Published in final edited form as:

Proteins. 2009 May 15; 75(3): 706–718. doi:10.1002/prot.22281.

Mechanical Stability and Differentially Conserved Physical-chemical Properties of Titin Ig-domains

Tzintzuni I. Garcia^{1,3}, Andres F. Oberhauser^{1,2,3}, and Werner Braun^{*,1,3}

¹Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX 77555, USA

²Department of Neuroscience and Cell Biology, University of Texas Medical Branch, Galveston, TX 77555, USA

³Sealy Center for Structural Biology and Molecular Biophysics University of Texas Medical Branch, Galveston, TX 77555, USA

Abstract

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 such as titin. We present a new approach to analyzing data from single-molecule force spectroscopy for different domains of the giant muscle protein titin. The region of titin found in the I-band of a sarcomere is composed of about 40 Ig-domains and is exposed to force under normal physiological conditions and connects the free-hanging ends of the myosin filaments to the Z-disc. Recent single-molecule force spectroscopy data show a mechanical hierarchy in the I-band domains. 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 of these Ig-domains are thought to share a fold and topology common to members of the Ig-like fold family, the sequences of neighboring domains vary greatly with an average sequence identity of only 25%. We examine in this study the relation of these unique mechanical stabilities of each I-band Ig domain to specific conserved physical-chemical properties of amino acid sequences in related Ig domains. We find that the sequences of each individual titin Ig domain are very highly conserved with an average sequence identity of 79% across species that are divergent as humans, chickens, and zebrafish. This indicates that the mechanical properties of each domain are well conserved and tailored to its unique position in the titin molecule. We used the PCPmer software to determine the conservation of amino acid properties in titin Ig domains grouped by unfolding forces into 'strong' and 'weak' families. We found two motifs unique to each family which may have some role in determining the mechanical properties of these Ig domains. A detailed statistical analysis of properties of individual residues revealed several positions that displayed differentially conserved properties in strong and weak families. In contrast to previous studies, we find evidence that suggests that the mechanical stability of Ig domains is determined by several residues scattered across the beta-sandwich fold, and force sensitive residues are not only confined to the A'-G region.

Keywords

titin; mechanical hierarchy; single-molecule atomic force microscopy; protein nanomechanics; PCPmer; Motif analysis

*Corresponding author: Werner Braun Department of Biochemistry & Molecular Biology University of Texas Medical Branch 301 University Boulevard Galveston, TX 77555-0857 Phone : +1 409 747 6810 Fax : +1 409 747 6000 E-Mail : webraun@utmb.edu.

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 titin¹⁻⁴. For example, titin based passive-elasticity is an important contributor to the diastolic wall stress of the myocardium⁵ or the storage of elastic strain in the passively elongating insect flight muscles⁶. Titin, a large rope-like protein composed of hundreds of sequentially arranged Ig and fibronectin type III (Fn3) domains (Fig. 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 titin⁷. 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 stretch⁸⁻¹⁰. 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 elasticity⁹⁻¹³. 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 data¹⁴ 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)^{2,4,10,15-18}.

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^{19,20}. 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^{21,22}. Further work revealed the existence of domain super-repeats in the I- and A-bands of titin²³⁻²⁶. 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 modules²⁷. Repetitive 28-residue motifs have been found in the PEVK^{28,29} region which is thought to play an important role in passive muscle tension as well as in signaling¹⁶. 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 linker³⁰. 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 linker³⁰. 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^{7,25}.

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 (<http://landau.utmb.edu:8080/WebPCPMer/>) 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 proteins³¹. Motifs defined by PCPMer can aid in locating functionally related regions of proteins that may have low sequence identity³². 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^{31,33}. 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 proteins³⁴. We also showed that a PCPMer analysis can find functionally important residues in surface exposed regions of viral proteins^{35,36}. 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 domains¹³. 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 unfolding³⁷⁻³⁹. In addition we found several individual positions in weak and strong domains that have a statistical difference in amino acid compositions with respect to hydrophobicity and size of amino acids. These positions are scattered across the beta-sandwich fold, and are not only confined to the A'-G region.

MATERIALS AND METHODS

Construction of I27 Polyproteins

We used a method based on Cys-Cys disulfide formation to generate I27 polyprotein⁴⁰. 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 Ni²⁺-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^{2,9,10,12,14,41-44}. 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^{9,10,43,44}. The force peaks in such a pattern are indicative of domain unfolding events and can be fit to the worm-like chain (WLC) equation⁴⁵ 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 algorithm⁴⁶. The structure of I1 (pdb code 1G1C47) from the I-band of titin was determined by X-ray crystallography, whereas that of I27 has both an NMR (pdb code 1TIT21) and crystallographic structure (pdb code 1WAA) which agree very well (C α 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 properties³¹. 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- E548. 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 4 and 7. The highlighting in figure 4 shows that a difference occurred in either vector while figure 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 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 1B). A representative Ig domