Toward the Rational Design of Mechanical Proteins

by

Tzintzuni Inde Garcia, B. S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas Medical Branch

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas Medical Branch

August, 2009

Dedication

To my mother Martha, who has always been there for me and through hard work and sacrifice brought us through some of the hardest times of our lives. To my step-father Raymond, who helped me become a good man, and helped my mother laugh again. To my wife Sara, who is my partner in adventure and exploration and who fills my life with laughter and love. To my son Lorenzo, who has taught me entirely new depths of love and patience and has much yet to teach me.

Acknowledgements

Graduate school was quite an adventure. Large and intimidating at first, it slowly became manageable with the indispensible help of faculty, staff, and friends many of whom have come and gone over the years. My mentors, Dr. Andres Oberhauser and Dr Werner Braun have been the best that I could have hoped for. Always ready and willing to listen to any problems and offer advice, they have built an excellent environment for research, study, and exploration in which any young mind would thrive. The members of my thesis committee have all been very helpful and provided useful insights throughout my time here. I would especially like to thank Deborah Botting for always being a dedicated and enthusiastic advocate and always being willing to help navigate the recondite maze of paperwork and bureaucracy that comes with being a student. I don't think the graduate school would be able to function without her. To all the program directors, administrators, and support staff that work very hard keeping this wonderful program running a very warm thank you. My family has always been supportive and loving and is ever becoming a larger part of my life as I embark on the new journey of fatherhood. I'd like to thank them all for all the support, love, and laughter they've given over the years.

Many collaborators have shared very interesting projects with us which have helped to expand the depth of my work. I'd like to thank Belinda Bullard, Vladimir

Benes, Mark C. Leake, and Wolfgang A. Linke for working with us on studying projectin and kettin; Eric Miller and Scott Hultgren for working with us on the E. coli pilus project; Dina N. Greene, Bryan R. Sutton, K. M. Gernert, and Guy M. Benian, for working with us to study the kinase domains in twichin and TTN-1; finally I owe thanks to Miguel Torres who's day job is teaching high school, but who gave up two summers to join us in working to build a database of allergenic protiens.

Thank go out to everyone at the Keck Center for Interdisciplinary Research for funding my fellowship for three years and being generally helpful.

And finally I want to thank all the authors of science fiction and fantasy that added their own fuel to the flames of my curiosity and ambition. I wish to acknowledge especially the works of Michael Chriton who's novel Jurassic Park first inspired me to see biology in a whole new way, Isaac Asimov who was just a generally inspirational figure, and Neal Stephenson who's books are just plain awesome.

Toward the Rational Design of Mechanical Proteins

Publication No._____

Tzintzuni Inde Garcia, Ph. D.

The University of Texas Medical Branch, 2009

Supervisors: Andres F. Oberhauser and Werner Braun

The biological functions of proteins have long been studied in a manner that has deprived us of a basic mode of inquiry: physical manipulation. In recent years theoretical and technological advances have made possible tools to directly manipulate single molecules mechanically. The atomic force microscope is a flexible and robust platform that allows us this ability. At the same time computational tools have become increasingly common companions and facilitators of theoretical and experimental science. Of the many mechanically important proteins titin and titin-like proteins are important on many levels. From a physiological understanding of the way mechanical strength is propagated from sarcomere to muscle tissue, to a theoretical understanding of the molecular mechanisms contributing to the mechanical design of proteins in general. We have used a combination of computational techniques on a basis of experimental evidence to make predictions and then test them experimentally to ultimately grow our body of knowledge concerning the mechanical design of proteins.

Table of Contents

Acknowledgments.....v

Abstract.....vii

Table of Contents.....viii

List of Tables.....xi

List of Figures.....xii

CHAPTER 1 – INTRODUCTION **1**

The Mechanical Function of Proteins.....1

Experimental Methods in Nanoscale Mechanics.....2

Computational Methods of Analysis.....3

CHAPTER 2 – MOTIFMATE: A LARGE-SCALE SEQUENCE ANALYSIS TOOL **5**

Section 1: Design of MotifMate.....6

Section 2: Characteristic Motifs for Families of Allergenic Proteins.....13

Introduction.....13

Methods.....15

Results.....17

Discussion.....32

Conclusions.....35

CHAPTER 3 – THE MECHANICAL FUNCTION OF PROTEINS **36**

Section 1: Overview of Mechanical Proteins.....37

Section 2: The Mechanical Properties of E. coli Type 1 Pili Measured by Atomic Force Microscopy Techniques.....39

Introduction.....39

Materials and Method.....44

Results.....46

Discussion 64

CHAPTER 4 – TITIN AND TITIN-LIKE PROTEINS **66**

Section 1: Overview of Titin and Titin-like Proteins 67

Section 2: The Molecular Elasticity of the Insect Flight Muscle Proteins Projectin and Kettin 69

Introduction 69

Results 70

Materials and methods 93

Supporting text 94

Additional Data 98

Section 3: Single-molecule force spectroscopy reveals a stepwise unfolding of C. elegans giant protein kinase domains 104

Introduction 104

Materials and Methods 107

Results 114

Discussion 128

Supporting Figures 133

CHAPTER 5 – FUNCTIONAL ANALYSIS OF THE TITIN I-BAND **136**

Section 1: Early Investigations 137

Introduction 137

PCPMer 138

Correlated Mutation Analysis 141

FANTOM 144

Steered Molecular Dynamics 149

Section 2: Mechanical Stability and Differentially Conserved Physical-chemical Properties of Titin Ig-domainsIntroduction 150

Introduction 150

Materials and Methods 155

Results 162

Discussion 178
Conclusion 181

CHAPTER 6 – EXPERIMENTAL TEST OF PREDICTIONS FROM SEQUENCE ANALYSIS FOR THE MECHANICAL STABILITY OF TITIN IG DOMAINS 182

Introduction 183
Materials and Methods 184
Results 187
Discussion 190
Conclusions 192

CHAPTER 7 – SUMMARY AND FUTURE DIRECTIONS 193

REFERENCES 196

VITA 227

Tzintzuni I. Garcia 227

Education 227

Publications 227

List of Tables

Table 2.2.1 - The most abundant Pfam A allergen families from SDAP. 19
Table 2.2.2 - Classification of allergens in the 12 Pfam families most populated with allergens 20
Table 2.2.3 - AutoMotifs for Allergens and Entire Pfam Families for Seed Storage Proteins, Bet v 1-related family, and Tropomyosin 28
Table 5.2.1 - Positions with significant t-values 176

List of Figures

Figure 2.1.1 - Data flow through the MotifMate system 7
Figure 2.1.2 - Browsing Families 9
Figure 2.1.3 - Family Detail 9
Figure 3.2.4 - Browse Proteins 10
Figure 2.1.5 - Protein Detail 11
Figure 2.1.6 - Summary 12
Figure 2.2.1 - PDB structures for allergens from the most abundant Pfam families 31
Figure - 3.2.1 Type 1 pilli and experimental setup 43
Figure 3.2.2 - Force-extension curves obtained after stretching Type 1 pili 48
Figure 3.2.3 - Force-extension patterns for P pili 51
Figure 3.2.4 - forced unraveling of the helical rod structure is fully reversible 54
Figure 3.2.5 - Simultaneous stretching of multiple Type 1 pili 56
Figure 3.2.6 - Monte Carlo simulation of Type 1 pili elasticity 59
Figure 3.2.7 - Effect of pili elastic properties on bond lifetime 63
Figure 4.2.1 - Flexibility of single projectin molecules. 73
Figure 4.2.2 - Force-extension relationships of projectin molecules. 76
Figure 4.2.3 - Force-extension relationships of recombinant kettin and N-terminal Sls fragments 78
Figure 4.2.4 - Measurements of the force dependence of the unfolding probability of projectin and kettin molecules. 82
Figure 4.2.5 - Refolding kinetics of projectin domains. 85
Figure 4.2.6 - Collapse of unfolded projectin domains under force. 89
Figure 4.2.S1 - Measurement of cantilever drift. 99
Figure 4.2.S2 - Additional collapse trajectories of unfolded projectin domains under force. 101
Figure 4.2.S3 - Effect of temperature on projectin domain unfolding forces. 102
Figure 4.2.S4 - The refolding of projectin domains is very robust. 103
Figure 4.3.1 - Expression, purification, and enzyme activity of recombinant kinase domains and tandem Ig domain segments from *C. elegans* giant proteins. 113
Figure 4.3.2 - 3D structures of *C. elegans* twitchin Ig and kinase domains and homology model for TTN-1 kinase 117
Figure 4.3.3 - Force-extension relationships of TTN-1 and twitchin Ig domains. 121
Figure 4.3.4 - Mechanical properties of *C. elegans* twitchin kinase 123
Figure 4.3.5 - Mechanical properties of *C. elegans* TTN-1 kinase 124
Figure 4.3.6 - Constant velocity steered molecular dynamics simulation of the mechanical unfolding of twitchin kinase. 126
Figure 4.3.S1 - Alignment of TTN1 on 1KOA. 133
Figure 4.3.S2 - Constant velocity SMD simulations of the mechanical unfolding of several protein domains. 134
Figure 4.3.S3 - Constant velocity SMD simulation of the mechanical unfolding of the homology model for TTN-1 twitchin kinase. 135

Figure 5.1.2 - Conservation pattern of I1 domains vs I-band domains..... 140
Figure 5.1.3 - Correlated mutation analysis..... 143
Figure 5.1.4 - Illustration of PATHWAY command..... 146
Figure 5.1.5 - PATHWAY for I27 linearization..... 147
Figure 5.1.6 - Rapid decay of native contacts..... 148
Figure 5.2.1 - Location and architecture of titin, and the structure of an Ig domain. 154
Figure 5.2.2 - Characterization of the mechanical stabilities of titin I-band domains.... 164
Figure 5.2.3 - Dot plot of the sequence identities of cardiac titin I-band Ig domains. ... 166
Figure 5.2.4 - Sequence alignments of seven I-band Ig domains for which the unfolding force is known. 168
Figure 5.2.5 - Best scoring window of motifs in each sequence 170
Figure 5.2.6 - Comparison of the scores for the six weak and strong motifs to find unique motifs. 173
Figure 5.2.7 - Positions found to be significantly different between the weak and strong families..... 177
Figure 6.1 - I27 V11Y mechanical and chemical stabilities..... 188
Figure 6.2 - Altered mechanical properties of I1 XF2..... 189

CHAPTER 1 – INTRODUCTION

The Mechanical Function of Proteins

Proteins have long been known to be the main functional components of all known forms of life. They carry out a grand array of functions including messaging, regulation, transcription, translation, protein folding/unfolding, translocation and transport of small and large molecules, sequestration of toxic or foreign molecules, and many more functions too numerous to mention here. The mechanical functions of proteins have long eluded detailed studies though they are responsible for some of the most basic cellular functions such as compartmentalization, division, and motility. Although there have been, in the past, some very clever studies aimed at determining the mechanical properties of single molecules based on scaling down macroscopic behavior, it is only recently that the available technology has allowed true single molecule mechanical studies.

Several types of protein folds are present in proteins that are "functionally mechanical," by which I mean they are naturally found to fulfill some mechanical role. Examples of these functionally mechanical proteins are: actin which is a structural protein that forms a rigid rod, elastin which is unstructured and provides an elastic matrix, myosin which is a motor protein that can be found in intra-cellular motile constructs. All of these proteins naturally oppose or apply some amount of force in the performance of their tasks.

Functionally mechanical proteins are composed of a wide array of secondary structural elements. Some like the fibronectin type III and immunoglobulin-like (Ig) domains are all-β domains, while spectrin is an all-α structure, and some like myosin are

a mixture of α and β structure. Generallly all-α structures are mechanically weaker than α/β or all-β structures. Elastin forms cross-links at periodic intervals to neighboring molecules to form fibrils with unstructured, elastic elements between linkage points.

Mutations in these proteins can have devastating effects on the macroscopic form and function of organs which can lead to decreased quality of life or death for those so affected. These studies could one day lead to a treatment for such diseases. Also, these mechanical studies are becoming important to nano-scale bio-engineering efforts.

Experimental Methods in Nanoscale Mechanics

It has only been due to technological advances in the past decade that single molecule techniques have exploded into a major field of interest. One of the most visually captivating techniques developed are fluorescence studies which can track individual protein assemblages as they move over DNA or actin fibers. These allow the researcher to better understand the programmatic behaviors of biological nano-machines as they encounter obstacles or other stressors.

More passive proteins require the researcher to act directly on them to elicit a response to applied force. Several techniques have been developed over the years. One is to simply use the flow of buffer to apply force to proteins bound to some substrate and use a marker such as a bead tethered to the other end to measure the response. Two other methods which involve beads are optical and magnetic tweezers. In both techniques a single molecule is tethered between two pairs of beads, one is generally held by suction via a glass pipette while the other is manipulated either by a magnetic field or a focused laser. The final method, which is the one employed for the experimental studies in this work, uses an atomic force microscope (AFM) to probe the mechanical properties of single protein molecules tethered between a stage and the AFM cantilever tip. Using an

AFM gives us the ability to measure forces in the piconewton (pN) range under physiological conditions with minimum difficulty in binding and isolating single molecules to probe.

Computational Methods of Analysis

While the continued development and refinement of these techniques has given us an unprecedented view at the mechanical workings of single molecules, the intra-molecular interactions are still beyond the resolution limit of current techniques. This is where computational modeling techniques can aid in our understanding of the mechanical design of proteins. The most clearly analogous computational technique is the all atom simulation of a protein being pulled apart at the N and C termini achieved by steered molecular dynamics (SMD). Molecular dynamics (MD) simulations have been used to study many aspects of protein behavior using an ever-evolving array of force field models and other degrees of complexity. SMD builds upon the basic MD simulation by adding an artificial force to one atom in the simulation which is then pulled away from a specified static atom at a constant velocity. The amount of force being applied to the mobile "SMD atom" is then recorded over the simulation. These simulations are useful in developing models and hypotheses about the intra-molecular interactions responsible for the mechanical stability of a protein of interest.

Simulation is only one computational method useful in this research. Other methods can also give us clues as to what parts of a protein are important to its mechanical properties and why that may be so. Sequence analysis techniques have long been used to identify related proteins by simple sequence similarity. More advanced methods try to detect distantly related proteins by functional motifs and other common patterns. PCPMer is a sequence analysis tool which identifies motifs based on the

physical-chemical properties of amino acids. When applied to a group of related proteins it can analyzes them for conservation patterns in their physical-chemical properties and finds characteristic conserved motifs for that group. This profile of motifs can then be compared to novel peptide sequences to detect similarities. We have used PCPMer in this capacity to create a database of the motif profiles of allergenic proteins with the intention of being able to provide some predictive power of the allergenicity of novel peptide sequences. We have also used this technique in a novel approach by comparing the motif profile of a group of mechanically strong proteins with that of a group of mechanically weak proteins to try to discern which parts of the domain are used to tune it to have a particular mechanical stability.

CHAPTER 2 – MOTIFMATE: A LARGE-SCALE SEQUENCE ANALYSIS TOOL

Major contributors to section 2 include: Ovidiu Ivanciuc, Miguel Torres, Catherine H. Schein, and Werner Braun.

Section 1: Design of MotifMate

PCPMer motif analysis is a powerful tool that aids in the investigation of protein functions and properties. It is desirable to apply it to large protein databases to screen for commonalities or differences between protein families. This has the potential to generate vast quantities of data. It is for this application that we designed and built MotifMate; an automated system that uses PCPMer to detect conserved property motifs in families of related proteins and store that information in a relational database then make it later available to users. The back end consists of a MySQL database for data storage and retrieval and a series of perl scripts that interface with PCPMer; feed it alignments and collect the output and storing it in the database. As new unprocessed information is entered into the database it is recognized automatically and processed in the background. The 'front end' is composed of a PHP-based website that interfaces with the database allowing on-line users access to the data. Key-word searches based on text-data associated with protein entries enable users to browse the data stored in Automotif along with basic browsing utilities for Proteins and Families. The data flow and user interface are detailed below in figures 2.1.1-6.

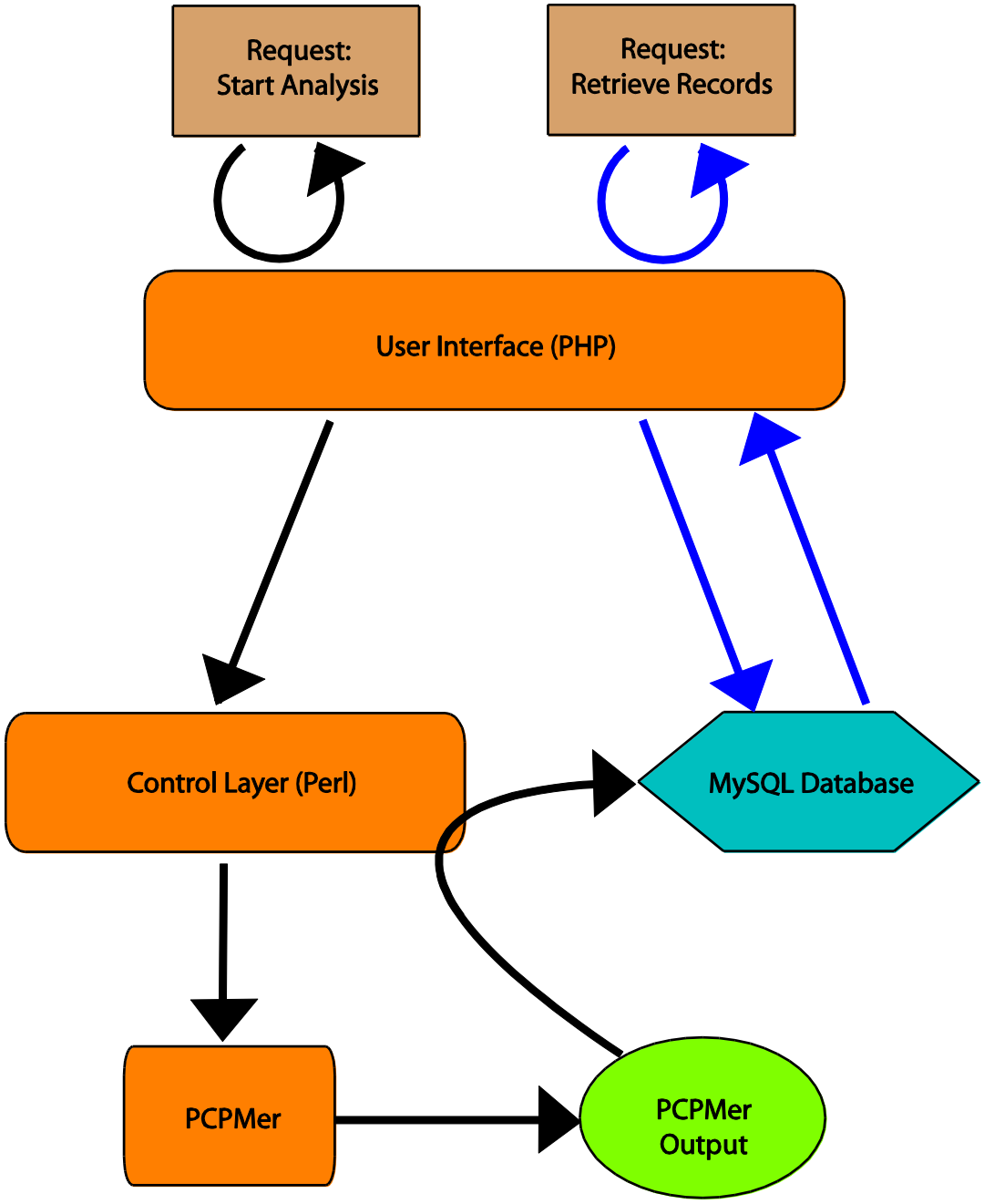

Figure 2.1.1 - Data flow through the MotifMate system

The two basic functions of MotifMate are detailed here. The black arrows describe the flow of data and commands when a request is received by the PHP user interface to begin

an analysis. The request initiates a perl script that can handle batch jobs and begins by retrieving pertinent information from the MySQL database including the list of proteins to analyze, requested settings for PCPMer, and paths to needed alignment files. All this information is then checked for errors, reformatted, and otherwise prepared for PCPMer. Finally the perl script executes PCPMer with the prepared data and waits for its termination upon which time the output data are read, parsed, and entered into the database. At this point if there are more jobs in the queue, the program moves on to the next one. The blue arrows describe the flow of data when a request is made for information stored in the database. In this case the PHP user interface can access the database directly and retrieve and display the required information.

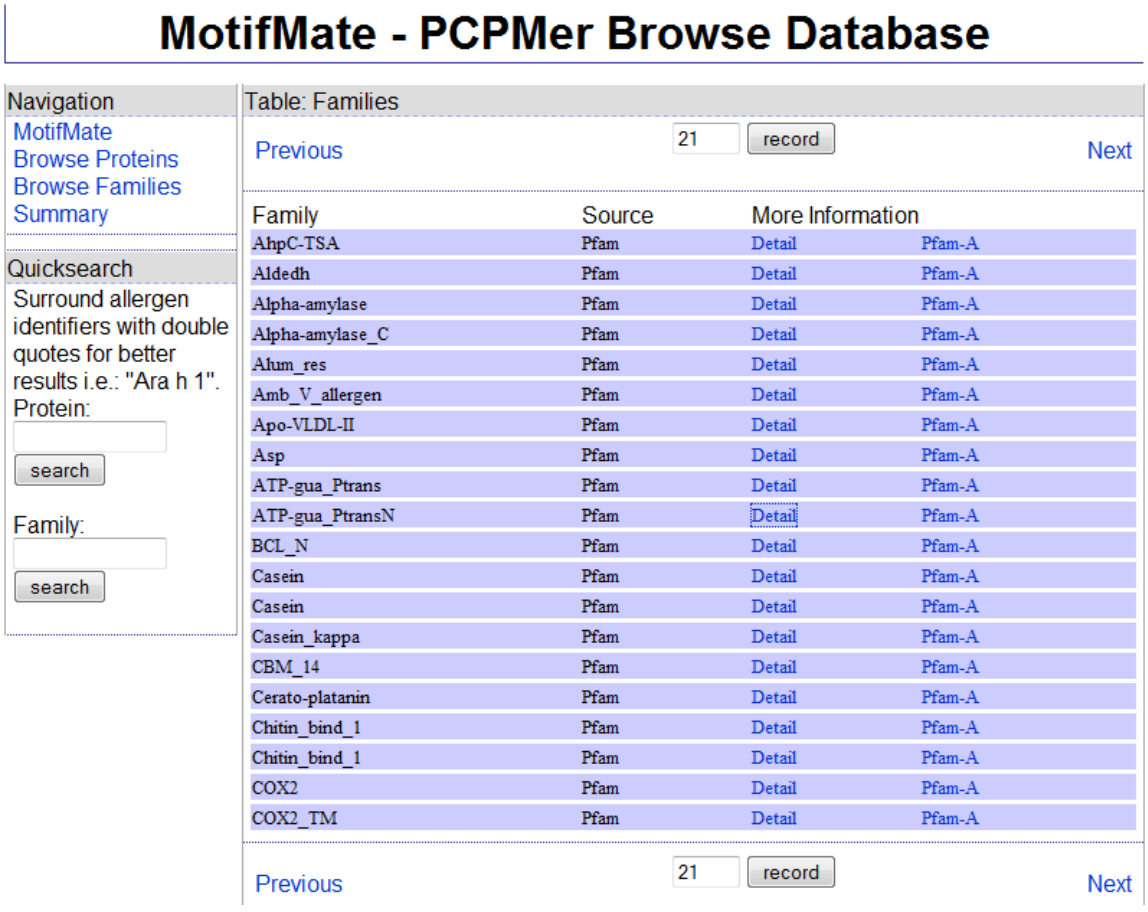

Figure 2.1.2 - Browsing Families

This screenshot depicts the MotifMate interface for browsing protein families stored in the database. Links are given to view a family in more detail and if available to the website from which the family information was originally obtained.

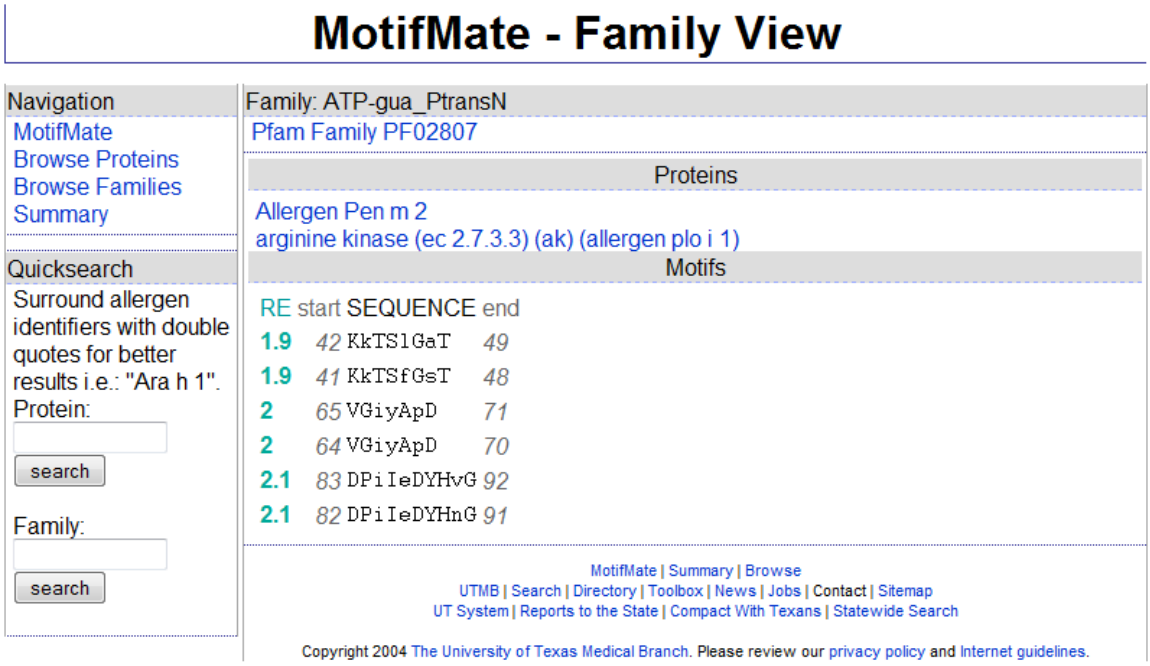

Figure 2.1.3 - Family Detail

This is an example of the details available in MotifMate for a family. The family name is given at the top followed by a link to the originating online database. Below that is a section that lists all the proteins in MotifMate that are in this family and links to their detailed descriptions. Finally a list of motif information that has been obtained for all proteins in this family is listed. The motifs are sorted by location so that similar motifs will tend to group together. The motif details include the maximum relative entropy score at which the motif was found, the starting and ending location in the protein sequence of the motif, and the amino acid sequence from the protein which covers the

motif. The amino acid sequence is given in upper case and lower case letters indicating positions in the motif which were significant to defining it.

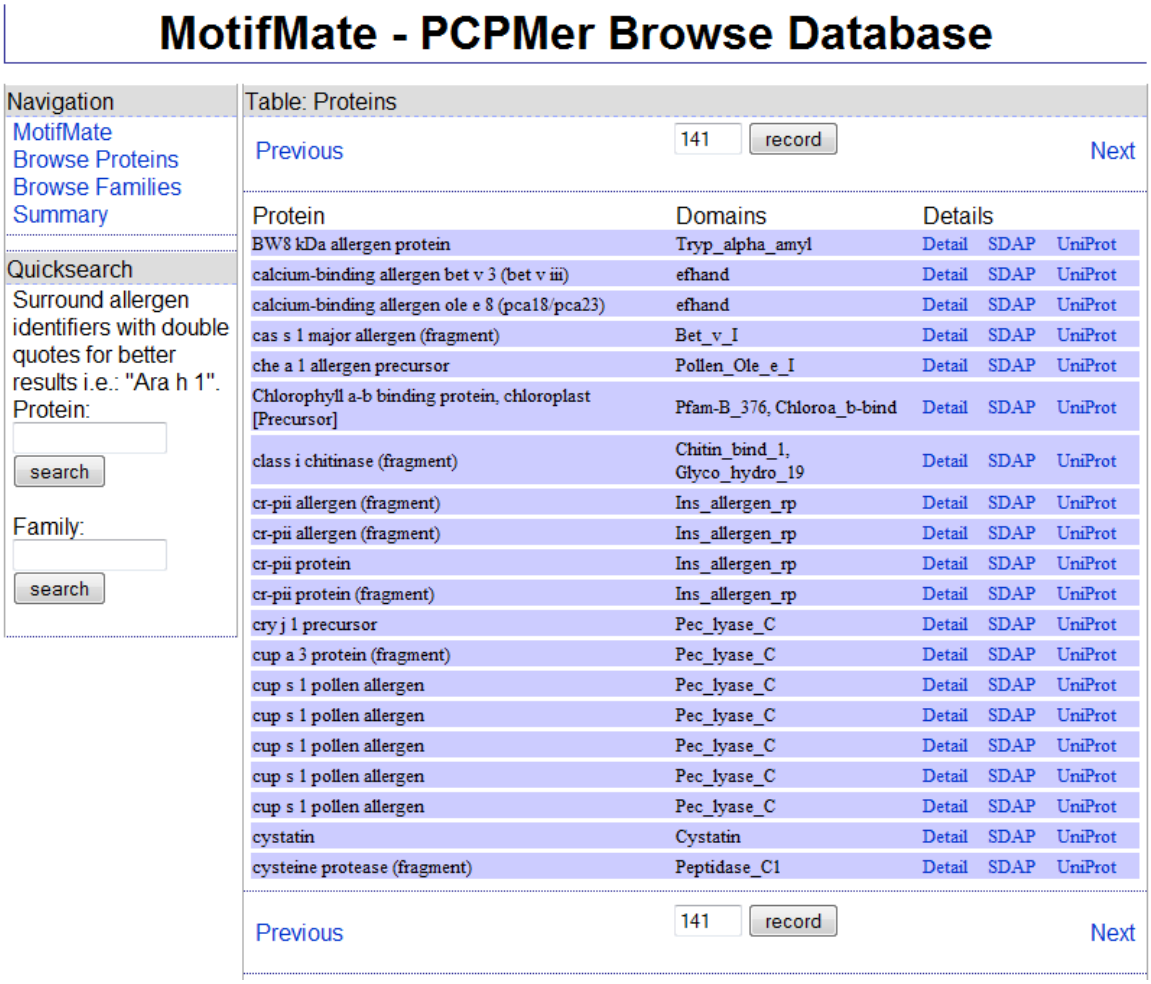

Figure 3.2.4 - Browse Proteins

Browsing the protein information contained in MotifMate gives a listing of protein names, the domains included in that protein, and links to more detailed information about the protein in MotifMate, SDAP, and UniProt. At the time of development the protein families were all taken from Pfam which detects common domains in proteins, therefore

a more general title of ‘families’ would be appropriate for the ‘Domains’ column in this view.

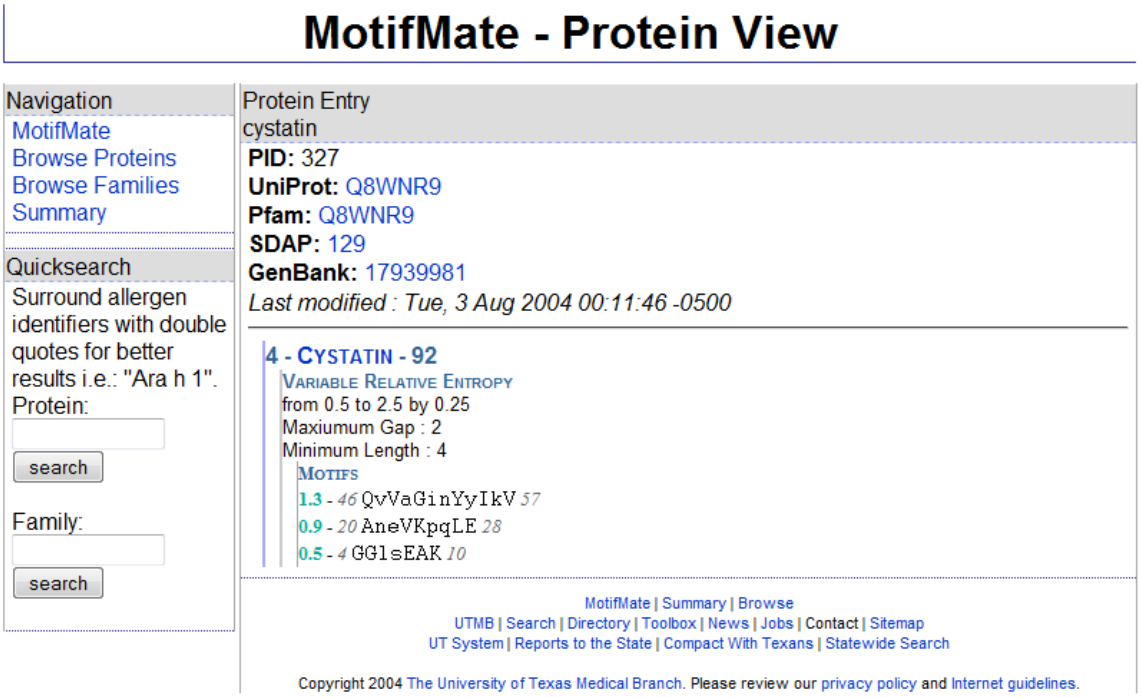

Figure 2.1.5 - Protein Detail

The details presented by MotifMate about a protein are listed here. The heading lists a common name for the protein followed by a list of identifiers for the protein in other web based protein databases. The protein entry will also list motif information for any families in which the protein is a member. In this case the protein is only a member of one family in the database and only residues 4-92 are included in that alignment. The first few lines of each PCPMer analysis record include the settings used to run PCPMer, in this case relative entropy settings from 0.5 to 2.5 were scanned in steps of 0.25 with a maximum allowed gap of 2 and a minimum required length of 4. Finally the detected

motifs are listed in decending order of relative entropy scores so that the most conserved motifs are listed first.

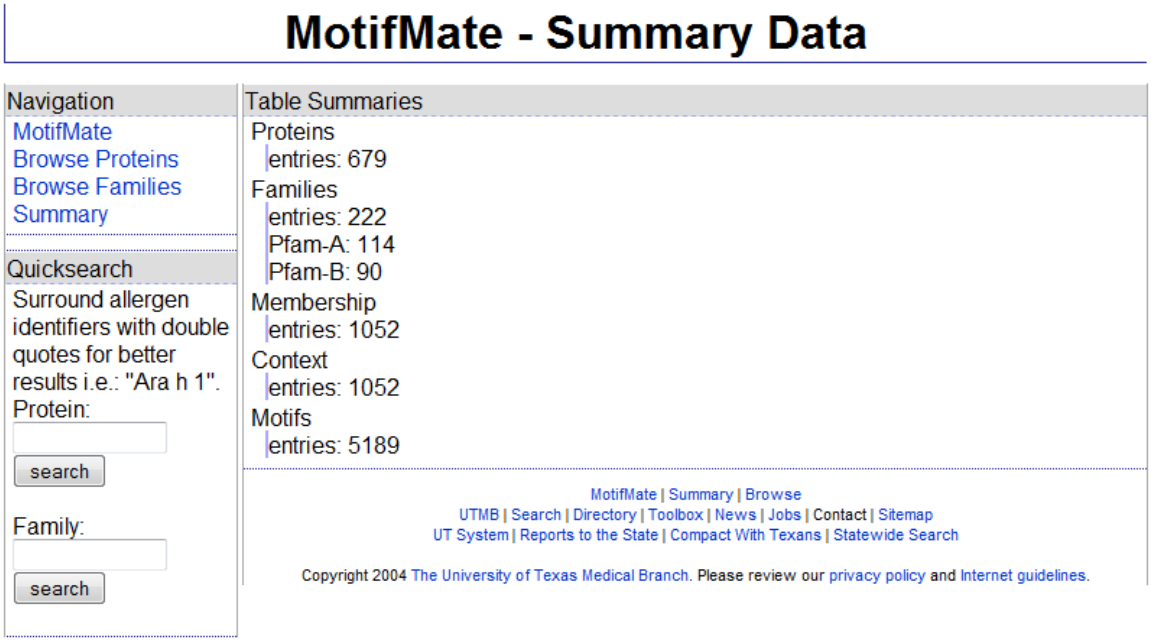

Figure 2.1.6 - Summary

This summarizes the different number of records in MotifMate. While the Proteins, Families, and Motifs entries are self explanatory, the Membership and Context require some explanation. A membership record indicates the membership of one segment of a protein in a multiple sequence alignment represented by a family. It is possible for one protein to have two identical domains and therefore be a member of a family two times, or to have two different domains and be a member of more than one family. A context record contains the details for PCPMer with which a given membership is analyzed. All the memberships in MotifMate currently have only been analyzed once using a single, general purpose set of PCPMer settings resulting in an equal number of membership and context entries.

Section 2: Characteristic Motifs for Families of Allergenic Proteins

INTRODUCTION

The possibility that proteins from novel foods, drugs, or genetically modified organism may exhibit cross-reactivity with known allergens is of utmost concern to regulatory agencies, food scientists and physicians [1]. Due to these considerations, it is important to be able to distinguish allergenic from non-allergenic proteins, and to predict potential IgE cross-reactivities [2-4]. Potential cross-reactive allergens often have very similar sequences [5,6]. Thus, one of the first questions in determining potential cross-reactive foods is the degree of similarity between allergens. Allergens are referred to by names assigned by the Allergen Nomenclature Sub-Committee of the International Union of Immunological Societies (IUIS, www.allergen.org), based on the species/genus name of the source and the order they were identified [7]. This nomenclature system is independent of the biochemical and structural nature of the protein, and the names do not readily identify structural and sequence based relationships among allergens. This means that, based on these names, one cannot easily identify the individual allergenic proteins in different organisms that could account for IgE cross-reactivity [8-11].

Bioinformatics approaches and allergenic databases are now well established to identify molecular similarities of proteins as an explanation for clinically observed cross-reactivity from very different sources [3,4,12-16]. The Structural Database of Allergenic Proteins (SDAP) [17,18] contains many sequence search tools that are seamlessly integrated in the design of the database. SDAP is user friendly and freely available on the Web to allergy researchers, food scientists and industrial engineers (http://fermi.utmb.edu/SDAP/). Allergy researchers can use SDAP primarily to determine food sources that might contain cross-reacting antigens. Regulators and industrial researchers can use the site tools to perform FASTA searches [19] of allergenic proteins or

sequence searches according to the WHO guidelines [20]. FASTA searches are also helpful in clustering related allergens or suggesting the appropriate nomenclature for novel allergenic proteins. For example, cross-reactions in individuals allergic to the birch pollen allergen Bet v 1 with several fruits are a well-documented example of the pollen-food syndrome [21,22], with symptoms ranging from local oral allergy syndrome to severe anaphylaxis. A FASTA search in SDAP quickly reveals that Bet v 1 has significant homology to the food allergens Pru av 1 from cherry (Bit score 160/Evalue $5.9e^{-35}$), Gly m 4 from soybean (Bit score 158/Evalue $3.1e^{-25}$) and Ara h 8 from peanut (Bit score 102/Evalue $4.3e^{-24}$) [23], which could account for the cross-reactions. Pollen cross-reactivity may extend across a large number of species, and even to species from different continents [24,25]. Similar cross-reactivities among allergens with a high degree of identity have been observed for profilins, lipid transfer proteins, calcium-binding proteins, and pathogenesis-related proteins [21,26-28]. Other examples include the ficus-fruit syndrome related to the similarity of cysteine proteases in tropical fruits [29] or the IgE-based cross-reactivity of shrimp with other crustaceans and even nonedible arthropods such as cockroaches or dust mites due to the similarity of the muscle protein tropomyosin in these organisms [30,31].

However, simple sequence similarity is not sufficient to conclusively predict IgE cross-reactivity. While short sequence elements can define an IgE epitope, short stretches of identical sequences are not long enough to predict with statistical significance cross-reactive IgE epitopes [32,33]. The statistical significance can be substantially increased if the sequence is a motif that is common to many related known allergens, and is not found in related proteins that are non-allergens. Here, we define specific sequence regions with common physicochemical properties, PCP-motifs [34,35] that may distinguish allergenic proteins.

Our work was predicated on previous studies which indicated that pollen and plant food allergens grouped to only a small number of all protein families [9,36,37]; most of these families also contain non-allergenic proteins as well. The first step was to obtain a comprehensive assignment of all known allergens according to an existing classification scheme for protein families, Pfam (Version 22.0, http://pfam.sanger.ac.uk/) [38]. These assignments have been made available on our SDAP web site. The major allergens belong to about 30 structural families, consistent with the results of others [39]. In order to discriminate the allergenic from the non-allergenic family members [4,12,16,40-43], we also determined common sequence motifs using our PCPMer program [44,45]. We show in three examples that motifs we defined as characteristic of allergens in a given Pfam coincided with previously determined IgE epitopes. The motifs thus represent a promising way to identify linear IgE epitopes that are likely to be responsible for IgE cross reactivities. All sequence motifs for the major Pfam families with allergens can be obtained from our web server MotifMate (http://born.utmb.edu/motifmate/summary.php). The motifs can now be analyzed in screening sequence data bases for potential IgE cross-reactivities [2,4,14,43,46,47], or used in conjunction with 3D structural information on allergens to shed new light on the molecular determinants of allergenicity [3,5,7,9,48].

METHODS

Assignment of Pfam domains to all allergens

All allergen sequences from SDAP were searched in the Pfam A (Version 22.0, http://pfam.sanger.ac.uk/) [38] database for the matching family. Whenever the TrEMBL or SwissProt accession number of the allergen sequence was known, the Pfam assignment was made based on the corresponding accession number. Otherwise we performed BLAST searches to find related proteins to the SDAP allergen entry. The Pfam database

has a collection of sequence alignments of related protein domains that were used to find Pfam domains for each allergen. Fragments of sequences without a significant match in Pfam where left unassigned. As a result of a direct match or individual BLAST searches, 594 out of 829 allergen protein sequences were grouped to their respective protein families and domains from Pfam A.

Generation of sequence motifs of allergens by MotifMate

MotifMate-PCP is a novel database and data mining tool developed by us to generate physical chemical property (PCP) motifs of allergens. PCP motifs were generated by our PCPMer web server (http://landau.utmb.edu:8080/WebPCPMer/). The motifs are based on the conservation of five physical-chemical descriptors E_1 to E_5 [49] in a multiple sequence alignment. The E_1-E_5 scale allows us to characterize motifs as protein regions where the side chains show conserved physico-chemical properties, such as hydrophobicity, size or alpha-helical propensity, rather than strict sequence identity. We have tested the PCPMer motifs in other protein families to locate functional important regions and as meaningful fingerprints to find distantly related proteins [35,40,44,50].

We generated two types of motifs: one set of motifs that represent a complete Pfam family containing allergenic proteins, i.e. these are motifs generated from the multiple sequence alignment as archived in the Pfam database, and a second set of motifs using only the allergenic proteins in a family (prepared using ClustalW). Using Perl scripts, multiple sequence alignments of all Pfam families containing allergens were downloaded to a MySQL database, PCP motifs were generated and stored in the MySQL database. Sequence alignments of only allergenic proteins in a Pfam family were manually generated with ClustalW [51]. In that phase the protein sequences were cut to the region of the known Pfam domains. In addition, the allergen proteins for each family

group were submitted to a pair-wise sequence search in SDAP to eliminate almost identical proteins or protein sequences from the same allergen source. Also, protein sequences with a sequence identity of only 20% or below to the other allergens from that group were eliminated.

RESULTS

Main Pfam Classes for Allergens

The allergens in SDAP group to only 130 of the 9318 protein families from Pfam A, and of these 31 contain multiple allergenic proteins (Table 2.2.1). A list of the allergens in the 12 Pfam families most populated with allergens is given in Table 2.2.2. The complete classification of allergens is available on our SDAP web server (http://fermi.utmb.edu/SDAP/). For each family, we determined motifs that were common to all members, and, using separate alignments of the known allergens, those motifs that were unique to allergenic proteins.

PF00234: Protease inhibitor/seed storage/LTP family

This domain (InterPro IPR003612) is found in plant lipid transfer proteins, seed storage proteins, and trypsin-alpha amylase inhibitors. The domain forms a four-helical bundle in a right-handed superhelix with a folded leaf topology, which is stabilized by disulfide bonds, and which has an internal cavity. Allergens from the lipid transfer protein (LTP) family are highly resistant to both heat treatment and proteolytic digestion, and are particularly important in the Mediterranean area [27,52]. Three-dimensional structures are known for three allergens from this family, namely Pru p 3 (2ALG, Figure 2.2.1A), Hor v 1 (1JTB), Zea m 14 (1MZM). The molecular determinants of allergenicity

for this family may be extracted from the known IgE epitopes, for Ara h 2 [53], Jug r 1 [54], Par j 1 [55], and Par j 2 [55]. The T-cell epitopes are known only for Ara h 2 [56].

Table 2.2.1 - The most abundant Pfam A allergen families from SDAP.

No	Pfam Code	Pfam domain	No Allergens
1	PF00234	Protease inhibitor/seed storage/LTP family	34
2	PF00235	Profilin	27
3	PF00036	EF hand	23
4	PF01357	Pollen allergen	19
5	PF00188	SCP-like extracellular protein	19
6	PF00407	Pathogenesis-related protein Bet v 1 family	16
7	PF00261	Tropomyosin	16
8	PF00190	Cupin	15
9	PF00061	Lipocalin/cytosolic fatty-acid binding protein family	12
10	PF03330	Rare lipoprotein A (RlpA)-like double-psi beta-barrel	12
11	PF00042	Globin	9
12	PF00544	Pectate lyase	9
13	PF00112	Papain family cysteine protease	8
14	PF00428	60s Acidic ribosomal protein	8
15	PF00082	Subtilase family	7
16	PF00314	Thaumatin family	7
17	PF01190	Pollen proteins Ole e 1 family	7
18	PF01620	Ribonuclease (pollen allergen)	7
19	PF00012	Hsp70 protein	6
20	PF00578	AhpC/TSA family	6
21	PF02221	ML domain	6
22	PF05922	Subtilisin N-terminal Region	6
23	PF00089	Trypsin	5
24	PF00113	Enolase, C-terminal TIM barrel domain	5
25	PF00187	Chitin recognition protein	5
26	PF00273	Serum albumin family	5
27	PF03952	Enolase, N-terminal domain	5
28	PF00151	Lipase	4
29	PF00197	Trypsin and protease inhibitor	4
30	PF00295	Glycosyl hydrolases family 28	4
31	PF01630	Hyaluronidase	4

Table 2.2.2 - Classification of allergens in the 12 Pfam families most populated with allergens

Allergen	Source	Allergen	Source	Allergen	Source
PF00234: Protease inhibitor/seed storage/LTP family					
Amb a 6	short ragweed	Ana o 3	cashew nut	Ara h 2	peanut
Ara h 6	peanut	Ber e 1	Brazil nut	Bra j 1	oriental mustard
Bra n 1	rapeseed	Cor a 8	hazelnut	Fag e 8kD	common buckwheat
Gly m 1	soybean	Hev b 12	rubber (latex)	Hor v 1	barley
Hor v 21	barley	Jug n 1	black walnut	Jug r 1	English walnut
Lyc e 3	tomato	Mal d 3	apple	Ory s TAI	rice
Par j 1	Parietaria judaica	Par j 2	Parietaria judaica	Pru ar 3	apricot
Pru av 3	sweet cherry	Pru d 3	European plum	Pru p 3	peach
Pyr c 3	pear	Ric c 1	Castor bean	Ses i 1	sesame
Ses i 2	sesame	Sin a 1	yellow mustard	Tri a gliadin	wheat
Tri a glutenin	wheat	Tri a TAI	wheat	Vit v 1	grape
Zea m 14	corn				
PF00235: Profilin					
Ana c 1	pineapple	Api g 4	celery	Ara h 5	peanut
Ara t 8	Mouse-ear cress	Bet v 2	birch	Cap a 2	bell pepper
Che a 2	lamb's-quarters	Cor a 2	hazelnut	Cuc m 2	muskmelon
Cyn d 12	Bermuda grass	Dau c 4	carrot	Gly m 3	soybean
Hel a 2	sunflower	Hev b 8	rubber (latex)	Lit c 1	litchi
Lyc e 1	tomato	Mal d 4	apple	Mer a 1	*Mercurialis annua*
Mus xp 1	banana	Ole e 2	olive	Par j 3	*Parietaria judaica*
Phl p 11	timothy	Phl p 12	timothy	Pru av 4	sweet cherry
Pru p 4	peach	Pyr c 4	pear	Tri a profilin	wheat
PF00036: EF hand					
Aln g 4	alder	Bet v 3	birch	Bet v 4	birch
Bos d 3	domestic cattle	Bra n 1	rapeseed	Bra n 2	rapeseed
Bra r 1	turnip	Che a 3	lamb's-quarters	Cyn d 7	Bermuda grass
Cyp c 1	common carp	Gad c 1	cod	Gad m 1	Atlantic cod
Hom s 4	human autoallergen	Jun o 4	prickly juniper	Ole e 3	olive
Ole e 8	olive	Phl p 7	timothy	Ran e 1	edible frog
Ran e 2	edible frog	Sal s 1	Atlantic salmon	Sco j 1	chub mackerel
Syr v 3	lilac	The c 1	Alaska pollock		
PF01357: Pollen allergen					
Ara t expansin	Mouse-ear cress	Cyn d 1	Bermuda grass	Cyn d 15	Bermuda grass
Cyn d 2	Bermuda grass	Dac g 2	orchard grass	Dac g 3	orchard grass
Gly m 2	soybean	Hol l 1	velvet grass	Lol p 1	rye grass
Lol p 2	rye grass	Lol p 3	rye grass	Ory s 1	rice
Pha a 1	canary grass	Phl p 1	timothy	Phl p 2	timothy
Poa p a	Kentucky blue grass	Tri a 3	wheat	Tri a ps93	wheat
Zea m 1	corn				
PF00188: SCP-like extracellular protein					
Cte f 2	cat flea	Dol a 5	yellow hornet	Dol m 5	white face hornet
Pol a 5	wasp	Pol d 5	Mediterranean paper wasp	Pol e 5	paper wasp

Pol f 5	Golden paper wasp	Pol g 5	wasp	Sol i 3	fire ant
Sol r 3	black fire ant	Ves f 5	hybrid yellowjacket	Ves g 5	German wasp
Ves m 5	Eastern yellowjacket	Ves p 5	Western yellowjacket	Ves s 5	Southern yellowjacket
Ves v 5	yellowjacket	Ves vi 5	yellowjacket	Vesp c 5	European hornet
Vesp m 5	giant asian hornet				

PF00407: Pathogenesis-related protein Bet v 1 family

Aln g 1	alder	Api g 1	celery	Ara h 8	peanut
Bet v 1	birch	Car b 1	hornbeam	Cas s 1	chestnut
Cor a 1	hazelnut	Dau c 1	carrot	Gly m 4	soybean
Mal d 1	apple	Pet c PR10	parsley	Pha v 1	kidney bean
Pru ar 1	apricot	Pru av 1	sweet cherry	Pyr c 1	pear
Tar o RAP	common dandelion				

PF00261: Tropomyosin

Ani s 3	herring worm	Cha f 1	crab	Chi k 10	midge
Cra g 1	Pacific oyster	Der p 10	European house dust mite	Hal d 1	abalone
Hel as 1	brown garden snail	Hom a 1	American lobster	Lep d 10	storage mite
Lep s 1	silverfish	Met e 1	greasyback shrimp	Mim n 1	scallop
Pan s 1	spiny lobster	Pen a 1	brown shrimp	Per a 7	American cockroach
Per v 1	tropical green mussel				

PF00190: Cupin

Ana o 1	cashew	Ara h 1	peanut	Ara h 3	peanut
Ara h 4	peanut	Ber e 2	Brazil nut	Cor a 11	hazelnut
Cor a 9	hazelnut	Fag e 1	common buckwheat	Gly m Bd28K	soybean
Gly m conglycinin	soybean	Gly m glycinin G1	soybean	Gly m glycinin G2	soybean
Jug n 2	black walnut	Jug r 2	English walnut	Ses i 3	sesame

PF00061: Lipocalin/cytosolic fatty-acid binding protein family

Aca s 13	flour mite	Blo t 13	dust mite	Bos d 2	domestic cattle
Bos d 5	domestic cattle	Can f 1	dog	Can f 2	dog
Equ c 1	domestic horse	Fel d 4	cat	Lep d 13	storage mite
Mus m 1	mouse	Rat n 1	rat	Tyr p 13	mould mite

PF03330: Rare lipoprotein A (RlpA)-like double-psi beta-barrel

Ara t expansin	Mouse-ear cress	Cyn d 1	Bermuda grass	Gly m 2	soybean
Hol l 1	velvet grass	Lol p 1	rye grass	Ory s 1	rice
Pha a 1	canary grass	Phl p 1	timothy	Poa p a	Kentucky blue grass
Tri a ps93	wheat	Zea m 1	corn		

PF00042: Globin

Chi t 1	midge	Chi t 2	midge	Chi t 3	midge
Chi t 4	midge	Chi t 5	midge	Chi t 6	midge
Chi t 7	midge	Chi t 8	midge	Chi t 9	midge

PF00544: Pectate lyase

Amb a 1	short ragweed	Amb a 2	short ragweed	Cha o 1	Japanese cypress
Cry j 1	Japanese cedar	Cup a 1	Arizona cypress	Cup s 1	common cypress
Jun a 1	Texas mountain cedar	Jun o 1	prickly juniper	Jun v 1	Eastern red cedar

PF00235: Profilin

Profilin (InterPro IPR002097) binds to monomeric actin in a 1:1 ratio and prevents the polymerization of actin into filaments. Three-dimensional structures for allergens in this class are available for Ara t 8 (3NUL), Bet v 2 (1CQA), and Hev b 8 (1G5U, Figure 2.2.1B).

PF00036: EF hand

This family collects calcium-binding proteins (InterPro IPR002048) that contain a common domain known as the EF-hand. The EF-hand motif has a twelve residue loop flanked on both side by a twelve residue alpha-helical domain. The proteins from this class may be signaling proteins (calmodulin, troponin C) or buffering/transport proteins (calbindin D9k). PDB structures are available for Bet v 4 (1H4B, Figure 2.2.1C), Che a 3 (1PMZ), and Phl p 7 (1K9U).

PF01357: Pollen allergen

This family (InterPro IPR007117, Pollen allergen/expansin, C-terminal) contains expansins, proteins that mediate cell wall extension in plants. Expansins allow wall polymers to slide by breaking hydrogen bonds that keel together the wall constituents. Grass pollen allergens are the main allergens from this family (Table 2.2.2). PDB structures are available for Phl p 1 (1N10) and Phl p 2 (1WHO, Figure 2.2.1D).

PF00188: SCP-like extracellular protein

This family (InterPro IPR001283, Allergen V5/Tpx-1 related) includes venom antigen 5 from wasps (Dol a 5 from the yellow hornet *Dolichovespula arenaria*, Dol m 5

from the white face hornet *Dolichovespula maculata*, Pol a 5 from the paper wasp *Polistes annularies*, Pol d 5 from the Mediterranean paper wasp *Polistes dominulus*, Pol e 5 from the paper wasp *Polistes exclamans*, Pol f 5 from the paper wasp *Polistes fuscatus*, Pol g 5 from the paper wasp *Polistes gallicus*, Ves f 5 from the downy yellowjacket *Vespula flavopilosa*, Ves g 5 from the German yellowjacket *Vespula germanica*, Ves m 5 from the Eastern yellowjacket *Vespula maculifrons*, Ves p 5 from the Western yellowjacket *Vespula pennsylvanica*, Ves s 5 from the Southern yellowjacket *Vespula squamosa*, Ves v 5 from the common yellowjacket *Vespula vulgaris*, Ves vi 5 from the wasp *Vespula vidua*, Vesp c 5 from the European hornet *Vespa crabo*, Vesp m 5 from the giant asian hornet *Vespa mandarina*) and venom antigen 3 (Sol i 3 from the fire ant *Solenopsis invicta* and Sol r 3 from the black imported fire ant *Solenopsis richteri*), which both are potent allergens that mediate allergic reactions to insect stings of the Hymenoptera family. The structure (1QNX, Figure 2.2.1E) and T-cell epitopes of Ves v 5 [57] are known.

PF00407: Pathogenesis-related protein Bet v 1 family

The most important allergen from this class (InterPro IPR000916) is the major white birch (*Betula verrucosa*) pollen antigen. Bet v 1, which is the main cause of type I allergies observed in spring. The Bet v 1 allergens are formed by 6 anti-parallel beta-strands and 3 alpha-helices. Four of the beta-strands dominate the global fold, and two of the helices form a C-terminal amphipathic helical motif. The family contains pathogenesis-related (PR-10) allergens [26], such as Aln g 1, Api g 1, Ara h 8, Bet v 1, Cor a 1, Dau c 1, Gly m 4, Mal d 1, Pru ar 1, Pru av 1, and Pyr c 1. PDB structures are reported for Api g 1 (2BK0), Bet v 1 (1BV1, Figure 2.2.1F), and Pru av 1 (1E09). The conformational IgE epitopes of Bet v 1 were identified [8].

PF00261: Tropomyosin

Tropomyosins (InterPro IPR000533) are alpha-helical proteins that form a coiled-coil structure of two parallel helices containing two sets of seven alternating actin binding sites. In striated muscles, tropomyosin regulates the muscle contraction by mediating the interactions between the troponin complex and actin. Allergies to crustaceans, such as shrimp, crab, crawfish and lobster, are mainly induced by tropomyosin [58]. IgE epitopes are known for the shrimp allergens Pen a 1 [30] and Pen i 1 [59].The high conservation of tropomyosin sequences among invertebrates explains why the cross-reactivity of allergens from shellfish and mollusks are often cross-reactive [60,61]. However, vertebrate tropomyosins are not known to be allergenic.

PF00190: Cupin

The cupin family (InterPro IPR006045) contains the conserved barrel domain of the cupin superfamily (cupa is the Latin term for a small barrel), and is comprised of 11S and 7S plant seed storage proteins. The IgE epitopes for 5 members of this family are reported in the literature: Ara h 1 [62], Ara h 3 [63], Fag e 1 [64], Gly m glycinin G1 [65], Gly m glycinin G2 [66].

PF00061: Lipocalin/cytosolic fatty-acid binding protein family

Lipocalins (InterPro IPR000566) are proteins that transport small hydrophobic molecules, such as lipids, retinoids, and steroids. The fold is an eight-strand antiparallel beta-barrel enclosing the binding site. The structures of several allergens from this family are known: Bos d 2 (1BJ7), Bos d 5 (1GXA, Figure 2.2.1G), Equ c 1 (1EW3), Mus m 1 (1MUP), Rat n 1 (2A2U).

PF03330: Rare lipoprotein A (RlpA)-like double-psi beta-barrel

The rare lipoprotein A (RlpA) fold (InterPro IPR005132) is found in bacterial and eukaryotic lipoproteins, and represents a double-psi beta-barrel fold. This domain may be found in the N-terminal part of several pollen allergens. The 3D structure of only one allergen, Phl p 1 (1N10, Figure 2.2.1H), is known.

PF00042: Globin

Globins (InterPro IPR000971) are heme-containing proteins involved in binding and/or transporting oxygen. Hemoglobin is a protein that in vertebrates transports oxygen from lungs to other tissues, containing two alpha and two beta chains with the characteristic three-dimensional globin fold. Monomeric and dimeric hemoglobins have been identified as major allergenic components in insects . The antigenic determinants of this family from *Chironomus thummi thummi* (midge) have been characterized as regions with dominant polar amino acids and high flexibility [67]. The global fold of the monomeric allergen Chi t 1 is shown in Fig. 2.2.1I (PDB code 1ECO).

PF00544: Pectate lyase

Pectate lyase (InterPro IPR002022) is an enzyme involved in the cleavage of pectate, which occurs during the maceration and rotting of plant tissue. This family contains several major pollen allergens, such as those from short ragweed (*Ambrosia artemisiifolia*), Amb a 1 and Amb a 2. The most common pollen allergen in Japan is Cry j 1, a glycoprotein from the Japanese cedar (*Cryptomeria japonica*). Other cedar allergens are Jun a 1 (*Juniperus ashei*, mountain cedar), Jun o 1 (*Juniperus oxycedrus*, prickly juniper), Jun v 1 (*Juniperus virginiana*, eastern red cedar). Pollen from several cypress species contains allergens homologous with pectate lyase, namely Cup a 1 (*Cupressus arizonica*, cypress), Cup s 1 (*Cupressus sempervirens*, common cypress), Cha o 1

(*Chamaecyparis obtuse*, Japanese cypress). The IgE epitopes are known for Cry j 2 [68] and Jun a 1 [25,69], and the T-cell epitopes were identified for Cha o 1 [70], Cry j 1 [71], and Cry j 2 [71]. The structure of one allergen for this family has been deposited in PDB: Jun a 1 (1PXZ, Figure 2.2.1J) [72]. All allergens from this family have similar sequences and there are significant cross-reactivities to food allergens [6]. Schwietz et al. studied the in vivo and in vitro cross-reactivity between pollen extracts of mountain cedar and 11 other Cupressaceae species, one Taxodiaceae species (Japanese cedar), one Pinaceae species, and an angiosperm, and found that the 12 Cupressaceae and the Japanese cedar are cross-reactive [10].

Sequence motifs characteristic of allergens may correlate with cross-reactivity

Proteins in the same Pfam class are homologous, are expected to share a similar 3D-structure, and often have common biochemical functions [38]. High overall sequence similarity is a good indicator of cross-reactivity [2]. However, as antibodies bind to surface patches of folded proteins, cross-reactivity may be better indicated by matching specific areas of the protein structure rather than just the global fold. To differentiate local sequence areas of known allergens, we first generated sequence motifs that are characteristic for the complete family of all those Pfam classes that contain allergens. These "Full-Pfam motifs" can be used to classify novel proteins, and to determine whether it belongs to a Pfam with many allergenic members. In addition, "Allrg-Pfam" motifs were defined that were derived from alignments of only the allergens within each protein family. This procedure allows us to distinguish allergen specific motifs from those that are common to all proteins in the family. All Full-Pfam sequence motifs are

publicly available on our MotifMate web server (http://born.utmb.edu/motifmate/summary.php).

Table 2.2.3 - AutoMotifs for Allergens and Entire Pfam Families for Seed Storage Proteins, Bet v 1-related family, and Tropomyosin

PF00234: Protease inhibitor/seed storage/LTP family (Subgroup B) – Seed sequence: Jug r 1

	Allergens Only		Entire Pfam Family	
No	*E*	Motif	*E*	Motif
1	1.7	1 CQYYLR 6		
2			0.5	10 RSGGYDED 17
3	1.8	26 CCQQLS 31	0.9	26 CCQQLSQI 33
4	2.0	37 CQCEGLR 43	0.5	37 CQCEGL 42
5	1.7	49 QQQQ 52		
6	1.8	59 **EMEEMV**QSA 67		
7			1.2	67 ARDLPKEC 74

PF00407: Pathogenesis-related protein Bet v 1 family – Seed sequence: Bet v 1

	Allergens Only		Entire Pfam Family	
No	*E*	Motif	*E*	Motif
1	2.0	6 ETETTSVIPA 15		
2			1.3	15 AARLFKA 21
3	2.0	31 PKVAP 35	1.2	25 DGDNLFPKVAP 35
4	2.0	42 **ENIEGNGGPGT**IK 54	1.8	46 **GNGGPG** 51
5	1.8	69 D**RV****D**EVD 75	1.5	68 KD**RV****D**EVD 75
6	1.7	81 YNYSV**I**EGGPI 91		
7	2.0	110 GGSILK 115	1.5	110 GGSILK 115
8	2.0	120 YHTKG 124	1.0	120 YHTKGD 125
9			0.7	129 KAEQVKASK 137
10	2.0	145 RAVESYLLAH 154	1.2	145 RAVESYLLAH 154

PF00261: Tropomyosin – Seed sequence: Pen a 1

	Allergens Only		Entire Pfam Family	
No	*E*	Motif	*E*	Motif
1	2.0	7 **END**LD 11		
2	1.8	14 QESL 17		
3	2.0	20 ANIQ 23	0.8	20 ANIQLV 25
4	2.0	33 NAEGEVA 39		
5			1.0	39 AALNRR 44
6	2.0	47 **LLEEDLERSEER** 58	1.0	54 **RSEER** 58
7	2.0	65 KLAEASQAADESERMRKVLE 84	1.4	62 ATTKLAEASQAADE 75
8			1.5	79 MRKVLENR 86
9	2.0	90 **DEERMDALENQ**LKEAR 105	1.5	90 **DEERM** 94
10			1.6	98 ENQLKEA 104
11	2.0	108 AEEADRKYDEVARKLAMVEADLERAEERAE 137	1.5	108 AEEADRKYDEVA 119
12			1.2	130 ERAEERAETGE 140
13	2.0	145 **ELEEELRVVG**NNLKSLEVSEEKANQRE 171	1.1	147 **EEELR** 151
14			1.0	155 NNLKS 159
15			1.1	166 KANQREEAYK 175
16	2.0	174 YKEQIKTL 181		
17	2.0	184 KLKAAEARA 192	1.4	186 KAAEARAEFAE 196
18	2.0	195 AERSV**QKLQKEVDRLEDELVNEKEKYK** 221	1.2	201 **KLQKEVDRLE** 210
19	2.0	225 **DELD** 228		

We next compared the motifs that were specific for the allergens with known IgE epitopes, to see if there was a correlation that could account for clinically significant cross-reactivities between allergens. Three major Pfam families were chosen: the seed storage proteins (a subset of PF00234), the pathogenesis-related protein Bet v 1 family (PF00407) and tropomyosin (PF00261). The motifs common to the allergen members of each family were compared to known IgE epitopes (Table 2.2.3). Motifs of Full-Pfam and Allrg-Pfam in equivalent positions in the Pfam domains are listed on the same line and referred to with the number in column 1. For the seed storage proteins, there are five motifs in Allrg-Pfam (numbered 1,3,4,5,6) and four Full-Pfam motifs (2,3,4,7). Motifs 1, 5 and 6 are unique to the allergens (Allrg-Pfam). A novel protein that contained some or all of the Full-Pfam motifs would probably be a member of this Pfam. If there was a significant match to motifs 1, 5 or 6 that characterized the allergenic proteins, it would be also flagged as potentially allergenic. The only representative of this family for which IgE epitopes have been determined is the walnut allergen Jug r 1. The epitope QGLRGEEMEEMV [54] partially overlaps with motif 6 that is characteristic of allergens (bold letters in table 2.2.3). This suggests that this common sequence could play a role in observed clinical cross-reactivities among allergens of this protein family [73-75].

Similarly, unique Allrg-Pfam motifs 1, 3-8 and 10 characterize allergens in the Bet v 1 family (Table 2.2.3). Here again, a conformational IgE epitope, 42ENIEGNGGPGT52 70R 72D 76H 86I 97K [8] correlates with sequences within these Allrg-Pfam motifs. The entire linear part of the IgE epitope is found in the Allrg-Pfam motif 4, and the individual residues 70R, 72D and 86I are in motifs 4 and 5. The cross-reactivites observed between allergens from this family [11,76,77] may be explained by the conservation of this physico-chemical profile for the Bet v1 IgE epitope across all these allergens. As in the first example, the experimentally documented IgE epitope sequence

correlates better with motifs derived from the known allergens than for those that characterize the whole Pfam class.

Numerous studies have related the similar structures of members of the tropomyosin family to clinically significant cross-reactions [60,61,78-81]. We previously demonstrated that tropomyosin allergens are difficult to discriminate from non-allergenic tropomyosins with the current web servers for allergenicity prediction [4]. In this work, we found that Allrg-Pfam and Full-Pfam motifs showed distinctions between the two groups. MotifMate identified 19 common motifs in the highly conserved sequences of tropomyosins. Five of these, 1, 2, 4, 16 and 19, are characteristic of the allergenic family members. We then mapped the sequences of 9 linear IgE motifs that were identified for the shrimp tropomyosin allergen, Pen a 1 [78]. While areas of the epitopes are found in motifs common to all tropomyosins, the sequences for the most part correlate with the Allrg-Pfam motifs that are specific for the allergenic tropomyosins. In particular, the Allrg-Pfam motifs 1 and 19 match well to epitope sequences. These three examples all indicate that distinguishing local areas of conserved physicochemical properties that are common to allergenic members of the same Pfam can be useful in predicting determinants of IgE reactivity, and potential cross-reactivity.

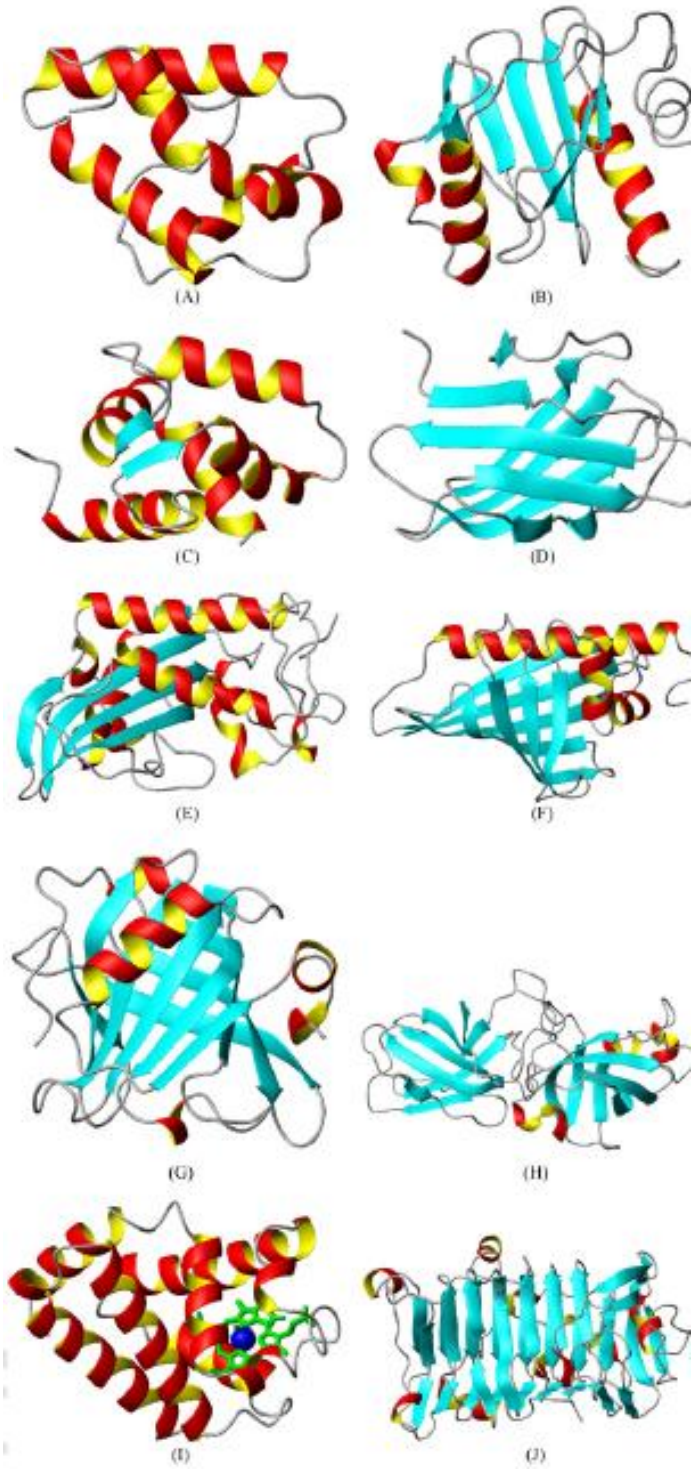

Figure 2.2.1 - PDB structures for allergens from the most abundant Pfam families

(A) Pru p 3 (PF00234, protease inhibitor/seed storage/LTP family; 2ALG); (B) Hev b 8, (PF00235, profilin; 1G5U); (B) Bet v 4, (PF00036, EF hand; 1H4B); (D) Phl p 2, (PF01357, pollen allergen; 1WHO); (E) Ves v 5, (PF00188, SCP-like extracellular protein; 1QNX): (F) Bet v 1, (PF00407, pathogenesis-related protein Bet v 1 family; 1BV1); (G) Bos d 5, (PF00061, lipocalin/cytosolic fatty-acid binding protein family; 1GXA); (H) Phl p 1, (PF03330, rare lipoprotein A (RlpA)-like double-psi beta-barrel; 1N10); (I) Chi t 1, (PF00042, globin; 1ECO); (J) Jun a 1, (PF00544, pectate lyase; 1PXZ).

DISCUSSION

One major goal of our SDAP database is to provide a rapid way for researchers to identify common features of allergenic proteins, as a basis for identifying proteins that could be expected to cause cross-reactions in patients. The sequence comparison tools in SDAP can be used for that purpose. However, grouping all the allergens in SDAP according to protein families within Pfam (Tables 2.2.1 and 2.2.2) now makes this determination even faster, and more accurate as distinct domains of the allergens are assigned to different Pfam. Further, this grouping allowed us to define a series of known 3D protein structures (Figure 2.2.1) to characterize the folds of the majority of allergens. We also derived new sequence specific motifs of proteins in those protein families with a large number of allergens and demonstrated that we also can generate specific motifs that can distinguish them from homologous but non-allergenic proteins in the same protein family (Table 2.2.3). Finally, we could show that specific motifs did indeed correlate with allergenicity, as they corresponded to experimentally determined linear IgE epitopes for three different examples.

However, the conserved motifs that are characteristic only for allergens from a Pfam family are not restricted to the set of IgE epitopes. These motifs may be buried, in which case they represent structurally important residues. Alternatively, the group of residues could give the necessary conformational flexibility to an antigenic site, thus distinguishing them from the rest of the family. The results reported in Table 2.2.3 demonstrate that allergens within a Pfam family have distinct conserved regions as compared to the entire Pfam family.

Our data correlate well with previous attempts to group allergenic proteins according to common sequences, structures [82], and functional classes [26,83,84]. Our finding demonstrate novel applications of allergen classifications [3,9,27,37,39], and allowed us to also analyze finer details of the sequences that correlate with allergenicity. As most of the known allergens can be grouped to only 31 Pfam, this indicates that bioinformatics approaches should be useful for predicting allergenicity for novel proteins. The MotifMate approach outlined here indicates further that sequence fingerprints of allergens and non-allergens within each Pfam family could provide a useful tool to predict cross reactivity of allergens with similar sequences.

These findings represent a considerable advance over the original decision tree for combining computational and experimental tests to determine whether a protein is a potential allergen [1,85,86]. There, cross-reactivity was predicted based on overall sequence similarity of 35% of sequence identity in a window of 80 residues, from FASTA alignments, or on identical regions, as short as six to eight residues, in the protein sequences. While the FAO/WHO procedure is available in SDAP [4,20], our results and those of others (e.g. [14]) indicated that too many non-allergenic proteins are detected by the suggested thresholds. We suggest that these guidelines be modified, to use more

sophisticated comparisons to the known allergen sequences, and particularly allergen specific motifs such as those we define here.

Others have also suggested that motif-based methods could identify allergenic proteins more specifically. The MEME protein motifs [41,47,87], for example, have average lengths of 50 residues [46], and are thus not as specific as the physical-chemical properties motifs we were able to extract. Our motifs correspond better to the length expected for epitopes. The MotifMate motifs can be used to filter large genomic databases directly, either before or after a preliminary classification to eliminate all sequences that do not belong to Pfam families in SDAP. In this way, a large number of sequences will be eliminated from the first step, without time-consuming computations.

The advantages of a motif based approach are clear from the examples presented above. Our MotifMate comparisons discriminated between allergenic and non-allergenic tropomyosins, a difficult task as allergenic tropomysosins, from mite (Der p 10) and shrimp (Pen a 1), are highly similar (80.28%, 228/284, E score 7.2e-73) to the mammalian homologs that are not allergenic [4]. Of four programs tested for their ability to distinguish 4 allergenic tropomysosins from 4 non-allergens, only WebAllergen [43] found that while all 8 proteins have five wavelet allergenic motifs [88] in common, the allergenic tropomyosins have several additional wavelet motifs that may distinguish them. Both Allermatch [89], which applies the FAO/WHO allergenicity guidelines [1,85,86] and AlgPred (a support vector machines classifier) [47] found all eight tropomyosins to be allergens, while MEME motifs [41] predicted all 8 to be non-allergens.

We should at this point note that PCPMer motifs, Allrg-Pfam and Full-Pfam, are numerical vectors based on the E_1-E_5 physico-chemical properties. They have been translated, for convenience, into representative amino acid sequences in Table 2.2.3.

They can be used, in combination with other SDAP tools, to compare the physicochemical properties of these motifs to those of novel proteins.

CONCLUSIONS

The identification of potential allergenic proteins is usually done by global sequence similarity searches. Tools to do overall similarity searching are now incorporated in SDAP. The classification of allergens into Pfam domains reveals the structural relationship between various allergens, thus providing a basis for identifying allergenic determinants. Our results show that allergens can be represented by a small fraction of possible protein families and folds. Out of the 9318 protein families from Pfam, only 130 families are currently listed for all allergens in SDAP, and 31 families contain more than 4 allergens. The most populated Pfam families are protease inhibitor/seed storage/lipid transfer protein, profilin, EF hand, group I pollen allergens, SCP-like extracellular protein, pathogenesis-related protein Bet v 1 family, tropomyosin, and cupins. Details for the Pfam classification of all allergens can be accessed from the SDAP web site (http://fermi.utmb.edu/SDAP/). The sequence motifs characteristic for Pfam classes are available via our web server MotifMate (http://born.utmb.edu/MotifMate/). Those motifs represent sequence-based fingerprints that characterize the major Pfam families with allergens. In addition we also showed, for three important Pfam classes that contain many allergens, how specific motifs correspond to known IgE epitopes. These allergen specific motifs are the basis of an original method to predict the potential allergenicity of novel proteins and clinical cross-reactivity.

CHAPTER 3 – THE MECHANICAL FUNCTION OF PROTEINS

Major contributors to section 2 include: Eric Miller, Scott Hultgren, and Andres Oberhauser.

Section 1: Overview of Mechanical Proteins

As discussed in the introduction**,** functionally mechanical proteins are those that naturally generate, transmit, or otherwise use mechanical forces to carry out their tasks. These proteins can be generally subdivided into two main groups: those that consume energy to generate force (myosin, kinesin, dynein, bacterial flagellar complex), and those that withstand mechanical forces to provide support or structure (actin, titin, spectrin, elastin). Thanks to recent advances in technology and single molecule techniques especially, we are now learning much about the programmatic behavior of these nano-scale molecular motors and structural elements. For example the complex behaviors of kinesin and dynein are becoming very well understood [90]. Actin, the ubiquitous scaffolding protein is a model of adaptability, finding its way into innumerable cell processes [91,92]. As more information becomes known about the activities of thes nano-mechanical engines, our concepts of cytoplasmic hydrodynamics are changing [93].

In order to study the molecular mechanisms involved, single molecule force spectroscopy and other nanomechanical techniques are used to study the forces, distances, motions, energies, and deformations involved in individual proteins or protein complexes, typically in the sub-micrometer and sub-nanonewton ranges. The number of proteins analyzed so far by SMFS is still relatively small (about 55 PDB structures) and they have been analyzed with differing degrees of detail. Although the molecular basis underlying the mechanical resistance of proteins is still unclear, several determinants have been identified through these studies: amino acid sequence, mechanical topology, unloaded unfolding rate constant, and pulling geometry. Some tendencies are already emerging:

a) Proteins have widely different mechanical stability when pulled in the N-C direction ranging from below than the limit of resolution of the AFM (typically ~10 pN; e.g. calmodulin) to ~330 pN (e.g., titin Ig domains). Interestingly, mechanical proteins that must resist force tend to be more mechanically stable than both non-mechanical and elastomeric proteins.

b) Unstructured and β-spiral proteins (e.g. elastin) are among the less mechanically stable proteins.

c) α-helical proteins (e.g., calmodulin, T4 lysozyme) have a relatively low mechanical stability although α-helical bundles (e.g., spectrin [94-96], myosin II tail [97]) and solenoids (e.g., ankyrin B [98]) are more stable.

d) β-stranded proteins tend to unfold at higher forces than α-helical ones.

e) The mechanical stability of most mechanical proteins tends to be determined by a mechanical clamp usually formed by a patch of highly localized mechanical hydrogen bonds [99]. However, in some cases the hydrophobic core contributes also to mechanical resistance [100]. In addition to secondary-structure based elasticity there is tertiary (e.g. ankyrin B solenoid [98]) and quaternary (e.g. myosin II tail [97], adhesive pili [101]) structure elasticity.

f) The mechanical stability and mechanical unfolding pathways depend also on the pulling geometry, which is affected by both the topology at the breakpoint and the point of application of the force. Hence, β-stranded proteins with a shear mechanical topology at the breakpoint (the force vector is orthogonal to the hydrogen bonds) are more mechanically stable than zipper β-stranded proteins (where the force vector is parallel to the hydrogen bonds). The points of application of the force to a protein are also relevant as they can substantially alter its mechanical stability [102,103] implying that proteins have "Achilles' heels".

g) The mechanical stability is a kinetic property which in general is not correlated with thermodynamic stability (ΔG) or with melting temperature (Tm = ΔG/ΔS) [104].

h) The mechanical stability can be modulated by ligand binding [105-107] and by disulfide bond formation [105,108-111].

The molecular structure of a protein, poses constraints on the location of the transition state in mechanical unfolding pathway. Tertiary interactions are thought to have shorter distances to their transition states than secondary structures, and they tend to be more brittle (i.e. they break at high forces and after small deformations) than secondary interactions, which are more compliant (breaking at low forces and after large deformations). Furthermore, tertiary interactions may require more time to equilibrate than secondary ones and therefore they often present hysteresis in the pulling-relaxation cycle (3). Most proteins show a high degree of connectivity and as a result, their unfolding seems to be highly cooperative. Due to the local action of the applied force, their mechanical stability tends to be related to localized molecular structures near the mechanical "breakpoint" rather than to the global structure [99,112]. Using simplified simulation methods a massive survey has been recently conducted, in proteins for which there is atomic structure, to identify these mechanical clamps and classify the available protein structures based on their mechanical stability [113].

Section 2: The Mechanical Properties of E. coli Type 1 Pili Measured by Atomic Force Microscopy Techniques

INTRODUCTION

Adhesion of many bacteria to host tissues is the first step in successful colonization and infection [114,115]. This adhesive interaction is typically mediated by pili,

which are long (~1 μm) rods composed of more than 1000 protein subunits (immunoglobulin (Ig)-like domains) that form a helical structure that is anchored to the outer bacterial membrane ([116]; Figure 3.2.1A). Pilus assembly in *Salmonella typhimurium*, *Bordetella pertussis*, *Klebsiella pneumoniae*, uropathogenic *Escherichia coli* (UPEC) and a multitude of other gram-negative pathogens requires a conserved chaperone-usher pathway to produce fibers important in gastroenteritis, whooping cough, pneumonia, urinary tract infections and a variety of other diseases [117,118]. Uropathogenic strains of *E. coli* use type 1 and P pili to colonize the bladder and kidney respectively. P and type 1 pili produced by UPEC have been shown important for numerous functions including mediating colonization, invasion and biofilm formation [119-121]. These pili, therefore, serve as a lifeline for UPEC. Without strong binding and the ability to withstand fluid forces in the bladder, the bacteria would be easily cleared in the urine. Shear forces in the bladder and kidney caused by fluid and urine flow are a major factor that uropathogenic *E. coli* must resist in order to persist in the urinary tract.

Type 1 pili are encoded by the *fim* gene cluster, *fimA-I* [122], and mediate binding to mannose receptors expressed on the surface of the bladder epithelia [123]. This triggers the invasion of UPEC into the superficial umbrella cells of the bladder where they multiply and form intracellular bacterial communities (IBCs; [120,121]). The formation of IBCs allows UPEC to evade host defenses and persist in the bladder epithelia and urinary tract [121,124]. The tips of type 1 pili are short and are comprised of only three proteins (FimF-H, where FimH is the adhesin protein and FimF and FimG are adaptor proteins; Figure 3.2.1A; [118]). The rigid pilus rod is comprised of repeating monomers of FimA subunits that form a helical quaternary structure that is 6-7 nm thick with a helical cavity 20Å wide and 3.1 subunits per turn [125]. The helical rod of the pilus has an interesting architecture in that none of the subunits that comprise them associate with one another

through covalent interactions (Figure 3.2.1B). Instead, the entire pilus is held together non-covalently by hydrogen-bonding networks and hydrophobic interactions [126]. Pilin subunits have an Ig-like fold, but they are missing their seventh (G) beta strand (Figure 3.2.1B). The absence of the C-terminal seventh beta strand results in a deep groove on the surface of the pilin that exposes its hydrophobic core. During pilus biogenesis, an N-terminal extension that is present on every structural subunit completes the Ig fold of its neighbor in a process termed donor-strand exchange [126]. The importance of this structural framework in pathogenesis is unknown.

Recent mechanical measurements done on P pili using optical tweezers show that P pili readily extend under an applied force [127-129]. Type 1 pili are structurally similar to P pili and are very important in IBC formation [120] and shear-dependent binding [130,131]; yet the mechanical properties of this ubiquitous type of pili are not known (type 1 pili are expressed by ~80% of all *E. coli* strains). Here we use single-molecule atomic force microscopy (AFM) techniques [132-134] to measure the mechanical properties of uropathogenic *E. coli* type 1 pili. We found that like P pili, the rods of type 1 pili are highly extensible. This dramatic extension is a result of unwinding the pilus rod's helical quaternary structure when exposed to mechanical stress. The forced unraveling of type 1 pili is also reversible, with helical rewinding taking place under considerable forces (~60 pN). These results demonstrate that type 1 pili are dynamic structures with spring-like properties under applied forces. The "spring forces" were also shown to be additive whereby the simultaneous unwinding of several pili required a much larger force, proportional to the number of pili being extended. To better understand the molecular origin of the elastic properties of the helical rod, we used a simple two-state kinetic model and Monte-Carlo simulation techniques. We show that this model closely reproduces the experimental data and provides a simple way to predict the mechanical

behavior of pili under a wide range of physiological forces. This model predicts that pili elasticity serves as a mechanism for extending the lifetime of the adhesin-receptor interaction and explains the mechanism by which bacteria remain bound under shear forces. Our results show that reversible unraveling of type 1 pili is essential for absorbing physiological shear forces encountered during urinary tract infections and that this mechanism might be essential for successful colonization and invasion of host tissues.

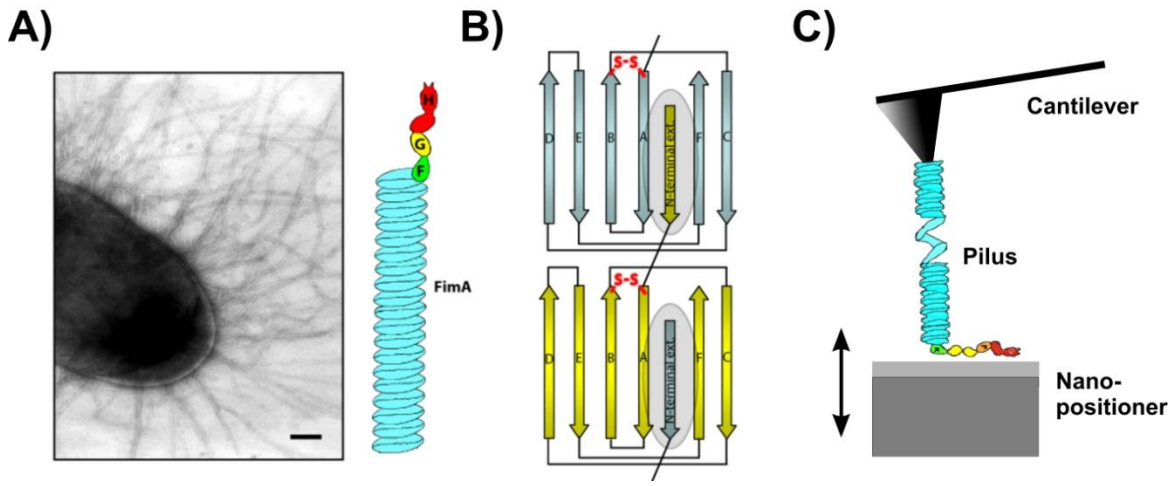

Figure - 3.2.1 Type 1 pilli and experimental setup

(A) Left: Negative-stain electron micrograph of uropathogenic E. coli (UTI89) cell expressing type 1 pili (scale bar = 200nm). Right: A cartoon diagram showing the structural subunits of type 1 pili. FimH is the adhesin at the distal end of the tip fibrillum followed by FimG and FimF. Oligos of FimA form the helical pilus rod structure that is visible in the micrograph. (B) 2D diagram of donor-strand exchange between FimA Ig-like domains. Each pilin structural subunits donates its N-terminal extension to complete its neighbor's fold and form a protein chain. The N-terminal extension is held in place through non-covalent interactions (shaded region) such as hydrogen-bonds and hydrophobic interactions. Each pilin subunit has one disulfide bond (red) between the A and B beta-strands, close to where the N-terminal extension ends. (C) Diagram of the single-molecule AFM. Purified pili were adsorbed onto a glass substrate and then stretched using the tip of a cantilever (see methods).

MATERIALS AND METHOD

Strains and plasmids

The *Escherichia coli* K-12 strains HB101 and ORN103 were as used as host strains for cloning and expression. The plasmid pPap5 carries the entire *pap* operon with its natural promoter [135] and plasmid pSH2 carries the entire *fim* operon with its natural promoter [136].

Whole Pili Purification

To induce pilus expression, the HB101 strain containing pPap5 was grown on TSA (tryptic soy agar) plates containing the appropriate antibiotic for 36 hours. The ORN103 strain containing pSH2 was grown in Luria broth with antibiotics for 48 hours at static conditions to induce pilus expression. Bound P pili were heat purified as described by Kuehn et al. [137]. Cells expressing Type 1 pili were pelleted and purified as described previously [118].

AFM

The mechanical properties of single pili was studied using a home-built single molecule AFM [132-134] that consists of a detector head (Digital Instruments) mounted on top of a single axis piezoelectric positioner with a strain gauge sensor (P841.10, Physik Instrumente). The P841 has a total travel of 15 μm and is attached to two piezo-electric positioners (P280.10A, Physik Instrumente) that are used to control the x and y position. This system has a z-axis resolution of a few nanometers and can measure forces in the range of 10-10,000 piconewtons (pN). The monitoring of the force reported by the cantilever, and the control of the movement of the piezoelectric positioners, are achieved by means of two data acquisition boards (PCI 6052E, PCI 6703, National Instruments)

and controlled by custom-written software (LabView; National Instruments and Igor, Wavemetrics). The spring constant of each individual cantilever (MLCT-AUHW: silicon nitride gold-coated cantilevers; Veeco Metrology Group, Santa Barbara, CA) was calculated using the equipartition theorem [138]. Cantilever spring constants varied between 20-50 pN/nm and rms force noise (1-kHz bandwidth) was ~10 pN. Unless noted, the pulling speed of the different force–extension curves was in the range of 1–3 nm/ms.

Pili mechanical measurements

In a typical experiment, a small aliquot of the purified pili (~1-50 µl, 10 µg/ml) was allowed to adsorb to a clean glass coverslip (for ~10 min) and then rinsed with PBS pH 7.4. Prior to use the glass coverslips were cleaned by sonication in acetone for 20 min following by boiling then for 10 min in 3N KOH and then 30% H_2O_2. Between each step, the coverslips were rinsed and sonicated with MilliQ water (>18.2 Mohm x cm). The coverslips were dried in a stream of N_2 gas. Segments of a pilus were then picked up randomly by adsorption to the cantilever tip, which was pressed down onto the sample for 1-2 seconds at forces of several nanonewtons and then stretched for several microns. The probability of picking up a pilus was typically kept low (less than one in 50 attempts) by controlling the amount of pili used to prepare the coverslips.

Analysis of Force extension curves

The elasticity of the stretched pili were analyzed using the worm-like chain (WLC) model of polymer elasticity [139,140]:

$$\mathrm{F}(x) = \frac{kT}{p}\left[\frac{1}{4}\left(1-\frac{x}{L_c}\right)^{-2} - \frac{1}{4} + \frac{x}{L_c}\right] \qquad (1)$$

where F is force, p is the persistence length, x is end-to-end length, L_c is contour length of the stretched protein. The adjustable parameters are the persistence length, p, and the contour length, L_c.

RESULTS

Stretching *E. coli* Type 1 pili using AFM

To study the mechanical properties of type 1 pili we used the AFM tip to pick up random segments of purified pili. Figure 3.2.2A-C shows typical force-extension curves obtained after stretching type 1 pili at a pulling speed of 1-3 nm/ms. One feature of these patterns is the presence of several distinct regions (vertical dashed lines, Figure 3.2.2A). The first region represents either the physical lifting of the pilus off the substrate or the natural elastic properties of the intact pilus rod. The second region most likely represents the unwinding of the pilus helical quaternary structure. This extension takes place at a constant force of ~60 pN (63 ± 17 pN, n=564, 94 different pili; Figure 3.2.2D) and seen as a plateau in the force-extension curve (dotted line). Interestingly, this force is within physiological levels of shear flow (~up to 90 pN/bacterium; [141]). We found that this plateau region can be very long reaching up to 4 microns in some cases, indicating that type 1 pili can elongate several times their unstretched length. In order to quantify the extensibility of individual type 1 pili, we used the WLC model (equation 1) that predicts the relationship between the extension of a polymer and the entropic restoring force generated [139,140]. The solid line shows the prediction of the WLC equation using a contour length L_c = 3.7 μm and a persistence length p = 1.1 nm. The average values are L_c ~2 μm (1.9 ± 0.7 μm, n=230) and p = 3.3 ± 1.6 nm (n=36). Once the rod has been completely unraveled, a larger force is then required to stretch the chain. This is seen as an increase in the slope of the force-extension curve (Figure 3.2.2A, region III). We

interpret the rupture force (height of the final peak; Figure 3.2.2A-C) as the detachment of the pilus from the AFM tip or substrate. The FimA subunits have an Ig-like fold, and several single molecule force spectroscopy studies have demonstrated that Ig-like domains unfold at forces of 50-300 pN [132,142,143]. Since FimA domains are linked head-to-tail via non-covalent interactions (Figure 3.2.1B), unfolding of a subunit will cause the whole pilus fiber to break. Hence, we cannot exclude the possibility that some of the rupture forces may result from the breakage of the connections between rod subunits. However, our data shows that the non-covalently interactions that link subunits together must be very strong since these can survive long extensions (several μm) and very high forces (up to 500 pN in some cases).

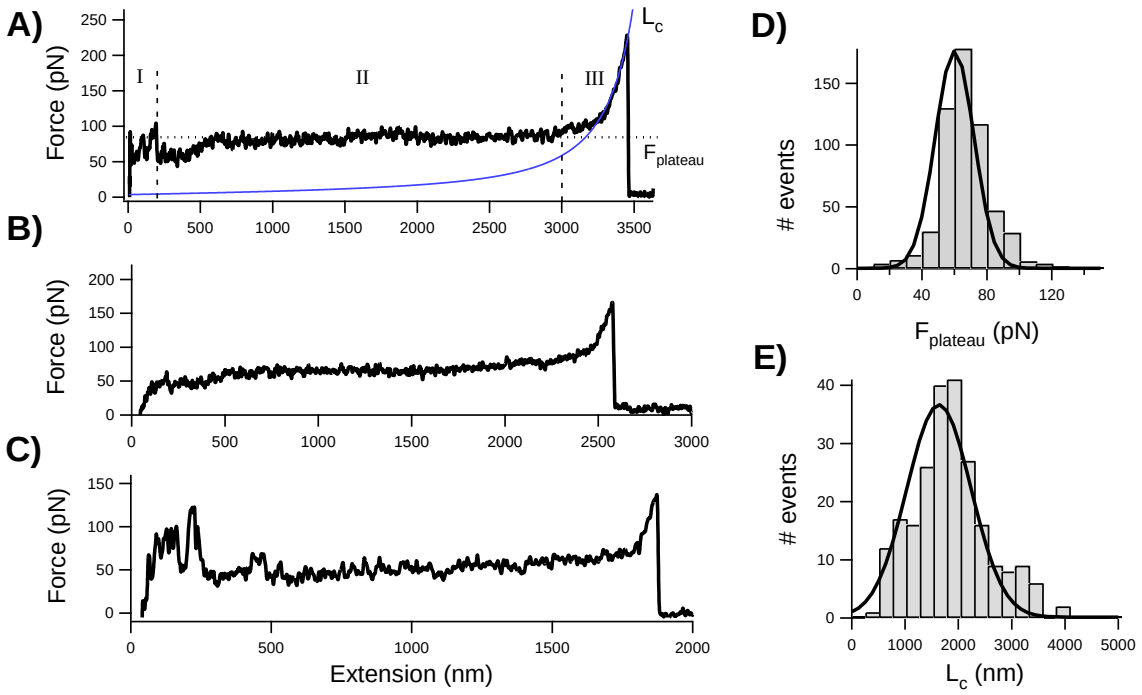

Figure 3.2.2 - Force-extension curves obtained after stretching Type 1 pili

(A-C) Examples of force-extension curves for Type 1 pili. There are three distinct regions of pili stretching which are marked by the vertical dotted lines (I, II, and III). The first region represents either the physical lifting of the pili or the natural elastic properties of the intact pili. The second region is a force plateau, $F_{plateau}$ (dashed line) that corresponds to the unwinding of the pilus rod at a constant force. The third region is the final stretching of the completely unraveled pilus rod. The continuous line shows the prediction of the WLC equation using a persistence length, *p*, of 1.2 nm. **(D and E)** Frequency histograms for the plateau force, $F_{plateau}$, and contour length, Lc. Gaussian fits gave a mean $F_{plateau}$ of ~60 pN (63 ± 17 pN, n=564, 94 different pili) and L_c ~2 μm (1.9 ± 0.7 μm, n=230).

Stretching *E. coli* P pili using AFM

We also measured the mechanical properties of *E. coli* P pili using AFM techniques. Since P pili have been studied recently using optical tweezers techniques [127], these experiments should also allow a direct comparison between the different techniques. The P and type 1 pilus rods have very similar structures and lengths [115,144,145]. FimA makes up the type 1 rod while PapA makes up the P pilus rod. PapA and FimA are 45% homologous over 95% of the protein sequence, so major differences in structural dynamics between the two pilus systems would be surprising. Figure 3.2.3 shows several examples of force-extension patterns obtained after stretching random stretches of P pili. We found that similar to type 1 pili, P pili are highly extensible with an average contour length, Lc, ~3 μm (2.9 ± 1.8 μm, n=130) and a plateau force, $F_{plateau}$, of ~35 pN (34 ± 14 pN, n=246, 48 different pili). One interesting feature of P pili force-extension curves is the presence of a 'hump' in the last elongation region (Figure 3.2.3A, dashed area). This may correspond to the simultaneous stretching of individual PapA subunits after unwinding of the rod helical structure [127]. These force-extension patterns for P pili are remarkably similar to those obtained with optical tweezers [127] which show that P pili extend at a constant force of ~30 pN (27 ± 2pN) with a clear 'hump' before the detachment force peak.

Our results show that type 1 and P Pili have similar but not identical mechanical properties. Both types of pili undergo a massive structural transition at high forces where the rod extends to several times its original length. Our data shows that the plateau region for P pili is seen at a lower forces than type 1 pili (35 vs. 60 pN), suggesting a weaker interaction between adjacent turns in the helical rod. This is consistent with scanning EM measurements of P and type 1 pili which suggests that the interaction

between single turns of P pilus helix is significantly weaker than type 1-pili [145]. Another difference is the lack of a 'hump' in the force-extension patterns of type 1. The structural significance for these differences need to be further investigated. Our results also suggest that physiological levels of shear flow (~up to 90 pN/bacterium; [141]) are likely to trigger unwinding the helical region of both types of pili.

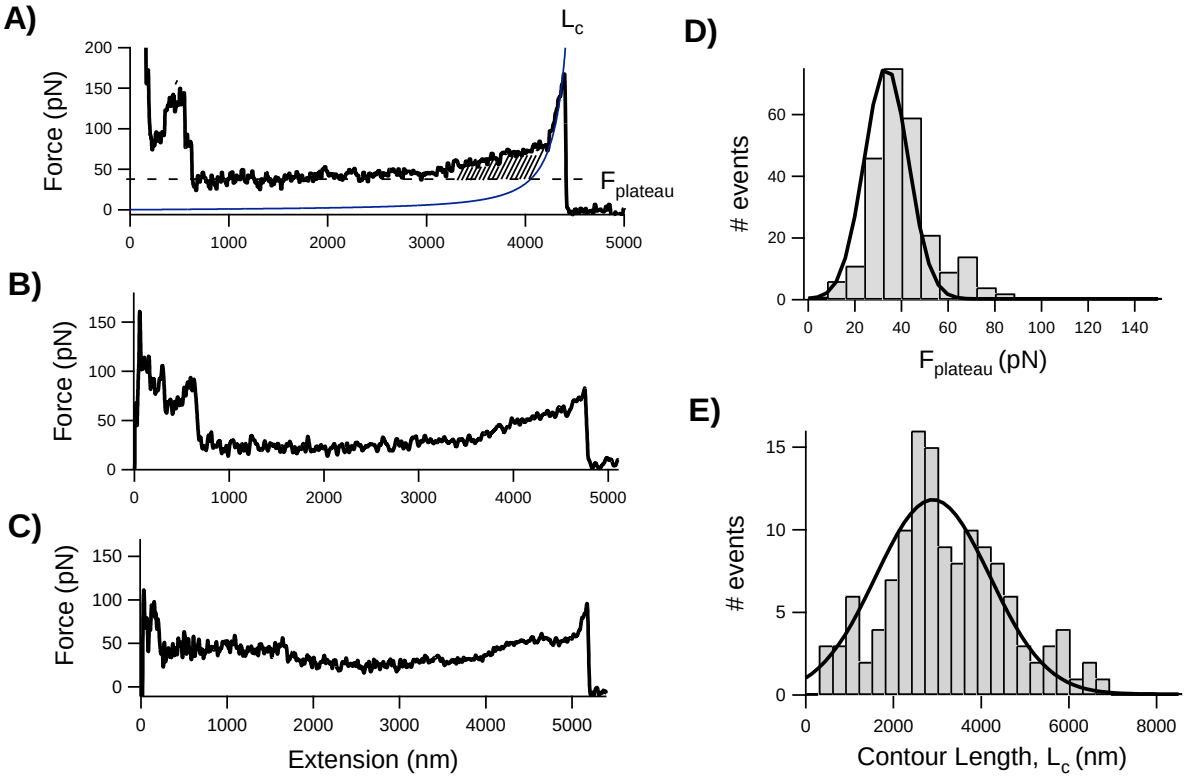

Figure 3.2.3 - Force-extension patterns for P pili

(A-C) Examples of force-extension curves for several P pili. In A), the solid line shows the prediction of the WLC equation using a persistence length, *p*, of 1.6 nm. (D and E) Frequency histograms for the plateau force, $F_{plateau}$, and contour length, Lc. The lines correspond to Gaussian fits which gave mean $F_{plateau}$ of ~35 pN (34 ± 14 pN, n=246, 48 different pili) and L_c ~3 μm (2.9 ± 1.8 μm, n=130).

The unraveling of the Type 1 helical rod structure is fully reversible

We found that type 1 pili could be stretched and relaxed repeatedly provided that we limited the extension so that the pilus did not detach from any of the attachment points (i.e., the AFM tip or the substrate). To measure the unraveling and refolding of the pilus quaternary structure, we used a double pulse protocol [128,142,146,147]. Figure 3.2.4 shows consecutive stretch/relaxation curves obtained on a type 1 pilus. These are a series of two pulling (black, forward arrow) and relaxing (red, backward arrow) cycles of a pilus (i and ii). The last trace shows the spontaneous detachment of the pili (Fig. 3.2.4iii). This recording has the typical force-extension pattern for type 1 pili and demonstrates that the previous traces corresponded to the reversible extension of a single pilus. The force-relaxation patterns in Figure 3.2.4i and Figure 3.2.4ii follow almost exactly the same trace as during pulling. The second extension-relaxation cycle, ii, starts at ~90 nm away from the coverslip to prevent picking up more pili. These data show that type 1 pili are able to quickly refold after being extended.

Our data demonstrates that the helical rod of type 1 pili is a truly elastic structure that can accommodate large increases in its length and quickly (< 1s) refold to its resting length. Our results also show that the refolding of the helical rod can take place under considerable forces. For example, Figure 3.2.4ii shows that most of the refolding takes place at forces of ~60 pN (we obtained the zero force baseline from Fig 3.2.4iii). At this force, denoted as $F_{Plateau}$, the helical rod structure is in equilibrium between the unraveled and folded conformations. Hence, very little energy is dissipated during the extension/relaxation cycles where most of the stretching energy is used back during the relaxation. These results are similar to force extension/relaxation studies done on P-pili [128], indicating that this is a conserved property of uropathogenic *E. coli* pili. This

observation is in contrast to most of other proteins investigated mechanically so far (such as titin, fibronectin or tenascin) which shows a large hysteresis in the unfolding and refolding curves [132,143,148]. Hence, *E. coli* type 1 pili are spring-like organelles designed to quickly extend and relax under force with very little hysteresis.

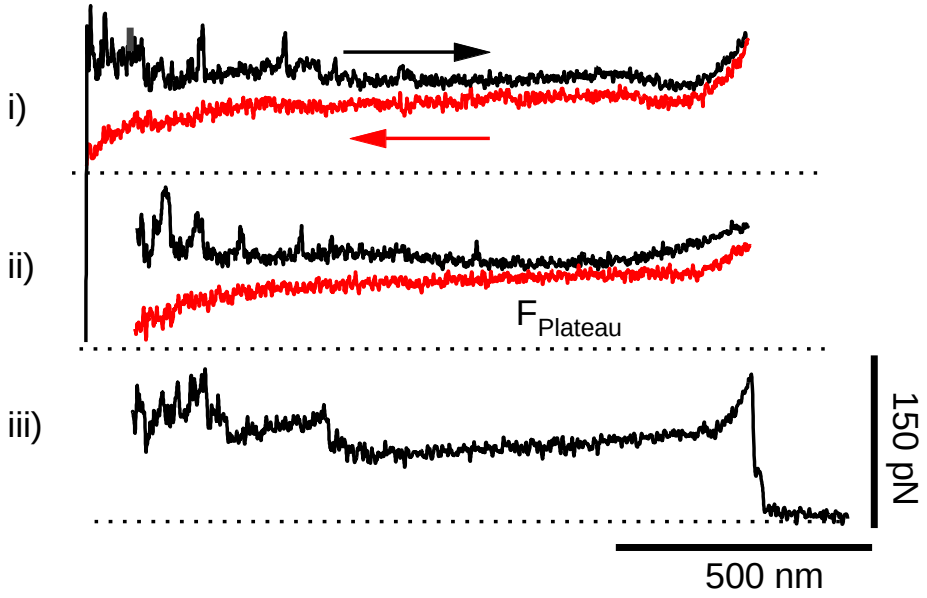

Figure 3.2.4 - forced unraveling of the helical rod structure is fully reversible

Consecutive force-extension and relaxation curves for a single Type 1 pilus using a double pulse stretching protocol (i-iii), in which the pilus was first extended (black traces) and then relaxed (red traces). The time delay between stretching pulses was 10 s. The second extension-relaxation cycle, ii, starts at ~90 nm away from the coverslip to prevent picking up more pili. The third trace, iii, corresponds to the spontaneous detachment of the pilus from the AFM tip.

Multimodal and stepwise unraveling

Bacteria bind host cells in a multivalent fashion, usually being tethered to the host surface by more than one receptor-adhesin interaction. Each *E. coli* bacterium can express about 100 type 1 pili on their surface (Figure 3.2.1A) making it likely that multiple pili attach simultaneously to host receptors. In order to mimic the effect of multiple pili binding we increased the concentration of type 1 pili adsorbed to glass coverslips. Figure 3.2.5A shows two typical recordings under these conditions. The characteristic feature is a stair-step pattern on the force-extension curve. Once a pilus detaches or breaks, the force on the cantilever drops and the pili that remain attached continue to unravel. Hence, the traces in Figure 3.2.5A correspond to the sequential detachment of many (~10) pili that connected the AFM tip to the substrate. The stair step patterns in these force-extension curves display multimodal properties. The force from baseline to each step was measured (F_{step}) and plotted as a frequency histogram (Figure 3.2.5B). This histogram shows multiple force peaks – 50, 112, 192, 272, 336 and 400 pN – reflecting the quantized nature of the pili detachment. This data shows that unraveling forces are additive, and may mimic what occurs *in vivo* when multiple pili from a single *E. coli* bind individual host receptors; unraveling forces add up and the group of pili bound can now withstand much greater forces (up to 600 pN when 10 pili are bound).

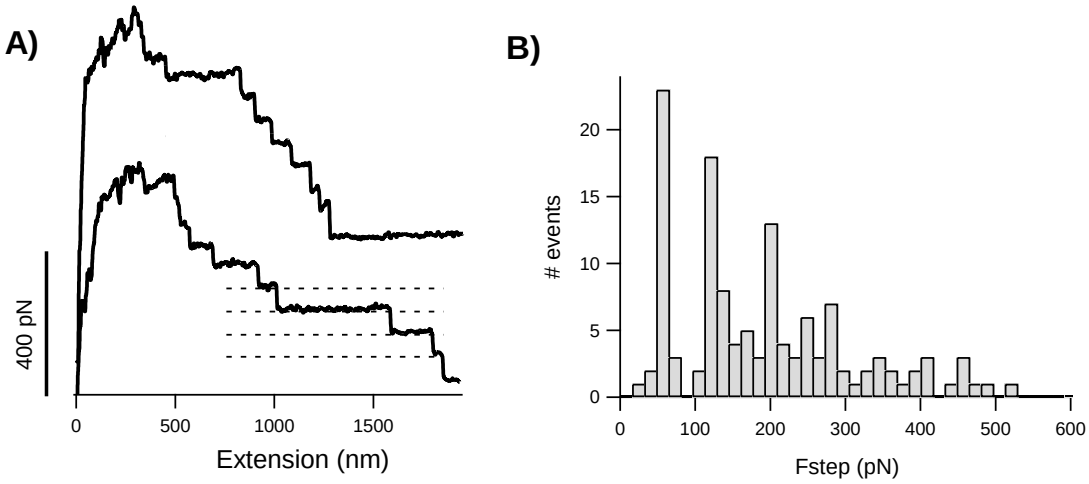

Figure 3.2.5 - Simultaneous stretching of multiple Type 1 pili

(A) To promote numerous pili binding to the cantilever, a large concentration of Type 1 pili was allowed to adsorb to the coverslips. Force-extension curves show a stair-step pattern, corresponding to the stretching and detachment of multiple pili. The number of pili being stretched was estimated by counting the number of back steps in the force-extension curves. **(B)** Frequency histogram for the force between steps, F_{step}, (dotted lines in bottom trace) reveals a multimodal force detachment pattern. The mean peaks are at found at 50, 112, 192, 272, 336 and 400 pN (n= 130 steps) with a mean F_{step} of 69 pN.

Monte Carlo simulations of pili extensibility

To better understand the molecular origin of the elastic properties of the helical rod, we used a simple two-state kinetic model and Monte-Carlo simulation techniques [148] as opposed to a sticky chain model as described previously [129]. We modeled the extensibility of the helical rod with an entropic elasticity as described by the WLC equation and a force dependent all-or-none unraveling of the individual turns. This model divides the helical rod into small folded segments of contour length l_F that can undergo an all-or-none transition into a stretched, unwound state of contour length l_U (Figure 3.2.6A). The increase in contour length upon unwinding of a turn is $\Delta L_c = l_U - l_F$. The ΔL_c was estimated from our AFM data. Figure 3.2.6C shows a force-extension curve obtained after stretching a type 1 pilus. The inset is a high resolution recording of the initial part of a force-extension which shows a clear sawtooth pattern. We interpret this sawtooth as the stepwise unwinding of consecutive turns. The continuous lines correspond to a family of curves generated with the WLC equation with a $\Delta L_c = 5$nm. We use this experimental value in our calculations.

The unraveling of the helical rod was modeled as a two state first-order (*Markov*) process, where the unwinding probability adjacent turns of the helix was $P_u = N_f * \alpha * \Delta t$ where N_f is the number of folded turns and Δt is the polling interval [132,146]. The winding probability was $P_w = N_u * \beta * \Delta t$ where N_u is the number of unwound turns. The rate constants for unwinding, α, and rewinding, β, are force dependent and are given by $\alpha(F) = \alpha_0 * \exp(F * \Delta x_u / kT)$ and $\beta(F) = \beta_0 * \exp(-F * \Delta x_f / kT)$ where F is the applied force and Δx_u and Δx_f are the unfolding and folding distances, which in a energy diagram correspond to the distance to the transition state (Figure 3.2.6B; [149]). α_0 and β_0 are the rate constants in the absence of an applied force and k and T have their usual meanings.

To simulate the extension of the helical rod, the force experienced by the pili during a stretching at a constant speed is calculated by using the WLC equation. We use this force value to compute the probability of unraveling of a turn using the Monte-Carlo approach. Figure 3.2.6D shows a Monte Carlo simulation of a force-extension curve obtained by stretching, at a constant speed (1 nm/ms), 245 subunits. As shown in Figure 3.2.6E, this simple model (green trace) closely reproduces the experimental force-extension data (orange trace). Furthermore, as shown in Figure 3.2.6F, this model accurately simulates the consecutive unwinding (black trace) and rewinding (red trace) of a single type 1 pilus (see Figure 3.2.4). Figure 3.2.6F shows a Monte-Carlo simulation of an force-extension (black trace) and force-relaxation (red trace) curves obtained by stretching/relaxing 320 subunits.

From these simulations we find that the parameters that best describe the experimental data are the following: $\alpha_o = 5x10^{-2}\ s^{-1}$; $\beta_o = 7\ x10^{2}\ s^{-1}$; $\Delta x_u = 0.2$ nm and $\Delta x_f = 0.5$ nm. We can use these kinetic parameters to estimate the free energies for the unbinding re-binding of adjacent pilin subunits. The free energies can be calculated from the rate constants α_0 and β_0 using Eyring rate theory: $\Delta G = kT*\ln(rate/A)$. Assuming a pre-exponential factor of $10^6\ s^{-1}$ [150] we estimate a $\Delta G_u = 17$ kT and a $\Delta G_f = 7$ kT. These values are similar to those estimated for P pili [129] and correspond to the energy of a typical protein-ligand bond (range: 5-30 kT; [151]).

In summary, a simple two-step Monte Carlo simulation accurately simulates the unwinding and rewinding of type 1 pili and offers an alternative model to the sticky chain model [129].

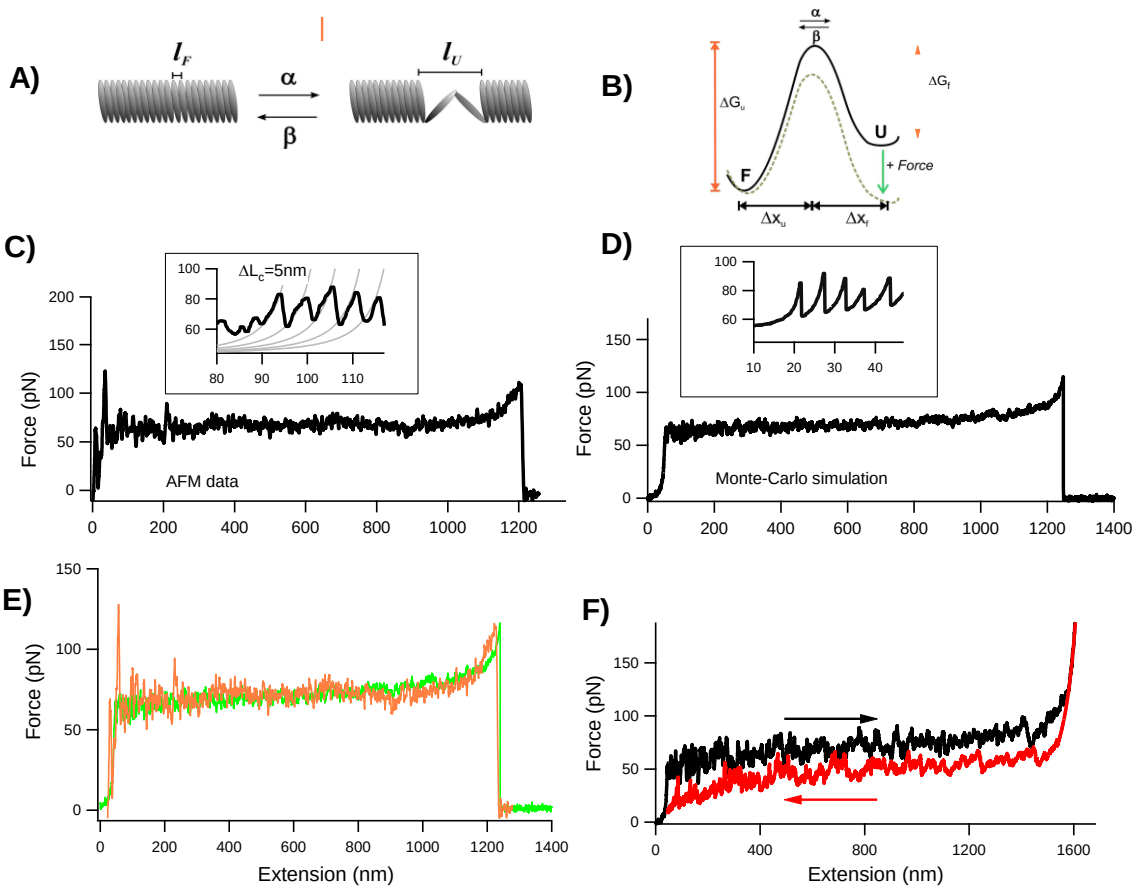

Figure 3.2.6 - Monte Carlo simulation of Type 1 pili elasticity

(A) A simple model for the reversible unwinding of the pili helical rod under a stretching force. This model divides the helical rod into small folded segments of contour length l_F that can undergo an all-or-none transition into a stretched, unwound state of contour length l_U. The increase in contour length upon unwinding of a turn is $\Delta L_c = l_U - l_F$, where $l_F \sim 3$ nm and $l_U \sim 8$ nm. **(B)** The unraveling of the helical rod was modeled as a two state Markovian process where the rate constants for unwinding, α, and refolding, β, are force dependent and are given by $\alpha=\alpha_o \exp(F\Delta x_u/kT)$ and $\beta=\beta_o \exp(-F\Delta x_f/kT)$ where F is the applied force and Δx_u and Δx_f are the unfolding and folding distances, α_0 and β_0 are the rate constants at zero force. **(C)** Experimental force-extension curve for a Type 1

pilus. The inset is a high resolution recording of the initial part of a force-extension which shows a clear sawtooth pattern. The grey lines correspond to a family of curves generated by using the WLC equation using a ΔL_c = 5nm. **(D)** Monte-Carlo simulation of a force-extension curve obtained by stretching, at a constant speed (1 nm/ms), 245 subunits. The kinetic parameters are: $\alpha_o = 5x10^{-2}\ s^{-1}$; $\beta_o = 7\ x10^2\ s^{-1}$; Δx_u= 0.2 nm and Δx_f = 0.5 nm. **(E)** Superimposition of the experimental trace (orange) and the simulated trace (green). **(F)** Monte-Carlo simulation of an force-extension (black trace) and force-relaxation (red trace) curves obtained by stretching/relaxing 320 subunits, using the same kinetic parameters as in D.

Pili extensibility can dramatically affect the lifetime of the bonds between bacteria pili and host receptors

Shear forces in the bladder and kidney caused by fluid and urine flow are a major factor that uropathogenic *E. coli* must subvert in order to persist in the urinary tract. Recent work by Thomas et al. [130,131] showed that *E coli* expressing type 1 pili bind tighter to target cells when under increased shear force and attributed this to force-driven conformational changes in the adhesin domain, FimH. However, an alternative scenario is that the receptor-ligand interaction could be modulated by force-driven elongation of the pilus rod, as proposed by Bullitt and Makowski [144]. To explore this idea we quantified the effect of pili elasticity on bond lifetime, using Monte-Carlo techniques as described by Oberhauser et al. [144]. According to this model a protein-ligand bond will break under an applied force as described by Bell [149]: $k_{off}(F) = k_{off}*\exp(F*d/kT)$, where k_{off} is the spontaneous off rate (at zero force) and d is the distance that will destabilize the bond and lead to failure. Although the parameters are not know for the specific binding affinities between FimH and mannose, we simulated the effect of pili elasticity on the lifetime of the receptor bonds under a stretching force using the P selectin-leukocyte rolling interaction as a model since it is a bond exposed to shear forces. For this bond the estimated rupture distance, d, is 0.04 nm and the off rate is $k_{off} = 0.95\ s^{-1}$ [152].

We assumed three extreme cases: A) a bond linked to a rigid rod made of 500 non-extensible turns where each turn can extend by only $\Delta L_c = 0.05$ nm (Figure 3.2.7A); B) a bond linked to a semi-rigid rod made of 400 non-extensible ($\Delta L_c = 0.05$ nm) plus 100 extensible ($\Delta L_c = 5$ nm) turns (Figure 3.2.7B); and C) a bond linked to an extensible rod made of 500 turns where the unwinding of each turn leads to an increase in contour length, $\Delta L_c = 5$ nm (Figure 3.2.7C). For these simulations we used the kinetic parameters estimated in the previous simulations (Figure 3.2.6). The pulling speed was 2 μm/s. As

the simulations show (Figures 3.2.7Ai, Bi, Ci), the elastic properties of the pili can have a dramatic effect on the lifetime of the receptor bond. A bond linked to a rigid rod tends break at high forces (~600 pN; Fig 3.2.7Aii) and survive relatively short times (~0.5 s; Fig 3.2.7Aiii). A bond linked to a semi-rigid rod tends break at lower forces (~450 pN; Fig 3.2.7Bii) and survive longer times (~1 s; Fig 3.2.7Biii). A bond linked to a fully extensible rod tends break at very low forces (~60 pN; Fig 3.2.7Cii) and survive much longer times (~up to 4s; Fig 3.2.7Ciii).

Although speculative, our simulation results clearly show that the elastic properties of the pili can have a dramatic effect on the lifetime of the receptor bond. We hypothesize that the dynamic extensibility of the long helical rod could allow the pili-receptor bond to persist over long extensions. This is important because longer persistence may lead to greater chances of activating signaling events either within the bacteria or host cell. These events may be the actin rearrangement within the host that allows bacterial invasion or the signaling of genes within the bacteria that are activated during intracellular replication. Maier et al. [153] attribute the force-dependent elongation of type IV pili as the ability to release tension generated during retraction without breaking the pilus. Thus, the same may be true for type 1 pili in controlling pilus-receptor interactions for shear-dependent binding and signaling.

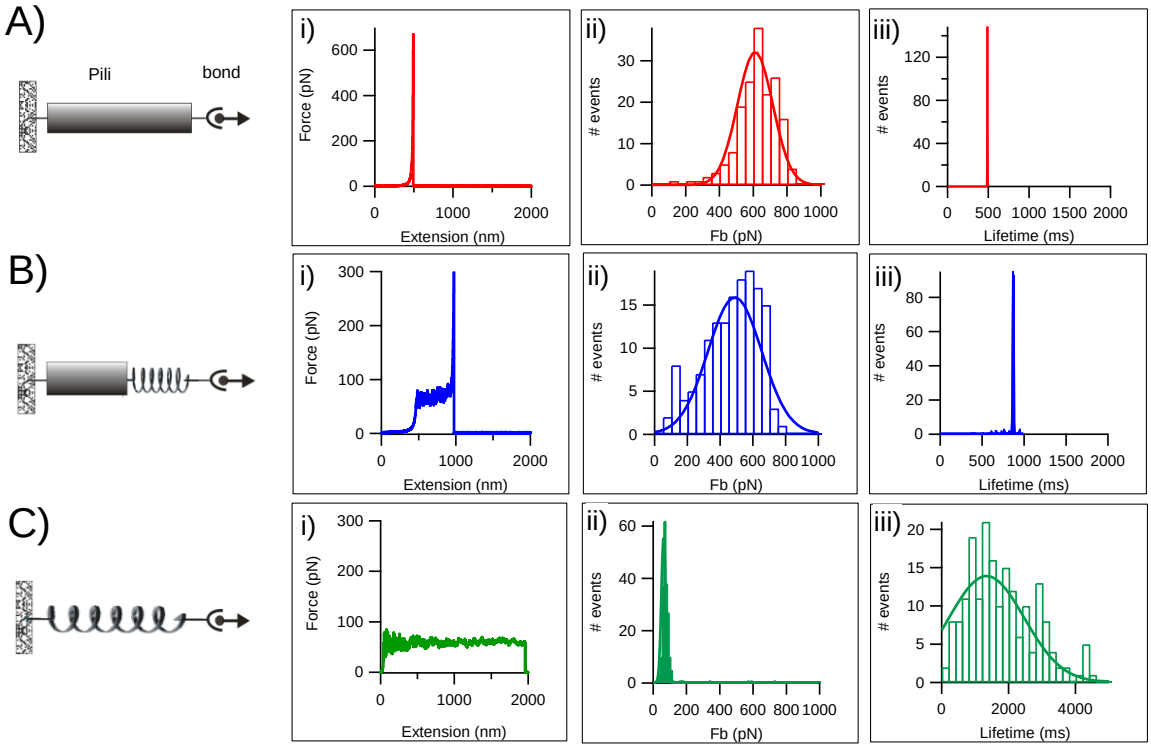

Figure 3.2.7 - Effect of pili elastic properties on bond lifetime

(A-C) Monte-Carlo simulations of the effect of pili mechanical properties on the lifetime of the receptor bonds under a stretching force. **A)** Simulation of a bond linked to a rigid rod made of 500 non-extensible turns where each turn can extend by only ΔL_c= 0.05 nm. The bond survives for 0.5 ± 0.3 s (ii) and breaks at 611 ± 148 pN, n=170 (iii). **B)** Simulation of a bond linked to a semi-rigid rod made of 400 non-extensible (ΔL_c= 0.05 nm) plus 100 extensible (ΔL_c= 5 nm) turns. The bond survives for 0.9 ± 0.3 s (ii) and breaks at 455 ± 231 pN, n=172 (iii). **C)** Simulation of a bond linked to an extensible rod made of 500 turns where the unwinding of each turn leads to an increase in contour length, ΔL_c= 5 nm. The bond survives for 1.9 ± 1.0 s (ii) and breaks at 62 ± 19 pN, n=176 (iii).

DISCUSSION

The first step in the encounter between a host and pathogen is the attachment of the pathogen to the host epithelium. These interactions begin a dynamic molecular cross-talk that ultimately determines the outcome of the infectious process. Bacterial entry into superficial umbrella cells lining the bladder lumen is a critical event in urinary tract infections. Bacterial entry into umbrella cells activates a complex genetic cascade leading to the formation of intracellular bacterial communities that undergo a defined maturation and differentiation program that is critical in disease. Type 1 and P pili are critical organelles and virulence factors for uropathogenic *E. coli*.

Our results demonstrate that pili are dynamic structures that function as molecular springs under applied forces. Using single-molecule AFM we found that the rods of both P pili and type 1 pili are highly extensible (they can be extended ~2-4 times their resting length). However, the required forces to unravel each of these pili are different (35 vs. 60 pN). This could possibly represent a specific adaptation to the biological niche in which type 1 pili promote bacterial colonization. The structural significance for this is to be further investigated. The forced unraveling of type 1 pili is also reversible, taking place under considerable forces (~60 pN). Together, these results suggest a conserved structural mechanism that is used in bacterial pathogenesis.

Importantly, type 1 pili elasticity provides a mechanism for extending the lifetime of its adhesin-receptor interaction. It has been shown that under shear flow bacteria can transition between an unbound, rolling, and stationary state [130,141]. Pilus unraveling provides a simple mechanism for explaining these transitional changes. When shear flow is low, drag forces are not enough to induce pilus unraveling and it acts as a rigid rod, promoting short-lived FimH-mannose interactions. As shear forces increase from

moderate to high flow, bacteria transition to rolling and stationary states. At high shear forces, the probability of unraveling also increases, promoting longer FimH-mannose interactions. Therefore, at moderate to high shear stress (drag forces <30pN) a rolling state would be expected where only a small population of the pili are unraveling under force and extending their bond lifetime. Additionally, the number of unraveling pili necessary to keep the bacterium stationary is only achieved under high shear stress.

Adhesin-receptor interactions are critical in the pathogenic cascade for binding and invasion; however, pilus dynamics may also play a major role in pathogenesis. Being able to unravel under applied forces, pili increase the lifetime of their adhesin-receptor interaction. Being able to refold under considerable force, pili are capable of acting like molecular springs and shock absorbers. And being able to achieve multimodal properties, pili are able to work in unison and withstand greater shear forces as they arise. The ability for these protein complexes to behave in such a manner is remarkable. These results not only lend a new understanding to how bacteria combat hospitable environments within the host, but also offer new insight into protein interactions and the functionality of protein complexes.

CHAPTER 4 – TITIN AND TITIN-LIKE PROTEINS

Major contributors in section 2 include: Belinda Bullard, Vladimir Benes, Mark C. Leake, Wolfgang A. Linke, and Andres F.Oberhauser. Major contributors in section 3 include: Dina N. Greene, R. Bryan Sutton, Kim M. Gernert, Guy M. Benian, and Andres F. Oberhauser.

Section 1: Overview of Titin and Titin-like Proteins

Titin is a rope-like protein that spans half the length of a sarcomere with opposite ends embedded in the Z-disc and M-line. There are well over 300 exons in human titin which form splice-isoforms that tailor the resting tension, length, elasticity, and other properties of the protein to a particular purpose. Titin isoforms are truly gigantic proteins varying in size up to 3-4MDa, and composed primarily of Fn3 and Ig domains arranged in repetitive patterns with several areas of unique sequence and a kinase domain thought to be mechanically activated. One of the large areas of unique sequence is rich in proline, glutamate, valine, and lysine which are arranged in tandem repeats. This PEVK region is thought to be unstructured and acts as an elastic element. Titin seems to have arisen around the time of early vertebrates and is well conserved in fish, birds, and mammals.

Invertebrates posses similar high molecular weight muscle proteins. Projectin and kettin are found in the flight muscles of *Drosophila* and are members of the titin protein superfamily [154]. Whereas a single titin molecule will stretch from the M-line to the Z-disc of a vertebrate sarcomere, in invertebrate muscles kettin is found in the I-band and projectin is usually found in the A-band. Projectin (800-1000kDa) has a similar composition to A-band titin including a kinase domain near the C-terminus and a region rich in proline, glutamate, valine, and lysine near the N-terminus. Kettin is a shorter protein (~540kDa) composed of 35 Ig domains similar to I-band titin.

Two titin-like proteins are also found in *C. elegans* and related species: twitchin (named for the phenotypic twitching caused by a mutant form) and TTN-1. Again, these proteins are comprised mostly of Ig/Fn3 repeats. Twitchin (~800kDa) has 30 Ig domains, 31 Fn3 domains, and a single kinase domain [155,156]. TTN-1 is significantly larger at

2.2MDa and contains 56 Ig domains and 11 Fn3 domains. Additionally TTN-1 contains regions predicted to be coiled-coil structures and several tandem repeats the largest of which is called PEVT. This PEVT region is similar in composition and repeat structure to the PEVK elastic element of vertebrate titin, and is thus thought to have a similar elastic function.

It is clear that these giant proteins have many important functions in the sarcomere, mechanically and otherwise. They are closely associated with every important structure in the sarcomere and have many binding partners. In some cases the protein-protein interactions are required to be very mechanically strong and resilient such as the binding into the M-line or Z-disc. In other cases it is thought that reversible domain unfolding serves as a safety mechanism protecting titin and the sarcomere from mechanical damage damage in case of extreme stretch during stress (e.g. hemodynamic overload) or pathological conditions (e.g. chronic heart disease) [157-163].

Section 2: The Molecular Elasticity of the Insect Flight Muscle Proteins Projectin and Kettin

INTRODUCTION

The success of insects as a major animal group may be attributed in part to the evolution of asynchronous flight muscles [164]. In asynchronous muscles there is asynchrony between muscle electrical and mechanical activity in that a single muscle action potential can trigger a series of contraction-relaxation cycles. In the indirect flight muscle (IFM), these oscillatory contractions are produced by delayed activation in response to stretch combined with the resonant properties of the thorax [165]. For example, some insect flight muscles can operate at very high frequencies (100-1000Hz; ref. 1). This rapid oscillatory contraction requires that the sarcomeres are stiff. This stiffness coupled with a stretch activation response allows insects' wings to beat hundreds of times per second. The molecular basis for this mechanism is not well understood but many proteins are emerging as contributing factors. Projectin and kettin form a mechanical link between the Z-discs and the ends of the thick filaments and are responsible for a large part of the passive elasticity of insect muscles [155,156,166] and may be responsible for the high relaxed stiffness as a prerequisite for the stretch activation response [167].

Projectin and kettin are members of the titin protein superfamily [154]. They are high molecular weight proteins found in invertebrate muscles. Kettin is in the I-band and projectin is in the A-band, except in IFM, where a large part of the molecule is in the short I-band. Projectin is an 800-1000 kDa protein consisting of long sections of repeated immunoglobulin (Ig) and fibronectin (FnIII) domains. There is also a kinase domain near the C-terminal end and a region rich in proline, glutamate, valine, and lysine (PEVK) near the N terminus. Immunofluorescence data indicate that, in the IFM, projectin

molecules are anchored at the Z-disc, extend over the I-band region and associate with myosin at the A-band edge [168,169]. Mutant forms of projectin have been shown to alter insect flight dynamics [170]. Kettin is an alternatively spliced product of the *Drosophila sls* gene. In contrast to projectin, kettin is a ~540-kDa protein and is made up of 35 repeating Ig domains. In the IFM, the molecule is anchored to the Z-disc, extends over the I-band running along the actin filament and then attaches to the thick filaments [171,172]. Kettin may have an essential function for sarcomere formation because adult fruit flies heterozygous for a kettin mutation cannot fly [173]. Laser-tweezers experiments suggest that kettin may have roles consistent with a provider of passive tension based both on entropic elasticity and folding of Ig domains [174].

The structure and location of both projectin and kettin suggest that they may have regions that are exposed to mechanical forces and hence contribute to the passive stiffness of the IFM sarcomere. Here we used single molecule atomic force microscopy (AFM) techniques [132,146,175] to examine the molecular elasticity of single projectin and kettin molecules. The results show that projectin and kettin Ig/FnIII domains refold much faster than titin domains, even under appreciable forces, which hints at the potential for a novel folding-based spring mechanism.

RESULTS

The flexibility of projectin and kettin molecules measured by EM techniques.

Projectin and kettin molecules are thought to function as elastic filaments in IFM fibers [176]. This elasticity may result from the flexibility of their tandemly arranged domains, the stretching of the PEVK region and also from the unfolding and refolding of individual Ig and FnIII domains [177]. We estimated the flexibility of intact projectin molecules by using rotary shadow electron microscopy techniques [166]. EM images of

Lethocerus projectin molecules show worm-like structures that appear to be flexible throughout their length (Fig. 4.2.1A). We found that the contour length, Lc (gray line in Fig. 4.2.1A), varied between ~50 and ~250 nm. The upper limit value is similar to the predicted length of a protein with 78 Ig and FnIII domains [178] (assuming 4nm/domain, [179]). The shorter molecules are degradation products and are useful here because they facilitate the estimation of the persistence length. According to models of polymer elasticity [180] the expected value of the end-to-end length, x, is related to the contour length of the polymer chain, Lc, and the flexibility of the polymer, measured by its persistence length, p:

$$\left\langle x^2 \right\rangle_{2D} = 4pL_c\left(1-\frac{2p}{L_c}\left(1-e^{\frac{-Lc}{2p}}\right)\right) \quad (1)$$

From these images we extracted the persistence length value by plotting x (black line in Fig. 4.2.1A), vs. Lc. The lines in Fig. 4.2.1B correspond to the prediction of Eq. 4.2.1 at three different values of p: 20, 30 and 40 nm. We found that a value of p = 30 nm best describes the experimental data. Similar analysis done for kettin molecules revealed a p= 45 nm (not shown). The higher p observed for kettin suggests that the Ig domain chains are less flexible than the Ig/FnIII domain chains in projectin. These values are higher than those for native skeletal titin and recombinant titin molecules estimated by EM (p = 13 nm and 9.8 nm, respectively; refs. [143,148]) but similar to those measured for titin using immunofluorescence microscopy and myofibril mechanics (p= 20-40 nm; ref. [181]). Structural and modeling techniques have shown that titin predicted to have an extended and relatively stiff conformation [182-184]. The high p values in projectin and kettin suggest that at least 6-8 domains must be directionally correlated in these molecules; this

long range order resulting in a relatively high stiffness. Thus, straightening of projectin or kettin Ig/FnIII domain chains during sarcomere extension would require little force.

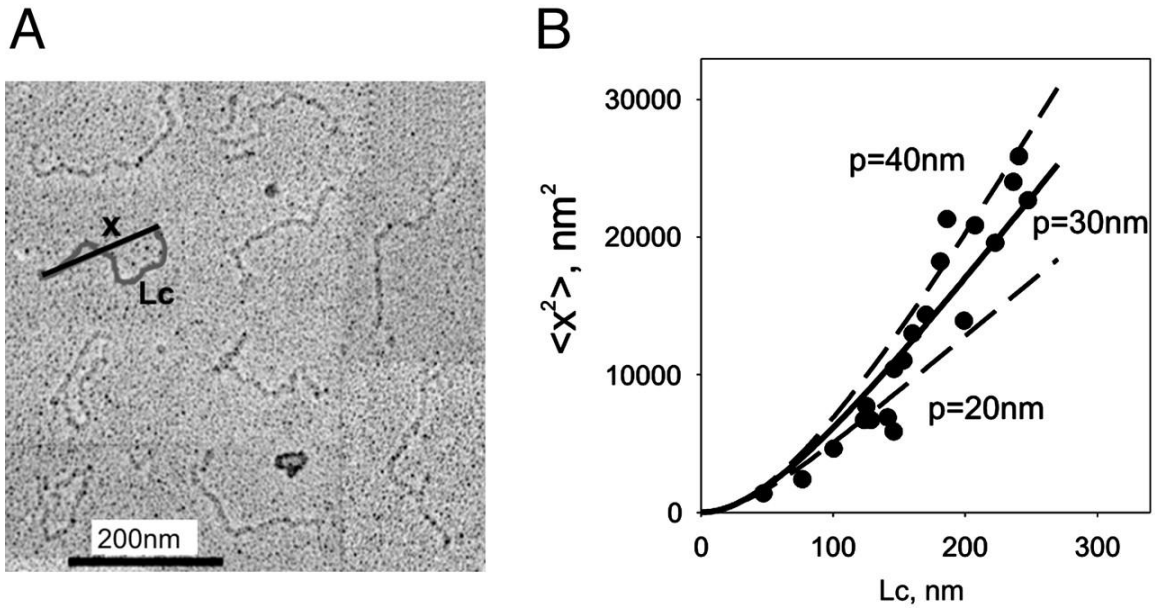

Figure 4.2.1 - Flexibility of single projectin molecules.

A) Rotary shadowed electron micrograph showing individual projectin proteins (micrograph courtesy of Dr. Kevin Leonard). The grey line shows the measured contour length, *L*c, and the black line the measured end-to-end distance, *x*. (**B)** Plot of the square of the average *x* as a function of the contour length (filled circles, data obtained from 20 molecules). The solid line is a nonlinear fit of equation 4.2.1, giving a value of *p*=30. The dashed lines correspond to fits of *p*=20 nm and *p*=40 nm.

Force-extension relationships of projectin molecules.

To measure the elastic properties of projectin we used single-molecule force spectroscopy techniques. Random segments of projectin were picked up by the AFM tip and then stretched with a pulling speed of ~0.5 nm/ms. The resulting force-extension curves (Fig. 4.2.2 A,B) showed sawtooth patterns which are characteristic of the unfolding of FnIII and Ig domains [132,146,175]. One striking feature of these sawtooth patterns is the presence of distinct levels of unfolding forces (dotted lines in Fig. 4.2.2Aa). Fig. 4.2.2D shows a histogram of unfolding forces measured for 36 projectin molecules. There are two prominent peaks, one centered at ~90 pN and a second at ~170 pN (n= 478 force peaks). Most of the projectin protein is arranged in a repeating pattern of Ig-FnIII-FnIII domains (Fig. 4.2.2). One simple explanation is that FnIII and Ig domains have a different mechanical stability and this would account for the two levels of unfolding forces observed in the sawtooth patterns.

We tested this hypothesis by analyzing the mechanical properties of a recombinant protein containing a small number of Ig and FnIII domains. Fig. 4.2.2E shows force-extension curves for a recombinant protein with 3 Ig and 4 FnIII domains (Ig24-FnIII-FnIII-Ig25-FnIII-FnIII-Ig26; top of Fig. 4.2.2E). Stretching this protein resulted in force extension curves (Fig. 4.2.2E) with equally spaced force peaks but with two distinct levels of unfolding forces, one at ~80 pN and the second at ~170 pN (Fig. 4.2.2F). We attribute the low force peaks to the unfolding of FnIII domains and the high force peaks to the unfolding of the Ig domains. Hence, these data shows that in the PIg24-PIg26 protein the FnIII and Ig domains are unfolding in a hierarchical pattern where the mechanically weaker FnIII domains unfold before the Ig domains.

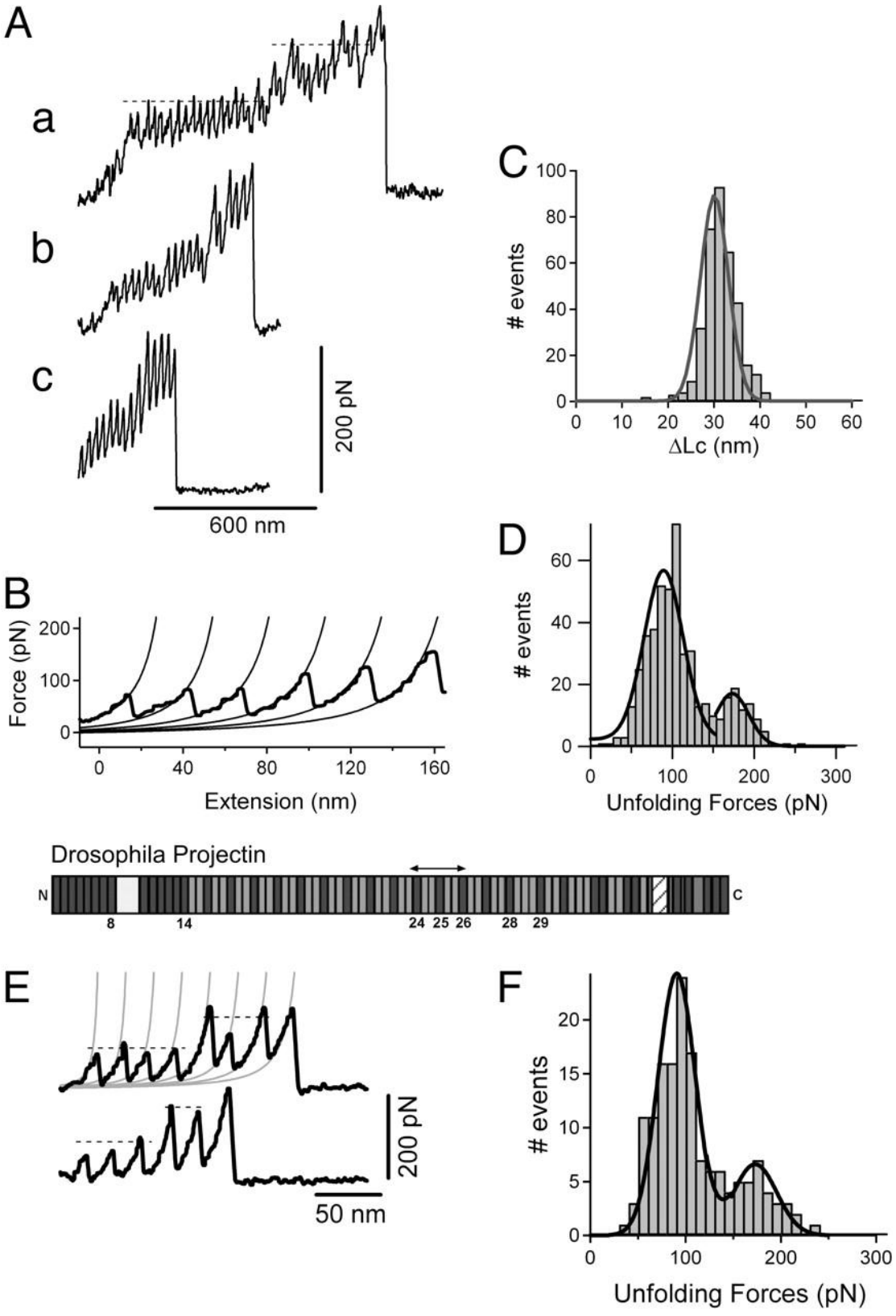
A
a
b
c
200 pN
600 nm
B
Force (pN)
200
100
0
0
40
80
120
160
Extension (nm)
C
events
100
80
60
40
20
0
0
10
20
30
40
50
60
ΔLc (nm)
D
events
60
40
20
0
0
100
200
300
Unfolding Forces (pN)
Drosophila Projectin
N
C
8
14
24 25 26
28
29
E
200 pN
50 nm
F
events
20
15
10
5
0
0
100
200
300
Unfolding Forces (pN)

Figure 4.2.2 - Force-extension relationships of projectin molecules.

A) Several examples of force-extension curves obtained after stretching single projectin molecules. **B)** To analyze the spacing between peaks in the sawtooth patterns we used the WLC model. The lines were generated using a p = 0.4 nm and ΔLc = 29 nm. **C)** Histogram of contour length increments observed upon unfolding, ΔLc (Gaussian fit: 30.1 ± 4.3 nm, n=362). **D)** Histogram of force peaks shows two main peaks (Gaussian fits: 89.1 pN and 172.3 pN; mean force peak value: 109.4 $\pm$ 41 pN, n=479). **E)** Mechanical properties of a projectin recombinant fragment containing 3 Ig and 4 FnIII domains (PIg24-PIg26). Top: Domain structure of *Drosophila* projectin. **F)** Histogram of unfolding forces for PIg24-PIg26 shows two peaks: at ~83 pN and at ~171pN (n=142).

Force-extension relationships of kettin and upstream Sls Ig domains.

We analyzed the force-extension spectra of three different recombinant kettin proteins (Fig. 4.2.3A): 1) a 3-Ig-domain fragment, SIg4-SIg6, from the 8 Ig segment in the N-terminal region of Sls upstream of kettin (these are part of a splice isoform that is in the M-line in IFM, rather than the I-band), 2) a 5-Ig-domain fragment from the actin binding region of kettin (KIg17-KIg21), and 3) a 2-Ig-domain fragment from the putatively elastic kettin region near the A-band edge (KIg34/35).

Fig. 4.2.3B shows examples of force-extension patterns obtained from the SIg4-SIg6 and KIg17-KIg21 fragments. There are clear differences in unfolding forces between these two proteins. The Ig domains in SIg4-SIg6 tend to unfold at lower forces (~100 pN) than the Ig domains in KIg17-KIg21 (~200 pN). In addition, in SIg4-SIg6, the first peak is of lower force than the last peak suggesting that the three Ig domains have different mechanical stabilities. This rising force effect is less pronounced in KIg17-KIg21, where all domains show similar unfolding forces.

Frequency histograms measured from SIg4-SIg6, KIg17-KIg21 and KIg34/35 show that the mean unfolding forces are 120 pN (123$\pm$24 pN, n=371), 190 pN (193$\pm$52 pN, n=104) and 250 pN (248$\pm$34 pN, n=192), respectively (Fig. 4.2.3C). Fig. 4.2.3D shows an interesting mechanical hierarchy: Ig domains closer to the N-terminus of Sls/kettin are mechanically less stable than those nearer the C-terminus

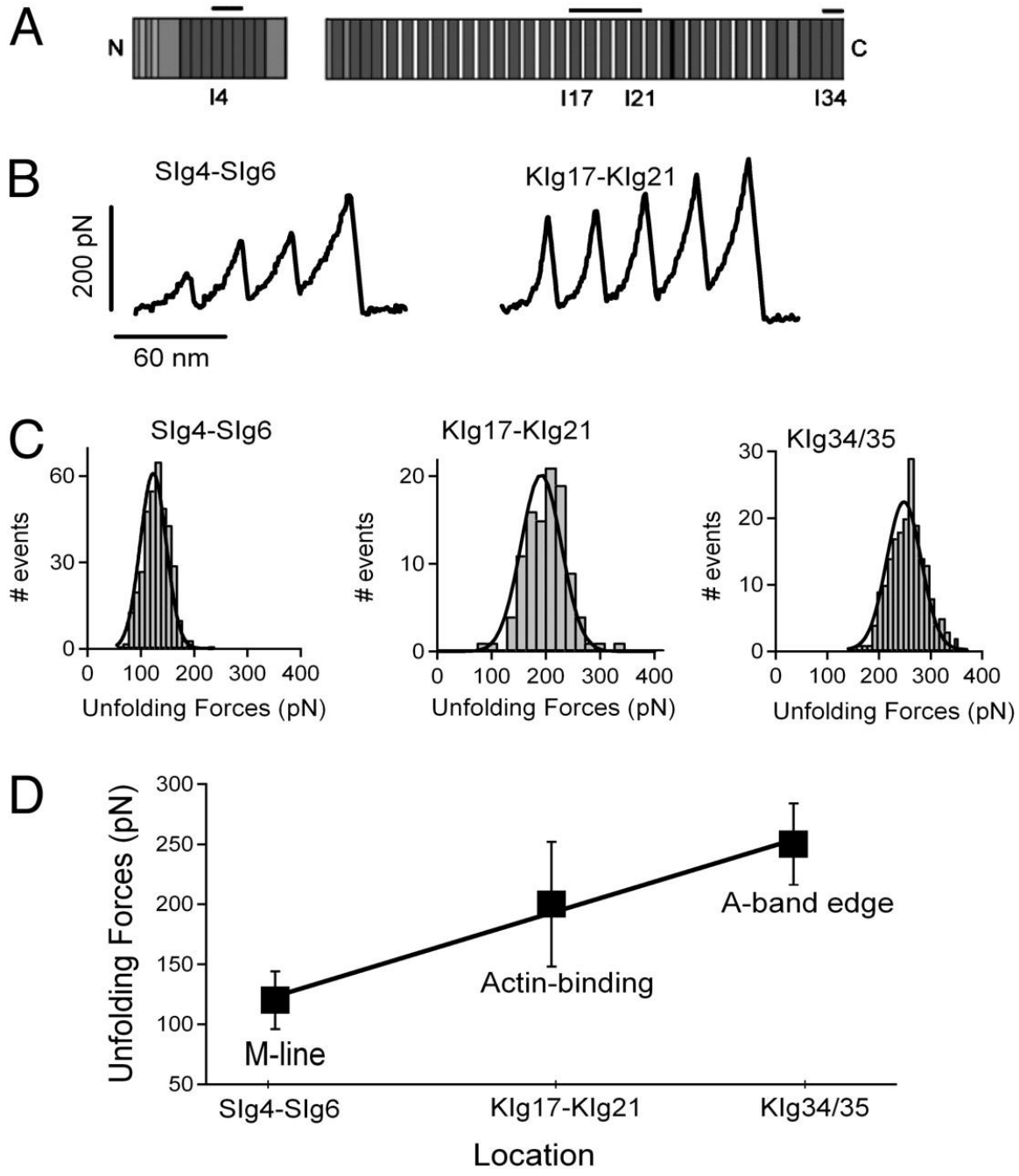

Figure 4.2.3 - Force-extension relationships of recombinant kettin and N-terminal Sls fragments.

A) Location of the SIg4-SIg6, KIg17-KIg21 and KIg34/35 in the Sls/kettin sequence. The two blocks of sequence are spliced products of the *Drosophila sls* gene; the longer protein is kettin. **B)** Examples of force-extension patterns obtained from the 3-Ig (left) and 5-Ig (right) fragment. **C)** Histogram of unfolding forces for the proteins SIg4-SIg6, KIg17-KIg21 and KIg34/35 show that their mean unfolding forces are 123±24 pN (n=371), 193±52 pN (n=104) and 248±34 pN (n=192). **D)** Plot of the mean unfolding forces of SIg4-SIg6, KIg17-KIg21 and KIg34/35 proteins vs. their location in the Sls and kettin sequence. Linear regression shows a slope of 4.4 pN/domain.

Force clamp unfolding of Kettin and Projectin.

We also used the force-clamp mode of the AFM for stretching single kettin and projectin molecules (Fig. 4.2.4). The advantage of this mode is that it is possible to measure the force dependence of the unfolding probability of single protein domains [175,185]. In these experiments we applied a force that increases linearly with time (Fig. 4.2.4A, C bottom traces) and observe the unfolding of single domains as a stepwise elongation of the proteins (Fig. 4.2.4A, C top traces).

Fig. 4.2.4A shows the stepwise elongation (top trace) of a single projectin molecule observed after increasing the force at a rate of 200 pN/s (bottom trace; the downward transients are caused by the feedback lag and can be used as markers for the unfolding events). In this example the protein first elongates slowly (we count 5 steps within the first 200 pN) and then there is rapid elongation of the protein during the next 70 pN (the large steps are due to the simultaneous unfolding of several domains). At ~ 300 pN the protein detached from the surface and this is seen as a sudden drop of the force due to the loss of the force-clamp. From these data we can calculate the distribution function for the probability of unfolding, P_u, as a function of the applied force. Fig. 4.2.4B shows the frequency of unfolding events (bars) and the unfolding probability distribution, P_u (squares) as a function of the force. The probability of unfolding changes from P_u = 0.1 to 0.9 over a wide range of forces (~150 pN; from 90 pN to 240 pN). In contrast, stretching a 5 Ig kettin fragment (KIg17-KIg21) with a force ramp (Fig. 4.2.4C) shows an unfolding probability that changes from P_u = 0.1-0.9 over a smaller force range of ~80 pN (90-170 pN; Fig. 4.2.4D, circles).

In order to analyze the data of Fig. 4.2.4B and 4.2.4D quantitatively, we used a simple two-state kinetic model for mechanical unfolding ([175,185]; see *Supporting Text*). In

this model, a protein is exposed to a force that increases linearly with time, simulating the conditions of our force-ramp experiment; the variables in model are the rate constant at zero force (α_0) and the distance the transition state (Δx_u). For the kettin fragment, KIg17-KIg21, values of $\alpha_0 = 9 \times 10^{-3}$ s^{-1} and $\Delta x_u = 0.17$ nm readily describe the data (solid line, Fig. 4.2.4D; equation 4.2.S2). For projectin the Pu vs. force data was best described by using two sets of parameters: $\alpha_{01} = 7 \times 10^{-2}$ s^{-1} and $\Delta x_{u1} = 0.1$ nm and $\alpha_{02} = 0.3 \times 10^{-3}$ s^{-1} and $\Delta x_{u2} = 0.2$ nm (solid line, Fig. 4.2.4B; equation 4.2.S3). The simplest interpretation is that in the case of projectin, FnIII domains have a higher unfolding rate constant at zero force ($\alpha_0 = 7 \times 10^{-2}$ s^{-1}) than the Ig domains ($\alpha_0 = 0.3 \times 10^{-3}$ s^{-1}) and this would result in higher unfolding probability for FnIII domains at a given force than for the Ig domains. The kettin construct has only Ig domains, which are more stable.

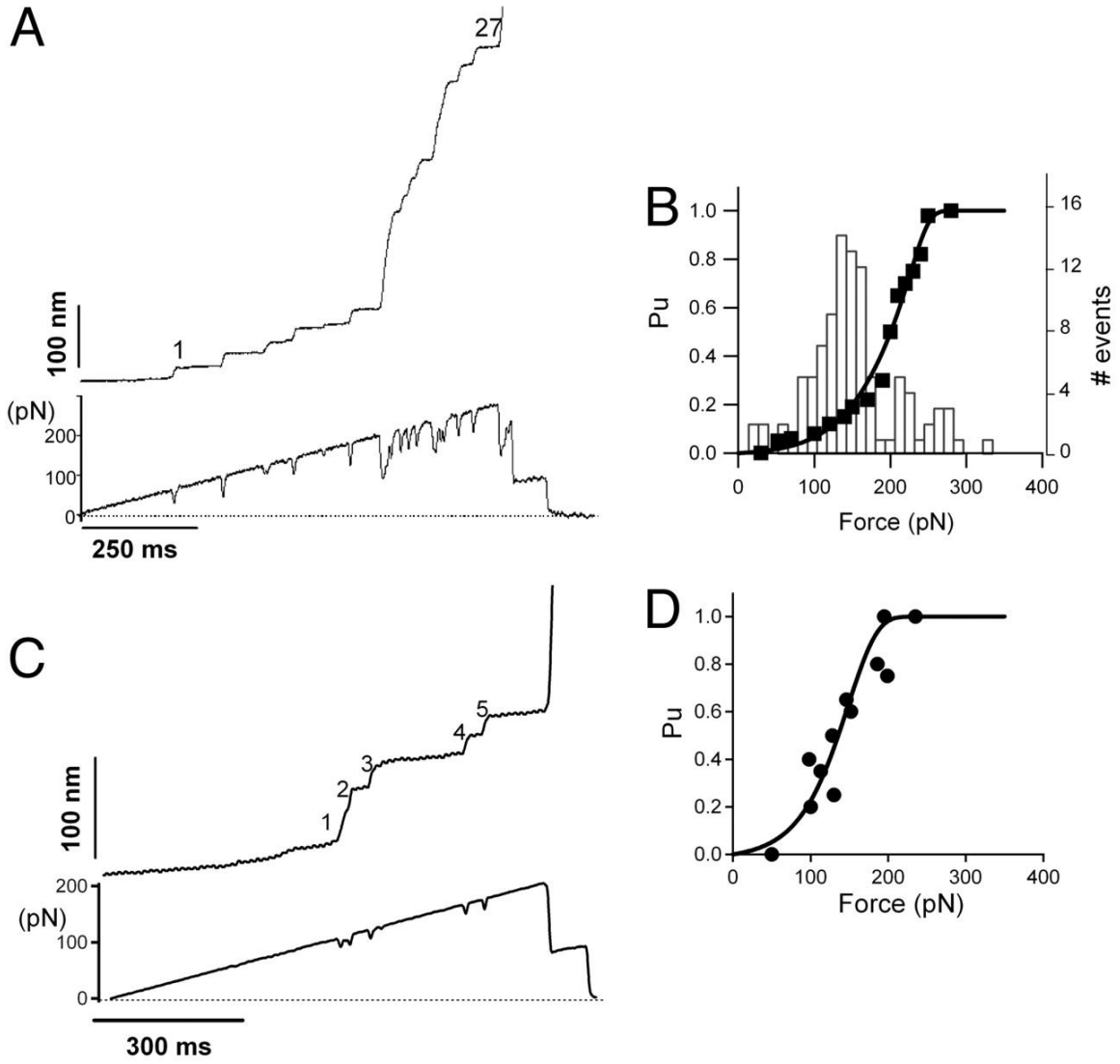

Figure 4.2.4 - Measurements of the force dependence of the unfolding probability of projectin and kettin molecules.

A, C) Stepwise unfolding of single native projectin and a 5 Ig kettin (KIg17-KIg21) fragment using the force-ramp method. The lower trace shows the time course of the force. **B)** Frequency histogram of unfolding forces (bars) and the unfolding probability, Pu (squares), as a function of the applied force measured for 6 projectin molecules with similar number of unfolding events (86 total steps). The pulling speed was 200 pN/s. The

lines correspond to the prediction of equation 4.2.S3 using α_{01} = 7 x 10^{-2} s^{-1} and Δx_{u1} = 0.1 nm and α_{02} = 0.3 x 10^{-3} s^{-1} and Δx_{u2} = 0.2 nm (continuous line). **D**) Plot of the unfolding probability, Pu, as a function of the applied force for the KIg17-KIg21 kettin fragment (circles, 19 steps from 4 experiments). The pulling speed was 150 pN/s. The line corresponds to the prediction of equation 4.2.S2 using α_0 = 9 x 10^{-3} s^{-1} and Δx = 0.17 nm.

Refolding of FnIII/Ig projectin domains.

In order to test if projectin domains refold after mechanical unfolding, we repeatedly stretched and relaxed single proteins. We found that single projectin molecules could be subjected to hundreds of stretching/relaxation cycles were the force-extension curves displayed similar patterns (Fig. 4.2.S4), demonstrating that domain unfolding is fully reversible and that, unlike mammalian titin domains [186], projectin domains can undergo multiple cycles of extension/relaxation with no signs of molecular fatigue or rundown.

To measure the refolding kinetics, we used a double pulse protocol in which the pulse interval was varied (Fig. 4.2.5A). Also the temperature was varied between 25^{o}C (room temperature) and 13^{o}C in these experiments. We found that 25^{o}C, the number of force peaks in the first and second stretching pulse was very similar, even at the shortest time interval between stretching pulses (~100 ms). This indicates that projectin domains refold in the millisecond time scale. In order to better resolve the refolding kinetics we repeated these experiments at the lower temperature of 13^{o}C (Fig. 4.2.5A). The first two traces correspond to the first and second stretching pulse, using a time delay of 100ms. Only ~50% of the domains were seen to refold. Increasing the time interval between stretching pulses allows a larger fraction of domains to refold, ~70% (delay of 1s) and 100% (delay of 10s), respectively. We observed that the number of projectin domains refolded recovered with a double exponential time course (Fig. 4.2.5B, triangles). About 68% of the recovery (at 13^{o}C) occurred at a fast rate (6 s^{-1}) whereas 32% occurred at a much slower rate (0.1 s^{-1}) indicative of two populations of domains refolding with very different rate constants. At 25^{o}C the fast rate increased by 2.5-fold to ~15 s^{-1} (Fig. 5B, circles).

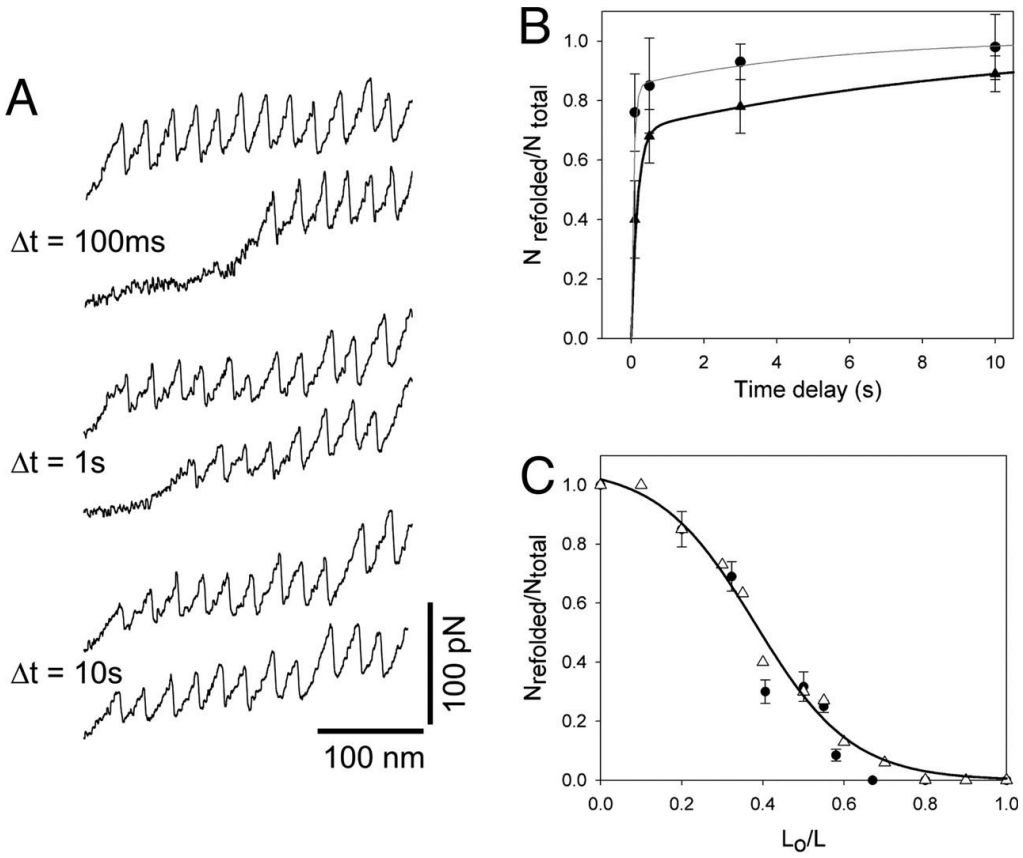

Figure 4.2.5 - Refolding kinetics of projectin domains.

A) The time interval between extensions affects the fraction of refolded domains (recorded at 13°C). **B)** Fraction of refolded domains as a function of the time delay between stretching pulses, measured at 13°C (triangles) and 25°C (circles). The solid lines are a two-exponential fit of the data to the function $N_{refolded}/N_{total} = A_1\left(1-e^{-\Delta t \bullet \beta_1}\right)+ A_2\left(1-e^{-\Delta t \bullet \beta_2}\right)$, where A_1= 0.7, A_2=0.3, β_1=6s^{-1} and β_2=0.1s^{-1} at 13°C, and A_1= 0.85, A_2=0.15, β_1=15s^{-1} and β_2=0.18s^{-1} at 25°C. **C)** Plot of the fraction of refolded domains, $N_{refolded}/N_{total}$, vs. the degree of relaxation, L_0/L, for 2 projectin molecules. The open triangles correspond to a Monte-Carlo simulation using an βo = 15 s^{-1} and Δx_f = 1.1 nm. The line is a polynomial fit to the Monte-Carlo simulation data.

The refolding of projectin domains depends on the degree of relaxation.

If a projectin segment was relaxed to only about one-half its length, the characteristic sawtooth pattern of the force extension curve (on re-stretch) disappeared. Fig. 4.2.5C shows a plot of the fraction of refolded domains, $N_{refolded}/N_{total}$, vs. the degree of relaxation, L_0/Lc. We used a three pulse protocol [187] to first completely unfold and extend the protein and obtain the contour length of the unfolded protein, Lc. Then the protein was rapidly relaxed to a length Lo for a fixed period of time (10 s). A second extension measures the number of domains that refolded during the relaxation period at that particular length, L_0. The open triangles correspond to a Monte-Carlo simulation of a two-state kinetic model with a folding distance, Δx_f = 1.1 nm. From this plot we can estimate how the applied force affects the refolding rate (see *Supplementary text*). The plot shows that ~20% of the domains can refold at L_0/Lc of 0.5 which corresponds to a force of ~20 pN. Hence from this analysis we can conclude that projectin domains can refold under large stretching forces.

Collapse of unfolded projectin domains under force.

In order to examine the effect of a mechanical force on domain refolding in more detail, we used force-clamp AFM techniques [175,185,188]. In these experiments we first unfolded and extended the protein at a high force and then relaxed the protein at lower forces. Fig. 4.2.6 and Fig 4.2.S2 (see *Supporting Information*) show several examples of native projectin molecules held at different force values.

In Fig. 4.2.6A, a projectin molecule was first unfolded and extended at a high force (97 pN). We observed 10 unfolding events. There was an initial large step elongation of ~100 nm upon application of force. (This initial phase most likely

corresponds to the length of the folded polypeptide chain plus a few already unfolded domains.) Then after ~12 sec the protein was relaxed to a force of 15 pN; before the protein reached its fully collapsed state there was a dramatic increase in the noise level with length fluctuations of up to 10 nm peak-to-peak. The source of this noise is not clear, but the phenomenon may reflect the transient formation of secondary structures or intermediate folded conformations, as suggested for ubiquitin domains [188]. Similar to ubiquitin refolding trajectories we also find three main phases: i) a fast phase (<200ms) corresponding to the elastic recoil of the unfolded polypeptide chain and accounting for ~60% of the unfolded length of the protein; ii) a slow phase (~1-8 nm/s) characterized by large fluctuations in end-to-end length (up to 10 nm in this example); and iii) again a fast phase (>100 nm/s) that corresponds to the final collapse of the polypeptide chain to its folded length. In order to test whether the domains are folded, we unfolded the protein by applying a second stretching pulse to 97 pN after 30 s (we count 9 steps). Hence this experiment clearly shows that projectin domains refolded at a force of 15 pN.

Fig. 4.2.6B shows another example. In this experiment the protein was first unfolded and extended at 124 pN. There was an initial large step elongation of ~200 nm upon application of force which was followed by 8 unfolding steps. Then the force was dropped to 30 pN; the polypeptide chain is seen to quickly (~200 nm/s) recoil and then slowly contract. In order to minimize the effect of drift, after ~30 s we dropped the force to ~5 pN (marked by the arrow). We then increased the force to 150 pN and counted a total of 12 steps indicating that domains refolded at the low force of 5 pN.

Fig. 4.2.6C shows that during the initial stages of the slow phase the domains do not fold. We first unfolded and extended the molecule at a high force of 115 pN (many steps are detected), then the protein was relaxed to a force of 48 pN and after 12 s the

polypeptide was extended again at a force of 100 pN. There are no detectable steps during this transition from low to high force indicating the absence of folded domains.

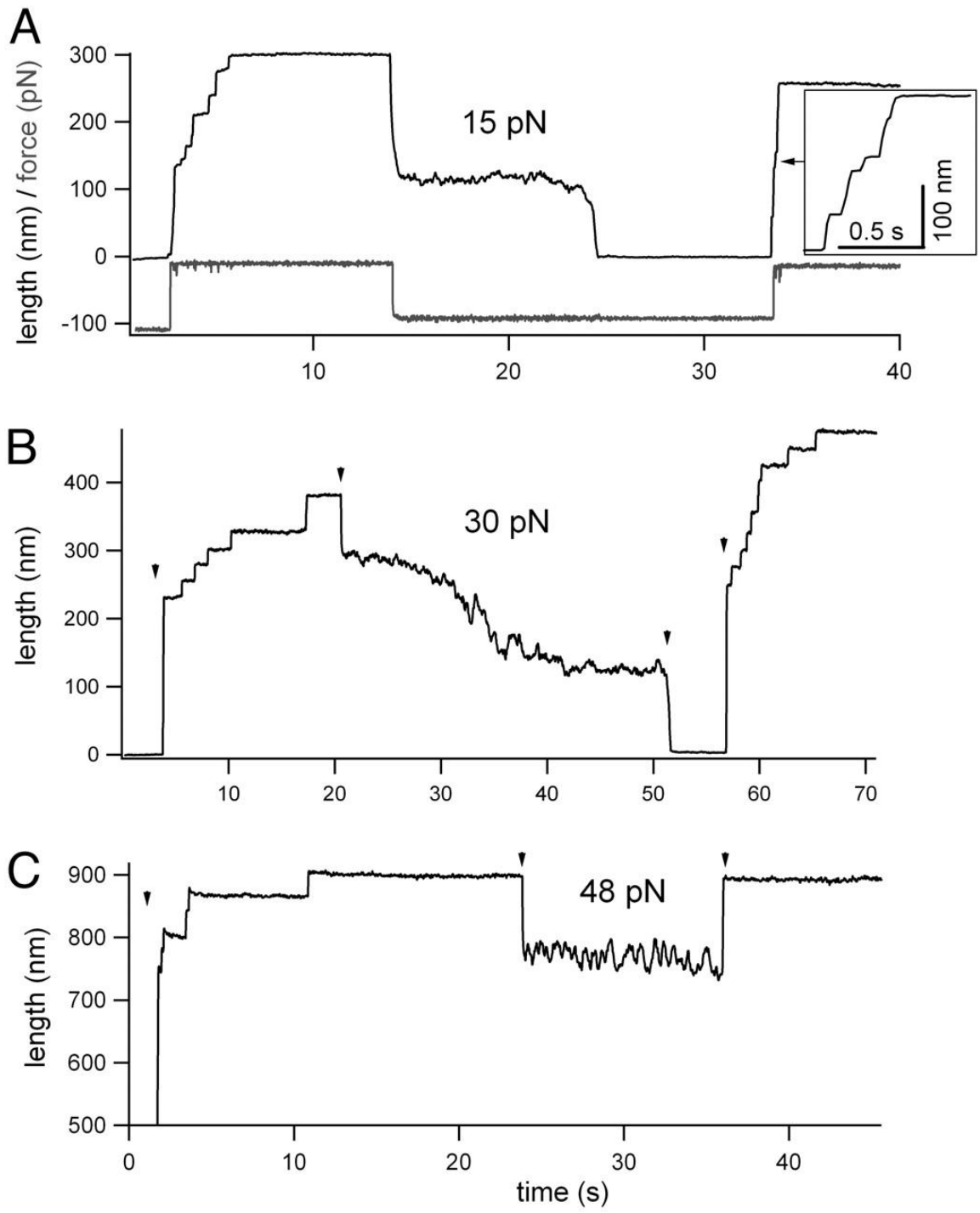

Figure 4.2.6 - Collapse of unfolded projectin domains under force.

A) Example of a projectin molecule that was first unfolded and extended at a high force (97 pN; the applied force trace is shown in the bottom trace), then relaxed to a force of 15 pN and extended again at a force of 97 pN. **B)** Example of a projectin molecule that was first unfolded and extended at 124 pN then relaxed to 30 pN and then to ~ 5 pN (marked by arrows). After ~ 5 s the force was increased to 150 pN. **C)** A projectin molecule was extended at a force of 115 pN then relaxed to 48 pN and after 12s extended again at a force of 100 pN.

Discussion

Here, we examined the mechanical properties of kettin and projectin and their individual domains by using single-molecule force spectroscopy. We found that kettin and projectin have different mechanical architectures, which are likely to be related to their different functions in the sarcomere.

Force-extension and force-clamp curves obtained from single projectin molecules revealed a complex elongation pattern, with some domains unfolding in a low force range (~50-150 pN) and a second population unfolding at higher forces (~150-250 pN). Similar experiments done on recombinant fragments revealed that the FnIII domains are mechanically weaker than the Ig domains. These differences seem to be well conserved, as studies on human titin also showed that FnIII domains unfold at lower forces than Ig domains [189]. These differences in mechanical strengths may result from different mechanical topologies. Both types of domain consist of two antiparallel beta-sheets packed against one another and they have a similar Greek key strand topology [190]. However, the domain types differ in the number of strands and the pattern of hydrogen bonds exposed to a mechanical force [99,191,192].

Two fragments of kettin and a fragment derived from a spliced isoform of Sls upstream of kettin displayed a strong mechanical unfolding hierarchy requiring 100 pN of force to unfold the weakest Ig domain and 250 pN for the most stable Ig domain. Interestingly, this mechanical hierarchy is correlated with the location of the Ig domains along the Sls and kettin sequence: Ig domains closer to the N-terminus of Sls/kettin are mechanically less stable than those nearer the C-terminus. Apparently, different regions of Sls/kettin in the sarcomere require different mechanical stabilities. The low unfolding force for SIg4-SIg6 is consistent with its position bound to myosin in the M-line in the

center of the sarcomere. In contrast, the property of KIg17-KIg21 to bind to actin may require high stability. The KIg 34/35 domains at the end of kettin may experience stretching forces since they are linked to the end of the thick filament [156] and hence may require a very high stability.

Another interesting finding is that projectin domains refold very fast at room temperature: ~85% of the domains refolded at a rate of 15 s^{-1}. This refolding rate is much faster than that of human cardiac titin Ig domains (~1 s^{-1}; refs [132,143,187]) or of native *Lethocerus* kettin Ig domains (~0.1 s^{-1}; ref. [174]), but it is comparable to that recorded from tenascin FnIII domains (~40 s^{-1}; ref. [146]). Also our data shows that the refolding of projectin domains follows a double exponential time course (Fig. 4.2.5B). This may result from different refolding kinetics of FnIII and Ig domains. To test this idea we included in our refolding analysis only the low force peaks (<100 pN) and found that they mainly contribute to the double exponential time course (see also Fig. 4.2.S4). This suggests at least two independent pathways for refolding of FnIII domains with different rate constants.

The speed at which projectin domains recover from the unfolded state was found to depend on the temperature. Fig. 4.2.5 showed that the fast refolding rate is 2.5-fold slower at 13°C than at 25°C. This data, which to our knowledge represents the first measurement of the temperature dependence of the mechanical refolding rate of a single protein, indicates a Q_{10} for the refolding rate of ~2.5. In contrast, domain unfolding is affected much less by temperature.

The mean unfolding forces are 97.4 $\pm$ 36.7 pN at 14°C and 74.7 $\pm$ 38.9 pN at 26°C (see *Supplementary text*). This translates into a Q_{10} for unfolding of 1.3. The low Q_{10} for unfolding in this temperature range is slightly higher than the temperature dependence of the mechanical unfolding of spectrin repeats (38; Q_{10} ~ 1) and bacteriorhodopsin helical

regions (39; Q_{10} ~1.2). These findings show that the temperature dependence of the forces driving mechanical refolding is higher than that for unfolding and probably reflects the cooperative nature of the folding mechanism.

Some insects have relatively high temperatures during flight (35-40^{o}C; refs. [164,193]); e.g. the temperature of the IFM in *Lethocerus* during flight is ~40°C. At this temperature projectin domains would refold at a rate of ~40 s^{-1} (estimated from our data considering a Q_{10} of ~2.5) — a value that is almost 2-fold higher than the natural wingbeat frequency of *Lethocerus indicus* (~25 Hz). In this estimate we assumed that the domains are refolding at zero force. However, we found that the domains still refold at higher forces and that the refolding rate slows down with the force level (Figs. 4.2.5-4.2.6). At high forces the refolding rate would be slower, following an exponential relationship with the force according to:

$$\beta(F) = \beta_o e^{-F\Delta x_F / kT} \qquad (2)$$

where β_o is the rate at zero force, Δx_F, is the folding distance and kT = 4.1 pN/nm. Using Δx_F = 1.1 nm (Fig. 4.2.5C), we calculate that at a physiological force of 5 pN [143] the refolding rate would be ~11 s^{-1} (at 40^{o}C). Although we do not know the exact number of domains in the extensible part of projectin, it has been estimated to be ~ 15 domains [176]. If we assume that only a fraction of these domains unfold during a stretching event (*e.g.* 5), we then calculate that during a single wing-beat about 50% of the domains would refold (probability of folding = number of unfolded domains*β(5pN)*Δt). Hence, at low forces projectin could function as a folding-based spring.

The observation that domains can refold under force is not a unique feature to projectin and kettin molecules, but is a property found in titin [186], kettin (can refold at forces of up to 30 pN; ref. [174]), I27 polyproteins (can refold at forces of ~15 pN; estimated from Fig. 4.2.5 in ref. 32) and ubiquitin (can refold at forces of up to 40 pN;

ref. [188]). In addition, recent molecular dynamics simulations [194,195] demonstrate that refolding can occur at high forces (~40 pN). In conclusion, our data show that projectin domains can refold sufficiently fast under relatively high forces at physiological temperatures, suggesting a robust refolding mechanism that may operate over a large range of sarcomere lengths. Given the structural and functional similarity of projectin and other proteins of the titin family, it is possible that this property may also be found in other titin-like proteins. This mechanism might be subjected to biochemical modulation through signaling pathways or molecular chaperones [196], providing a novel way of modulating muscle passive elasticity. Future experiments may show whether this mechanism operates *in vivo*.

MATERIALS AND METHODS

Proteins

Native projectin (800-1000 kDa) was isolated from *Lethocerus indicus* leg muscle as previously described [172]. Cloning and expression of *Drosophila* projectin and kettin fragments: Projectin sequences coding for PIg24 to PIg26 (accession number AF047475) and the kettin sequences coding for KIg17 to KIg21 (accession number AJ245406) were obtained by PCR using *Drosophila* genomic DNA. Plasmid construction, protein expression, and protein purification are described in detail in *Supporting Text*.

Rotary shadowed EM

Projectin and kettin in 50% glycerol were sprayed onto freshly cleaved mica and rotary shadowed at 7° with platinum and palladium, and at 90° with carbon. Replicas were floated on distilled water and picked up on uncoated copper grids. Micrographs were taken at 100 kV, 40,000x magnification in a Philips EM 400.

AFM

To study the mechanical properties of kettin and projectin we used a home-built single molecule atomic force microscope (AFM) as previously described ([146,175,176]; see *Supporting Text*).

SUPPORTING TEXT

Cloning and expression of projectin and kettin fragments, Projectin sequences coding for PIg24 to PIg26 (accession number AF047475) was obtained by PCR using *Drosophila* genomic DNA. The corresponding protein sequence is from VPVTGEPLPSKD at the N-terminus to TANSVTISWKPP at the C-terminus, with a molecular weight of 80 kDa. The DNA was subcloned into the pETM11 expression vector, which has a 6His tag at the N-terminus, and the vector was transformed into *E. coli* BL21(DE3)pRARE cells (Stratagene). Expressed protein from a 2 l culture was in inclusion bodies. The inclusion body pellet was washed with 50 mM K-phosphate pH 7.5, 0.1% Triton X-100, 2 mM DTT and then taken up in 10 ml 8 M urea, 50 mM Na-phosphate pH 7.5. The soluble fraction was dialyzed against 50% glycerol, 20 mM Tris pH 8.0, 1 mM DTT, 1 mM EDTA, then against the same buffer with 25% glycerol and 50 mM NaCl, and finally against this buffer without glycerol. At each stage, insoluble protein was removed by centrifugation.

Kettin sequence coding for KIg17 to KIg21 (accession number AJ245406) was cloned and expressed by the same method as used for the projectin fragment. The sequence at the N-terminus of the protein is DAPISPPHFTAE and at the C-terminus is TSGTLKCTGGKT; the molecular weight is 71 kDa. The protein was expressed in *E. coli* BL21(DE3)pJY2 cells (Stratagene); soluble protein was purified from the cell lysate on a (Ni-NTA)-agarose column (Qiagen) and then on a Mono-Q column (Pharmacia). A construct coding for three Ig domains in the Sls sequence upstream of kettin (SIg4 to

SIg6) (accession number AJ544075) was obtained by PCR. The corresponding protein sequence is from SDSEMASDIEPI at the N-terminus to FLNIRGSGLPAS at the C-terminus, with a molecular weight of 38 kDa. The construct was subcloned into the pET8c vector and expressed and purified by the same method as used for KIg17-KIg21. The kettin fragment KIg34/35 was cloned and expressed as described by Kulke et al. [156,164]. Proteins were concentrated with a Millipore centrifugal concentrator.

AFM.

To study the mechanical properties of kettin and projectin we used a home-built single molecule atomic force microscope (AFM). The spring constant of each individual cantilever (MSCT-AUHW, sharpened silicon nitride gold-coated cantilevers; Veeco Metrology Group, Santa Barbara, CA) was calibrated using the equipartition theorem [165] and varied between 20-80 pN/nm depending on the type of cantilever. With this system it is possible to measure the force as a function of the extension of the protein (force-extension mode), or measure the elongation of the protein at a constant force (force-feedback mode). The step time response of our force-feedback system was ~20 ms. Unless noted in the text, the pulling speed of all force–extension curves was in the range of 0.4–0.6 nm/ms.

Single protein recordings.

In a typical experiment, a small aliquot of the purified protein (~10-50 µl, 10 µg/ml) was allowed to adsorb to a clean glass coverslip (for ~10min) and then rinsed with PBS pH 7.4. Segments of the proteins were then picked up randomly by adsorption to the cantilever tip, by pressing it down onto the sample for 1-2 seconds at forces of several nanonewtons, and stretched for several hundred nanometers. The probability of picking up a protein was typically kept low (less than one in 50 attempts) by controlling the

amount of protein used to prepare the coverslips. In order to study the effect of low temperature on refolding kinetics we used a simple device where we enclosed the AFM inside a refrigerator, filled with lead bricks (to increase the heat capacity). We then turned on the refrigerator and let the AFM equilibrate at low temperatures (7-10^{o}C). After ~1h the power was turned off because the mechanical noise introduced by the cooling system interferes with the AFM measurements. The temperature increase was typically very slow (~8^{o}C/hr) during the duration of the experiment (20-30 min).

Analysis of force extension curves.

The elasticity of the stretched proteins was analyzed using the worm-like chain (WLC) model of polymer elasticity [156,196],

$$\mathrm{F}(x) = \frac{kT}{p}\left[\frac{1}{4}\left(1-\frac{x}{Lc}\right)^{-2} - \frac{1}{4} + \frac{x}{Lc}\right] \qquad \text{(eq. 4.2.S1)}$$

where F is force, p is the persistence length, x is end-to-end length, Lc is contour length of the stretched protein, k is Boltzmann's constant, and T is absolute temperature. The adjustable parameters are the persistence length, (which defines the flexibility), and the contour length. The change in contour length was used to calculate the size of the folded domain.

Estimation of the unfolding rate from force-ramp experiments.

In order to analyze the data of Fig. 4.2.4B and 4.2.4D quantitatively, we used a simple two-state kinetic model for mechanical unfolding [166,167]. In this model, a protein is exposed to a force that increases linearly with time, simulating the conditions of our

force-ramp experiment. According to this model the cumulative probability, Pu(F), that an unfolding event has occurred at a force lower than or equal to F, is given by,

$$P_u(F) = 1 - e^{-\frac{\alpha_o}{a}\int_0^F e^{\frac{f \cdot \Delta x_u}{kT}} df} \qquad \text{(eq. 4.2.S2)}$$

where *a* is the rate of change of the applied force (*a* = 200 pN/s in our experiments), α_0 is the rate of unfolding at zero force, Δx_u is the distance to the transition state and the other symbols have their usual meaning. For the kettin fragment, KIg17-KIg21, values of α_0 = 8 x 10^{-3} s^{-1} and Δx_u = 0.17 nm readily describe the data (continuous line, Fig. 4.2.4D). In the case of native projectin we used an equation that describes the cumulative probability of the unfolding of two independent populations of domains with different unfolding rates, α_{o1} and α_{o2} [154],

$$P_u(F) = 2 - e^{-\frac{\alpha_{o1}}{a}\int_0^F e^{\frac{f \cdot \Delta x_{u1}}{kT}} df} - e^{-\frac{\alpha_{o2}}{a}\int_0^F e^{\frac{f \cdot \Delta x_{u2}}{kT}} df} \qquad \text{(eq. 4.2.S3)}$$

Using this equation we estimate, α_{o1} = 0.3x10^{-3} s^{-1} and α_{o2} = 7 x 10^{-2} s^{-1} using a Δx_{u1} = 0.2 nm and Δx_{u2} = 0.1 nm (solid line in Fig. 4.2.4B).

Estimation of the refolding distance using a three pulse protocol.

Determination of the folding distance, Δx_f, involves measuring how much the folding rate constant depends on the applied force. To measure the folding distance, we used a three pulse protocol [168] to first completely unfold and extend the protein and obtain the contour length of the unfolded protein, Lc. Then the protein was rapidly relaxed to a length Lo for a fixed period of time (10 s). A second extension then allowed us to count the number of domains that refolded during the relaxation period at that

particular length, L_o. From this plot we can estimate how the applied force affects the refolding rate [168]. In the experiment shown in Fig. 4.2.5C the protein is allowed to fold under an applied force that depends on the ratio L_0/Lc. Hence we can write,

$$\frac{N_{refolded}}{N_{total}} = 1 - e^{-tk_f^o (Lo/Lc)} \qquad \text{(eq. 4.2.S4)}$$

where $k_f(L_0/Lc)=k_f^0\exp(-F(L_o/Lc)\Delta x_f/kT)$. Since during a refolding experiment the contour length, Lc, is known, we can calculate the force, $F(L_0/Lc)$, that strains the protein at length L_o using the WLC equation.

Monte Carlo simulations.

The folding/unfolding of a domain was modeled as a two state Markovian process where the probability of unfolding was $P_u=N_f*\alpha*\Delta t$ where N_f is the number of folded domains and Δt is the polling interval [168-171]. The folding probability was $P_f=N_u*\beta*\Delta t$ where N_u is the number of unfolded domains. The rate constants for unfolding, α, and refolding, β, are given by $\alpha=\alpha_o\exp(F\Delta x_u /kT)$ and $\beta=\beta_o\exp(-F\Delta x_f /kT)$ where F is the applied force and Δx_u and Δx_f are the unfolding and folding distances.

ADDITIONAL DATA

a) Cantilever Drift

An important problem in force-clamp experiments like those shown in Fig 4.2.6 and Fig 4.2.S2, is cantilever drift and this can lead to significant errors in the force measurements. We found a small fraction (about 1 in 20) of cantilevers had exceptionally low drift and we used these for force-clamp experiments. Figure 4.2.S1 shows the measurement of the drift from 4 different cantilevers (n=21 measurements; Kc, 30-35 pN/nm). As shown by Figure 4.2.S1A the drift over ~ 1 min is in general quite random;

some cantilevers tend to have negative or positive slope with random fluctuations. However, on average the drift during the first 60 s tends to have a positive slope of 0.17 pN/s (Fig 4.2.S1B); hence after 60 sec the cantilever would have drifted on average by ~ 10 pN. This effect would certainly affect the force the protein is subjected to at very long time scales (> 1 min).

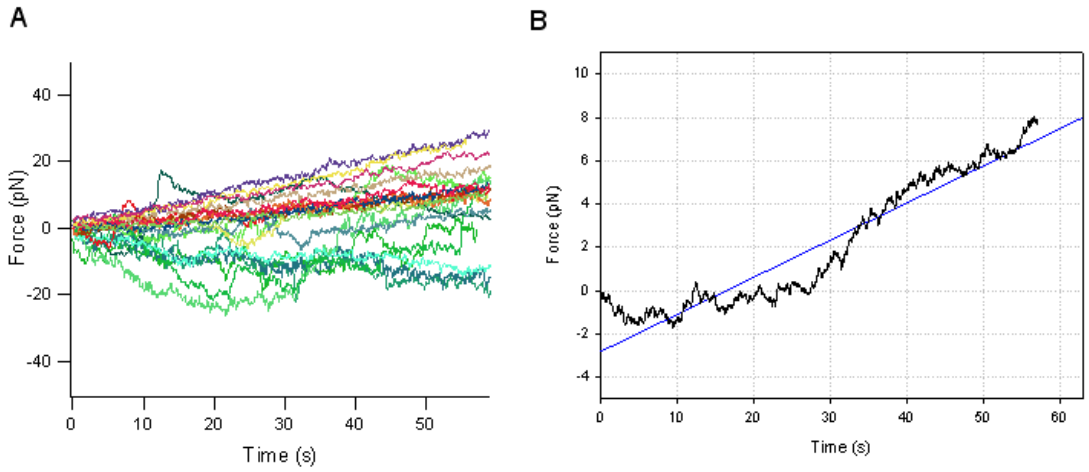

Figure 4.2.S1 - Measurement of cantilever drift.

A) Force as a function of time for 4 different cantilevers (n=21 measurements; Kc, 30-35 pN/nm). B) Average cantilever drift; the slope of the force vs. time plot is 0.17 pN/nm.

As discussed by Fernandez and Li [172], one way to independently measure the actual value of the quenched force is to measure the magnitude of the elastic recoil observed immediately after relaxing to unfolded polypeptide chain to the lower force. Since we know the length of the unfolded chain, we used the worm-like-chain equation to estimate the actual force after initial relaxation. We have used this method to estimate the force after relaxation in Fig 4.2.6 and in Fig 4.2.S2.

b) Additional Force-clamp data

Figure 4.2.S2 shows additional examples of force clamp data obtained on a longer time scale (> 1 min) than those shown in Figure 4.2.6. In Fig. 4.2.S2A, a projectin molecule was first unfolded and extended at a high force (95 pN). We observed 1 step and then 6 steps (because of the compressed time scale these are seen as one large step) corresponding to the unfolding of 7 domains. There was an initial large step elongation of ~200 nm upon application of force. (This initial phase most likely corresponds to the length of the folded polypeptide chain plus a few already unfolded domains.) Then the protein was relaxed to a force of 35 pN; before the protein reached its fully collapsed state there was a dramatic increase in the noise level with length fluctuations of up to 50 nm peak-to-peak. Three phases are distinguishable, i) a fast phase (<100ms) corresponding to the elastic recoil of the unfolded polypeptide chain and accounting for ~20% of the unfolded length of the protein; ii) a slow phase (~5 nm/s) characterized by large fluctuations in end-to-end length (up to 50nm); and iii) again a fast phase (350nm/s) that corresponds to the final collapse of the polypeptide chain to its folded length.

In the experiment shown in Fig. 4.2.S2B the protein was first unfolded and extended at 68 pN (8 unfolding steps preceded by a large elongation of ~120 nm; see inset on the left) and then the force was dropped to ~5 pN; the polypeptide chain is seen to quickly (270nm/s) contract to its original end-to-end length. After 6 s a force of 68 pN was applied again (marked by arrow) and we observed 4 steps indicating that during this time 4 out of 8 domains were able to refold under force. Then the force was lowered to 38 pN and we observed the polypeptide chain collapsing in fast and very slow (here, 4.5 nm/s) phases.

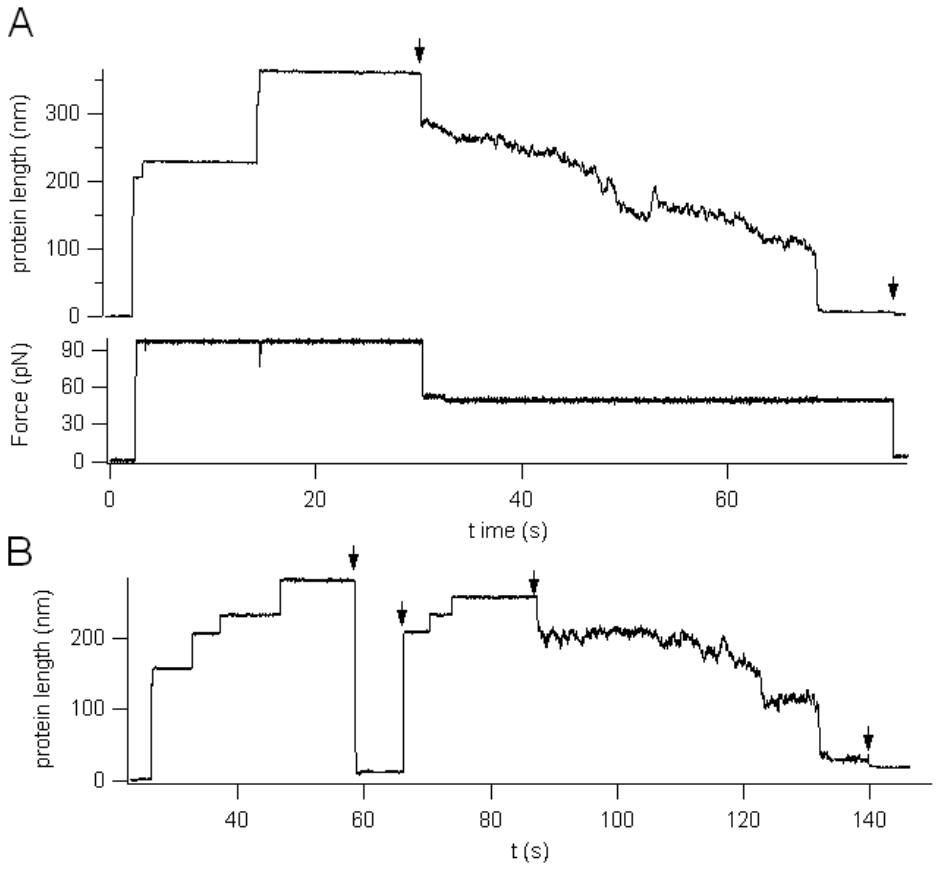

Figure 4.2.S2 - Additional collapse trajectories of unfolded projectin domains under force.

We used force-clamp AFM to examine the effect of a mechanical force on the folding of projectin domains. A) A projectin molecule was first unfolded and extended at a high force (95 pN). We observe several steps corresponding to the unfolding of 7 domains. Then the protein was relaxed to a force of 35 pN (calculated from the WLC equation) and then to 15pN. B) A projectin molecule was extended at 68 pN (8 unfolding steps) and then the force was dropped to ~5 pN (marked by arrow); after 6 s the force was stepped to 73 pN (4 steps are detected) and after ~15 s then the force was lowered to 38 pN and

finally to 0 pN. The inset shows the initial response to a stretching force in a log time scale.

c) Effect of temperature on unfolding forces

Figure 4.2.S3 shows unfolding force histograms obtained at 26°C (red trace) and 14°C (blue trace). The mean unfolding forces are 97.4 ± 36.7 pN at 14°C and 74.7 ± 38.9 pN at 26°C. Hence lowering the temperature by ~10°C increases the unfolding forces by ~23 pN. This translates into a Q_{10} for unfolding of 1.3. In contrast the Q_{10} for refolding is almost twice as large (Q_{10} = 2.5).

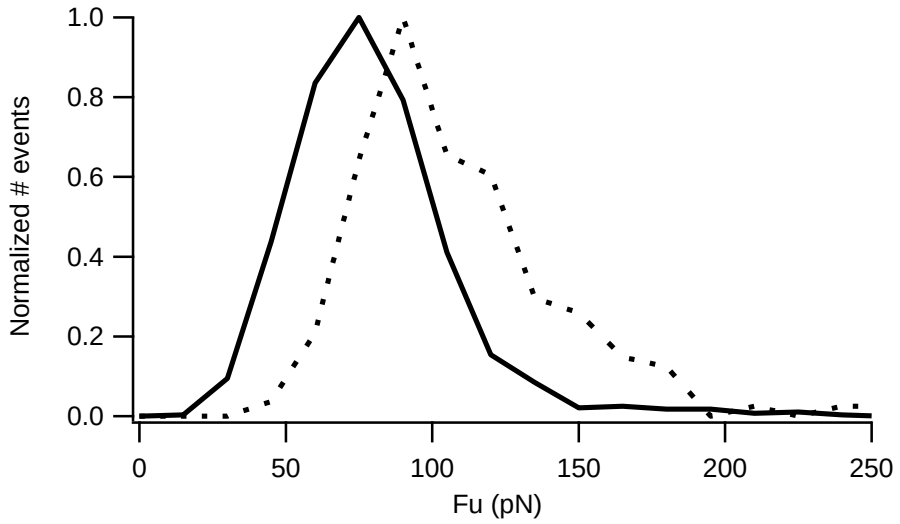

Figure 4.2.S3 - Effect of temperature on projectin domain unfolding forces.

Normalized unfolding force histograms obtained at 26°C (continuous trace) and 14°C (dashed trace). The mean unfolding forces are 97.4 ± 36.7 pN (n=328) at 14°C and 74.7 ± 38.9 pN at 26°C (n = 846). These data where obtained using the same cantilever.

d) Refolding of projectin domains.

Figure 4.2.S4 shows a typical experiment in which a single projectin molecule remained attached to an AFM tip allowing for repeated extension and relaxation cycles (up to 28 cycles in this experiment over a period of ~9 min). After each extension, the molecule was allowed to relax completely (the relaxation traces are not shown). In this experiment we waited 15 s between each stretching pulse. Consecutive force-extension curves display similar patterns, demonstrating that domain unfolding is fully reversible and that projectin domains can undergo multiple cycles of extension/relaxation with no signs of molecular fatigue or rundown.

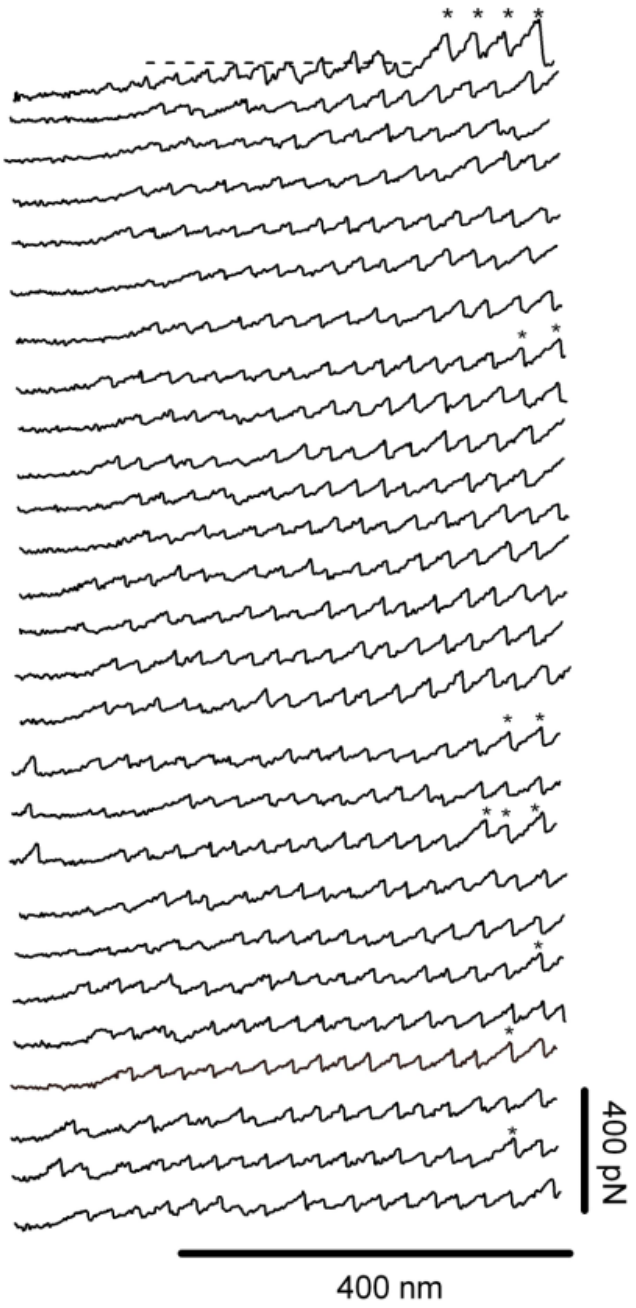

Figure 4.2.S4 - The refolding of projectin domains is very robust.

A series of force curves collected from a single molecule over approximately 9 minutes. The "high-force" peaks are marked by an asterisk.

Section 3: Single-molecule force spectroscopy reveals a stepwise unfolding of C. elegans giant protein kinase domains

INTRODUCTION

All animals have complex mechanisms to regulate their overall muscle mass. The functional unit of a muscle cell is the sarcomere, a miniature machine comprised of overlapping, interacting filaments. During muscle activity, the sarcomeres act in unison, undergoing rapid contraction/relaxation, transducing work throughout the muscle. The three filaments that give the sarcomere its integrity are thin filaments, thick filaments, and titin filaments. These filaments, together with their multi-protein attachment structures, the M-lines and Z-disks, must undergo assembly and disassembly during myofibril growth and maintenance. Although progress has been made in understanding this intricate system, the precise mechanisms that control these assembly/disassembly processes remain central questions in muscle biology. Previous research has implicated signaling from human titin kinase to the building of myofibrils [197]. When in its active conformation, titin kinase transmits a phosphorylation cascade that ultimately leads to the expression of genes involved in myofibril assembly. Further, mutations that inhibit the kinase from initiating this signaling pathway result in one type of a human myopathy [197].

In *C. elegans* there are two titin-like proteins, twitchin and TTN-1. These giant muscle proteins, like those found in other species, are comprised primarily of multiple copies of immunoglobulin (Ig) and fibronectin type-III (Fn) domains [198]. Twitchin (~800kDa), named for its characteristic twitching mutant phenotype, has 30 Ig domains, 31 Fn domains, and a single kinase domain [199,200]. Upon its discovery, twitchin was hypothesized to be an important regulator of muscle contraction, mainly due to its motility phenotype and MLCK-like kinase domain. Later, its role in inhibiting the rate of

relaxation was identified in studies of twitchin from *Aplysia* and *Mytilus* [201,202]. TTN-1 is significantly larger than twitchin, with a molecular weight of 2.2 MDa. In addition to 56 Ig domains and 11 Fn domains, TTN-1 has several regions predicted to be coiled-coil, and two regions consisting of tandem repeats [203]. The largest of these tandem repeat regions, called PEVT, is similar in amino acid composition and tandem repeat structure to the main elastic element of vertebrate titin called the PEVK region. Thus, this region of nematode TTN-1 is hypothesized to be elastic. Similar to mammalian titin kinase, TTN-1 and twitchin both have a single kinase and regulatory domain located near the C-terminal end of the giant molecule [203,204]. In the primary sequence, the inhibitory/regulatory region is located just downstream of the catalytic core (Figure 4.3.1A & B) and in 3-D space (Figure 4.3.2A & B) the domain is wedged in between the two subdomains of the catalytic core, making extensive contact with the active site, blocking substrate entry [205,206]. The kinases with the highest homology, the vertebrate smooth muscle and non-muscle myosin light chain kinases (MLCKs), are also autoinhibited by a sequence just C-terminal of the catalytic core. In the case of the MLCKs, binding of the autoinhibitory sequence to Ca^{+2}/calmodulin causes a conformational change sufficient to allow access to its substrate [207]. Although the giant kinases can also bind to Ca^{+2}/calmodulin, this alone is not sufficient to activate the enzymes [208]. In developing muscle, the combination of Ca^{+2}/calmodulin and phosphorylation of a key tyrosine residue can activate vertebrate titin kinase [209]. However, the required tyrosine phosphorylating activity is not found in mature muscle. Thus, to date, the precise mechanism(s) resulting in the conformational changes that relieve the kinase of this autoinhibition remain a mystery.

One hypothesis is that the giant kinases may act as force sensors; that the forces generated from the contraction/relaxation cycles of muscle activity are sufficient to

unleash the regulatory domain from the catalytic core [210]. Thus, after experiencing a certain threshold of force, the kinase would become active. Grater et al. [211] used molecular dynamics simulations to pull the human titin kinase from its amino and carboxy termini to simulate the strain the molecule would undergo during muscle activity. The authors found that not only can the kinase withstand expected forces, but that the kinase domain is also positioned in perfect orientation within the "molecular spring" of the titin molecule, making the kinase an ideal force sensor. They conclude that the strain on the autoinhibitory domain leads to an ordered sequence of conformational changes that open the catalytic cleft, while maintaining the structural integrity of the enzyme. This remarkable stability is a function of the orientation of the molecule with respect to the pulling force; the pulling geometry of activation is such that the beta sheets most responsible for the force resistance are located parallel to the pulling force, but the beta-sheets responsible for exposing the active site are oriented perpendicular to the force. This simple, yet elegant, difference in the orientation of beta-sheets sets the stage for the ability of the kinase to simultaneously remove the autoinhibitory region from the catalytic core, while maintaining the structural integrity of the active site. The molecular dynamics simulations support a model in which human titin kinase will sense force and pass along the message by phosphorylation of its substrate.

In order to begin to test this hypothesis experimentally, we recombinantly expressed twitchin and TTN-1 kinases, the human titin kinase homologs, from *C. elegans*. Using single-molecule atomic force microscopy (AFM) we analyzed the mechanical strength of the kinase and their flanking Ig/Fn domains, along with a tandem repeat of five Ig domains that immediately follow the kinase domains in the primary structure. Our results show that these kinase domains have remarkably high mechanical stability. The kinase domain unfolds at a force range of ~30-150pN, a value that is

similar to the unfolding range of nematode Ig and Fn domains (40-180pN). Further, in contrast to the Ig/Fn domains, which unfold in an all-or-none highly cooperative fashion, the unfolding of the kinase is sequential, first an unwinding of the autoinhibitory region, followed by the biphasic rupture of the catalytic core. These data provide support for the hypothesis that the kinase domains of the giant muscle proteins function as effective force sensors.

MATERIALS AND METHODS

AFM

The mechanical properties of single proteins were studied using a home-built single molecule AFM as previously described [101,147,176,187,211,212]. The spring constant of each individual cantilever (MLCT-AUHW: silicon nitride gold-coated cantilevers; Veeco Metrology Group, Santa Barbara, CA) was calculated using the equipartition theorem [175]. The cantilever spring constant varied between 10-50 pN/nm and rms force noise (1-kHz bandwidth) was ~10 pN. Unless noted, the pulling speed of the different force–extension curves was in the range of 0.4–0.6 nm/ms.

Single Protein Mechanics

In a typical experiment, a small aliquot of the purified proteins (~1-50 µl, 10-100 µg/ml) was allowed to adsorb to a clean glass coverslip (for ~10 min) and then rinsed with PBS pH 7.4. Proteins were picked up randomly by adsorption to the cantilever tip, which was pressed down onto the sample for 1-2 seconds at forces of several nanonewtons and then stretched for several hundred nm.

Analysis of Force extension curves

The elasticity of the stretched proteins were analyzed using the worm-like chain (WLC) model of polymer elasticity (22, 23):

$$F(x) = \frac{kT}{p}\left[\frac{1}{4}\left(1-\frac{x}{L_c}\right)^{-2} - \frac{1}{4} + \frac{x}{L_c}\right]$$

where F is force, p is the persistence length, x is end-to-end length, and L_c is contour length of the stretched protein. The adjustable parameters are the persistence length, p, and the contour length, Lc.

Monte-Carlo simulations

The folding and unfolding of a domain was modeled as a two state Markovian process where the probability of unfolding was $P_u=N_f*\alpha*\Delta t$ where N_f is the number of folded domains and Δt is the polling interval [132,146,175]. The folding probability was $P_f=N_u*\beta*\Delta t$ where N_u is the number of unfolded domains. The rate constants for unfolding, α, and refolding, β, are given by $\alpha=\alpha_o\exp(F\Delta x_u/kT)$ and $\beta=\beta_o\exp(-F\Delta x_f/kT)$ where F is the applied force and Δx_u and Δx_f are the unfolding and folding distances.

Homology Modeling of TTN1 Kinase and Regulatory Domains

Molecular modeling of *C. elegans* TTN-1 was completed using Modeler (version 7v7; University of California San Francisco; http://salilab.org/modeler/) for model construction and SYBYL (version 7.0; Tripos, Inc., http://www.tripos.com) for analysis and refinement. The TTN-1 model was built and minimized based on the existing structural information available from *C. elegans* twitchin (PDB entry 1KOA) (10). Initial sequence alignments were obtained from ClustalW (using TTN-1, human titin, *C. elegans*, *Aplysia*, and *Mytilus* twitchin, chicken smooth muscle MLCK, and *Drosophila* myosin MLCK, http://searchlauncher.bcm.tmc.edu/multi-align/multi-align.html), NCBI

BlastP (http://www.ncbi.nlm.nih.gov/), CPHmodel (using TTN-1 and *C. elegans* twitchin, www.cbs.dtu.dk), SwissModel (using TTN-1 and *C. elegans* twitchin, http://swissmodel.expasy.org//SWISS-MODEL.html), CDDomain (http://www.ncbi.nlm.nih.gov/Structure/ cdd/wrpsb.cgi, hits = KOA, KOB, 1TKI), and 3Djigsaw (using TTN-1 and *C. elegans* twitchin, http://www.bmm.icnet.uk/~3djigsaw/). The alignment was optimized manually after each round of modeling using visual structural comparison of the previously published structural data of *C. elegans* twitchin (PDB entry 1KOA (10)), *Aplysia* twitchin (PDB entry 1KOB)(10), and human titin (PDB entry 1TKI) (13). The final alignment is shown as Supplementary Figure 4.3.S1). The proteins showed ~50% (for *C. elegans* and *Aplysia* twitchin) and ~40% (for titin) similarity against TTN-1 residues 15907–16373 (Figure 4.3.1B), the approximate boundaries of TTN-1 kinase and the flanking Ig in the full-length giant polypeptide.

We specifically studied the sequence alignment at the terminal beta sheets and the regulatory alpha helices, as these are the regions predicted to be most important in the force activation mechanism (14), and are also the regions with some of the highest variability (7, 8). The N-terminal beta sheets, beta C1-C3, are located at residues: 24-33 (betaC1), 46-53 (beta C2), and 48-56 (beta C3) in TKI; 53-62 (betaC1), 65-72 (beta C2), and 77-85 (beta C3) in KOB; 5942-5951 (betaC1), 5954-5961 (beta C2), and 5966-5974 (beta C3) in KOA; 28-37 (betaC1), 40-47 (beta C2), and 52-60 (beta C3) in TTN-1. The C-terminal beta sheets, beta C10-11 and beta R1 are located at residues: 174-182 (betaC10), 193-197 (betaC11), and 328-336 (betaR1) in TKI; 204-212 (betaC10), 224-228 (betaC11), and 362-370 (betaR1) in KOB; 6093-6100 (betaC10), 6113-6117 (betaC11), and 6253-6260 (betaR1) in KOA; 179-186 (betaC10), 199-203 (betaC11), and 334-341 (betaR1) in TTN-1. Alpha R1 and alpha R2 are located at residues: 292-306 (alphaR1) and312-318 (alphaR2) in TKI, 321-335 (alphaR1) and 344- 351(alphaR2) in

KOB, 6211-6225 (alphaR1) and 6234-6241 (alphaR2) in KOA, and 292-307 (alphaR1) and 315-323 (alphaR2) in TTN-1. All protein structure images were generated by PyMOL version 1.1beta1 (http://www.pymol.org).

Steered Molecular Dynamics Simulations

The giant kinase structures were analyzed by SMD (steered molecular dynamics) as implemented in NAMD (25, 26). The CHARM22 force field was employed throughout. The structural coordinates for each kinase structure (1KOA and the homology model for TTN-1) were solvated in a 65Å x 65Å x 65Å box. Eighteen Na^+ ions were added, corresponding to a concentration of 0.1M. The system was then minimized with 1000 steps of conjugate gradient minimization from an initial temperature of 310K. This was followed by a 400ps MD simulation to equilibrate the entire system (protein, water, and ions). The backbone RMSD was evaluated at the completion of the equilibration step. The SMD protein-ion-water system contained ~30,000 atoms. Forces were applied by restraining a fixed termini point harmonically and moving the SMD atom with constant velocity (0.5A/ps) along a predetermined vector. Kinase domains were stretched at a constant speed of 0.5Å ps^{-1} until their extension exceeded 99% of the contour length. The trajectories were recorded every 2 fs and analyzed with VMD. Coulombic forces were restricted using the switching function from 10 Å to a cutoff at 12 Å. A spring constant (κ) of $10k_BT/ Å^2$ was used during each simulation. We ran three simulations of the extension of twitchin kinase and the homology model for TTN-1 kinase domains with similar results.

Cloning and expression of TTN-1 and Twitchin constructs

The TTN-1 and twitchin constructs were amplified from the *C. elegans* random primed cDNA library RB2 (kindly provided by Robert Barstead, Oklahoma Medical

Research Foundation, Oklahoma City, OK) using the primer pairs listed below. Each construct was first subcloned into the cloning vector bluescript pKS (+), sequenced and subcloned into the expression vector pET 28, fusing an in frame 6-his tag to the N-terminus.

TTN-1 Ig (38-42):	5' GGTACGGATCCAGACTCACTATGGACGGAG,
	3' GGACTGAATTCTTAGCAACAAGTCTTAGACAATCCCATATC;
TTN-1 FnKinIg:	5' GGTACGAATTCGAGGACAAATATGCAATTGGTATTC,
	3' GGATCAAGCTTTTAGCAACACTTCTCGATGACAGCTGGAG;
twc Ig (26-30):	5' GGTAC GAGCTC GCCTTCTGGGATCGATCTGAAGC,
	3' GGACTTCTAGATTAGCAACAGACAAGGAGAAGAGC;
twc FnKinIg:	5' GGTACGGATCCGACTCTGGAACTGTTAATGTC,
	3' GGACTCTCGAGTTAGCAACATGGCTCGAATTTGAGTGGTTC

The following restriction sites were inserted and used for the subcloning procedure: TTN-1 Ig (38-42) 5' BamHI, 3' EcoRI; TTN-1 FnKinIg 5' EcoRI, 3'HindIII; twc Ig (26-30) 5'SacI, 3'XbaI; twc FnKinIg 5'BamHI, 3'XhoI.

The recombinant plasmids were transformed into E. coli BL21 (DE3) RIL (Stratagene). The cells were grown to mid-log phase in the presence of 25microg/ml kanamycin and 34 microg/ml chloramphenicol at 37°C. Protein expression was induced with 0.5 mM IPTG and continued overnight at 23°C. The cells were harvested and resuspended in 20mM TRIS, 500mM NaCl, and 5mM imidazole (pH7.9) with the addition of Roche complete EDTA-free protease inhibitor pellets. Lysis was completed by passage through a French press at 1000 psi. Detergents PEI (to 0.1%) and NP40 (to 0.01%) were added to the cell free extract. The soluble fraction was collected and loaded onto a pre-charged Novagen nickel column. The column was first washed with 25 column volumes of 20mM TRIS, 750mM NaCl, 5mM imidazole, 0.01% NP40 (pH7.9) followed by a second wash with 5 column volumes of 20mM TRIS, 750mM NaCl, 20mM imidazole, 0.01% NP40 (pH7.9). The proteins were eluted with 20mM TRIS,

500mM NaCl, and 1M imidazole. Protein purity of ~95% was confirmed using coomasie brilliant blue staining of 12% SDS PAGE.

Kinase Assays

The purified kinases were dialyzed against 20mM TRIS (pH8.0), 20mM NaCl, and 5mM betaME. The enzymes were added to reaction buffer (20mM TRIS pH7.4, 10mM magnesium acetate, 0.05% triton, 0.2mg/ml BSA) to a final concentration of 30 pg/microliter. The model substrate was a derivative of chicken smooth muscle regulatory myosin light chain (kMLC 11-23) with the sequence KKRARAATSNVFS [208] (synthesized by the Microchemical Facility, Emory University). The peptide substrate was added to the reaction mixture in excess (0.2mg/ml). The reactions were initiated by the addition of 400microM gamma^{32}P-ATP (0.25 microCi/microliter). Catalysis occurred for 10 minutes at 30°C. A portion of the reaction (25 microliters/40 microliters total volume) was removed and spotted onto Whatman P81 filters. The filters were washed in 75mM phosphoric acid. Once dry, the washed filters were placed into scintillation vials, the counts were measured, and the specific activity was calculated by the following equation:

$$\frac{(CPM_{sample} - CPM_{blank})}{((\text{Specific Activity } ^{32}\text{P-ATP})(\text{reaction time})([\text{enzyme}])(\text{reaction volume/spot volume}))}$$

Each reaction was done in triplicate and performed on freshly purified protein.

As expected from earlier results, specific activities were the highest for the reactions with the Fn-TTN-1Kin-Ig and significantly lower for the Fn-Twc kinase-Ig reactions [203,204]. The specific activity of the enzymatic labeling was normalized to reactions without the addition of kinase.

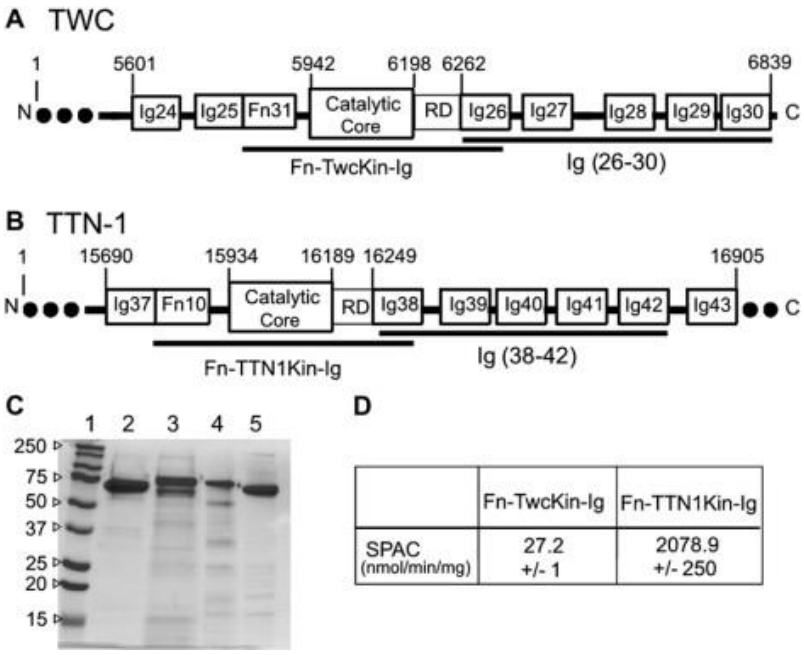

Figure 4.3.1 - Expression, purification, and enzyme activity of recombinant kinase domains and tandem Ig domain segments from *C. elegans* giant proteins.

A, B) Twitchin and TTN-1 recombinant protein constructs for AFM experiments. The kinase constructs have an autoregulated kinase domain flanked by Fn and Ig domains. The Ig constructs consist of 5 tandem Ig domains that immediately follow the kinase domain. The numbers denote the position of the amino acid in the full-length polypeptide. C) 12% SDS PAGE stained with coomasie brilliant blue shows that the proteins were isolated to > 95% purity. In each lane, 2 micrograms of protein was loaded. The lanes are as follows: 1) molecular weight standard 2) Fn-TwcKin-Ig, 3) Fn-TTN1Kin-Ig, 4) Twc Ig 26-30, 5) TTN-1 Ig 38-42. D) The purified TTN-1 and twitchin kinases retain phosphotransferase activity in vitro. Using ^{32}P-ATP and a model peptide as substrates the specific activity values (SPAC) obtained are similar to those previously published [203,204]. Enzyme activities were determined in triplicate on freshly purified protein.

RESULTS

3D structures of C. elegans twitchin and TTN-1 kinase domains

Crystal structures for both *C. elegans* and *Aplysia* twitchin kinases [205,206] and human titin kinase [209] have been previously solved. However, the structure of TTN-1 kinase is unknown. In the interest of developing a better understanding of how the structural similarities among the enzymes might relate to their activation mechanism(s), we developed a homology model of TTN-1 kinase catalytic core and autoinhibitory region (Figure 4.3.2B). Overall, the predicted structure is very similar to the *C. elegans* twitchin kinase. The catalytic region has 2 lobes, a smaller, beta rich lobe, and a larger alpha helical lobe (Figure 4.3.2B, grey and light green, respectively). The regulatory tail (in red) wraps between the 2 lobes, nesting itself into the active site. This inhibited conformation would be maintained as the native structure during catalytic arrest. When the muscle cells require the catalytic activity of the giant protein kinases it is necessary that the regulatory domain be removed for optimal catalysis to be achieved. Closer examination of how the regulatory domain interacts with the catalytic core reveals secondary structure elements that are believed to be the fundamental basis for the force activation hypothesis [210].

The molecular dynamic simulations by Grater et al. (2005) predicted the N and C terminal beta sheets (Figure 4.3.2B, dark blue and dark green plus red beta strand, respectively) to be the primary mechanical elements responsible for the force resistance [210]. Similar to the known crystal structures, the homology model of TTN-1 kinase maintains the alignment of these important substructures. The C-terminal beta sheet consists of 3 beta strands: betaC10, betaC11, and betaR1 ("C" for catalytic; "R" for regulatory) (Figure 4.3.2B; dark green plus red beta strand). Grater et al. (2005) showed

that during activation, the C-terminal beta sheets are positioned perpendicular to the pulling force and are expected to undergo the initial rupture that leads to the activation of the kinase. When comparing the crystal structures to the model, there is a general conservation along betaC11 visualized by comparing the hydrogen-bonding pattern and the side chain interactions. betaR1 is sandwiched between betaC10 and betaC11 and continues to reveal the consistent bonding pattern between the enzymes.

According to the force activation model, the N-terminal beta sheet (betaC1-C3, Figure 4.3.2B, dark blue) is the region that lies parallel to the pulling force to maintain the active site integrity during force activation [210]. This sheet is strikingly similar in all four enzymes. Almost all of the backbone interactions are maintained. The side chain interactions, which further stabilize the beta sheets, are also consistent between the kinases. A conserved lysine residue in the N-terminal betaC3 sheet (K82 KOB, K5971 KOA, K57 TTN1, K53 TKI) interacts with the regulatory tail alphaR2 via van der Waals interactions between the side chain and isoleucine/leucine and valine. This lysine also interacts directly with the active site through electrostatic interactions with an aspartic acid and glutamic acid [205,206]. Linked intimately to the N-terminal beta strands, the alphaR2 helix is a tightly packed 3_{10} helix buried beneath the betaC1-C3 sheets, the substrate binding site, and the portion of the protein that lies directly upstream of the helix (the alphaR1 helix and the linker between alphaR1 and alphaR2 helices). The face of the helix that is in contact with the beta strands is notably similar in each of the enzymes; a string of hydrophobic interactions from the non-polar side chains of the beta strands contacts the non-polar face of the helix (with the exception of one serine present only in TTN-1 and *C. elegans* twitchin). On the alternate side of the helix, the residues interact with both the substrate-binding site and the longer of the regulatory helices,

alphaR1. Many of the interactions are conserved between all four kinases, allowing the alphaR2 helix to retain the overall stability and topology of the enzymes.

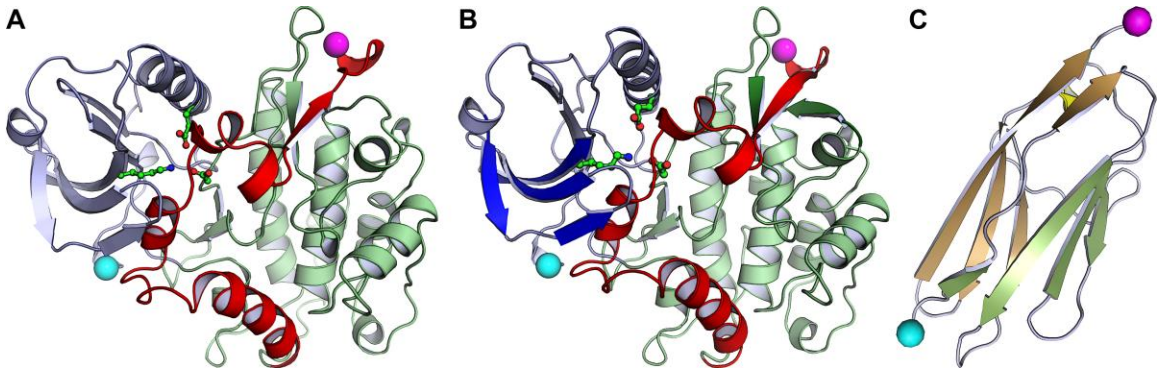

Figure 4.3.2 - 3D structures of *C. elegans* twitchin Ig and kinase domains and homology model for TTN-1 kinase.

A) Twitchin kinase is composed of 3 subdomains: an alpha-helical rich large lobe (green), a small lobe of mainly beta sheets (grey), and the autoregulatory tail (red) (structure taken from [206]). The autoregulatory tail is situated between the 2 lobes making extensive contact with the active site. To orient the active site a conserved lysine residue helping to neutralize the ATP binding pocket is shown in a ball and stick interpretation. B) 3D homology model of TTN-1 kinase. The N-terminal beta sheet (strands betaC1-C3) predicted to be parallel to the pulling force is colored dark blue. This sheet is thought to stabilize the kinase upon activation by force. The C-terminal beta sheet is predicted to be perpendicular to the pulling force and is composed of betaC10 (colored dark green), betaR1 (colored red), and betaC11 (colored dark green) strands. The autoinhibitory domain is colored red and includes the betaR1 strand and alphaR1-R2 helices. Shown in ball and stick representation is a key lysine residue from the N-terminal beta strands interacting directly with the ATP binding pocket of the active site. C) Twitchin Ig domain 26, which immediately follows the twitchin kinase domain (structure taken from [206]). The two beta sheets, characteristic of an Ig fold are depicted in green and brown. All three structures have their N terminal alpha carbons marked with cyan spheres and their C terminal alpha carbons marked with magenta spheres.

Force-extension relationships of C. elegans TTN-1 and twitchin Ig domains

For titin-like proteins, the organization is such that a set of five tandem Ig domains immediately follows the kinase catalytic core and its regulatory sequence (Figure 4.3.1A & B). If the kinase were to act as a force sensor, we would expect these domains to withstand forces greater than the kinase itself. To test this hypothesis we used single-molecule AFM techniques to analyze the mechanical properties of recombinant proteins containing 5 tandem Ig domains from twitchin and TTN-1 (Figure 4.3.1A-C). Random segments of these proteins were picked up by the AFM tip and then stretched with a pulling speed of ~0.5 nm/ms. The resulting force-extension curves showed a sawtooth-like pattern, characteristic of the unfolding of Ig domains (Figure 4.3.3A & D) (15, 27). To analyze the spacing between peaks in the sawtooth patterns we used the worm-like chain (WLC) model for polymer elasticity, which predicts the entropic restoring force (F) generated upon the extension (x) of a polymer [138,146]; see methods). The thin lines in Fig. 4.3.3A and 4.3.3D correspond to fits of the WLC equation to the curve that precedes each force peak. We found that the separation between force peaks for both proteins is ~30 nm (30.6 nm $\pm$ 3.2 for the twitchin protein; n= 53 force peaks; Fig. 4.3.3C). This value corresponds very well with the expected increase in contour length of a 95 amino acid (aa) Ig domain: 95 aa x 0.35 nm (length of an aa) – 3 nm (size of folded Twc Ig domain 26; (10) and Figure 4.3.2C) = 30.3 nm.

Unfolding force histograms show that the Ig domains from twitchin and TTN1 unfold at a similar range of forces and similar average unfolding force values (~90 pN and ~85 pN, respectively; Fig 4.3.3B & E). The unfolding forces of these Ig domains from twitchin and TTN-1 fall within the range or are somewhat weaker than those reported for Ig domains from vertebrate titin [99] and insect projectin [213]. It is noteworthy that the recordings from the 5 Ig domains of twitchin show an ascending pattern of peak

heights suggesting a hierarchy in mechanical stabilities for these Ig domains (Figure 4.3.3A). In addition, we found that mechanical unfolding of TTN1 Ig domains is fully reversible and that they refold to their native states with a rate at zero force of ~2 s^{-1} (not shown).

Our results show that the TTN-1 and twitchin Ig domains unfold at ~90 pN at a fixed pulling speed of 0.5 nm/ms. However, during normal muscle contraction cycles these domains may experience a wide range of stretching speeds. Since the mechanical stability may not be the same at different pulling speeds, as shown for titin domains [139,211] we studied the rate-dependency of the stability of Ig domains from TTN-1 and twitchin. Figure 4.3.3F shows the relationship between the unfolding force and the pulling speed (0.1-5nm/ms) for TTN-1 Ig domains. A tenfold decrease in pulling speed decreases the unfolding forces by only 20 pN indicating that these domains are mechanically stable over a wide range of pulling speed. The continuous lines correspond to the result of Monte Carlo simulations of two-state unfolding of this sequence at the corresponding range of pulling rates [139,143,147,211]. We found that, by using a combination of the rate constant at zero force, k_u^o, of $5x10^{-2}$ s^{-1} and the unfolding distance Δx_u between the folded state and transition state of 0.35nm, we can adequately describe the unfolding force histogram (Fig. 4.3.3E) as well as the speed dependency of unfolding forces (Fig. 4.3.3F). These values are comparable to those used for several other mechanically stable beta-strand rich domains [99,147,212,214]. These results are consistent with the hypothesis that these Ig domains are designed to resist stretching forces.

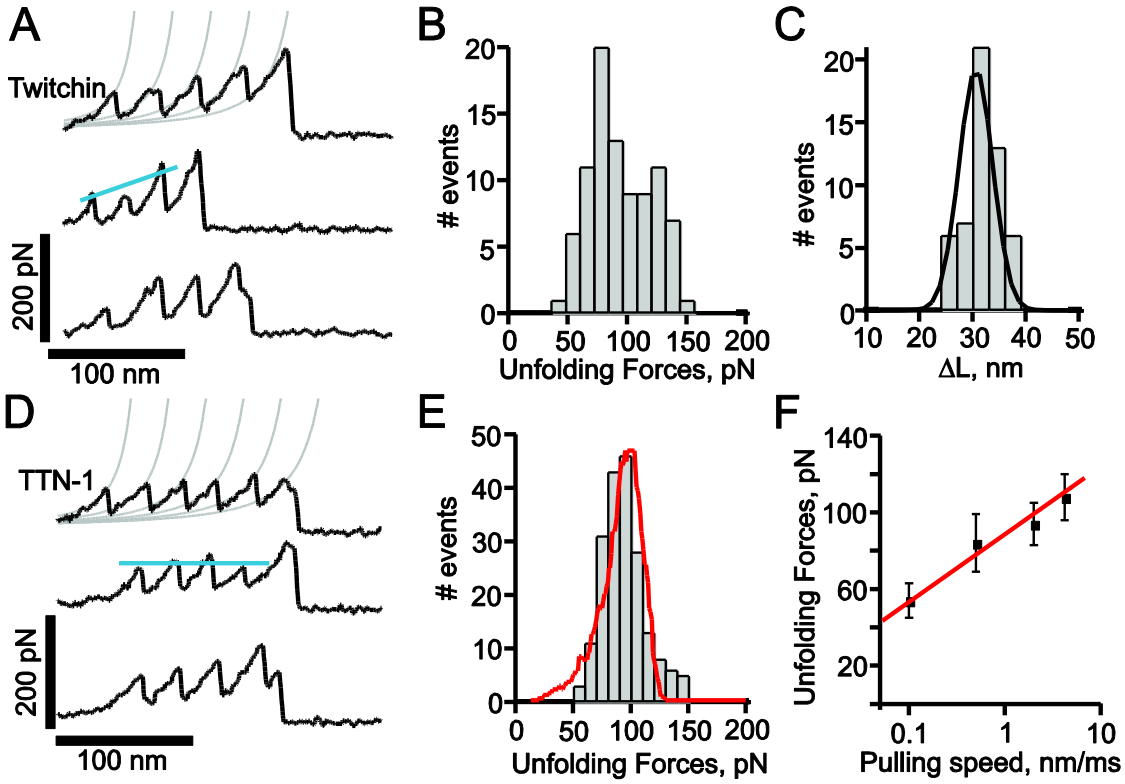

Figure 4.3.3 - Force-extension relationships of TTN-1 and twitchin Ig domains.

A, D) Several examples of force-extension curves obtained after stretching twitchin (A) and TTN-1 (D) Ig domains. The grey lines were generated with the WLC equation using a persistence length of 0.4 nm and contour length increments, ΔL, of 30 nm. B, E) Unfolding force histograms for twitchin and TTN-1 domains. The mean force peak values are 93 ± 25 pN, (n=88 peaks) and 85 ± 22 pN (n=193 peaks), respectively. In E) the red line corresponds to a Monte-Carlo simulation of TTN-1 Ig using $k_u^o = 5x10^{-2}\ s^{-1}$ and Δx_u = 0.35 nm at a pulling speed of 500 nm/s. C) Histogram of contour length increments observed upon unfolding of twitchin Ig domains shows one main peak centered at ~30nm (Gaussian fit: 30.6 ± 3.2 nm). F) The unfolding forces of TTN-1 Ig domains depend on pulling speed. The experimental data (black symbols) can be well described by Monte-Carlo simulations (red line) using $k_u^o = 5x10^{-2}\ s^{-1}$, Δx_u = 0.35 nm.

Force-extension relationships of C. elegans giant kinases and flanking domains

The protein kinase domains of twitchin and TTN-1 are autoinhibited, containing an endogenous regulatory sequence situated between the two subdomains, making extensive contact with residues essential for ATP binding, substrate recognition and catalysis. Unlike their closest homologs, the MLCKs, binding to Ca^{+2}/calmodulin does not relieve autoinhibition. It has been hypothesized that the giant kinases may act as force sensors [210]; that the forces generated from the contraction/relaxation cycles of muscle activity are sufficient to unleash the regulatory domain from the catalytic core and activate the kinase. The release from a stretching force would restore the inhibited conformation of the kinase.

We used single-molecule AFM to analyze the mechanical strength of the kinase and their flanking Ig/Fn domains. To be sure that the proteins were in their native conformations, the kinase activity of the proteins was measured. Although the proteins are autoinhibited, they have modest activity in vitro [203,204]. We found that our recombinant kinase constructs have specific activities comparable to those previously published (Figure 4.3.1D) [203,204].

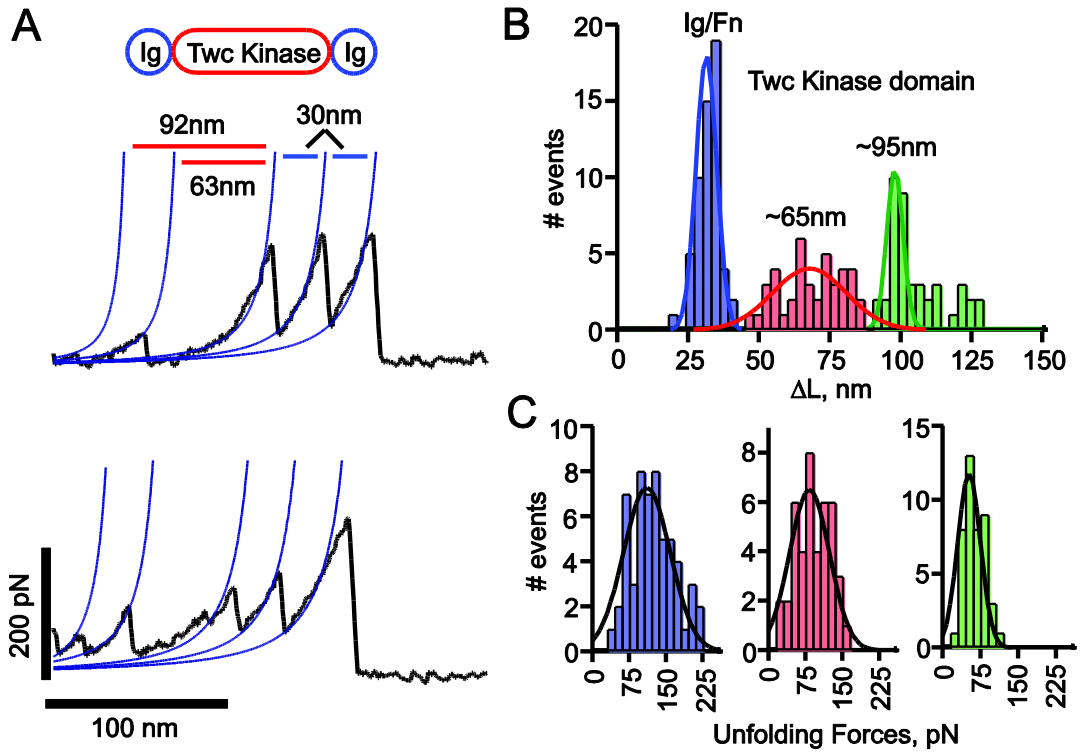

Figure 4.3.4 - Mechanical properties of *C. elegans* twitchin kinase.

A) Two examples of force-extension curves obtained for the Fn-Twc kinase-Ig construct. The two small force peaks correspond to the stepwise unfolding of the Twc kinase domain and the last two peaks to the unfolding of the flanking Ig/Fn domains. B) Histogram of increases in contour length increments observed upon unfolding, ΔL, of Fn-Twc kinase-Ig. There are peaks at ~30 nm, 65 nm and 95 nm (Gaussian fits: 31±5 nm, 67±18 nm and 97±10 nm, *n*=142), which correspond to the unfolding of Ig/Fn domains (blue bars), and kinase domain (red and green bars). C) Unfolding force distributions for Ig/Fn domains (blue bars) and kinase domain (red and green bars); the respective force peaks are at 111±67 pN (n=57), 83±57 pN (n=41) and 52±11 pN (n=43).

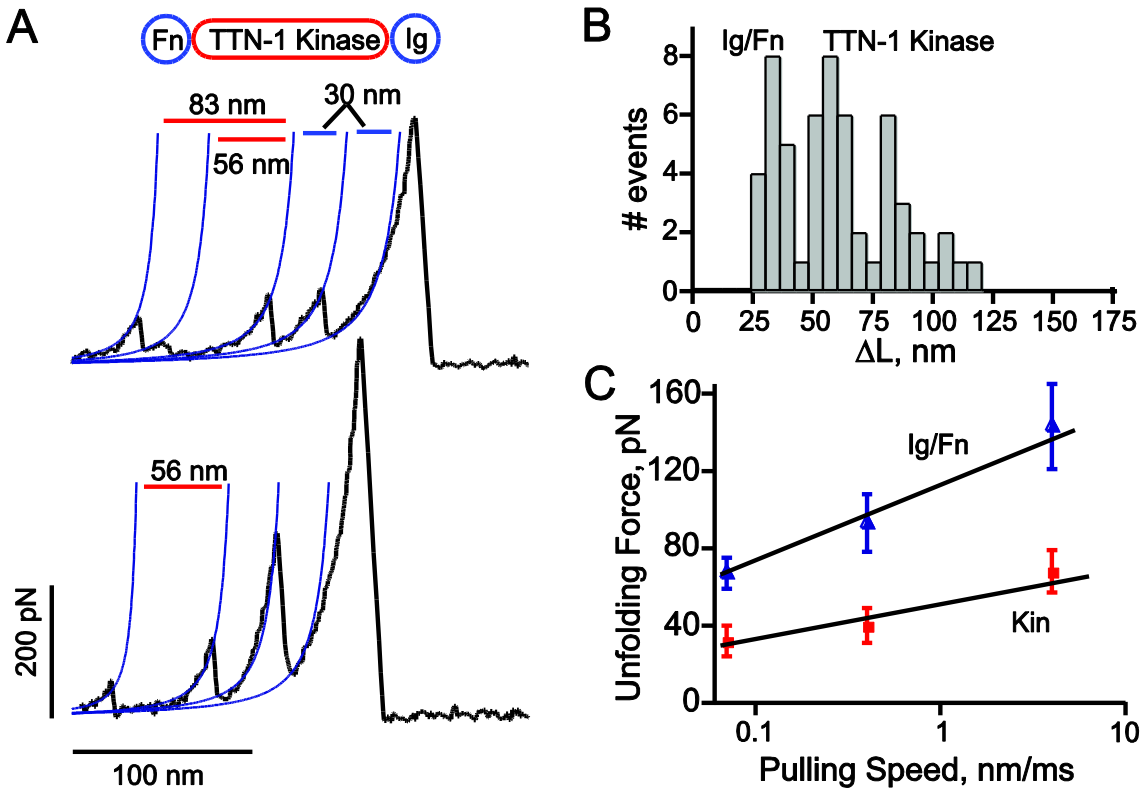

Figure 4.3.5 - Mechanical properties of *C. elegans* TTN-1 kinase.

A) Two examples of force-extension curves obtained for the Fn-TTN1 kinase-Ig construct. B) Histogram of increases in contour length increments observed upon unfolding, ΔL, of the Fn-TTN-1 kinase-Ig construct (16 molecules). There are peaks at ~30nm, 65nm and 95nm (n=35). (Gaussian fits: 29±16 nm, 63±15 nm and 95±13 nm). C) Plot of the average unfolding force versus the pulling rate for TTN-1 Ig/Fn domains and the kinase domain. For the kinase, we analyzed the first force peak with a ΔL of ~95nm. The continuous lines correspond to the result of Monte Carlo simulations of two-state unfolding at the corresponding pulling rates. The parameters used for the Monte-Carlo simulation are: $k_u^o = 1.4 \times 10^{-2}\ s^{-1}$, $\Delta x_u = 0.35$nm for the Ig/Fn domains and $k_u^o = 4 \times 10^{-2}\ s^{-1}$. $\Delta x_u = 0.6$nm for the TTN-1 kinase.

For the force measurements, dilute solutions of the recombinant kinase proteins were non-specifically attached to a glass coverslip. Random segments of the proteins were picked up by the AFM tip and then stretched with a pulling speed of 0.5 nm/ms. Figure 4.3.4A shows examples of force–extension curves obtained after stretching single Fn-Twc kinase-Ig molecules. We typically observed recordings with multiple force peaks before the last detachment peak. In the Fn-Twc kinase-Ig construct the Ig and Fn domains have ~95 aa and should therefore contribute to an increase in contour length, ΔL, of ~30nm (Figure 4.3.3). Hence, we attribute the two force peaks before the detachment peak as the unfolding of Fn and Ig domains and the initial two force peaks to the sequential unfolding of the kinase domain, which has a contour length of ~95 nm. Figure 4.3.4B shows a histogram of increases in contour length increments observed upon unfolding of Fn-Twc kinase-Ig. There are peaks at ~30 nm, 65 nm and 95 nm, which correspond to the unfolding of Ig/Fn domains (blue bars), and kinase domain (red and green bars), respectively. The corresponding unfolding forces are 111 ± 67 pN (n=57), 83 ± 57 pN (n=41) and 52 ± 11 pN (n=43) (Figure 4.3.4C). The small peak has a contour length of ~30nm, a length that would be expected to fit the unwinding of the ~100 aa small lobe. The second peak gives a contour length of ~65nm, fitting nicely to the unwinding of the ~190 aa large lobe. The regulatory domain probably unfolds at very low forces (<10pN), forces indistinguishable from noise on the AFM. Hence, we interpret the data shown in Figure 4.3.4 as the stepwise unfolding of the two lobes of the kinase. The smaller lobe, composed mainly of beta sheets will unfold first, leaving the larger, alpha helical lobe intact. This hypothesized order of rupture, and lack of any unfolding peak from the regulatory domain, is supported by our SMD simulations (Figure 4.3.6).

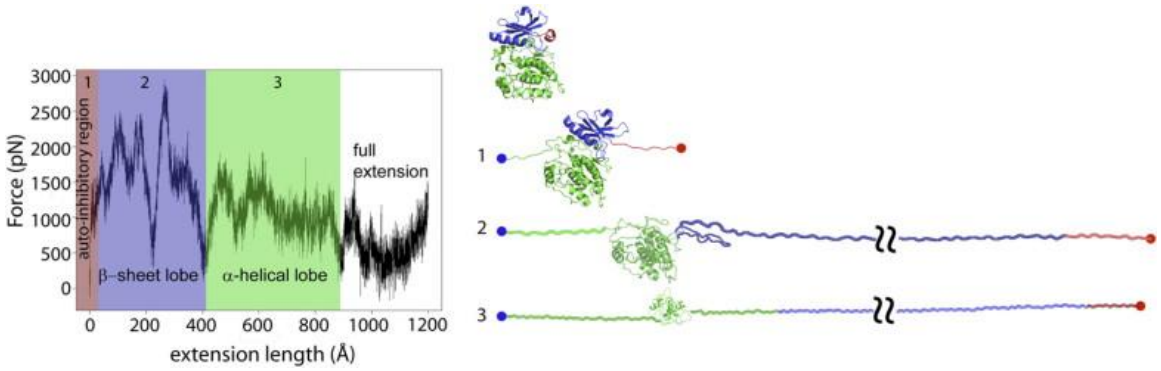

Figure 4.3.6 - Constant velocity steered molecular dynamics simulation of the mechanical unfolding of twitchin kinase.

Left) Force-extension curve obtained from SMD simulations by stretching the twitchin kinase domain (1KOA) between its C terminus and its N terminus at a pulling speed of 0.5Å/ps. The total simulation time was 2.4ns using 33,428 total atoms including 18 Na^+ and 9,305 water molecules. The fixed atom was Tyr 5915 and the SMD atom Arg 6261. Right) 4 snapshots of twitchin kinase stretched from its termini taken at no extension (rest), after 65Å (1), 340Å (2) and 639Å (3) of extension. At rest, the kinase domain is in a closed conformation. The active site is occupied by the autoinhibitory region (red), which makes extensive contact with the catalytic site, blocking substrate binding. 1) At low forces the regulatory tail will unravel reversibly and expose the active site to its substrates. 2) At high forces the kinase begins to unfold and the integrity of the active site is disrupted. The small lobe (blue), made mainly of beta sheets, unravels first followed by the unfolding of the alpha helical rich large lobe (green).

The unfolding pattern of TTN-1 Fn-Kinase-Ig parallels that of twitchin Fn-Kinase-Ig. The complete unfolding of TTN-1 kinase has four force peaks (Figure 4.3.5A) corresponding first to the biphasic rupture of the catalytic core, followed by the unfolding of Ig and Fn domains. There are peaks at ~30nm, 65nm and 95nm (Figure 4.3.5B). To further characterize the mechanical unfolding of the kinase domains, we analyzed the unfolding kinetics by doing experiments at different pulling speeds. Figure 4.3.5C shows a plot of the average unfolding force versus the pulling rate for TTN-1 Ig/Fn domains and the kinase domain. For the kinase, we analyzed the first force peak that gives a ΔL of ~30nm, which interpret as the unfolding of the small, beta-sheet rich lobe. The continuous lines correspond to the result of Monte Carlo simulations of two-state unfolding at the corresponding pulling rates. The parameters used for the Monte-Carlo simulation are: $k_u^o = 1.4x10^{-2}\ s^{-1}$, $\Delta x_u = 0.35nm$ for the Ig/Fn domains and $k_u^o = 4x10^{-2}\ s^{-1}$, $\Delta x_u = 0.6nm$ for the TTN-1 kinase. The unfolding rate constant at zero force, k_u^o, and unfolding distance between the folded and the transition state, Δx_u, for the Ig and Fn beta-sandwich domains are similar to that for TTN-1 Ig domains 38-42 (Figure 4.3.3F). However, the transition state distance Δx_u for the kinase alpha helical lobe is considerable larger than in the beta-sandwich domains (0.6nm vs. 0.35nm), suggesting different transition state structures.

From these data, we can conclude that the TTN-1 and twitchin kinase domains show a significant mechanical resistance and they unfold at forces similar to those for Ig/Fn beta-sandwich domains (30-150pN). These results further support the force activation hypothesis for the giant protein kinases.

DISCUSSION

The kinase domain of human titin has been postulated to act as a force sensor [197,210]. The kinase domain is thought to be autoinhibited at rest, but during muscle activity, catalysis by the kinase results in a phosphorylation cascade that ultimately causes the expression of genes important in myofibril maintenance and growth. Human titin spans half of the sarcomere, from the M-line to the Z-disk [215]. The segment of titin in the A-band ending at the M-line is fixed, whereas the portion of titin in the I-band varies its length in response to the state of muscle contraction [216]. The kinase domain, located at the M-line, is in an ideal position and orientation to sense the mechanical strain that occurs during the contraction/relaxation cycle of muscle activity. In *C. elegans* striated muscle, the proteins of greatest similarity to human titin are TTN-1 located in the I-bands [203], and twitchin located at the non-M-line portions of A-bands [217]. Given the high degree of similarity of the kinase catalytic cores and the conservation of organization of surrounding Ig and Fn domains, it is reasonable to hypothesize that force activation might also occur in TTN-1 and twitchin.

If the kinase domains were to act as force sensors, we would expect them to withstand stretching forces. Here we used single-molecule force spectroscopy to test the mechanical properties of *C. elegans* TTN-1 and twitchin kinases. The proteins were recombinantly expressed and shown to retain activity in their inhibited form. We found that the mechanical stabilities of the kinase domains are compatible with their postulated function. The kinases unfold in two clearly resolvable steps at forces of ~50pN and ~80pN. These unfolding forces are lower than most beta-strand rich domains, range ~80-250pN [99,139,187,213,214,218] but similar to alpha-helix rich proteins, range ~30-100pN [103,219,220].

Our results show that the TTN-1 and twitchin kinase domains unfold in a biphasic, stepwise fashion, indicated by the presence of two peaks in the force-extension recordings. The first force peak probably represents the unwinding of the smaller lobe of the catalytic core, corresponding to the breakage of beta sheets. Following the initial rupture is a second peak, probably representing the rupture of the larger, alpha helical lobe of the catalytic core. The small peak has a contour length ΔL of ~30nm, a length that is consistent with the unwinding of the ~100 aa small lobe. The second peak gives a contour length of ~65nm, which may correspond to the unwinding of the ~190 aa large lobe. The regulatory domain most likely unfolds at very low forces (<10pN), which is below the resolution of our AFM.

To further understand the molecular origin for the mechanical unfolding of the *C. elegans* twitchin kinase we carried out steered molecular dynamics (SMD) simulations. Molecular dynamics simulations have been extensively used to examine the mechanical unfolding of a wide variety of proteins and in general, there is a good correlation with single-molecule data [105,140,219,221]. To validate the accuracy of our *in silico* experiments we performed SMD simulations on other protein domains, such as titin I27, ubiquitin and synaptotagmin C2A. The I27, ubiquitin and C2A results are very similar to those published previously ([132,139,140]; Supplemental Figure 4.3.2). The magnitude of the forces observed in the SMD simulations does not directly correspond to those measured with AFM. This is partially because the pulling speeds are several orders of magnitude different. However, the simulations are qualitatively consistent with the AFM results. For example, similar to our AFM results twitchin Ig26 unfolds at slightly smaller forces than I27. In addition, as shown by AFM data [222] the C2A domain unfolds at much lower forces (~50pN; [222] than I27 or ubiquitin (~200pN) and does not show an initial force-extension burst.

Figure 4.3.6 shows constant velocity SMD simulation of the mechanical unfolding of the twitchin kinase domain. The force-extension curve was obtained from SMD simulations by stretching the *C. elegans* twitchin kinase domain (1KOA) between its C terminus and its N terminus at a pulling speed of 0.5Å/ps. On the right four snapshots of twitchin kinase are shown taken at no extension (rest), after 6.5nm (1), 34nm (2) and 64nm (3) of extension. At rest, the kinase domain is in a closed conformation. The active site is occupied by the autoinhibitory region (red), which makes extensive contact with the catalytic site, blocking substrate binding. At low forces, the regulatory tail will unravel reversibly and expose the active site to its substrates (snapshot 1; red region in the force-extension plot). The kinase is in an open, active conformation and downstream signaling can occur. During muscle activity, low forces may result from the repeated contraction/relaxation of the sarcomeres. At high forces, the kinase begins to unfold and the integrity of the active site is disrupted (snapshot 2; blue region in the force-extension plot). The small lobe (blue), made mainly of beta sheets, unravels first followed by the unfolding of the alpha helical rich lobe (green; snapshot 3; green region in the force-extension plot). Eventually, with sustained high forces, the enzyme will completely unravel (white region in the force-extension plot). We also performed constant velocity SMD simulations of the mechanical unfolding of the model for TTN-1 kinase (Supplemental Figure 4.3.3). The unfolding trajectory for TTN-1 kinase is nearly identical to that of the Twitchin kinase indicating a similar stepwise mechanical unfolding pathway, as seen in the AFM data (Fig. 4.3.5).

Grater et al. suggested a similar model from results of molecular dynamics simulations [210]. However, in their studies they predicted that only human titin, but not *Aplysia* twitchin, would be likely to function as a force sensor. The simulations show that when force is applied to human titin the termini of the protein shift away from the

kinase, the autoinhibitory region detaches from the active site, and that this release is accompanied by a shift in the two lobes of the catalytic core that results in an active site that is accessible to substrates. For *Aplysia* twitchin it was predicted that although applied force would remove the inhibitory region, the lobes of the catalytic core remain static and the enzyme would still be inactive. Our data argues that, in *C. elegans,* both TTN-1 and twitchin kinases meet the requirements of an enzyme likely to be activated by mechanical forces.

The most common feature of all the giant titin-like muscle proteins is the presence of multiple copies of Ig and Fn domains, either as tandem Ig domains or as super-repeats of Ig and Fn domains [198]. One major challenge in muscle biology is to understand how these modular domains function both at the individual and group scale, and how their mechanical properties vary to suit the type of muscle, or even the location of a domain within a single sarcomere. Single molecule experiments have revealed the strength of Ig domains of titin and titin-like proteins to vary between 50-300pN, depending on the position of the domain within the sarcomere [99,213]. We studied the mechanical properties of five tandem Ig domains from *C. elegans* TTN-1 and twitchin. These 5 Ig domains are located in the C-terminal regions of the endogenous proteins and immediately follow the conserved kinase domain. We hypothesize that this region is functionally important because the architectural arrangement of the kinase domain plus 5 Igs is conserved in all of the giant titin-like proteins. Our data shows that these Ig domains, at least in the nematode, are slightly weaker (~93 pN for twitchin and ~85 pN for TTN-1) than the average Ig domain and those found in other muscle types such as cardiac titin (~200 pN) [211], insect kettin (~125-250 pN) and insect projectin (~109 pN) [213].

TTN-1 can be regarded as a twitchin/titin hybrid. At the sequence level, the TTN-1 kinase catalytic core is more similar to the kinase catalytic core of twitchin (54.4%

identical) than it is to the kinase of human titin (39.2% identical). What makes TTN-1 "titin-like" is its enormous 2.2 MDa size and the presence of several regions consisting of tandem repeats that are likely to act as molecular springs (these are not found in twitchin, only in human titin). Using known crystal structures of the kinase domains of *C. elegans* twitchin (Figure 4.3.2A) [206], *Aplysia* twitchin [206], and human titin [209], we built a molecular model of TTN-1 kinase (Figure 4.3.2B) to further compare the proteins and analyze the subtle characteristics that underlie the force activation hypothesis. When visually comparing the TTN-1 model to the twitchin and titin crystal structures, TTN-1 seems most structurally related to twitchin (Figure 4.3.2A &B), however there does exist a few varying regions that seem to be upheld only between the TTN-1 model and the structure of human titin. The topology and the bond-structure that holds the regulatory domain in intimate contact with the catalytic core seem closely related for all four kinases. It seems likely that, if the force activation hypothesis is true, all of these enzymes could adhere to such an on/off mechanism.

In conclusion, we have provided evidence that in response to mechanical force, two giant titin-like protein kinase domains unfold in a step-wise manner. The first step is likely to be the movement of the autoinhibitory domain from the catalytic pocket, without complete unfolding of the domain. With the application of increased force to the protein, the kinase domain will rupture, in a biphasic manner. These data bolster the hypothesis that the autoinhibited giant kinases in muscle cells may be activated in response to the forces generated during each contraction/relaxation cycle. Further experiments are underway to test if applying small forces to these enzymes, forces sufficient only to remove the autoinhibitory region (Figure 4.3.6), would indeed result in the activation of these enzymes.

SUPPORTING FIGURES

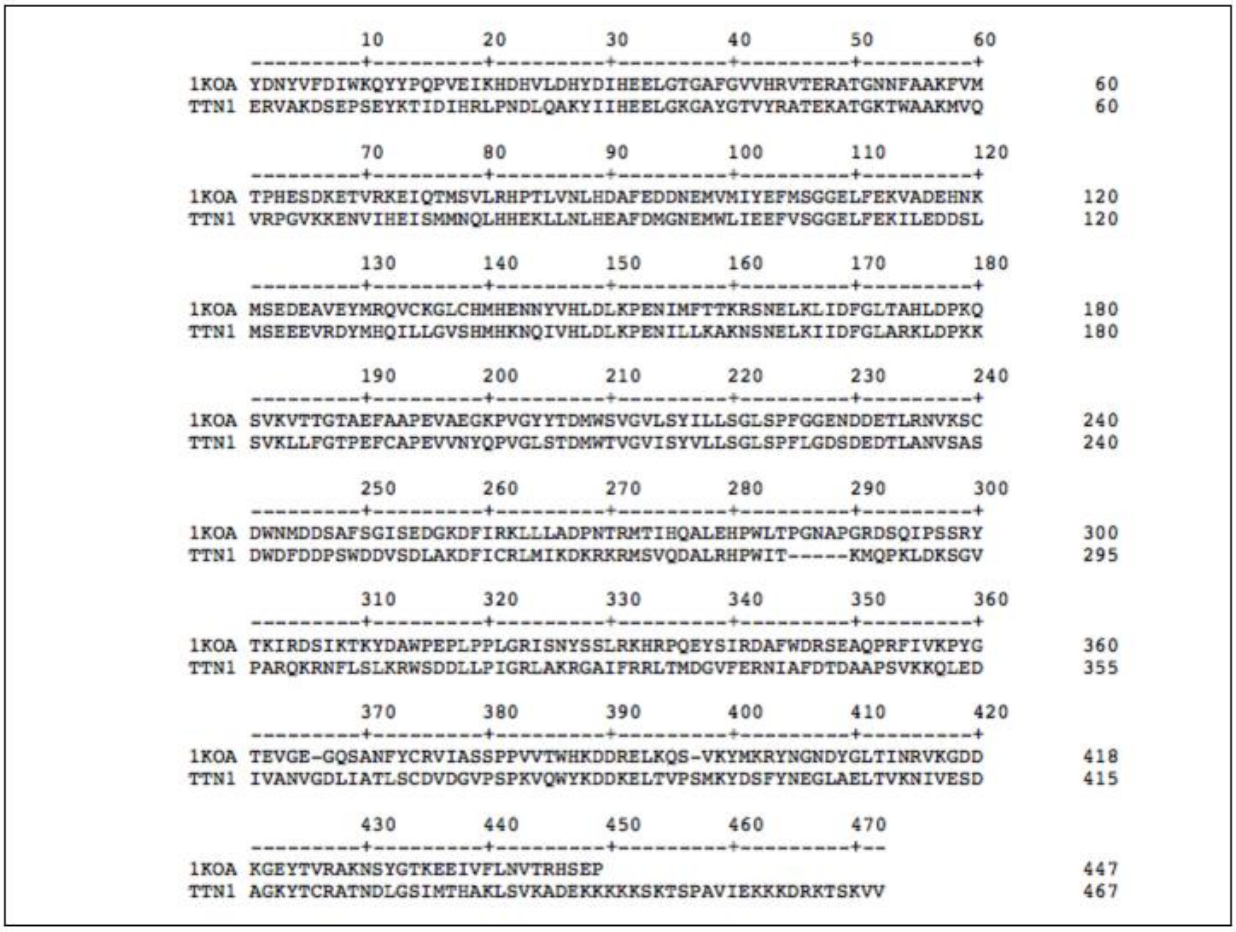

Figure 4.3.S1 - Alignment of TTN1 on 1KOA.

Alignment used for building the initial homology model of TTN-1. Note that the arbitrary start at 1 corresponds to twitchin (KOA) residue Y915 and TTN-1 residue E15907.

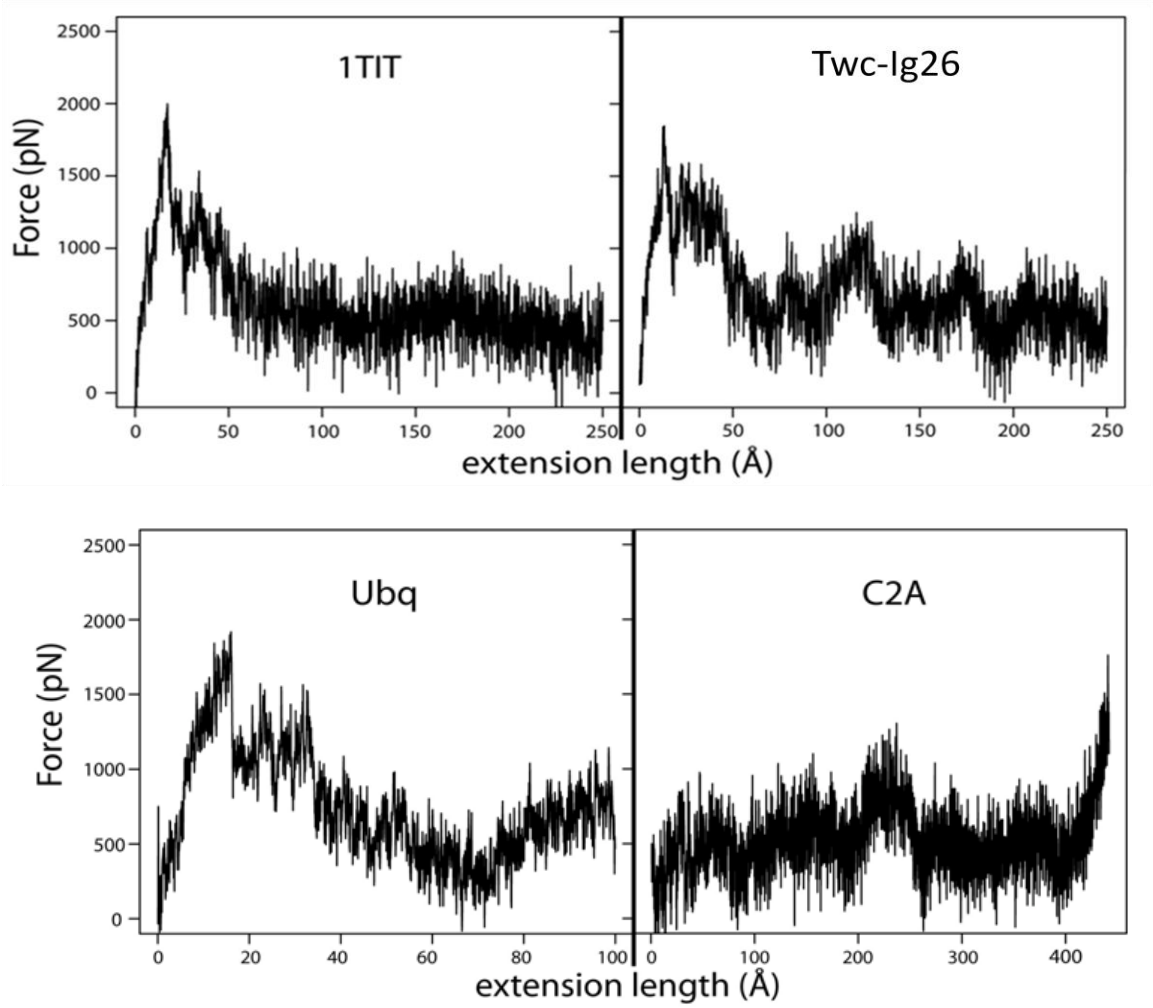

Figure 4.3.S2 - Constant velocity SMD simulations of the mechanical unfolding of several protein domains.

To validate the accuracy of our *in silico* experiments we performed SMD simulations on other protein domains, such as titin I27, ubiquitin and synaptotagmin C2A. Our SMD results are very similar to those published previously (Supplemental Figure 2).

I27 domain of titin (1TIT.pdb). 6875 total atoms in simulation. Water sphere composed of 1830 water molecules with 6 Na^+. Pulling speed of 0.1A/ps for a total simulated time of 1.25ns. Fixed atom = Leu1 and SMD atom = Leu 89

Twc-Ig26 domain (1KOA.pdb). 6879 total atoms. 0.1A/ps for a total simulation time of 1.25ns. Water sphere including 5Cl^- and 1781 water molecules. Fixed atom = Ala 6264, SMD atom = His6358

Ubiquitin (1UBQ.pdb). 0.5 ns total simulation time. 6670 atoms, 1813 waters. Pulling speed of 0.1A/ps. Fixed atom = Met1. SMD atom = Gly76

C2A (1RSY.pdb). 5ns total simulation time. 0.1A/ps pulling speed 11,052 atoms total including 4 Na^+ and 2 Cl^-, 3025 water molecules. Fixed atom = Leu 142, SMD atom = Leu 262

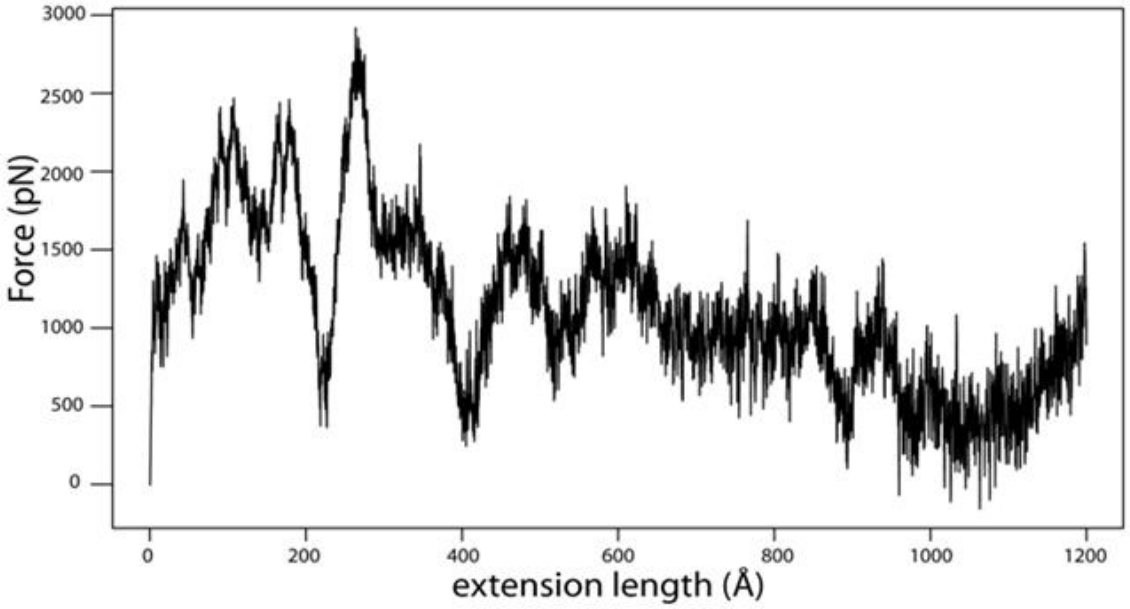

Figure 4.3.S3 - Constant velocity SMD simulation of the mechanical unfolding of the homology model for TTN-1 twitchin kinase.

Pulling speed 0.5A/ps. 2.4ns total simulation time. 33460 total atoms + 18 Na^+, 9305 waters. Fixed atom= Glu 1 SMD atom = Asp 342. The unfolding trajectory is nearly identical to that of that of twitchin kinase indicating a similar stepwise mechanical unfolding pathway, as seen in the AFM data (Fig. 5).

CHAPTER 5 – FUNCTIONAL ANALYSIS OF THE TITIN I-BAND

Major contributors to section 2 include: Werner Braun, and Andres F. Oberhauser

Section 1: Early Investigations

INTRODUCTION

With the advent of single molecule techniques which can probe the mechanical properties of macromolecules, new questions about the molecular basis of mechanical stability have been raised. In this regard there have been many experimental investigations of titin Ig domains and especially the 27th Ig domain from the N2B splice isoform of titin. These domains are small (~90 residues) 'beta sandwich' domains consisting of two antiparallel beta sheets in which the N and C terminal beta strands (A-A', and G) are adjacent and terminate on opposite ends of the domain in 3D space (figure 5.1.1). This topology in which the A and G strands are pulled in an antiparallel direction contributes to the strong mechanical design and resilience of the domain.

Previous SMD studies have shown that the hydrogen bonds between the A' and G strands are all stressed at once when a tensile force is applied in opposite directions to the N and C termini. Once those bonds are broken the recorded force on the SMD atom drops significantly indicating that this event is the major energetic barrier to the mechanical unfolding of these domains. This was an important observation that led to the hypothesis that the hydrogen bonding between these two strands was the main determinant of mechanical stability. This view would not last long as future studies published new data that indicated a much more complex model was needed.

Prior to the present work no sequence or property based analysis had been published on titin with the goal of determining determinants of mechanical stability. The only computational tools applied had been simulations of varying complexity. In this project we have focused on the mechanical properties of the seven domains from the I-

band of titin for which mechanical data is available. We have applied several computational methods to examine the molecular mechanisms behind mechanical stability including correlated mutation analysis, PCPMer motif analysis, in silico unfolding using the FANTOM energy minimizer, and SMD.

PCPMER

Initial applications of PCPMer to this data set used a traditional approach of searching for motifs that could tell us something about the important functional regions of the protein. This began with the collection of titin sequences from every available source. Once this large set of sequences were collected and organized, they were grouped in varying ways for alignment and PCPMer analysis.

The 1st (I1) and 27th (I27) domain of the I-band of N2B titin had both been mechanically characterized, and the structures were available for both. Because of the availability of this information we focused on these two domains. We used PCPMer to analyze alignments of the entire I-band, just I1 sequences, and just I27 sequences. This produced a set of motifs characteristic of each alignment some of which overlapped with motifs in other alignments.

The motifs designated by PCPMer consist of a series of residue positions which score above some threshold value, are within some maximum gap distance of one another (usually 2), and have a total length above a defined minimum (usually 4). A high scoring position indicates that that position had a high degree of conservation in one of the physical-chemical property indices used by PCPMer. A motif may therefore be comprised of several significantly conserved positions as well as several non-conserved positions.

One interesting result was a difference in conservation patterns for specific I-band domains when compared to the pattern of conserved residue properties for the entire N2B I-band (figure 5.1.2). In both domains there is a striking alternating pattern between many of the conserved positions from each group. The positions that are well conserved in the I-band alignment are largely those that compose the hydrophobic core of the protein. It is interesting then that the most well conserved positions in each domain-specific alignment largely ignore the hydrophobic core residues. In part this is due to an inequity in the way PCPMer scores conservation.

The more rare residues can achieve much higher scores if they are conserved if compared to a more common residue. For example a completely conserved tryptophan would score much higher than a completely conserved leucine. This is not necessarily an undesired behavior because the tryptophan would be a much more interesting residue to consider than the leucine. Even given this consideration, the residues that are well conserved in the domain-specific alignments indicate a degree of specialization for each of them. Several of the positions highlighted in the I1 domain form important inter-sidechain hydrogen bonds and salt bridges.

This initial PCPMer analysis begins to reveal that important differences exist in the domains of the I-band which depart from the generic Ig domain skeleton. Those differences could be the responsible determinants for the differences in mechanical properties of I-band domains. A later approach is detailed in section 2 of this chapter which was able to provide much more specific predictions about these molecular determinants.

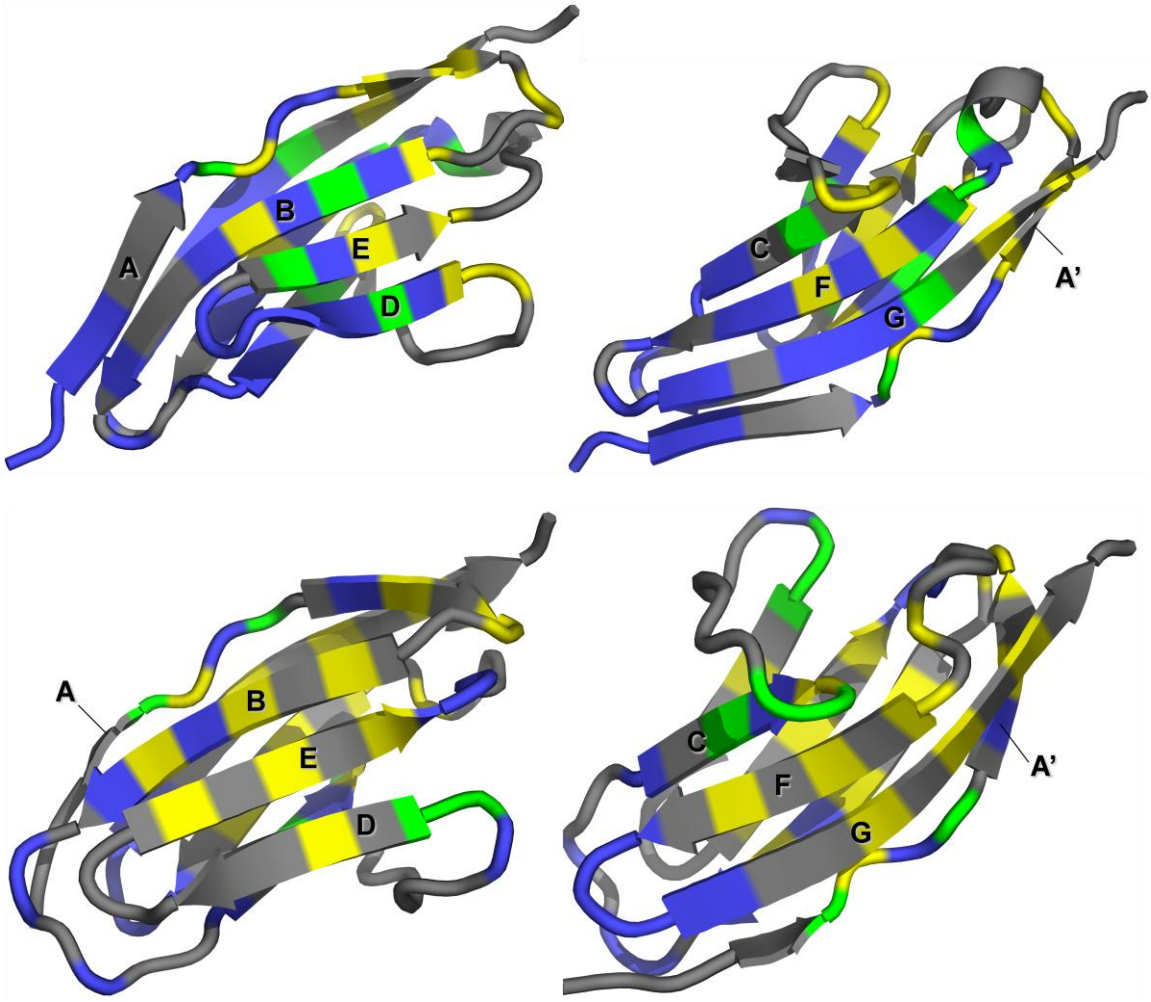

Figure 5.1.2 - Conservation pattern of I1 domains vs I-band domains

Positions that are well conserved in the I-band are highlighted in blue, those that are well conserved in just I1 domains are highlighted in yellow. Any overlapping positions are shown in green. Top) Front and back views of the I1 domain. Bottom) Front and back views of the I27 domain.

CORRELATED MUTATION ANALYSIS

This technique is usually designed to identify intraprotein residue-residue contacts from primary sequence. In this application we were instead trying to discover important residue-residue interactions that could be responsible for mechanical stability. For this purpose we generated an alignment of titin I-band Ig domains for which unfolding forces were known. This alignment included sequences from seven Ig domains from the I-band from up to nine species.

The basic method is to treat each pair of residues in a sequence as a point in physical chemical property space. A correlation coefficient (r-value) is calculated for the set of these points defined by the same pair of residues from all individual sequences in the alignment (i.e. residues 1 and 2 from all sequences). In this way the physical-chemical properties of each pair of columns in an alignment are used to calculate the r-value according to the following formula:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}$$

Where x and y are physical chemical property index values for a pair of residues and n is the number of residue pairs under consideration. Columns with completely conserved residues or gaps are disregarded.

A web-based analysis tool was created to allow easier inspection of correlation scores and the columns in the alignment between which a correlation is detected (figure 5.1.3). This was written entirely in PHP and can analyze any alignment of proteins using a variety of property indices to detect correlations.

The strongest correlations are shown in figure 5.1.3 for the indices of sidechain volume and charge. These r-values are not very strong, and the correlated residues they

highlight do not suggest a clear interaction. This could suggest that paired interactions are just too simplistic a model to accurately represent these proteins. CMA methods are generally used to predict long-range residue contacts within a peptide chain and even after nearly two decades of study success has been limited [223]. Although several methods for reducing background noise are being pursued, the technique does not yet seem to be sufficiently sensitive for our purposes.

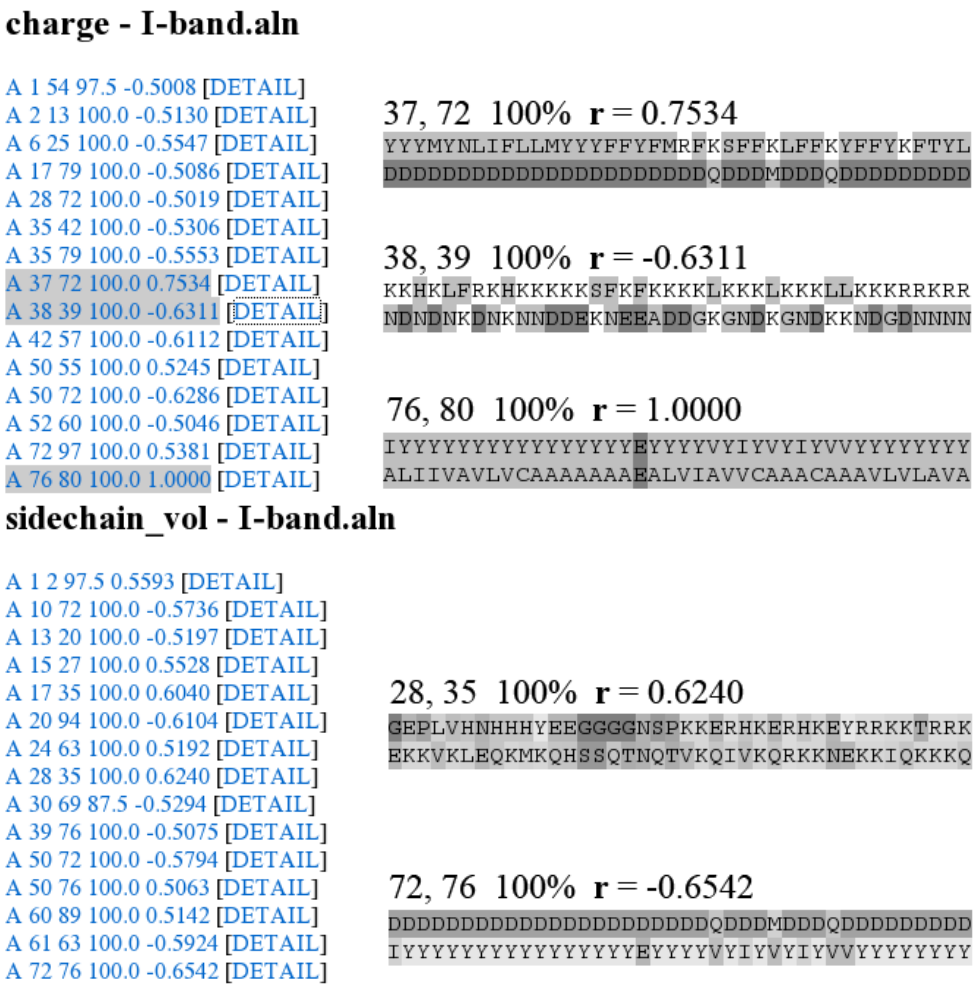

Figure 5.1.3 - Correlated mutation analysis

Screenshots of the output screen from the CMA PHP script. Each result page is headed by the property index used and the alignment file used. A listing follows for all correlations with an r-value above 0.5000. Each line represents one pair of columns which are listed first. These are followed by the fraction of sequences which contributed to the score (gaps are excluded), and the r-value itself. Finally each line has a [DETAIL] link which displays detailed information about the columns that were compared. The highest scoring correlations detected for the I-band Ig domains with respect to charge and sidechain volume.

FANTOM

The **Fa**st **N**ewton-Raphson **To**rsion Angle **M**inimizer is used to energy-minimize protein structures for homology modeling or other similar purposes. One of its functions is called PATHWAY and finds a low energy pathway between two protein conformations. This was initially developed for relatively benign protein movements such as the opening and closing of an enzyme catalytic site. We were interested to apply it to better understand the mechanical stability of titin I-band Ig domains.

In order to use PATHWAY two conformations of the same protein are required. The program then determines a linear path through torsion angle space that would connect the two conformations (figure 5.1.3). At a user specified number of intermediate steps, the conformation is minimized in a manner that is constrained to not move forward or backward along the path in torsion angle space. The minimized structures can be recorded along with energetic and conformational information. For our purposes we were interested in starting from the native folded state and moving to a linearized state. Starting with the folded NMR structure of I27 (pdb 1TIT) we created a linearized form of the protein using the editing capabilities of the PyMOL structure viewer. Using the native and linearized forms of I27 we ran PATHWAY.

The PATHWAY command shows promising results in replicating features observed by force spectroscopy (figures 5.1.5 and 5.1.6). Unfolding pathways show an initial resistance to unfolding followed by a rapid increase in conformational energy correlating to extended, disordered conformations. Large variations in the radius of gyration and a recollapse toward the end are also similar to behaviors observed during incomplete relaxation experiments. Although it is not yet possible to directly connect the results of unfolding pathways from FANTOM to force spectroscopy it is hoped that with

some modifications this tool could become a computationally inexpensive tool to study protein mechanical design.

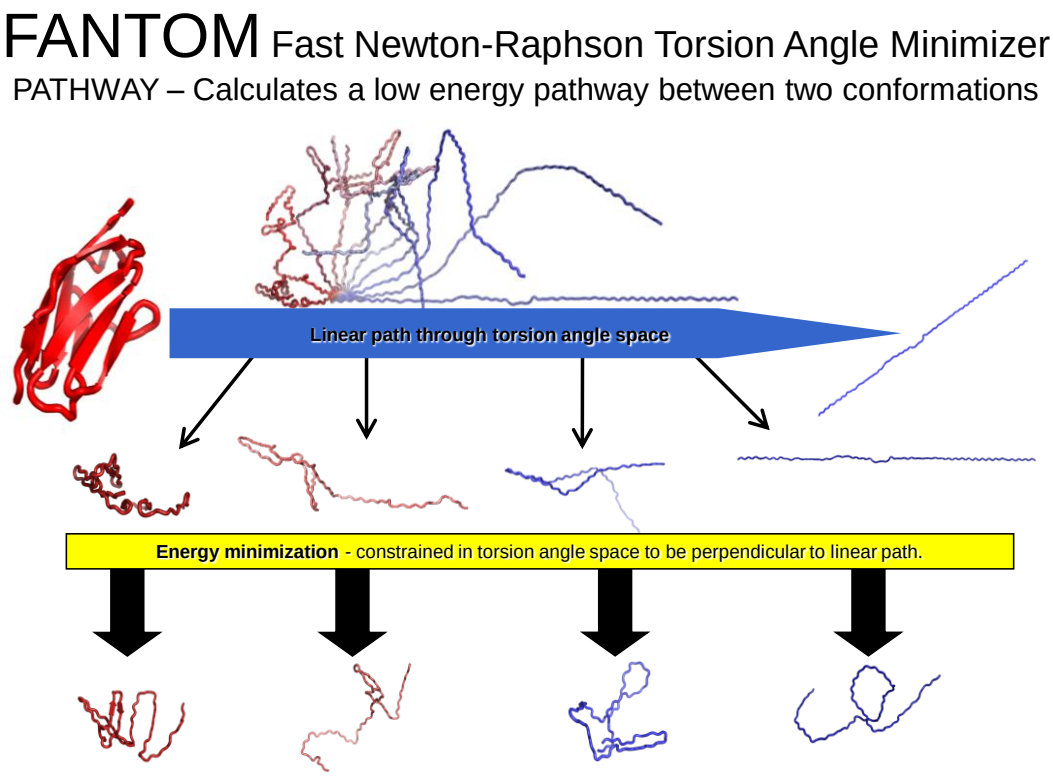

Figure 5.1.4 - Illustration of PATHWAY command

The pathway command first determines a linear vector through torsion angle space between two conformations. The user specifies a number of equidistant intermediate conformations which are then minimized but are constrained to be perpendicular to the original vector. The minimized states form a low energy pathway between conformations.

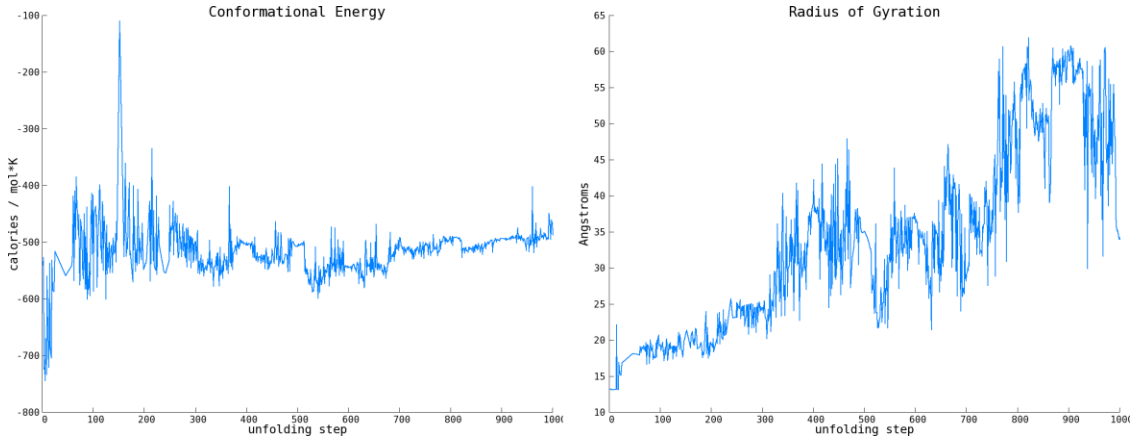

Figure 5.1.5 - PATHWAY for I27 linearization

The resulting pathway exhibits some expected features. The conformational energy plot shows a low energy valley near the folded state of the protein followed by a steep rise as the pathway passes through an energy barrier. The energy of the protein rises once more toward the end of the pathway as increasingly linear conformations are explored. The radius of gyration also has interesting results. The very beginning of the pathway is characterized by a flat line before a sudden steep increase in the radius of gyration indicating that the native fold is able to resist and recover from the perturbations at first. After a steady increase the radius of gyration explores a large range between ~25 and ~45 Angstroms after which there is another sudden increase.

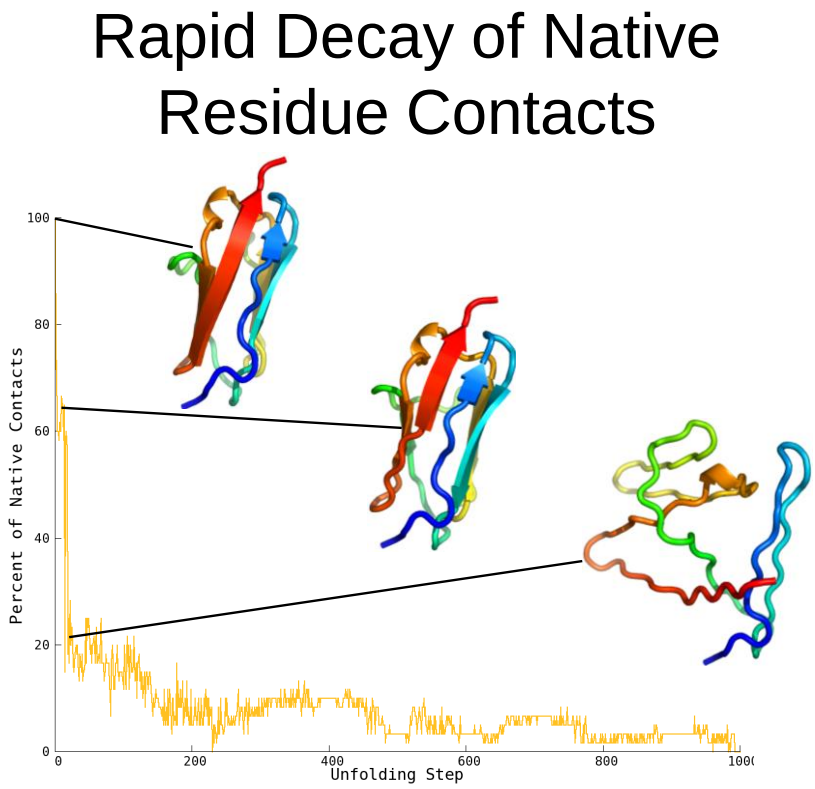

Figure 5.1.6 - Rapid decay of native contacts

Approximately eighty percent of the native residue-residue contacts are lost in the first 50 steps of the low energy pathway. We also see evidence of an initial resistance to unfolding in the first ~10 steps as in the plots of radius of gyration and conformational energy.

STEERED MOLECULAR DYNAMICS

Steered molecular dynamics (SMD) is the final computational tool we used to examine the mechanical design of Ig domains. With the continued trend toward more powerful and less expensive computers, all atom MD and SMD simulations have become feasible and within reach for researchers without the need for large computer clusters. This technique has been applied to titin Ig domains before [99,224,225] and led to the important observation that the rupture of the A' to G hydrogen bonding patch represents the highest barrier to unfolding for I-band Ig domains. Although computational power has improved, the length of time simulated must remain much shorter than the average pulling speeds in experimental techniques. The main implication is that the recorded forces of unfolding will be greater by an order of magnitude. Molecular dynamics simulations are of course not perfect representations. It follows that SMD is not a perfect representation, but some useful results have been achieved. Although the exact forces of unfolding are currently not possible to predict, the general hierarchy of mechanical stability seems to hold in some cases. We used SMD as a supplemental tool to examine the mechanical unfolding of various domains of interest. Some of these results are given in chapter 4 section 3.

Section 2: Mechanical Stability and Differentially Conserved Physical-chemical Properties of Titin Ig-domainsIntroduction

INTRODUCTION

The mechanisms that determine mechanical stabilities of protein folds remain elusive. Our understanding of these mechanisms is vital to both bio-engineering efforts and to the better understanding and eventual treatment of pathogenic mutations affecting mechanically important proteins. In the past the stability of protein folds has been measured mainly by thermodynamic experiments that give data on the behavior of an ensemble of protein molecules. In this work we seek the determinants of the mechanical properties of the immunoglobulin-like (Ig) domains of the giant muscle protein titin by single molecule techniques and comparative sequence analysis.

It is now well established that the passive elasticity of striated muscle is mainly regulated by titin [160,163,226,227]. For example, titin based passive-elasticity is an important contributor to the diastolic wall stress of the myocardium [228] or the storage of elastic strain in the passively elongating insect flight muscles [167]. Titin, a large rope-like protein composed of hundreds of sequentially arranged Ig and fibronectin type III (Fn3) domains (Fig. 5.2.1), functions as a molecular spring and ensures the return of the sarcomere to its initial dimensions after muscle relaxation. The ends of each titin molecule are anchored into the side and median walls (Z-disc and M-line respectively) of a sarcomere so that the ~4MDa protein spans half its length. The well over 300 exons found in the titin gene give an indication of the vast number of possible splice isoforms of titin [229]. Because of the availability of mechanical data on individual domains in the constitutive segments of titin, in this study we focus on the N2-B splice isoform of titin. The region of N2-B titin found in the I-band of a sarcomere is composed of ~40 Ig domains plus unique sequence

in the PEVK and N2B segments. Both elements are known to contribute to the extensibility and passive force development of relaxed muscle fibers during stretch [160,177,230]. Single molecule experiments with optical tweezers and AFM have demonstrated that titin behaves as an unusual spring where reversible domain unfolding plays an important role in its elasticity [132,143,161,186,230]. Domain unfolding has a dramatic effect on the extensibility of titin because of the large gain in length: unfolding increases the length of each domain by 7-fold (from ~4 to 28nm). Our recent data [214] show that titin Ig domains can refold under relatively high forces, indicating a very robust refolding mechanism that can operate over a large range of sarcomere lengths. Reversible Ig-like domain unfolding is thought to serve as a safety mechanism that protects titin and the sarcomere from mechanical damage in case of extreme stretch during stress (e.g. hemodynamic overload) or pathological conditions (e.g. chronic heart disease) [157-163].

Early sequence analysis of titin concentrated on the functional identification of domains within the titin sequence. These studies detected hundreds of type I and type II repeats which show strong homology to fibronectin type III and immunoglobulin-like domains respectively [182,231]. The earliest reported sequence alignments of Ig domains from the I-band of titin revealed features that defined domain and linker subclasses of these domains [179,232]. Further work revealed the existence of domain super-repeats in the I- and A-bands of titin [183,233-235]. Additional studies revealed conserved surface patterns of amino acids common to a large fraction of titin proteins as potential myosin binding sites within their fn3 modules [184]. Repetitive 28-residue motifs have been found in the PEVK [236,237] region which is thought to play an important role in passive muscle tension as well as in signaling [158]. The inter-domain regions between Ig domains show a well-conserved EPP motif which plays a significant role in maintaining a tight junction between neighboring Ig domains that lack a longer linker [238]. It has also been proposed

that multi-domain super-repeats exists where longer 3-residue linkers between domains allow for the flexibility of segments otherwise connected by stiffer Ig domain junctions lacking the C-terminal 3-residue linker [238]. These studies on various isoforms and regions of titin have all provided useful insights into the many functions of the titin molecule based on somewhat overt sequence conservation motifs [229,234].

Here we apply a new differential sequence motif analysis to understand the difference in the mechanical stabilities of titin domains. This new approach is based on conservation of physical chemical properties of amino acids rather than on the conservation of individual amino acid residues. With this method we can identify common and unique motifs in different Ig domains that have a large sequence variation. The analysis is performed with the sequence analysis tool PCPMer that detects motifs based on the conservation of physical-chemical properties of amino acids. It is capable of detecting similarities in structure or function in otherwise very different proteins [35]. Motifs defined by PCPMer can aid in locating functionally related regions of proteins that may have low sequence identity [44]. For example we successfully used this approach to locate regions in the apurinic/apyrimidinic endonuclease (APE1) far from the active site that modulate substrate binding and processivity and related nucleases [35,239]. Physical chemical property (PCP) motifs can also be used to select among different sequence alignments to potential templates in 3D homology modeling projects and can suggest biochemical functions for hypothetical proteins [34]. We also showed that a PCPMer analysis can find functionally important residues in surface exposed regions of viral proteins [45,240]. Here we will apply this powerful utility in a novel way that could become a powerful method of identifying common sequence properties among protein folds with similar mechanical stability.

Recent single-molecule force spectroscopy data show a mechanical hierarchy in the I-band domains [143]. Domains near the C-terminus in this region unfold at forces 2-3 times greater than domains near the beginning of the I-band. Though all Ig domains are thought to share a common fold and topology, the sequences of neighboring domains vary greatly with sequence identities in the range of 25%. We found, however, that the sequences of Ig domains with identical positions within the titin molecule are highly conserved across widely diverse species such as humans, chickens, and zebrafish. This implies that the mechanical properties of each domain are well conserved functions. The question we address here is; what are the molecular determinants for the unique mechanical stability of each I-band Ig domain? We have used sequence analysis techniques to search for properties common in weak domains as opposed to those found in strong ones and have found several interesting results. In contrast to other studies, we use experimental mechanical data to classify our sequences instead of sequence similarities. The clearest overall trend we found was that residues in the stronger domains are larger than in weaker domains. Our analysis also revealed new sequence motifs that are unique to families of different strengths. We consider these unique motifs most likely to be important in determining the mechanical stabilities of each domain. Further, these unique motifs exist mostly in the A, A′, and B β-strands supporting their importance in resisting mechanical unfolding [99,192,241]. In addition we found several individual positions in weak and strong domains that have a statistical difference in amino acid compositions with respect to hydrophobity and size of amino acids. These positions are scattered across the beta-sandwich fold, and are not only confined to the A'-G region.

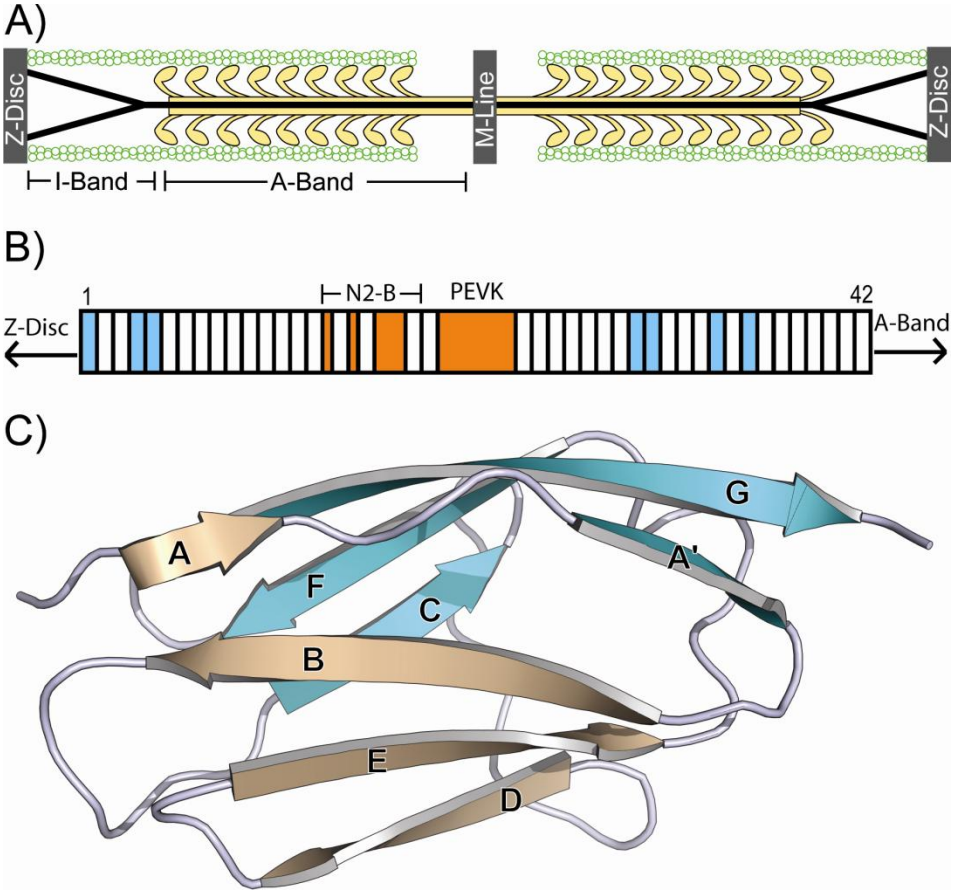

Figure 5.2.1 - Location and architecture of titin, and the structure of an Ig domain.

A) Single titin molecules (black lines) span half the length of the sarcomere. B) The domain architecture of the I-band of the N2-B titin isoform is composed almost entirely of Ig-like domains and is exposed to tensile forces in vivo. The domains used in our analysis are highlighted in cyan (I1, I4, I5, I27, I28, I32, I34). C) The structure of I27 (pdb code 1TIT [179]) is given as an example of the beta-sandwich, Ig-like domain repeated approximately 40 times in the I-band of human cardiac N2-B titin. The repeats occur head-to-tail producing an extended chain similar to beads on a necklace. The topology of the domains lends itself to mechanical resistance when its N and C termini are pulled in opposite directions.

MATERIALS AND METHODS

Construction of I27 Polyproteins

We used a method based on Cys-Cys disulfide formation to generate I27 polyprotein [242]. Standard molecular biology techniques were used to mutate the third residue in I27 to a cysteine and append an Arg-Ser-Cys-Cys sequence to the C-terminal end. The protein was expressed in *E. coli* and purified by Ni^{2+}-affinity chromatography. The protein was then concentrated to ~4 mg/ml in a volume of 1 ml. This was incubated at 37˚ C for 67 hours to promote the formation of polyproteins with multiple number of I27 domains (up to 20 as judged by gel electrophoresis). Dilute solutions of the polyprotein (~0.1 mg/ml) were suitable for AFM studies.

Single Molecule Force Spectroscopy

In this technique the atomic force microscope (AFM) is used to apply end to end tensile force to a single molecule on the scale of forces near and above those that may be experienced natively by that molecule [132,133,160,161,214,230,243-245]. In such an experiment a protein is tethered between a glass coverslip and a cantilever and stretched over several hundred nanometers. As the molecule is stretched it is elongated and the slack is taken up. Once taut, the force-bearing interactions are stressed and the force being applied to the molecule increases. As the force builds up to some threshold the force-bearing interactions break and the length of the backbone is introduced into the end-to-end length of the molecule. The force experienced by the molecule then drops sharply. As the extension continues the force builds back up to unfold the next folded domain. This repeating pattern of slowly building forces followed by sharp drops produced by domain unfolding is called a saw-tooth pattern and can give much useful information about the mechanical properties of the protein [161,230,244,245]. The force peaks in such a pattern are

indicative of domain unfolding events and can be fit to the worm-like chain (WLC) equation [140] that describes the extension of polymers under a mechanical stretching force. From fitting each peak to this equation, the size of the domain protected by the force bearing bonds can be estimated by measuring the increases in contour length between fit models. In addition, by constructing a frequency histogram for the force peaks we can determine the average unfolding force (or mechanical stability) at a given pulling speed for each titin Ig domain.

Multiple Sequence Alignment

Searching for titin domain sequences from different species posed a challenge as there are many hundreds of titin forms arising from differential splicing. Additionally the titin gene may not be entirely represented in the sequenced genomes of all species. We used a BLAST search to find well matching sequences for domains with the same position within titin sequences from different species. Since the I-band domains show high sequence diversity among domains at different positions, false matches could be reliably ruled out. Among high scoring matches we checked if they were properly positioned in their respective overall sequence; i.e. that the sequence for Ig_n came before that of Ig_{n+1} by about the same distance in all species. For example segments of a given titin sequence that match I1 should be separated from the start of sequences that match I2 by roughly the same number of amino acids across species. Additionally, the high degree of conservation among domains from different species yet with identical positions gave very low BLAST e-values (1e-50 to 1e-20) that were easily distinguishable from e-values of neighboring Ig domains with significantly higher e-values (1e-10 to 1e-3).

Once the sequences were gathered we align them to match structurally equivalent areas. To this end we began with a structural alignment of I1 to I27 conducted using the

CE algorithm [246]. The structure of I1 (pdb code 1G1C [247]) from the I-band of titin was determined by X-ray crystallography, whereas that of I27 has both an NMR (pdb code 1TIT [179]) and crystallographic structure (pdb code 1WAA) which agree very well (Cα rmsd 1.2Å as calculated by CE). I1 has a slightly longer sequence with two single-residue insertions and one four-residue insertion; all of which are in loop or turn regions allowing for a very nice overlap of common residue positions. The sequence alignments generated by CE from I1 to both structures of I27 are identical. Next the rest of the sequences were aligned to the I27 and I1 sequences using the clustal/w algorithm. Finally the alignments were merged by hand preserving the information gained from the structural alignment.

PCPMer

The PCPMer software package (available at: landau.utmb.edu:8080/WebPCPMer) is used to analyze protein sequence alignments of related proteins in order to detect conserved physical-chemical properties [35]. Briefly, PCPMer uses a 5-dimensional space to describe the physical-chemical properties of each amino acid. When processing a multiple sequence alignment, PCPMer records, in a global profile, the distribution in each of 5 orthogonal vectors at each position of the alignment as a mean and standard deviation. In order to define motifs PCPMer compares the observed distributions of amino acid in a multiple sequence alignment to the *a priori* distribution of amino acids in a representative set of protein sequences by means of a relative entropy measure. Positions above a threshold level are flagged as significant and sequential clusters of these significant positions that satisfy minimum length and maximum gap criteria are defined as conserved property motifs.

Analysis of the Properties of 'Weak' and 'Strong' Titin Ig Domain

In this study, PCPMer has been applied in a novel way. First the seven domains for which characteristic unfolding forces are known were divided into two families; a '*weak*' family containing domains with an average unfolding force of less than 200pN and a '*strong*' family with an average unfolding force above 200pN. The *weak* family is comprised of sequences for Ig domains 1, 4, and 5 from the titin I-band. Based on its sequence and structural characteristics, I1 has more in common with the Ig domains of the skeletal tandem. Its inclusion in our *weak* family strengthens our analysis because it introduces more background variability while maintaining elements that may be responsible for a lower force of unfolding. The *strong* family contains Ig domains 27, 28, 32, and 34. We make the assumption that the unfolding forces of these domains are conserved functions across species and so alignments were made containing sequences for these domains from up to ten different species (*H. sapiens*, *P. troglodytes*, *R. norvegicus*, *M. musculus*, *C. familiaris*, *T. nigroviridis*, *D. rerio*, *G. gallus*, *O. cuniculus*, *B. taurus*). Sequences were not available from all species for all 7 domains. Three domains had 8 sequences, three had 7 sequences, and one had 6. The *strong* family contained 28 sequences and the *weak* family contained 23. PCPMer was then employed to search for conserved property motifs and record global, conserved-property profiles for both of these alignments.

PCPMer detected a set of motifs within each family. In order to determine which of these motifs might be unique, they were then scored against the sequences of each family. PCPMer has a built-in scoring function that can report the top scoring windows for a motif against a given sequence. In order to describe this scoring function it is necessary to give more background on the PCPMer concept of a motif. A motif is defined by PCPMer as a contiguous sequence of at least a minimum length with

significant positions separated by no more than a defined maximum gap. A position is considered significant if at least one of the five physical chemical property vectors shows conservation above a threshold. This conservation is calculated by a relative entropy calculation which measures the similarity of one discrete distribution to another. In this case the distribution of the property vector values of amino acids in an alignment is compared to those that would be expected by the natural frequency of amino acids. Any position with a relative entropy score above a given threshold is considered significant. This motif data is recorded for each position as the distribution as a mean and standard deviation and the relative entropy for each of the five property vectors.

The scores, then, are calculated for each window by comparing each amino acid in the test sequence against the observed distribution in each significant property vector for that position in the motif. Only the vectors with relative entropy scores above the threshold are considered, so each significant position can contribute between one and five fitness scores to the overall score for the motif at that window. We first calculate a Z-value for each significant vector:

$$Z = \frac{E_{aa}^{v} - \mu^{v}}{w_{\sigma} + \sigma^{v} + \varepsilon}$$

Where E_{aa}^{v} is the value of amino acid *aa* in vector *v*, μ^{v} is the motif mean in vector *v*, w_{σ} is a standard deviation weighting factor, σ^{v} is the motif standard deviation in vector *v*, and ε is a small value to avoid division by zero. These Z-values are then scaled between 0 and 1 into a partial score as follows:

$$S_{p} = \frac{1}{1 + Z^{2}}$$

A length-normalized total score for the window is then calculated:

$$S = \frac{\sum_{i=1}^{N_p} S_{p,i}}{N_p}$$

Where N_p is the number of partial scores and $S_{p,i}$ is the *i*th partial score.

The top three scoring windows in each sequence were collected and aggregated by window position *i*. The scores at each window were then normalized by overall sum of scores for each motif. The score for a motif *m* matching at a given position *i* in a test set of sequences is given by:

$$S_i^m = \frac{\left(\sum_{j=1}^{N_i^m} s_j^m \right) N_i^m}{\sum_{k=1}^{N^m} s_k^m}$$

Where *m* designates a particular motif, *s* is the value of a recorded score, N^m is the number of recorded scores over all sequences in the test set for motif *m*, and N_i^m is the number of recorded scores at position *i* over all sequences in the test set for motif *m*.

Position Specific Comparison of Conserved Properties

The amino acid composition at each individual position of both families was in addition analyzed for differences in the distributions of the five quantitative descriptors E1- E5 [49]. We calculated the mean values and standard deviations for E1-E5 in each of the weak and strong family at each amino acid position, and compared the statistical

significance of the difference of the mean values by a t-test. We considered t-values above an absolute value of 3.55 as a significant difference, as they are in the 0.001 confidence interval. These residue positions were then further examined in the context of the sign of the t-value and the physical meaning of each vector. Preliminary analysis indicated that among the five descriptors the two descriptors E1 and E2 showed the most pronounced differences, i.e. the t-values indicating significant differences in the occurrence of descriptors are higher for E1 and E2. For example, the number of t-values larger than 5 is 10 in E3-E5 with values up to 7.7 as compared to 18 t-values in E1-E2 with values up to 30.1. We therefore present here only the results for descriptors E1 and E2. Positions that had a significant difference in one or both of the E1 and E2 vectors are highlighted in figures 5.2.4 and 5.2.7. The highlighting in figure 5.2.4 shows that a difference occurred in either vector while figure 5.2.7 illustrates the direction of these changes.

RESULTS

Human Cardiac Titin Ig Domains Show a Mechanical Hierarchy

Titin (also known as connectin) is the third major filament in the sarcomere (Figure 5.2.1A). The giant ~3 MDa molecule is composed primarily of immunoglobulin-like (Ig) and fibronectin type III domains (Fn3) in repetitive patterns which vary by sarcomere region. In the cardiac N2-B titin isoform, the I-band is composed primarily of Ig domains with several segments of unique sequence (Figure 5.2.1B). A representative Ig domain (pdb code 1TIT [179]) is shown in Figure 5.2.1C with the seven beta strands labeled. In recent years single molecule force spectroscopy data have revealed that the titin I-band has a complex mechanical structure [132,143,243].

Atomic force spectroscopy carried out on segments of the I-band show that domains in this region have a variety of unfolding forces ranging from ~50pN to ~300pN [132,143]. Several domains from this region have been studied in detail by constructing polyproteins comprised of a given domain repeated several times. In contrast to the extension recording of native segments of the I-band, an extension of a polyproteins composed of multiple repeats of the 27th I-band Ig domain results in a sawtooth pattern with force peaks which have similar force values (figure 5.2.2A).

The data collected from many such force-extensions produces a histogram of unfolding events centered around a mean unfolding force of 220 pN (figure 5.2.2B). We show this example to illustrate how the recordings are analyzed. However, all the data shown in figure 5.2.2C was taken from Li *et. al.* 2002, and 2003. Seven I-band Ig domains have been characterized in similar fashion [143,248] (Figure 5.2.2C) to reveal a mechanical hierarchy in the I-band in which domains proximal to the beginning of the

protein unfold at lower forces when compared to distal domains. This force hierarchy has also been observed in other proteins such as fibronectin, projectin and kettin [147,214,230].

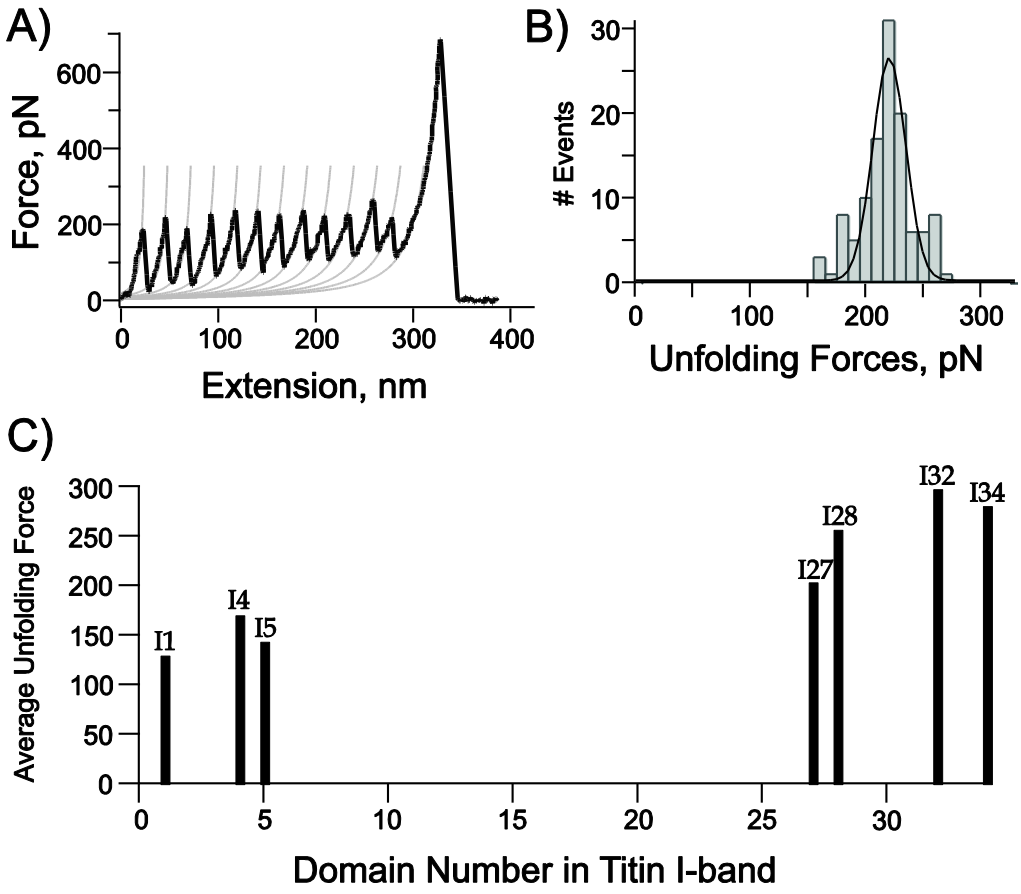

Figure 5.2.2 - Characterization of the mechanical stabilities of titin I-band domains.

A) A force-extension of a polyprotein consisting of multiple repeats of identical I27 Ig domains. The lines correspond to fits to the worm-like chain equation. B) Histogram of the measured unfolding forces for this polyprotein; the mean value is 220+/-20pN (n=115). C) Plot of the unfolding force versus the location of different titin Ig domains in the I-band of human cardiac N2-B. Seven I-band Ig domains have been characterized to reveal a mechanical hierarchy in the I-band in which the proximal domains unfold at significantly lower forces when compared to distal domains (Data from Li et. al. 2002 and 2003). The mean unfolding force for both groups are shown as red (strong family) and blue (weak family) lines.

Sequence-Diverse Domain Set

In addition to having widely varying unfolding forces, the approximately 40 Ig-like domains that make up the I-band of human cardiac titin have widely varying sequences with an the average sequence identity of only 25% possibly to prevent misfolding between neighboring domains [249]. Out of 780 possible pairs of sequences, only 17 pairs have a sequence identity higher than 40%, and only two pairs (I12, I13 and I33, I34) are sequential and may be in danger of co-aggregation [249,250]. The sequence variation between these different Ig domains is illustrated in figure 5.2.3. Immediately visible is a short region of parallel diagonal lines of high sequence identity indicative of possible gene duplication [250,251]. The dot plot indicates that domains I23-I26 are very similar to domains I28-I31, but also that domains I23, I27, and I31 are very similar to one another. The similarity of I23, I27, and I31 was first noted in a phylogenetic study of titin and related proteins [233]. This forms an interesting pattern of 9 domains in which the first, fifth, and ninth domain are 40-60% identical and the 2^{nd}-4^{th} domains are similar to the 6^{th}-8^{th} domains.

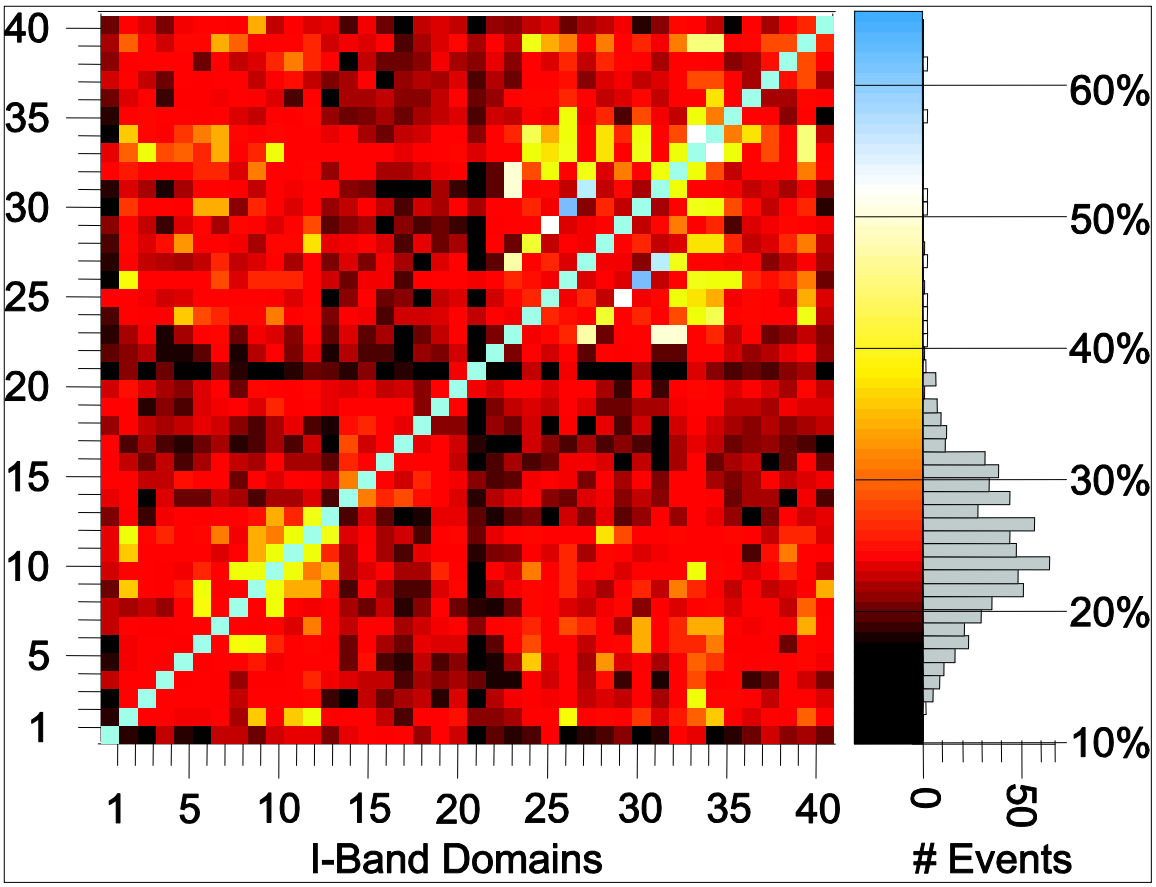

Figure 5.2.3 - Dot plot of the sequence identities of cardiac titin I-band Ig domains.

Sequence identities of 10% or less begin at black and proceed in a smooth gradient to bright red then yellow as the identities approach 40%. Identities of 40% and over are emphasized by a shift through white to blue at identities of 65%. The 100% identity diagonal squares are colored a pale cyan. At least three self-similar groupings become obvious from this plot. The groups consist of domains I2 to I13 (I), domains I14 to I20 (II), and domains I23 to I39 (III). Domains I1, I21, and I22 each seem quite dissimilar to other domains and each other. Group III contains a large number of high sequence identity pairs and may be an example of gene duplication.

Due to the high sequence diversity of Ig domains at different positions, it is not obvious that common sequence properties can be found within the I-band domains with similar mechanical stabilities, and what features are significantly different between the domains I1, I4 and I5 of the '*weak*' family (Fig. 5.2.4A) and the four domains I27, I28, I32, and I34 of the '*strong*' family (Fig. 5.2.4B). As is detailed in the methods section the two families are compared position by position using a student's t-test to determine statistically significant differences in conserved physical-chemical properties. A physical-chemical property is conserved if at least one of the five vectors E1 to E5 has similar values within each of the two families, and a position is differentially conserved if the average values of a property is significantly different in the weak and the strong family as compared to the standard deviations within each family. These positions are highlighted in Fig. 5.2.4. A position highlighted in cyan indicates a significant change in the average E1 value between families, while one highlighted in red indicates a change in E2. Positions highlighted in violet indicate significant changes in both properties. A change such as that in the violet-highlighted position 55 is clearly visible. In the strong family there is an absolutely conserved Gly, whereas in the weak family Gly, Lys, and Asp are present. Likewise a change in charge in position 24 where only Glu is present in the strong family, but Lys and Arg are present in the weak family. In other cases the alterations in property conservation are much more subtle. For example, position 42 contains primarily Leu and Ile residues in the *strong* family but primarily Val and Ile residues in the *weak* family.

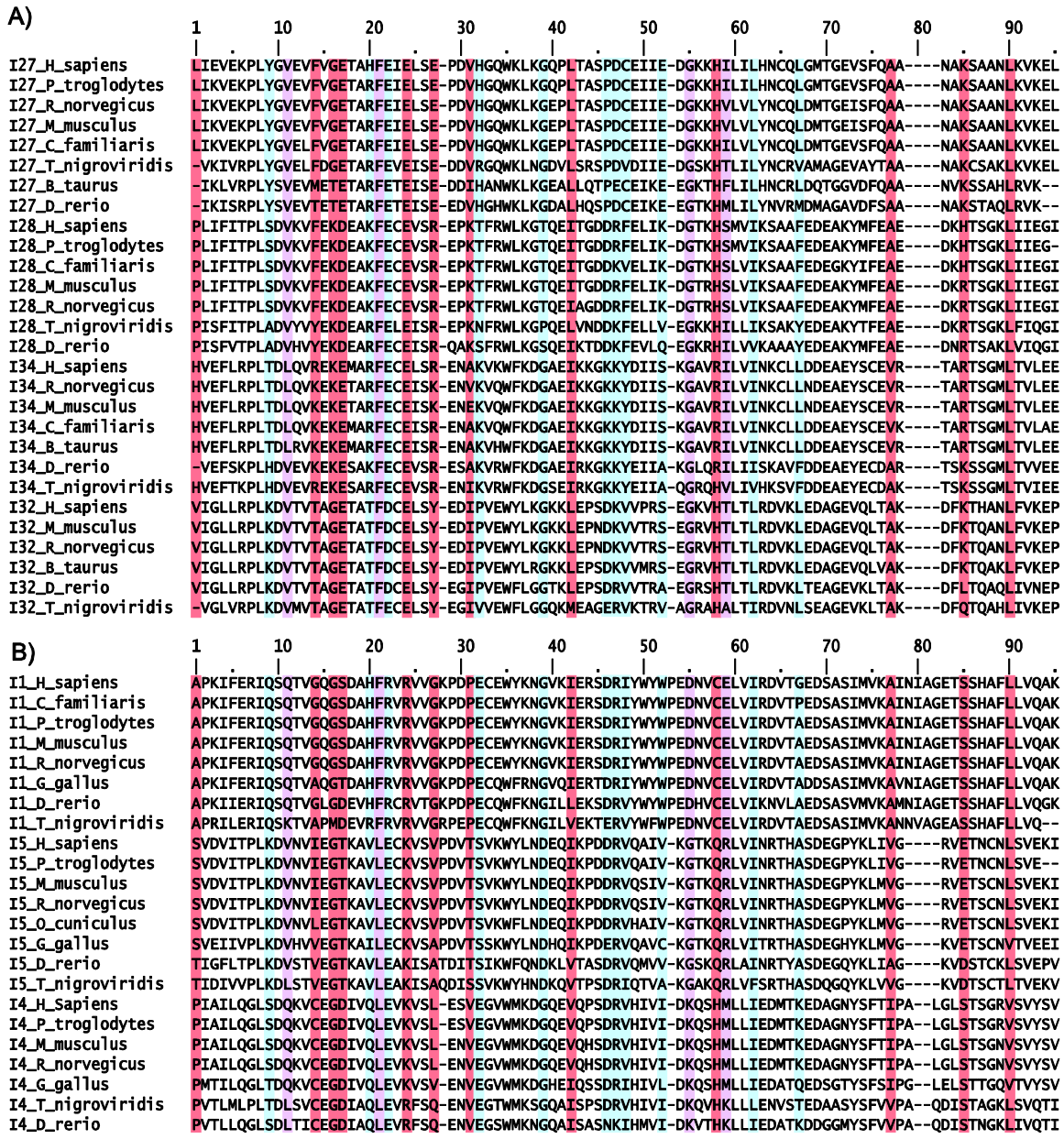

Figure 5.2.4 - Sequence alignments of seven I-band Ig domains for which the unfolding force is known.

The domains are grouped by rupture force into two families: A) a strong family consisting of I27, I28, I32, and I34 and B) a weak family consisting of I1, I4, and I5.

Differentially Conserved Motifs

The motifs detected by PCPMer are numerical representations of sequentially occurring residues with well conserved properties. Six motifs were detected in each family designated *w1-w6* for the *weak* family motifs and *s1-s6* for the *strong* family motifs. As mentioned in the methods section we also scored each set of motifs against the sequences in both families to distinguish between motifs that are common to both families and those that are unique for each family. As an analysis of these results Figure 5.2.5 illustrates the distribution of motifs (5.2.5a: *weak* motifs; 5b: *strong* motifs) as they match in the human sequences of both families. The motifs generated by each family match well in their home families. The matches across families begin to reveal which motifs are unique or common to both families. Motifs *w2*, *w4*, *w5*, *s3*, s5, and *s6* are all well conserved across families indicating that they may be common structural motifs. The other motifs *w1*, *w3*, *w6*, *s1*, *s2*, and *s4* had their best matches scattered around sequences of the opposing family indicating that some of these may be unique to their respective families, or that they are detecting common structural elements in alternate positions of the domain. Although this gives an initial indication of the uniqueness of motifs, this treatment is qualitative and a more extensive quantitative analysis is desirable. We therefore carried out a more detailed analysis of motif matches to sequences from all the species we had as described in the methods section.

Figure 5.2.5 - Best scoring window of motifs in each sequence

This figure illustrates the conservation of some motifs across families and the uniqueness of others. A) Motifs of the weak family are shown in unique colors (w1: red, w2: orange, w3: yellow, w4: green, w5: blue, w6: violet) highlighting their best matching window for each human sequence in the weak and strong families. B) Motifs of the strong family (s1: red, s2: orange, s3: yellow, s4: green, s5: blue, s6: violet) highlight their best matching window in each of the human sequences in both families. Here we can easily see the conservation of motifs w2, w4, w5 and s3, s5, and s6 across both families.

Figure 5.2.6 summarizes this quantitative analysis of how well motifs matched in all sequences and at the locations they did so (see Methods). As expected each motif scored well at the position in the sequence alignment at which it was originally defined. We will refer to this position as the 'native' position for a motif. Some motifs also had low yet not insignificant scores at locations in the sequence alignment other than their native position. Because these motifs show significant similarity to multiple locations in an alignment we refer to them as exhibiting 'degeneracy' in their matches. We kept the top three scores per sequence when scoring each motif so there is also a low level of background noise that is easily discernible from legitimate matches. We call a motif 'unique', if it has only high scores at its native position. Two motifs in the weak family were found to be unique, motif *w1* that corresponds structurally to the A′-B turn and most of the B strand, and motif *w3,* a short stretch of only 4 residues in the C-D loop and D strand. Both motifs are spatially well separated. No motif was detected that covered the corresponding area in the *strong* family. The motifs *w2*, *w4*, *w5*, and *w6* scored also reasonably highly in the *strong* family sequences and are thus considered common structural elements. Motifs *w4* and *w6* occur in strands G and E respectively and seemed to degenerately match other regions of similar secondary structure elsewhere in the sequence as well as their native positions.

Among the motifs derived from the strong family, *s1* had the lowest score by screening it against sequences from the weak family. This unique motif that has no corresponding motif in the weak family encompasses the A′ strand and the connecting region between A and A′ strands. The intermediately scoring motifs *s2*, *s4*, and *s6* all overlap at least in part with some motifs in the *weak* family, but still show slight differences in the conservation pattern. Motifs *s3* and *s5* show degeneracy in the locations they match and are thus considered common structural motifs among the weak

and strong family. Since motif s2 matched only marginally better than s1 and did not show degeneracy it is also likely an important region for determining the mechanical stability of the domain. Motif *s4* natively encompasses strand E along with turn and loop regions on either side. It matches its native location in the *weak* family, but also matches a region of similar structure starting just before the B strand. Motif *s6* shows degeneracy in both *strong* and *weak* families. In the *strong* family it additionally shows some similarity to the location of motif *s4*, and in the *weak* family it matches well at its native location as well as locations that motif *s4* matches.

In summary, most motifs detected were common to both families and were thus considered structural motifs that represent the common fold of all Ig domains. However, each family also contained two unique motifs that could not be matched to the sequences of the opposite strength. We consider these unique motifs most likely to be important in determining the mechanical stabilities of each domain. The degeneracy of certain motifs in recognizing general secondary structural elements certainly illustrates the ability of PCPMer to recognize structural motifs based on common physical/chemical properties. That unique motifs were detected between mechanically differentiated subsets of the Ig domain type is an exciting finding. Further, these unique motifs exist mostly in the A, A′, and B beta strands supporting their importance in resisting mechanical unfolding [99,192,241].

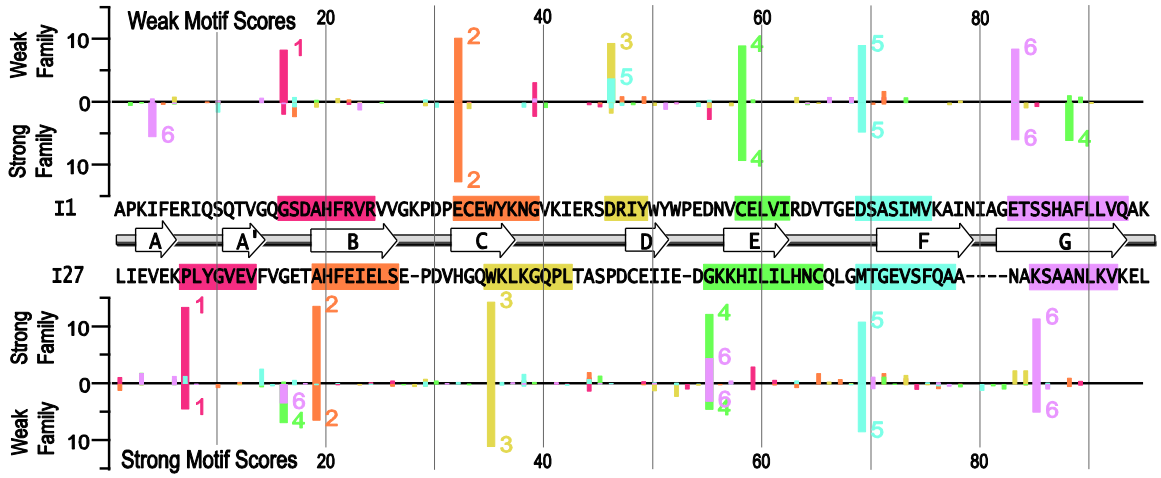

Figure 5.2.6 - Comparison of the scores for the six weak and strong motifs to find unique motifs.

In the top part all six weak motifs are matched against all sequences and scores are given above the zero line for sequences in the weak family (averaged as described in the Methods part) and below the zero line for sequences in the strong family. Each weak motif is highlighted in a different color in the representative human sequence of I1 and the colors of the score bars are characteristic for each motif. The middle part of the figure gives the secondary structure of the Ig domains with the representative human sequence of I1 for the weak family and I27 for the strong family. The lower part represents the scores of the six motifs derived from the strong family matched individually through the sequences of the strong family and weak family with scores given as in the top portion.

Residue by Residue Differences

Each residue position was compared using a t-test based method as described in the methods section. This approach resulted in many positions being flagged as having differently conserved properties. Vector E1 is most closely correlated with hydrophilicity scales, while E2 is most closely related to measures of sidechain size [49]. Table 5.2.1 lists residues with t-scores in at least one vector above a cutoff of 3.55 though both scores are given for completeness. There are 19 significant differences in the E1 vector and 21 differences in the E2 vector so neither seems to dominate the analysis. Of the 95 positions in the alignment 34 have significant differences in one or both vectors. Of these, 22 occur in 10 linear pairs or triplets. Out of those 10 groups, 5.2.7 have a significant difference in one vector common to the group. This could be pure coincidence, but would seem to indicate local selection for a given property.

The most prominent differences of the vectors E1 and E2 in the two families are mapped onto the 3D structure of the I27 domain. Figure 5.2.7a displays significant differences in hydrophilicity, and figure 5.2.7b shows differences in the side chain size. The color code denotes the change in properties by comparing sequences of the *strong* family relative to the *weak* family. For example a red residue in figure 5.2.6a indicates a position in which the property distribution has shifted towards hydrophilic values in sequences of the *strong* family as compared to sequences of the *weak* family. It is clear from figure 5.2.7a that no overall trend toward hydrophobic or hydrophilic side chains is present in either inward or outwardly facing residues. Figure 5.2.7b on the other hand shows a strong overall trend toward larger residues in the strong family. This could be due to increased van-der Waals interactions and other forces between side-chains that increase the overall strength of these domains or that the increased sidechain length in exterior surfaces helps to shield hydrogen bonds between β-strands and thus allow them

to persist longer. One interesting observation to note from figure 5.2.7 is that most of the differences cluster spatially and mostly in one face of the β-sandwich. This result seems to support the notion that several small changes work together to alter the overall mechanical properties of the domain.

Position	Residue	E1	E2	Position	Residue	E1	E2
1	A	**-3.67**	**-4.18**	32	E	**-4.13**	0.72
2	P	**-3.61**	-3.21	39	G	**-4.28**	1.73
4	I	2.25	**-3.91**	42	I	1.68	**-4.83**
5	F	**3.56**	-0.27	46	D	**-4.01**	0.75
9	Q	**-4.07**	1.53	47	R	**7.48**	**3.78**
11	Q	**-5.47**	**5.46**	48	I	**4.99**	-2.39
12	T	-0.44	**-3.78**	52	W	**14.44**	-0.89
14	G	0.26	**-5.87**	55	D	**-4.90**	**6.26**
16	G	1.03	**-5.16**	58	C	1.55	**-4.18**
17	S	2.34	**-10.03**	59	E	**-4.23**	**17.21**
20	H	**4.85**	-3.14	62	I	**4.60**	-1.70
21	F	**-6.56**	**-6.57**	67	G	**-15.59**	-0.86
22	R	**4.14**	**3.77**	72	S	2.32	**-8.10**
24	R	1.25	**30.13**	77	A	**3.64**	2.41
25	V	-1.17	**-3.94**	78	I	2.16	**-8.16**
27	G	2.83	**-8.83**	85	S	-1.83	**-7.72**
31	P	-1.88	**-5.22**	91	L	-1.07	**-4.32**

Table 5.2.1 - Positions with significant t-values

This shows the positions in which significant differences were detected in conserved properties between *strong* and *weak* families. The 'position' column indicates residue position in the alignment, while the 'residue' column gives the residue at that position in I1 as a reference guide. The E1 and E2 columns give the R-values in the E1 and E2 vectors respectively with values above the 3.55 cutoff for a 0.001 confidence interval in bold.

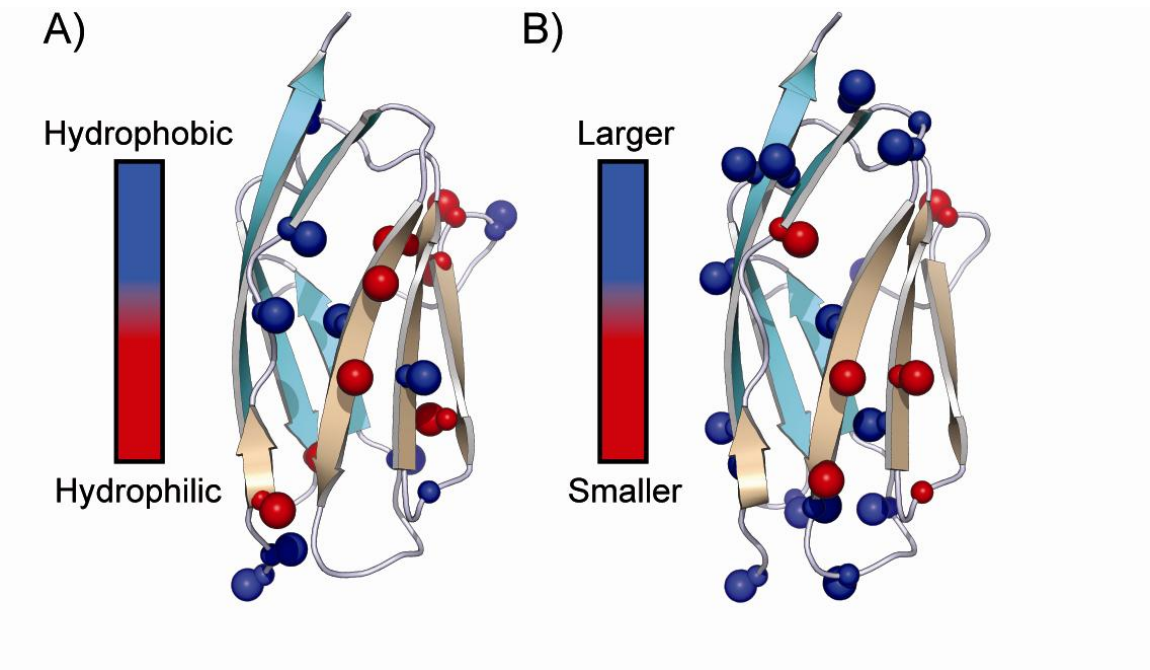

Figure 5.2.7 - Positions found to be significantly different between the weak and strong families.

Residues with significant property shifts from the weak to strong families are displayed on the structure of I27 (pdb code 1TIT [179]) with a small Cα sphere and large Cβ sphere. A) The first vector E1 correlates well with descriptors of hydrophilicity/hydrophobicity. Red colored residues indicate a shift from hydrophobic values toward hydrophilic in the strong family, while blue indicate the reverse shift. There are more hydrophobic shifts with many of these in loops or extended regions of the domain. B) The second vector E2 describes the size of the side-chains. Red colored residues indicate a shift from larger side-chain values toward smaller ones in the strong family, while blue indicate the reverse shift. More residues increase in the size of side-chains when we compare sequences from the strong family versus the weak family.

DISCUSSION

Titin Ig domains of different species but at identical positions show a remarkable conservation across vertebrates (Fig. 5.2.4). This point is further confirmed with a BLAST search against the UniProt database using human Ig sequences. The BLAST expectation values for the best matching titin domains of mammals are in the range of 1e-50 to 1e-40 as compared to a human reference domain, approximately 1e-38 to 1e-30 for the chicken species, and 1e-29 to 1e-21 for zebrafish and puffer fish. Other titin-like proteins from invertebrates and organisms like the ascidians also match but show much less conservation with values beginning in the range of 1e-8. The high degree of conservation in vertebrate species was in contrast to the high degree of sequence variability between neighboring domains in any given titin sequence which shared an average of only 25% sequence identity. The expectation values rapidly drop off when comparing one domain to the others in the same titin sequence rapidly reaching the 1e-8 to 1e-3 range.

Titin plays a mechanical as well as signaling role [161,211]. The almost absolute conservation of these domains implies a strong conservation of both signaling and mechanical functions. We predict the mechanical hierarchy found in the I-band domains is one function maintained over evolutionary time. This hierarchy is also found in the differentially expressed domains of titin [252]. In this work the authors found that the constitutive Ig domains I91-I98 unfold at higher forces than the differentially expressed I65-I70. One interesting feature of these AFM recordings is that there is a rising force pattern, indicating that both segments have a mixture of 'weak' and 'strong' domains. However at present we cannot include these data in our analysis because the characteristic unfolding forces for the individual domains are not known. Mechanical

unfolding hierarchies are also found in fibronectin [147] and the titin-like proteins projectin and kettin from *Drosophila* flight muscle [214].

The hierarchical unfolding of titin Ig domains may have the result of considerably decreasing the affinity of titin for its ligands. At the same time, the application of a mechanical force may trigger the exposure of new binding sites (or cryptic signaling sites) that were buried between domains or in the fold [147,161]. Hence, Ig domain unfolding in titin may modulate its resting length and elasticity, and also its ligand-binding properties.

The hierarchy of mechanical stability is mirrored in well-defined sequence positions that are specific for the weak and strong family (Fig. 5.2.6) and common sequence features that might be responsible for the common Ig fold. The C, D, E, and F β-strands form a greek key structural motif around the conserved tryptophan in the C strand. The regions of the protein around the E and F strands are well conserved as shown by the overlapping and common motifs there. On the other hand, the weak motifs *w1* and *w3* are clearly specific to the weak family sequences and *w3* is found partially in the D strand and incorporates residues largely distant from the conserved tryptophan core. Interestingly, no large differences were found in the motifs of the G strand. It seems that the majority of the larger differences are found in the A, A′, and B strands.

Further we found several differentially conserved properties at residue positions scattered over the domain. These differences seemed mostly to agree with the results of the motif studies, though several subtle point differences were detected in regions where large scale differences were not detected such as the G strand and the otherwise more conserved greek-key strands.

Based on our detailed sequence analysis we propose a model where side chains, in addition to backbone hydrogen bonds, play an important role in the unique mechanical

stability of titin domains. This notion has already been suggested to explain the difference in mechanical stability between titin domains I27 and I28 [241,244]. We suggest that long polar side chains work to protect key hydrogen bonds between beta-strands which oppose the sliding of these two β-sheets during mechanical unfolding. We also predict that larger solvent exposed side-chains may contribute to the mechanical stability by forming salt-bridges or other types of interactions; these side-chain interactions may affect the elasticity and robustness of the β-sheets along the directions stressed during mechanical unfolding. We are currently testing these predictions experimentally by mutagenesis and protein engineering.

The study of the mechanical stability of titin Ig domains has been rich with fascinating findings which continually expand our knowledge and allow us to delve into ever more detail. Steered molecular dynamics simulations of the stretching of Ig domain I27 identified a patch of backbone hydrogen bonds between the A'G to play a key in its mechanical stability [99,192]. Proline mutagenesis experiments aimed at breaking these key bonds showed that there are other important interactions [241]. Recently Sharma et al [253] shuffled large segments between I27 and I32 in an attempt generate hybrid domains that would share mechanical properties of their parent domains. They found that the A′-G patch may not be the only structural region responsible for the mechanical stability of titin Ig domains. In addition phi-value analysis of the mechanical unfolding of the I27 domain found that the key event is not a simple case of loss of hydrogen bonding interactions between the A′ and G-strands alone [254]. Consistent with these findings we found other regions in titin Ig domains which may contribute to determining its mechanical stability.

CONCLUSION

We have grouped several titin Ig domains by the forces required to unfold them into a strong and a weak family and used computational tools to compare conserved properties between the two groups. We found differences in amino acid property conservation between those groups. This approach was designed to further examine the contributing factors to the mechanical properties of these domains. We found several striking differences between the *weak* and *strong* families. Each family contained 2 motifs unique to that family, and several residue positions with differentially conserved properties scattered across the domain structure.

This novel application of PCPMer may prove to be a powerful tool to compare subsets of related protein families with different mechanical stability or other different key functions. By carefully constructing the input alignments to contain individual sequences with differences primarily in one key function we have found differentially conserved properties related to that key function. We envision that this knowledge should prove useful to fine-tune the mechanical properties of titin Ig domains or other mechanically important domains which function as mechanical force sensors.

CHAPTER 6 – EXPERIMENTAL TEST OF PREDICTIONS FROM SEQUENCE ANALYSIS FOR THE MECHANICAL STABILITY OF TITIN IG DOMAINS

INTRODUCTION

Previously we have used computational methods to detect differentially conserved properties in groups of titin domains with different average unfolding strengths (chapter 5 section 2). We found several residue positions in the titin Ig fold in which there were significant differences in the physical-chemical properties present between our mechanically weak and strong groups. Based on those results we have made several mutations to the I1 and I27 domains in order to alter their mechanical strengths.

Prior studies have shown this sort of manipulation to be highly complex and difficult [255]. The earliest such published work sought to shuffle sections of the I27 and I32 domains to alter the mechanical properties of the host domain, and probe the significance of different sections of the structure. The results were mixed and unexpected but they began to describe the importance of certain areas of the structure and how adapted to the specific domain they were. A second published study by the same group expanded the project by shuffling even more sections between the two proteins. I27 and I32 share a 41.6% sequence identity and are in the same subclass of titin Ig domains, and yet many of the shuffling attempts resulted in weakend and thermodynamically unstable domains. Encouragingly, however, some mutants did behave as predicted and most either supported or did not contradict the results of our previous motif analysis. These results indicate a very high degree of specificity of design present in the titin Ig domains.

Our own efforts to manipulate the mechanical properties of I1 and I27 were directed by the results of our computational study of the titin Ig domains. We first created several mutants of I27 via site directed mutagenesis (V11Y, E12N) in an I27 template designed for disulfide polymerization. The goal of these early efforts was to

build up point mutants to insert groups of mutations that reflected motifs or groups of differentially conserved positions. This incremental mutation strategy was ultimately abandoned due to the large amount of work required to introduce mutations and check for the desired changes at each step. The proteins developed by these early efforts were also less stable, they tended to aggregate over time. The reduced stability made them very difficult to work with when attempting cysteine polymerization.

After trying several modifications to the basic cysteine polymerization technique we concluded that it is best suited for stable, robust proteins and ultimately abandoned it as a strategy for the novel mutants with unknown thermodynamic stabilities. We also abandoned the iterative mutagenesis approach in favor of synthetic DNA services. The turnaround from the point of designing and ordering a gene to delivery was about a week compared to several weeks of lab work and several sequencing rounds.

We ordered the wild type I1 and a mutant form of I1 (XF2) which was designed to strengthen the domain both of which were designed for insertion into the pAFM cassette. The XF2 set of mutations focused on the A' and G strands and the EF loop, as well as the very beginning of the A strand. Explicitly the mutations we made are: Q11V, G14T, G67L, F89N, L91F, Q93K, and A94E. Our results indicate that we increased the unfolding force of I1 XF2 by ~30-50 pN.

MATERIALS AND METHODS

Cysteine polymerization

This strategy involves expressing a monomer Ig domain with N and C terminal cysteine residues in high quantities and allowing or inducing polymerization by disulfide bond formation between monomers. The desired outcome of this procedure is the formation of large homogenous poly-proteins. Proteins were expressed in BL21 cells in

1-2 L volumes of LB broth with an appropriate antibiotic agent. Cell lysis was carried out in a cold room using a french press. Proteins contained an N-terminal 6xHis tag and were purified by nickel affinity using either gravity or HPLC. A concentration step was carried out via spin concentration tubes with a pore size smaller than 10 nm to achieve the final desired concentration of 2-5 mg/mL for the polymerization reactions to occur. At this point several methods for the polymerization step were tried.

Warm Incubation – The protein sample is incubated at 37 C for 48-72 hours. This is only an option for stable proteins such as wild type I27.

Glutathione induction – An experimental approach that involved incubation with a buffer containing 1xPBS, 1mM GSH, 0.2mM GSSG. No significant increase in polyprotein formation was observed.

Stepwise DTDP polymerization – Monomers are first reduced by incubating 200-500 μL of protein with 10mM DTT for at least 30 minutes at room temperature. Then the DTT is removed by applying the sample to a NAP5 column and eluting to a final volume of 1mL in phosphate buffered saline at pH 7.4. A 3.8mM stock of 2,2-dithiodipyridine (DTDP) in dimethylformamide (DMF) was prepared and then used to achieve a 2:1 molar ratio of DTDP to inactive protein with 2 reduced cysteines. The reaction takes about 5 minutes at room temperature and can be followed by measuring absorbance at 343nm over time. Next the active and inactive protein is mixed in an alternating, stepwise manner. First 1 part active to 2 parts inactive protein are mixed and allowed to react for 30 minutes at room temperature, then 2 parts active are added and the mixture is incubated again for 30 minutes, followed by 2 parts inactive and so on.

pAFM cassette

This is a second strategy for producing polyproteins for our experiments. The pAFM cassette is a synthetic polyprotein gene containing 8 I27 domains with unique restriction enzyme sites between them. Since longer constructs had sometimes been problematic in the past we designed inserts to replace the central 4 I27 domains with a single test domain so that the final construct would have two I27 domains on each side of the experimental domain. This arrangement has the advantage of introducing a control into each poly-protein.

Ig mutants

Mutants created in the lab were carried out using the Qiagen quickchange site directed mutagenesis kit.

Synthetic DNA was ordered from DNA 2.0.

Single Molecule Force Spectroscopy

This technique is described in chapter 5 section 2.

RESULTS

Of the several attempts to produce mutants in I27 only one reached the stage of being manipulated on the AFM. In general mutations to protein structures are dangerous in terms of causing undesired destabilization. These mutations were designed to lower the mechanical stability of I27 and so were targeting areas that were hoped to be important to the structure. Unfortunately all of the mutations adversely affected the overall stability of protein fold causing the domain to become less soluble and complicating or blocking attempts to use cysteine disulfide polymerization. The V11Y mutant was one which, though also less stable, was successfully polymerized and analyzed with the AFM. We were unable to record enough data for this mutant to produce a histogram but an example single molecule force spectroscopy trace for this mutant is superimposed on one of wild type I27 showing that the mechanical stability was not significantly altered by the V11Y mutation (fig 6.1).

The I1 mutant XF2 was expressed in the pAFM construct such that a single XF2 domain was flanked by two wild type I27 domains on either side (figure 6.2 B). By measuring the force peaks of dozens of single molecule force spectroscopy studies of XF2 we are able to construct a histogram of unfolding peak forces (fig 6.2 D). The gray histogram is comprised of the unfolding peaks of I27 poly-proteins and is superimposed on the red histogram which is comprised of the peaks from the XF2 pAFM construct. I27 has an unfolding force of about 200 pN and is clearly the main peak visible in the red histogram. There is a smaller peak visible at around 150 pN which is due to the unfolding of the XF2 domain. The wild type unfolding force of the I1 domain was reported as 127 pN in Li and Fernandez 2003 from which we borrow figure 6.2 E.

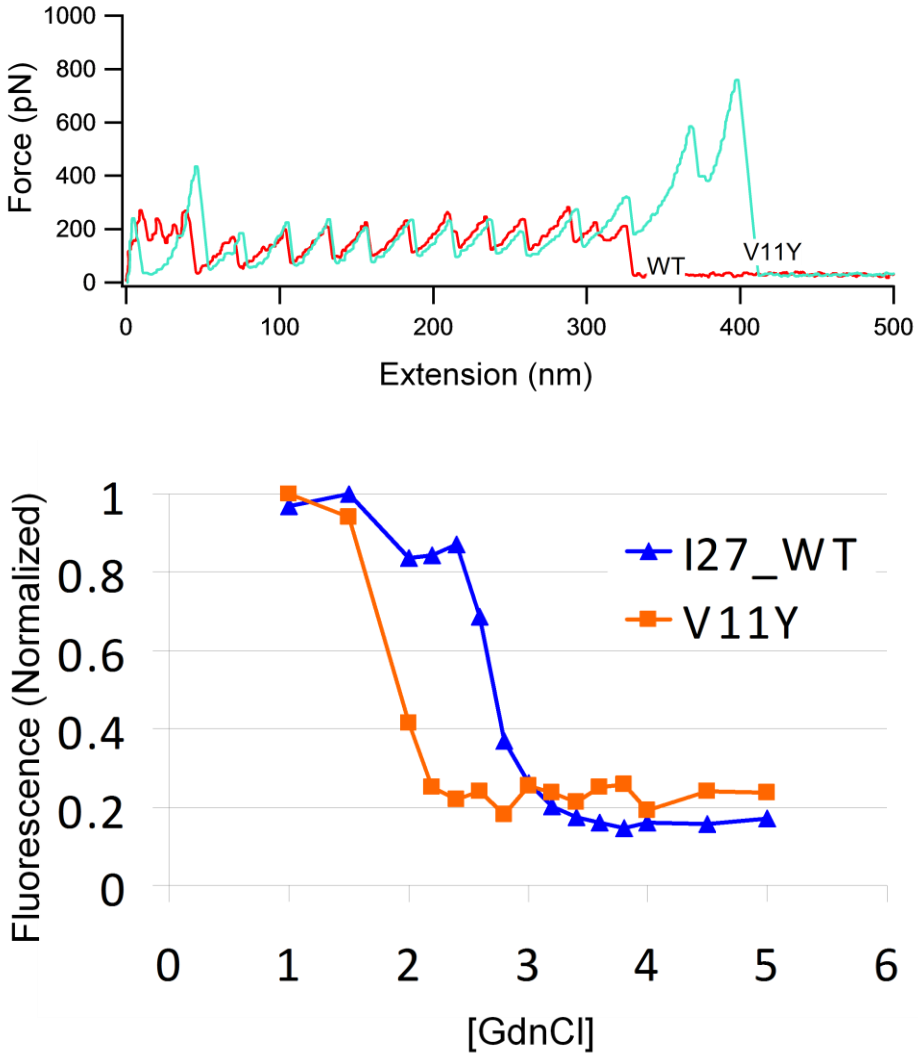

Figure 6.1 - I27 V11Y mechanical and chemical stabilities

The I27 mutant V11Y is less chemically stable than the wild type but has approximately the same mechanical stability. Top) A single molecule force spectroscopy trace for polyproteins of wild type I27 and the V11Y mutant are superimposed. Bottom) Tryptophan fluorescence of wild type and mutant I27 as a function of guanadinium chloride concentration.

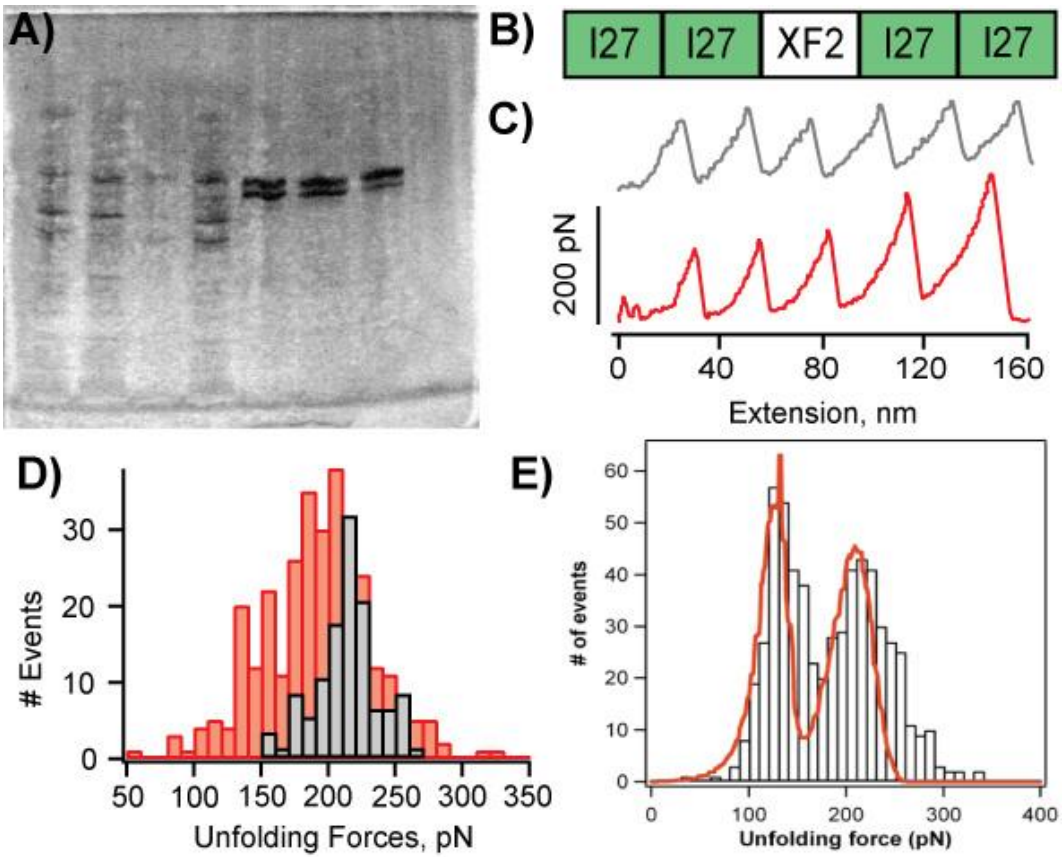

Figure 6.2 - Altered mechanical properties of I1 XF2

The I1 mutant XF2 possesses novel mechanical properties. A) A gel showing in lanes 1-4 pre-induction whole cell lysates and in lanes 5-8 purified protein for: I1 wild type, I1 XF2, I27w1, and I27w3 experimental domains in the pAFM construct. B) The final configuration of I1 XF2 in pAFM in which XF2 is flanked on both ends by two wild type I27 domains. C) Single molecule force spectroscopy recordings of a purely I27 polyprotein (grey) and XF2 in pAFM (red). D) A histogram of I27 unfolding events (grey) overlapping a histogram of XF2 in pAFM unfolding events (red). E) A histogram of unfolding events from a $(I1\text{-}I27)_4$ construct from REF. The unfolding force of I1 is 127pN as published in Li and Fernandez 2003 whereas the unfolding force of XF2 is 150pN.

DISCUSSION

The result that the I27 V11Y point mutant did not affect the unfolding force of the domain indicates that the mutation did not significantly disrupt any interactions responsible for the mechanical stability. Prior studies have shown that mutations at this position can affect the unfolding force of the domain [241,256] which implies that our selection of tyrosine maintained the mechanical integrity of the domain. While the mechanical integrity of the domain was maintained by this mutation, the chemical stability was reduced. One possible explanation is that we did not explicitly consider the effect of these mutations on chemical stability in our computational methodology. We relied instead on the profiles of our weak and strong protein families to maintain sufficient information to preserve the chemical stability of the domain. Another possible explanation is that the residue at this position is required to interact with surrounding residues in a manner that was not possible with the native residues and that in fact it clashed with the native residues. In other words, several mutations are required to work in groups to alter the mechanical stability of the domain and not adversely affect the chemical stability. It is possible that the single mutation clashed with the local structure destabilizing the domain.

The I1 XF2 mutant contains seven point mutations as listed in the introduction and was an attempt to test the hypothesis that groups of point mutants in key locations were responsible for the mechanical stability of the domain. As compared with previous efforts, this was a surgical attempt to affect only properties of the domain in key locations which would affect the mechanical properties in a predictable way while leaving the structure of the domain in tact. The increase in unfolding force is an exciting result supporting our hypothesis and computational approach. The specific mutations selected were based on a larger set of differentially conserved residue positions. The selection of

the subset is something that should still be refined. It includes some of the larger differences and more intellectually interesting mutations but was otherwise a matter of educated guesswork based on structural hints. For example the structure of I1 includes several salt bridges and other side chain interactions that bend the ABED β-sheet outward with respect to the structure of I27. These sorts of interactions were avoided as much as possible. We also focused on the A-G region that has long been thought to be the structure responsible for the opposition of mechanical forces in titin Ig domains.

CONCLUSIONS

Previous efforts to manipulate the mechanical properties of titin Ig domains have cast broad nets, making mutations and substitutions in a largely exploratory nature. The goal of this work has been to better understand the mechanical design of titin I-band domains and intelligently manipulate the mechanical stabilities of those domains. By building upon previously existing computational techniques and applying them in novel ways we discovered a set of residue positions in the titin I-band Ig domains with differentially conserved physical-chemical properties between 'weak' and 'strong' domains. Based on these differences we produced several mutations to I1 and I27 domains. While the point mutant to I27 did not affect the mechanical strength, the more ambitious XF2 mutant was successful in increasing the mechanical strength of the I1 domain. These results are very encouraging and may lead to further understanding of the mechanical design of titin Ig domains and the mechanical tuning therof.

CHAPTER 7 – SUMMARY AND FUTURE DIRECTIONS

We set out in this project to probe the mechanical design of titin Ig domains and to expand computational methods for probing the mechanical functions of proteins. In the process we explored the properties of several mechanical proteins including bacterial pili, projectin, kettin, TTN-1, twitchin, and titin. We provided experimental evidence that supports the hypotheis that the kinase domains in twitchin and TTN-1 are activated by mechanical force. We demonstrated the ability of pili to easily unwind under moderate forces over large distances, absorbing the energy of a tensile force, and thus allowing the bacterium to maintain an attachment.

We developed and applied computational techniques in new ways to analyze the mechanical design of titin Ig domains. PCPMer and other similar motif analysis tools have classically been applied to characterize protein families and detect proteins related by functional motifs. We used it to characterize protein families which we defined by the function of mechanical strength as opposed to sequence similarity. We then processed the raw data in a novel manner to find key differences in the sequences. This technique proved fruitful in this application and we expect that it can be successfully applied to answer many more questions in biology.

Our results regarding the I1 mutant XF2 are very encouraging, though further experimentation would be required to better disentangle the interplay of factors governing the mechanical stability of these domains. It would be desireable to make smaller subsets of the XF2 mutations in order to find which are the most important. Further study of the chemical stability of XF2 would also be desireable. Further expansion of the original data set upon which we based our computational work would also be a useful step which could reduce the amount of background noise and better target important residues. Finally an application of this technique to a completely

different set of mechanical proteins would begin to elucidate any commonalities of design for mechanical proteins.

REFERENCES

1. WHO. in *Codex ad hoc intergovernmental task force on foods derived from biotechnology* (World Health Organization, Yokohama, 2003).
2. Aalberse, R. C. Assessment of allergen cross-reactivity. *Clin Mol Allergy* **5**, 2 (2007).
3. Breiteneder, H. & Mills, C. Structural bioinformatic approaches to understand cross-reactivity. *Mol Nutr Food Res* **50**, 628-632 (2006).
4. Schein, C. H., Ivanciuc, O. & Braun, W. Bioinformatics approaches to classifying allergens and predicting cross-reactivity. *Immunol Allergy Clin North Am* **27**, 1-27 (2007).
5. Aalberse, R. C. & Stadler, B. M. In silico predictability of allergenicity: from amino acid sequence via 3-D structure to allergenicity. *Mol Nutr Food Res* **50**, 625-627 (2006).
6. Bonds, R. S., Midoro-Horiuti, T. & Goldblum, R. A structural basis for food allergy: the role of cross-reactivity. *Curr Opin Allergy Clin Immunol* **8**, 82-86 (2008).
7. Chapman, M. D., Pomes, A., Breiteneder, H. & Ferreira, F. Nomenclature and structural biology of allergens. *J Allergy Clin Immunol* **119**, 414-420 (2007).
8. Mirza, O. et al. Dominant epitopes and allergic cross-reactivity: complex formation between a Fab fragment of a monoclonal murine IgG antibody and the major allergen from birch pollen Bet v 1. *J Immunol* **165**, 331-338 (2000).

9. Jenkins, J. A., Griffiths-Jones, S., Shewry, P. R., Breiteneder, H. & Mills, E. N. Structural relatedness of plant food allergens with specific reference to cross-reactive allergens: an in silico analysis. *J Allergy Clin Immunol* **115**, 163-170 (2005).

10. Schwietz, L. A., Goetz, D. W., Whisman, B. A. & Reid, M. J. Cross-reactivity among conifer pollens. *Ann Allergy Asthma Immunol* **84**, 87-93 (2000).

11. Aalberse, R. C., Akkerdaas, J. & van Ree, R. Cross-reactivity of IgE antibodies to allergens. *Allergy* **56**, 478-490 (2001).

12. Brusic, V. & Petrovsky, N. Bioinformatics for characterisation of allergens, allergenicity and allergic crossreactivity. *Trends Immunol* **24**, 225-228 (2003).

13. Thomas, K. et al. In silico methods for evaluating human allergenicity to novel proteins: International Bioinformatics Workshop Meeting Report, 23-24 February 2005. *Toxicol Sci* **88**, 307-310 (2005).

14. Hileman, R. E. et al. Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *Int Arch Allergy Immunol* **128**, 280-291 (2002).

15. Zorzet, A., Gustafsson, M. & Hammerling, U. Prediction of food protein allergenicity: a bioinformatic learning systems approach. *In Silico Biol* **2**, 525-534 (2002).

16. Furmonaviciene, R. et al. An attempt to define allergen-specific molecular surface features: a bioinformatic approach. *Bioinformatics* **21**, 4201-4204 (2005).

17. Ivanciuc, O., Schein, C. H. & Braun, W. Data mining of sequences and 3D structures of allergenic proteins. *Bioinformatics* **18**, 1358-1364 (2002).

18. Ivanciuc, O., Schein, C. H. & Braun, W. SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res* **31**, 359-362 (2003).

19. Pearson, W. R. Using the FASTA program to search protein and DNA sequence databases. *Methods Mol Biol* **25**, 365-389 (1994).

20. Schein, C. H., Ivanciuc, O. & Braun, W. *Structural Database of Allergenic Proteins (SDAP)* (eds. J., M. S., W., B. A. & M., H. R.) (ASM Press, Washington DC, 2006).

21. Egger, M. et al. Pollen-food syndromes associated with weed pollinosis: an update from the molecular point of view. *Allergy* **61**, 461-476 (2006).

22. Mittag, D. et al. Birch pollen-related food allergy to legumes: identification and characterization of the Bet v 1 homologue in mungbean (Vigna radiata), Vig r 1. *Clin Exp Allergy* **35**, 1049-1055 (2005).

23. Mittag, D. et al. A novel approach for investigation of specific and cross-reactive IgE epitopes on Bet v 1 and homologous food allergens in individual patients. *Mol Immunol* **43**, 268-278 (2006).

24. Midoro-Horiuti, T. et al. Molecular cloning of the mountain cedar (Juniperus ashei) pollen major allergen, Jun a 1. *J Allergy Clin Immunol* **104**, 613-617 (1999).

25. Midoro-Horiuti, T. et al. Major linear IgE epitopes of mountain cedar pollen allergen Jun a 1 map to the pectate lyase catalytic site. *Mol Immunol* **40**, 555-562 (2003).

26. Midoro-Horiuti, T., Brooks, E. G. & Goldblum, R. M. Pathogenesis-related proteins of plants as allergens. *Ann Allergy Asthma Immunol* **87**, 261-271 (2001).

27. Breiteneder, H. & Mills, C. Nonspecific lipid-transfer proteins in plant foods and pollens: an important allergen class. *Curr Opin Allergy Clin Immunol* **5**, 275-279 (2005).

28. Weber, R. W. Cross-reactivity of pollen allergens: recommendations for immunotherapy vaccines. *Curr Opin Allergy Clin Immunol* **5**, 563-569 (2005).

29. Hemmer, W., Focke, M., Gotz, M. & Jarisch, R. Sensitization to Ficus benjamina: relationship to natural rubber latex allergy and identification of foods implicated in the Ficus-fruit syndrome. *Clin Exp Allergy* **34**, 1251-1258 (2004).

30. Ayuso, R., Reese, G., Leong-Kee, S., Plante, M. & Lehrer, S. B. Molecular basis of arthropod cross-reactivity: IgE-binding cross-reactive epitopes of shrimp, house dust mite and cockroach tropomyosins. *Int Arch Allergy Immunol* **129**, 38-48 (2002).

31. Reese, G. et al. Structural, immunological and functional properties of natural recombinant Pen a 1, the major allergen of Brown Shrimp, Penaeus aztecus. *Clin Exp Allergy* **36**, 517-524 (2006).

32. Goodman, R. E. Practical and predictive bioinformatics methods for the identification of potentially cross-reactive protein matches. *Mol Nutr Food Res* **50**, 655-660 (2006).

33. Ladics, G. S., Bardina, L., Cressman, R. F., Mattsson, J. L. & Sampson, H. A. Lack of cross-reactivity between the Bacillus thuringiensis derived protein Cry1F

in maize grain and dust mite Der p7 protein with human sera positive for Der p7-IgE. *Regul Toxicol Pharmacol* **44**, 136-143 (2006).

34. Ivanciuc, O. et al. Using property based sequence motifs and 3D modeling to determine structure and functional regions of proteins. *Curr Med Chem* **11**, 583-593 (2004).

35. Mathura, V. S., Schein, C. H. & Braun, W. Identifying property based sequence motifs in protein families and superfamilies: application to DNase-1 related endonucleases. *Bioinformatics* **19**, 1381-1390 (2003).

36. Breiteneder, H. & Radauer, C. A classification of plant food allergens. *J Allergy Clin Immunol* **113**, 821-830; quiz 831 (2004).

37. Radauer, C. & Breiteneder, H. Pollen allergens are restricted to few protein families and show distinct patterns of species distribution. *J Allergy Clin Immunol* **117**, 141-147 (2006).

38. Finn, R. D. et al. Pfam: clans, web tools and services. *Nucleic Acids Research* **34**, D247-D251 (2006).

39. Radauer, C., Bublin, M., Wagner, S., Mari, A. & Breiteneder, H. Allergens are distributed into few protein families and possess a restricted number of biochemical functions. *J Allergy Clin Immunol* **121**, 847-852 e7 (2008).

40. Schein, C. H., Ivanciuc, O. & Braun, W. Common physical-chemical properties correlate with similar structure of the IgE epitopes of peanut allergens. *J Agric Food Chem* **53**, 8752-8759 (2005).

41. Stadler, M. B. & Stadler, B. M. Allergenicity prediction by protein sequence. *Faseb J* **17**, 1141-1143 (2003).

42. Bjorklund, A. K., Soeria-Atmadja, D., Zorzet, A., Hammerling, U. & Gustafsson, M. G. Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins. *Bioinformatics* **21**, 39-50 (2005).

43. Riaz, T., Hor, H. L., Krishnan, A., Tang, F. & Li, K. B. WebAllergen: a web server for predicting allergenic proteins. *Bioinformatics* **21**, 2570-2571 (2005).

44. Schein, C. H., Zhou, B., Oezguen, N., Mathura, V. S. & Braun, W. Molego-based definition of the architecture and specificity of metal-binding sites. *Proteins* **58**, 200-210 (2005).

45. Schein, C. H., Zhou, B. & Braun, W. Stereophysicochemical variability plots highlight conserved antigenic areas in Flaviviruses. *Virol J* **2**, 40 (2005).

46. Marti, P. et al. Allergen motifs and the prediction of allergenicity. *Immunol Lett* **109**, 47-55 (2007).

47. Saha, S. & Raghava, G. P. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res* **34**, W202-209 (2006).

48. Oezguen, N. et al. Comprehensive 3D-modeling of allergenic proteins and amino acid composition of potential conformational IgE epitopes. *Mol Immunol* **45**, 3740-3747 (2008).

49. Venkatarajan, M. S. & Braun, W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *Journal of Molecular Modeling* **7**, 445-453 (2001).

50. Schein, C. H., Ozgun, N., Izumi, T. & Braun, W. Total sequence decomposition distinguishes functional modules, "molegos" in apurinic/apyrimidinic endonucleases. *BMC Bioinformatics* **3**, 37 (2002).

51. Chenna, R. et al. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31**, 3497-3500 (2003).

52. Salcedo, G., Sanchez-Monge, R., Diaz-Perales, A., Garcia-Casado, G. & Barber, D. Plant non-specific lipid transfer proteins as food and pollen allergens. *Clin Exp Allergy* **34**, 1336-1341 (2004).

53. Stanley, J. S. et al. Identification and mutational analysis of the immunodominant IgE binding epitopes of the major peanut allergen Ara h 2. *Arch Biochem Biophys* **342**, 244-253 (1997).

54. Robotham, J. M., Teuber, S. S., Sathe, S. K. & Roux, K. H. Linear IgE epitope mapping of the English walnut (*Juglans regia*) major food allergen, Jug r 1. *Journal of Allergy and Clinical Immunology* **109**, 143-149 (2002).

55. Asturias, J. A., Gomez-Bayon, N., Eseverri, J. L. & Martinez, A. Par j 1 and Par j 2, the major allergens from Parietaria judaica pollen, have similar immunoglobulin E epitopes. *Clin Exp Allergy* **33**, 518-524 (2003).

56. Glaspole, I. N., de Leon, M. P., Rolland, J. M. & O'Hehir, R. E. Characterization of the T-cell epitopes of a major peanut allergen, Ara h 2. *Allergy* **60**, 35-40 (2005).

57. Bohle, B. et al. Characterization of the human T cell response to antigen 5 from Vespula vulgaris (Ves v 5). *Clin Exp Allergy* **35**, 367-373 (2005).

58. Reese, G., Ayuso, R. & Lehrer, S. B. Tropomyosin: an invertebrate pan-allergen. *Int Arch Allergy Immunol* **119**, 247-258 (1999).

59. Shanti, K. N., Martin, B. M., Nagpal, S., Metcalfe, D. D. & Rao, P. V. Identification of tropomyosin as the major shrimp allergen and characterization of its IgE-binding epitopes. *J Immunol* **151**, 5354-5363 (1993).

60. Chu, K. H., Wong, S. H. & Leung, P. S. Tropomyosin Is the Major Mollusk Allergen: Reverse Transcriptase Polymerase Chain Reaction, Expression and IgE Reactivity. *Mar Biotechnol (NY)* **2**, 499-509 (2000).

61. Jeong, K. Y., Hong, C. S. & Yong, T. S. Allergenic tropomyosins and their cross-reactivities. *Protein Pept Lett* **13**, 835-845 (2006).

62. Shin, D. S. et al. Biochemical and structural analysis of the IgE binding sites on ara h1, an abundant and highly allergenic peanut protein. *J Biol Chem* **273**, 13753-13759 (1998).

63. Rabjohn, P. et al. Molecular cloning and epitope analysis of the peanut allergen Ara h 3. *J Clin Invest* **103**, 535-542 (1999).

64. Yoshioka, H., Ohmoto, T., Urisu, A., Mine, Y. & Adachi, T. Expression and epitope analysis of the major allergenic protein Fag e 1 from buckwheat. *J Plant Physiol* **161**, 761-767 (2004).

65. Beardslee, T. A., Zeece, M. G., Sarath, G. & Markwell, J. P. Soybean glycinin G1 acidic chain shares IgE epitopes with peanut allergen Ara h 3. *Int Arch Allergy Immunol* **123**, 299-307 (2000).

66. Helm, R. M. et al. A soybean G2 glycinin allergen. 2. Epitope mapping and three-dimensional modeling. *Int Arch Allergy Immunol* **123**, 213-219 (2000).

67. Baur, X. et al. Structure, antigenic determinants of some clinically important insect allergens: chironomid hemoglobins. *Science* **233**, 351-354 (1986).

68. Tamura, Y. et al. Analysis of sequential immunoglobulin E-binding epitope of Japanese cedar pollen allergen (Cry j 2) in humans, monkeys and mice. *Clin Exp Allergy* **33**, 211-217 (2003).

69. Midoro-Horiuti, T. et al. Structural basis for epitope sharing between group 1 allergens of cedar pollen. *Mol Immunol* **43**, 509-518 (2006).

70. Sone, T. et al. Identification of human T cell epitopes in Japanese cypress pollen allergen, Cha o 1, elucidates the intrinsic mechanism of cross-allergenicity between Cha o 1 and Cry j 1, the major allergen of Japanese cedar pollen, at the T cell level. *Clin Exp Allergy* **35**, 664-671 (2005).

71. Sone, T. et al. T cell epitopes in Japanese cedar (Cryptomeria japonica) pollen allergens: choice of major T cell epitopes in Cry j 1 and Cry j 2 toward design of the peptide-based immunotherapeutics for the management of Japanese cedar pollinosis. *J Immunol* **161**, 448-457 (1998).

72. Czerwinski, E. W., Midoro-Horiuti, T., White, M. A., Brooks, E. G. & Goldblum, R. M. Crystal structure of Jun a 1, the major cedar pollen allergen from Juniperus ashei, reveals a parallel beta-helical core. *J Biol Chem* **280**, 3740-3746 (2005).

73. Comstock, S. S., McGranahan, G., Peterson, W. R. & Teuber, S. S. Extensive in vitro cross-reactivity to seed storage proteins is present among walnut (Juglans) cultivars and species. *Clin Exp Allergy* **34**, 1583-1590 (2004).

74. Robotham, J. M. et al. Ana o 3, an important cashew nut (Anacardium occidentale L.) allergen of the 2S albumin family. *J Allergy Clin Immunol* **115**, 1284-1290 (2005).

75. Goetz, D. W., Whisman, B. A. & Goetz, A. D. Cross-reactivity among edible nuts: double immunodiffusion, crossed immunoelectrophoresis, and human specific igE serologic surveys. *Ann Allergy Asthma Immunol* **95**, 45-52 (2005).

76. Kazemi-Shirazi, L. et al. Quantitative IgE inhibition experiments with purified recombinant allergens indicate pollen-derived allergens as the sensitizing agents responsible for many forms of plant food allergy. *J Allergy Clin Immunol* **105**, 116-125 (2000).

77. Wensing, M. et al. IgE to Bet v 1 and profilin: cross-reactivity patterns and clinical relevance. *J Allergy Clin Immunol* **110**, 435-442 (2002).

78. Ayuso, R., Lehrer, S. B. & Reese, G. Identification of continuous, allergenic regions of the major shrimp allergen Pen a 1 (tropomyosin). *Int Arch Allergy Immunol* **127**, 27-37 (2002).

79. Fernandes, J. et al. Immunoglobulin E antibody reactivity to the major shrimp allergen, tropomyosin, in unexposed Orthodox Jews. *Clin Exp Allergy* **33**, 956-961 (2003).

80. Wild, L. G. & Lehrer, S. B. Fish and shellfish allergy. *Curr Allergy Asthma Rep* **5**, 74-79 (2005).

81. Zhang, Y., Matsuo, H. & Morita, E. Cross-reactivity among shrimp, crab and scallops in a patient with a seafood allergy. *J Dermatol* **33**, 174-177 (2006).

82. Aalberse, R. C. Structural biology of allergens. *J Allergy Clin Immunol* **106**, 228-238 (2000).

83. Breiteneder, H. & Ebner, C. Molecular and biochemical classification of plant-derived food allergens. *J Allergy Clin Immunol* **106**, 27-36 (2000).

84. Ebner, C., Hoffmann-Sommergruber, K. & Breiteneder, H. Plant food allergens homologous to pathogenesis-related proteins. *Allergy* **56 Suppl 67**, 43-44 (2001).

85. WHO. in *Report of a joint FAO/WHO expert consultation* (World Health Organization, Geneva, 2000).

86. WHO. in *Report of a joint FAO/WHO expert consultation* (World Health Organization, Geneva, 2001).

87. Mari, A. Importance of databases in experimental and clinical allergology. *Int Arch Allergy Immunol* **138**, 88-96 (2005).

88. Li, K. B., Issac, P. & Krishnan, A. Predicting allergenic proteins using wavelet transform. *Bioinformatics* **20**, 2572-2578 (2004).

89. Fiers, M. W. et al. Allermatch, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. *BMC Bioinformatics* **5**, 133 (2004).

90. Gennerich, A. & Vale, R. D. Walking the walk: how kinesin and dynein coordinate their steps. *Curr Opin Cell Biol* **21**, 59-67 (2009).

91. Chesarone, M. A. & Goode, B. L. Actin nucleation and elongation factors: mechanisms and interplay. *Curr Opin Cell Biol* **21**, 28-37 (2009).

92. Hofmann, W. A. Cell and molecular biology of nuclear actin. *Int Rev Cell Mol Biol* **273**, 219-263 (2009).

93. Brangwynne, C. P., Koenderink, G. H., MacKintosh, F. C. & Weitz, D. A. Cytoplasmic diffusion: molecular motors mix it up. *J Cell Biol* **183**, 583-587 (2008).

94. Batey, S., Randles, L. G., Steward, A. & Clarke, J. Cooperative folding in a multi-domain protein. *J Mol Biol* **349**, 1045-1059 (2005).

95. Law, R. et al. Cooperativity in forced unfolding of tandem spectrin repeats. *Biophys J* **84**, 533-544 (2003).

96. Rief, M., Pascual, J., Saraste, M. & Gaub, H. E. Single molecule force spectroscopy of spectrin repeats: low unfolding forces in helix bundles. *J Mol Biol* **286**, 553-561 (1999).

97. Schwaiger, I., Sattler, C., Hostetter, D. R. & Rief, M. The myosin coiled-coil is a truly elastic protein structure. *Nat Mater* **1**, 232-235 (2002).

98. Lee, G. et al. Nanospring behaviour of ankyrin repeats. *Nature* **440**, 246-249 (2006).

99. Lu, H., Isralewitz, B., Krammer, A., Vogel, V. & Schulten, K. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophys J* **75**, 662-671 (1998).

100. Ng, S. P. et al. Mechanical unfolding of TNfn3: the unfolding pathway of a fnIII domain probed by protein engineering, AFM and MD simulation. *J Mol Biol* **350**, 776-789 (2005).

101. Miller, E., Garcia, T., Hultgren, S. & Oberhauser, A. F. The mechanical properties of E. coli type 1 pili measured by atomic force microscopy techniques. *Biophys J* **91**, 3848-3856 (2006).

102. Brockwell, D. J. et al. Pulling geometry defines the mechanical resistance of a beta-sheet protein. *Nat Struct Biol* **10**, 731-737 (2003).

103. Carrion-Vazquez, M. et al. The mechanical stability of ubiquitin is linkage dependent. *Nat Struct Biol* **10**, 738-743 (2003).

104. Carrion-Vazquez, M. et al. Mechanical design of proteins studied by single-molecule force spectroscopy and protein engineering. *Prog Biophys Mol Biol* **74**, 63-91 (2000).

105. Ainavarapu, S. R., Li, L., Badilla, C. L. & Fernandez, J. M. Ligand binding modulates the mechanical stability of dihydrofolate reductase. *Biophys J* **89**, 3337-3344 (2005).

106. Cao, Y., Balamurali, M. M., Sharma, D. & Li, H. A functional single-molecule binding assay via force spectroscopy. *Proc Natl Acad Sci U S A* **104**, 15677-15681 (2007).

107. Junker, J. P., Hell, K., Schlierf, M., Neupert, W. & Rief, M. Influence of substrate binding on the mechanical stability of mouse dihydrofolate reductase. *Biophys J* **89**, L46-48 (2005).

108. Bhasin, N. et al. Chemistry on a single protein, vascular cell adhesion molecule-1, during forced unfolding. *J Biol Chem* **279**, 45865-45874 (2004).

109. Carl, P., Kwok, C. H., Manderson, G., Speicher, D. W. & Discher, D. E. Forced unfolding modulated by disulfide bonds in the Ig domains of a cell adhesion molecule. *Proc Natl Acad Sci U S A* **98**, 1565-1570 (2001).

110. Wiita, A. P., Ainavarapu, S. R., Huang, H. H. & Fernandez, J. M. Force-dependent chemical kinetics of disulfide bond reduction observed with single-molecule techniques. *Proc Natl Acad Sci U S A* **103**, 7222-7227 (2006).

111. Wiita, A. P. et al. Probing the chemistry of thioredoxin catalysis with force. *Nature* **450**, 124-127 (2007).

112. Bustamante, C., Chemla, Y. R., Forde, N. R. & Izhaky, D. Mechanical processes in biochemistry. *Annu Rev Biochem* **73**, 705-748 (2004).

113. Sulkowska, J. I. & Cieplak, M. Stretching to understand proteins - a survey of the protein data bank. *Biophys J* **94**, 6-13 (2008).

114. Langermann, S. et al. Prevention of mucosal Escherichia coli infection by FimH-adhesin-based systemic vaccination. *Science* **276**, 607-611 (1997).

115. Bullitt, E. & Makowski, L. Bacterial adhesion pili are heterologous assemblies of similar subunits. *Biophys J* **74**, 623-632 (1998).

116. Gong, M. & Makowski, L. Helical structure of P pili from Escherichia coli. Evidence from X-ray fiber diffraction and scanning transmission electron microscopy. *J Mol Biol* **228**, 735-742 (1992).

117. Soto, G. E. & Hultgren, S. J. Bacterial adhesins: common themes and variations in architecture and assembly. *J Bacteriol* **181**, 1059-1071 (1999).

118. Jones, C. H. et al. FimH adhesin of type 1 pili is assembled into a fibrillar tip structure in the Enterobacteriaceae. *Proc Natl Acad Sci U S A* **92**, 2081-2085 (1995).

119. Mulvey, M. A. et al. Induction and evasion of host defenses by type 1-piliated uropathogenic Escherichia coli. *Science* **282**, 1494-1497 (1998).

120. Anderson, G. G. et al. Intracellular bacterial biofilm-like pods in urinary tract infections. *Science* **301**, 105-107 (2003).

121. Justice, S. S. et al. Differentiation and developmental pathways of uropathogenic Escherichia coli in urinary tract pathogenesis. *Proc Natl Acad Sci U S A* **101**, 1333-1338 (2004).

122. Hultgren, S. J., Normark, S. & Abraham, S. N. Chaperone-assisted assembly and molecular architecture of adhesive pili. *Annu Rev Microbiol* **45**, 383-415 (1991).

123. Abraham, S. N., Sun, D., Dale, J. B. & Beachey, E. H. Conservation of the D-mannose-adhesion protein among type 1 fimbriated members of the family Enterobacteriaceae. *Nature* **336**, 682-684 (1988).

124. Mulvey, M. A., Schilling, J. D. & Hultgren, S. J. Establishment of a persistent Escherichia coli reservoir during the acute phase of a bladder infection. *Infect Immun* **69**, 4572-4579 (2001).

125. Brinton, C. C., Jr. The structure, function, synthesis and genetic control of bacterial pili and a molecular model for DNA and RNA transport in gram negative bacteria. *Trans N Y Acad Sci* **27**, 1003-1054 (1965).

126. Sauer, F. G., Pinkner, J. S., Waksman, G. & Hultgren, S. J. Chaperone priming of pilus subunits facilitates a topological transition that drives fiber formation. *Cell* **111**, 543-551 (2002).

127. Jass, J. et al. Physical properties of Escherichia coli P pili measured by optical tweezers. *Biophys J* **87**, 4271-4283 (2004).

128. Fallman, E., Schedin, S., Jass, J., Uhlin, B. E. & Axner, O. The unfolding of the P pili quaternary structure by stretching is reversible, not plastic. *EMBO Rep* **6**, 52-56 (2005).

129. Andersson, M., Fallman, E., Uhlin, B. E. & Axner, O. A sticky chain model of the elongation and unfolding of Escherichia coli P pili under stress. *Biophys J* **90**, 1521-1534 (2006).

130. Thomas, W. E., Trintchina, E., Forero, M., Vogel, V. & Sokurenko, E. V. Bacterial adhesion to target cells enhanced by shear force. *Cell* **109**, 913-923 (2002).

131. Thomas, W. et al. Catch-bond model derived from allostery explains force-activated bacterial adhesion. *Biophys J* **90**, 753-764 (2006).

132. Rief, M., Gautel, M., Oesterhelt, F., Fernandez, J. M. & Gaub, H. E. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science* **276**, 1109-1112 (1997).

133. Fisher, T. E., Oberhauser, A. F., Carrion-Vazquez, M., Marszalek, P. E. & Fernandez, J. M. The study of protein mechanics with the atomic force microscope. *Trends Biochem Sci* **24**, 379-384 (1999).

134. Fisher, T. E. et al. Single molecular force spectroscopy of modular proteins in the nervous system. *Neuron* **27**, 435-446 (2000).

135. Lindberg, F. P., Lund, B. & Normark, S. Genes of pyelonephritogenic E. coli required for digalactoside-specific agglutination of human cells. *Embo J* **3**, 1167-1173 (1984).

136. Orndorff, P. E. & Falkow, S. Organization and expression of genes responsible for type 1 piliation in Escherichia coli. *J Bacteriol* **159**, 736-744 (1984).

137. Kuehn, M. J., Heuser, J., Normark, S. & Hultgren, S. J. P pili in uropathogenic E. coli are composite fibres with distinct fibrillar adhesive tips. *Nature* **356**, 252-255 (1992).

138. Florin, E.-L., Rief, M., Lehmann, H. , Ludwig, M., Dornmair, C., Moy V.T., and Gaub, H.E. Sensing specific molecular interactions with the atomic force microscope. *Biosensors and Bioelectronics* **10**, 895-901 (1995).

139. Bustamante, C., Marko, J. F., Siggia, E. D. & Smith, S. Entropic elasticity of lambda-phage DNA. *Science* **265**, 1599-1600 (1994).

140. Marko, J. F. & Siggia, E. D. Stretching DNA. *Macromolecules* **28**, 8759–8770 (1995).

141. Thomas, W. E., Nilsson, L. M., Forero, M., Sokurenko, E. V. & Vogel, V. Shear-dependent 'stick-and-roll' adhesion of type 1 fimbriated Escherichia coli. *Mol Microbiol* **53**, 1545-1557 (2004).

142. Oberhauser, A. F., Marszalek, P. E., Carrion-Vazquez, M. & Fernandez, J. M. Single protein misfolding events captured by atomic force microscopy. *Nat Struct Biol* **6**, 1025-1028 (1999).

143. Li, H. et al. Reverse engineering of the giant muscle protein titin. *Nature* **418**, 998-1002 (2002).

144. Bullitt, E. & Makowski, L. Structural polymorphism of bacterial adhesion pili. *Nature* **373**, 164-167 (1995).

145. Hahn, E. et al. Exploring the 3D molecular architecture of Escherichia coli type 1 pili. *J Mol Biol* **323**, 845-857 (2002).

146. Oberhauser, A. F., Marszalek, P. E., Erickson, H. P. & Fernandez, J. M. The molecular elasticity of the extracellular matrix protein tenascin. *Nature* **393**, 181-185 (1998).

147. Oberhauser, A. F., Badilla-Fernandez, C., Carrion-Vazquez, M. & Fernandez, J. M. The mechanical hierarchies of fibronectin observed with single-molecule AFM. *J Mol Biol* **319**, 433-447 (2002).

148. Tskhovrebova, L. & Trinick, J. Flexibility and extensibility in the titin molecule: analysis of electron microscope data. *J Mol Biol* **310**, 755-771 (2001).

149. Bell, G. I. Models for the specific adhesion of cells to cells. *Science* **200**, 618-627 (1978).

150. Chakrapani, S. & Auerbach, A. A speed limit for conformational change of an allosteric membrane protein. *Proc Natl Acad Sci U S A* **102**, 87-92 (2005).

151. Puvanendrampillai, D. & Mitchell, J. B. L/D Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein-ligand complexes. *Bioinformatics* **19**, 1856-1857 (2003).

152. Alon, R., Chen, S., Puri, K. D., Finger, E. B. & Springer, T. A. The kinetics of L-selectin tethers and the mechanics of selectin-mediated rolling. *J Cell Biol* **138**, 1169-1180 (1997).

153. Maier, B., Koomey, M. & Sheetz, M. P. A force-dependent switch reverses type IV pilus retraction. *Proc Natl Acad Sci U S A* **101**, 10961-10966 (2004).

154. Tskhovrebova, L. & Trinick, J. Titin: properties and family relationships. *Nat Rev Mol Cell Biol* **4**, 679-689 (2003).

155. Bullard, B., Leake, M. C., & Leonard, K. in *Nature's versatile engine: Insect flight muscles inside and out.* (ed. Vigoreaux, J.) (Landes: Bioscience / Springer, Austin, TX, 2005).

156. Kulke, M. et al. Kettin, a major source of myofibrillar stiffness in Drosophila indirect flight muscle. *J Cell Biol* **154**, 1045-1057 (2001).

157. Granzier, H. L. & Labeit, S. The giant protein titin: a major player in myocardial mechanics, signaling, and disease. *Circ Res* **94**, 284-295 (2004).

158. Labeit, S., Kolmerer, B. & Linke, W. A. The giant protein titin. Emerging roles in physiology and pathophysiology. *Circ Res* **80**, 290-294 (1997).

159. Linke, W. A. Stretching molecular springs: elasticity of titin filaments in vertebrate striated muscle. *Histol Histopathol* **15**, 799-811 (2000).

160. Linke, W. A. & Grutzner, A. Pulling single molecules of titin by AFM--recent advances and physiological implications. *Pflugers Arch* **456**, 101-115 (2008).

161. Linke, W. A. Sense and stretchability: the role of titin and titin-associated proteins in myocardial stress-sensing and mechanical dysfunction. *Cardiovasc Res* **77**, 637-648 (2008).

162. Makarenko, I. et al. Passive stiffness changes caused by upregulation of compliant titin isoforms in human dilated cardiomyopathy hearts. *Circ Res* **95**, 708-716 (2004).

163. Tskhovrebova, L. & Trinick, J. Properties of titin immunoglobulin and fibronectin-3 domains. *J Biol Chem* **279**, 46351-46354 (2004).

164. Josephson, R. K., Malamud, J. G. & Stokes, D. R. Asynchronous muscle: a primer. *J Exp Biol* **203**, 2713-2722 (2000).

165. Pringle, J. W. The Croonian Lecture, 1977. Stretch activation of muscle: function and mechanism. *Proc R Soc Lond B Biol Sci* **201**, 107-130 (1978).

166. Bullard, B. & Leonard, K. Modular proteins of insect muscle. *Adv Biophys* **33**, 211-221 (1996).

167. White, D. C. The elasticity of relaxed insect fibrillar flight muscle. *J Physiol* **343**, 31-57 (1983).

168. Saide, J. D. Identification of a connecting filament protein in insect fibrillar flight muscle. *J Mol Biol* **153**, 661-679 (1981).

169. Saide, J. D. et al. Characterization of components of Z-bands in the fibrillar flight muscle of Drosophila melanogaster. *J Cell Biol* **109**, 2157-2167 (1989).

170. Moore, J. R., Vigoreaux, J. O. & Maughan, D. W. The Drosophila projectin mutant, bentD, has reduced stretch activation and altered indirect flight muscle kinetics. *J Muscle Res Cell Motil* **20**, 797-806 (1999).

171. Lakey, A. et al. Kettin, a large modular protein in the Z-disc of insect muscles. *Embo J* **12**, 2863-2871 (1993).

172. van Straaten, M. et al. Association of kettin with actin in the Z-disc of insect flight muscle. *J Mol Biol* **285**, 1549-1562 (1999).

173. Hakeda, S., Endo, S. & Saigo, K. Requirements of Kettin, a giant muscle protein highly conserved in overall structure in evolution, for normal muscle function, viability, and flight activity of Drosophila. *J Cell Biol* **148**, 101-114 (2000).

174. Leake, M. C., Wilson, D., Bullard, B. & Simmons, R. M. The elasticity of single kettin molecules using a two-bead laser-tweezers assay. *FEBS Lett* **535**, 55-60 (2003).

175. Oberhauser, A. F., Hansma, P. K., Carrion-Vazquez, M. & Fernandez, J. M. Stepwise unfolding of titin under force-clamp atomic force microscopy. *Proc Natl Acad Sci U S A* **98**, 468-472 (2001).

176. Bullard, B., Linke, W. A. & Leonard, K. Varieties of elastic protein in invertebrate muscles. *J Muscle Res Cell Motil* **23**, 435-447 (2002).

177. Linke, W. A. et al. I-band titin in cardiac muscle is a three-element molecular spring and is critical for maintaining thin filament structure. *J Cell Biol* **146**, 631-644 (1999).

178. Ayme-Southgate, A., Bounaix, C., Riebe, T. E. & Southgate, R. Assembly of the giant protein projectin during myofibrillogenesis in Drosophila indirect flight muscles. *BMC Cell Biol* **5**, 17 (2004).

179. Improta, S., Politou, A. S. & Pastore, A. Immunoglobulin-like modules from titin I-band: extensible components of muscle elasticity. *Structure* **4**, 323-337 (1996).

180. Rivetti, C., Guthold, M. & Bustamante, C. Scanning force microscopy of DNA deposited onto mica: equilibration versus kinetic trapping studied by statistical polymer chain analysis. *J Mol Biol* **264**, 919-932 (1996).

181. Linke, W. A., Stockmeier, M. R., Ivemeyer, M., Hosser, H. & Mundel, P. Characterizing titin's I-band Ig domain region as an entropic spring. *J Cell Sci* **111 (Pt 11)**, 1567-1574 (1998).

182. Amodeo, P., Fraternali, F., Lesk, A. M. & Pastore, A. Modularity and homology: modelling of the titin type I modules and their interfaces. *J Mol Biol* **311**, 283-296 (2001).

183. Gautel, M. The super-repeats of titin/connectin and their interactions: glimpses at sarcomeric assembly. *Adv Biophys* **33**, 27-37 (1996).

184. Muhle-Goll, C. et al. Structural and functional studies of titin's fn3 modules reveal conserved surface patterns and binding to myosin S1--a possible role in the Frank-Starling mechanism of the heart. *J Mol Biol* **313**, 431-447 (2001).

185. Schlierf, M., Li, H. & Fernandez, J. M. The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques. *Proc Natl Acad Sci U S A* **101**, 7299-7304 (2004).

186. Kellermayer, M. S., Smith, S. B., Granzier, H. L. & Bustamante, C. Folding-unfolding transitions in single titin molecules characterized with laser tweezers. *Science* **276**, 1112-1116 (1997).

187. Carrion-Vazquez, M. et al. Mechanical and chemical unfolding of a single protein: a comparison. *Proc Natl Acad Sci U S A* **96**, 3694-3699 (1999).

188. Fernandez, J. M. & Li, H. Force-clamp spectroscopy monitors the folding trajectory of a single protein. *Science* **303**, 1674-1678 (2004).

189. Rief, M., Gautel, M., Schemmel, A. & Gaub, H. E. The mechanical stability of immunoglobulin and fibronectin III domains in the muscle protein titin measured by atomic force microscopy. *Biophys J* **75**, 3008-3014 (1998).

190. Clarke, J., Cota, E., Fowler, S. B. & Hamill, S. J. Folding studies of immunoglobulin-like beta-sandwich proteins suggest that they share a common folding pathway. *Structure* **7**, 1145-1153 (1999).

191. Klimov, D. K. & Thirumalai, D. Native topology determines force-induced unfolding pathways in globular proteins. *Proc Natl Acad Sci U S A* **97**, 7254-7259 (2000).

192. Paci, E. & Karplus, M. Unfolding proteins by external forces and temperature: the importance of topology and energetics. *Proc Natl Acad Sci U S A* **97**, 6521-6526 (2000).

193. Roberts, S. P. & Harrison, J. F. Mechanisms of thermal stability during flight in the honeybee apis mellifera. *J Exp Biol* **202 (Pt 11)**, 1523-1533 (1999).

194. Best, R. B. & Hummer, G. Comment on "Force-clamp spectroscopy monitors the folding trajectory of a single protein". *Science* **308**, 498; author reply 498 (2005).

195. Kirmizialtin, S., Huang, L. & Makarov, D. E. Topography of the free-energy landscape probed via mechanical unfolding of proteins. *J Chem Phys* **122**, 234915 (2005).

196. Bullard, B. et al. Association of the chaperone alphaB-crystallin with titin in heart muscle. *J Biol Chem* **279**, 7917-7924 (2004).

197. Lange, S. et al. The kinase domain of titin controls muscle gene expression and protein turnover. *Science* **308**, 1599-1603 (2005).

198. Witt, C. C. et al. Cooperative control of striated muscle mass and metabolism by MuRF1 and MuRF2. *Embo J* **27**, 350-360 (2008).

199. Ferrara, T. M., Flaherty, D. B. & Benian, G. M. Titin/connectin-related proteins in C. elegans: a review and new findings. *J Muscle Res Cell Motil* **26**, 435-447 (2005).

200. Gotthardt, M. et al. Conditional expression of mutant M-line titins results in cardiomyopathy with altered sarcomere structure. *J Biol Chem* **278**, 6059-6065 (2003).

201. Benian, G. M., Kiff, J. E., Neckelmann, N., Moerman, D. G. & Waterston, R. H. Sequence of an unusually large protein implicated in regulation of myosin activity in C. elegans. *Nature* **342**, 45-50 (1989).

202. Benian, G. M., L'Hernault, S. W. & Morris, M. E. Additional sequence complexity in the muscle gene, unc-22, and its encoded protein, twitchin, of Caenorhabditis elegans. *Genetics* **134**, 1097-1104 (1993).

203. Probst, W. C. et al. cAMP-dependent phosphorylation of Aplysia twitchin may mediate modulation of muscle contractions by neuropeptide cotransmitters. *Proc Natl Acad Sci U S A* **91**, 8487-8491 (1994).

204. Siegman, M. J. et al. Phosphorylation of a twitchin-related protein controls catch and calcium sensitivity of force production in invertebrate smooth muscle. *Proc Natl Acad Sci U S A* **95**, 5383-5388 (1998).

205. Flaherty, D. B. et al. Titins in C.elegans with unusual features: coiled-coil domains, novel regulation of kinase activity and two new possible elastic regions. *J Mol Biol* **323**, 533-549 (2002).

206. Lei, J., Tang, X., Chambers, T. C., Pohl, J. & Benian, G. M. Protein kinase domain of twitchin has protein kinase activity and an autoinhibitory region. *J Biol Chem* **269**, 21078-21085 (1994).

207. Hu, S. H. et al. Insights into autoregulation from the crystal structure of twitchin kinase. *Nature* **369**, 581-584 (1994).

208. Kobe, B. et al. Giant protein kinases: domain interactions and structural basis of autoregulation. *Embo J* **15**, 6810-6821 (1996).

209. Wilmann, M., Gautel, M. & Mayans, O. Activation of calcium/calmodulin regulated kinases. *Cell Mol Biol (Noisy-le-grand)* **46**, 883-894 (2000).

210. Heierhorst, J. et al. Substrate specificity and inhibitor sensitivity of Ca2+/S100-dependent twitchin kinases. *Eur J Biochem* **242**, 454-459 (1996).

211. Grater, F., Shen, J., Jiang, H., Gautel, M. & Grubmuller, H. Mechanically induced titin kinase activation studied by force-probe molecular dynamics simulations. *Biophys J* **88**, 790-804 (2005).

212. Mayans, O. et al. Structural basis for activation of the titin kinase domain during myofibrillogenesis. *Nature* **395**, 863-869 (1998).

213. Phillips, J. C. et al. Scalable molecular dynamics with NAMD. *J Comput Chem* **26**, 1781-1802 (2005).

214. Bullard, B. et al. The molecular elasticity of the insect flight muscle proteins projectin and kettin. *Proc Natl Acad Sci U S A* **103**, 4451-4456 (2006).

215. Best, R. B., Fowler, S. B., Toca-Herrera, J. L. & Clarke, J. A simple method for probing the mechanical unfolding pathway of proteins in detail. *Proc Natl Acad Sci U S A* **99**, 12143-12148 (2002).

216. Sharma, D. et al. Single-molecule force spectroscopy reveals a mechanically stable protein fold and the rational tuning of its mechanical stability. *Proc Natl Acad Sci U S A* **104**, 9278-9283 (2007).

217. Furst, D. O., Osborn, M., Nave, R. & Weber, K. The organization of titin filaments in the half-sarcomere revealed by monoclonal antibodies in immunoelectron microscopy: a map of ten nonrepetitive epitopes starting at the Z line extends close to the M line. *J Cell Biol* **106**, 1563-1572 (1988).

218. Tskhovrebova, L., Trinick, J., Sleep, J. A. & Simmons, R. M. Elasticity and unfolding of single molecules of the giant muscle protein titin. *Nature* **387**, 308-312 (1997).

219. Moerman, D. G., Benian, G. M., Barstead, R. J., Schriefer, L. A. & Waterston, R. H. Identification and intracellular localization of the unc-22 gene product of Caenorhabditis elegans. *Genes Dev* **2**, 93-105 (1988).

220. Obermann, W. M. et al. The structure of the sarcomeric M band: localization of defined domains of myomesin, M-protein, and the 250-kD carboxy-terminal region of titin by immunoelectron microscopy. *J Cell Biol* **134**, 1441-1453 (1996).

221. Best, R. B., Li, B., Steward, A., Daggett, V. & Clarke, J. Can non-mechanical proteins withstand force? Stretching barnase by atomic force microscopy and molecular dynamics simulation. *Biophys J* **81**, 2344-2356 (2001).

222. Li, L., Wetzel, S., Pluckthun, A. & Fernandez, J. M. Stepwise unfolding of ankyrin repeats in a single protein revealed by atomic force microscopy. *Biophys J* **90**, L30-32 (2006).

223. Ashkenazy, H., Unger, R. & Kliger, Y. Optimal data collection for correlated mutation analysis. *Proteins* **74**, 545-555 (2009).

224. Gao, M., Wilmanns, M. & Schulten, K. Steered molecular dynamics studies of titin I1 domain unfolding. *Biophys J* **83**, 3435-3445 (2002).

225. Lee, E. H., Gao, M., Pinotsis, N., Wilmanns, M. & Schulten, K. Mechanical strength of the titin Z1Z2-telethonin complex. *Structure* **14**, 497-509 (2006).

226. Lange, S., Ehler, E. & Gautel, M. From A to Z and back? Multicompartment proteins in the sarcomere. *Trends Cell Biol* **16**, 11-18 (2006).

227. Sotomayor, M. & Schulten, K. Single-molecule experiments in vitro and in silico. *Science* **316**, 1144-1148 (2007).

228. Granzier, H. L. & Irving, T. C. Passive tension in cardiac muscle: contribution of collagen, titin, microtubules, and intermediate filaments. *Biophys J* **68**, 1027-1044 (1995).

229. Bang, M. L. et al. The complete gene sequence of titin, expression of an unusual approximately 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. *Circ Res* **89**, 1065-1072 (2001).

230. Oberhauser, A. F. & Carrion-Vazquez, M. Mechanical biochemistry of proteins one molecule at a time. *J Biol Chem* **283**, 6617-6621 (2008).

231. Labeit, S. et al. A regular pattern of two types of 100-residue motif in the sequence of titin. *Nature* **345**, 273-276 (1990).

232. Witt, C. C. et al. A survey of the primary structure and the interspecies conservation of I-band titin's elastic elements in vertebrates. *J Struct Biol* **122**, 206-215 (1998).

233. Kenny, P. A., Liston, E. M. & Higgins, D. G. Molecular evolution of immunoglobulin and fibronectin domains in titin and related muscle proteins. *Gene* **232**, 11-23 (1999).

234. Marino, M. et al. Poly-Ig tandems from I-band titin share extended domain arrangements irrespective of the distinct features of their modular constituents. *J Muscle Res Cell Motil* **26**, 355-365 (2005).

235. Mrosek, M. et al. Molecular determinants for the recruitment of the ubiquitin-ligase MuRF-1 onto M-line titin. *Faseb J* **21**, 1383-1392 (2007).

236. Ma, K., Kan, L. & Wang, K. Polyproline II helix is a key structural motif of the elastic PEVK segment of titin. *Biochemistry* **40**, 3427-3438 (2001).

237. Greaser, M. Identification of new repeating motifs in titin. *Proteins* **43**, 145-149 (2001).

238. von Castelmur, E. et al. A regular pattern of Ig super-motifs defines segmental flexibility as the elastic mechanism of the titin chain. *Proc Natl Acad Sci U S A* **105**, 1186-1191 (2008).

239. Izumi, T., Schein, C. H., Oezguen, N., Feng, Y. & Braun, W. Effects of backbone contacts 3' to the abasic site on the cleavage and the product binding by human apurinic/apyrimidinic endonuclease (APE1). *Biochemistry* **43**, 684-689 (2004).

240. Negi, S. S., Kolokoltsov, A. A., Schein, C. H., Davey, R. A. & Braun, W. Determining functionally important amino acid residues of the E1 protein of Venezuelan equine encephalitis virus. *J Mol Model* **12**, 921-929 (2006).

241. Li, H., Carrion-Vazquez, M., Oberhauser, A. F., Marszalek, P. E. & Fernandez, J. M. Point mutations alter the mechanical stability of immunoglobulin modules. *Nat Struct Biol* **7**, 1117-1120 (2000).

242. Dietz, H. et al. Cysteine engineering of polyproteins for single-molecule force spectroscopy. *Nat Protoc* **1**, 80-84 (2006).

243. Leake, M. C., Grutzner, A., Kruger, M. & Linke, W. A. Mechanical properties of cardiac titin's N2B-region by single-molecule atomic force spectroscopy. *J Struct Biol* **155**, 263-272 (2006).

244. Rounsevell, R., Forman, J. R. & Clarke, J. Atomic force microscopy: mechanical unfolding of proteins. *Methods* **34**, 100-111 (2004).

245. Forman, J. R. & Clarke, J. Mechanical unfolding of proteins: insights into biology, structure and folding. *Curr Opin Struct Biol* **17**, 58-66 (2007).

246. Shindyalov, I. N. & Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng* **11**, 739-747 (1998).

247. Mayans, O., Wuerges, J., Canela, S., Gautel, M. & Wilmanns, M. Structural evidence for a possible role of reversible disulphide bridge formation in the elasticity of the muscle protein titin. *Structure* **9**, 331-340 (2001).

248. Li, H. & Fernandez, J. M. Mechanical design of the first proximal Ig domain of human cardiac titin revealed by single molecule force spectroscopy. *J Mol Biol* **334**, 75-86 (2003).

249. Wright, C. F., Teichmann, S. A., Clarke, J. & Dobson, C. M. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature* **438**, 878-881 (2005).

250. Han, J. H., Batey, S., Nickson, A. A., Teichmann, S. A. & Clarke, J. The folding and evolution of multidomain proteins. *Nat Rev Mol Cell Biol* **8**, 319-330 (2007).

251. Bjorklund, A. K., Ekman, D. & Elofsson, A. Expansion of protein domain repeats. *PLoS Comput Biol* **2**, e114 (2006).

252. Watanabe, K. et al. Molecular mechanics of cardiac titin's PEVK and N2B spring elements. *J Biol Chem* **277**, 11549-11558 (2002).

253. Sharma, D., Cao, Y. & Li, H. Engineering proteins with novel mechanical properties by recombination of protein fragments. *Angew Chem Int Ed Engl* **45**, 5633-5638 (2006).

254. Best, R. B. et al. Mechanical unfolding of a titin Ig domain: structure of transition state revealed by combining atomic force microscopy, protein engineering and molecular dynamics simulations. *J Mol Biol* **330**, 867-877 (2003).

255. Balamurali, M. M. et al. Recombination of protein fragments: a promising approach toward engineering proteins with novel nanomechanical properties. *Protein Sci* **17**, 1815-1826 (2008).

256. Fowler, S. B. & Clarke, J. Mapping the folding pathway of an immunoglobulin domain: structural detail from Phi value analysis and movement of the transition state. *Structure* **9**, 355-366 (2001).

VITA

Tzintzuni I. Garcia

Tzintzuni was born on May 11, 1980 in Mexico in the small town of Ucaréo in the state of Michoacán to Rodrigo Garcia and Martha A Lopez. His parents soon relocated to San Antonio, Texas where he grew up. He attended college at Texas A&M University at Corpus Christi before going on to graduate school at the University of Texas Medical Branch at Galveston. He was married to Sara M Volk Ph.D., and the couple have recently welcomed their first son into the world.

Education

B.S. January 2002, Texas A&M University – Corpus Christi, Corpus Christi, Texas

Publications

Bullard B, Garcia T, Benes V, Leake MC, Linke WA, Oberhauser AF. *The molecular elasticity of the insect flight muscle proteins projectin and kettin.* Proc Natl Acad Sci U S A. 2006 Mar 21;103(12):4451-6. Epub 2006 Mar 14.

Miller E, Garcia T, Hultgren S, Oberhauser AF. *The mechanical properties of E. coli type 1 pili measured by atomic force microscopy techniques.* Biophys J. 2006 Nov 15;91(10):3848-56. Epub 2006 Sep 1.

Greene DN, Garcia T, Sutton RB, Gernert KM, Benian GM, Oberhauser AF. *Single-molecule force spectroscopy reveals a stepwise unfolding of Caenorhabditis elegans giant protein kinase domains.* Biophys J. 2008 Aug;95(3):1360-70. Epub 2008 Apr 4.

Garcia TI, Oberhauser AF, Braun W. *Mechanical stability and differentially conserved physical-chemical properties of titin Ig-domains.* Proteins. 2008 Sep 25;75(3):706-718. [Epub ahead of print]

Ivanciuc O, Schein CH, Garcia T, Oezguen N, Negi SS, Braun W. *Structural analysis of linear and conformational epitopes of allergens.*Regul Toxicol Pharmacol. 2008 Dec 14. [Epub ahead of print]

Ivanciuc O, Garcia T, Torres M, Schein CH, Braun W. *Characteristic motifs for families of allergenic proteins.*Mol Immunol. 2009 Feb;46(4):559-68. Epub 2008 Oct 31.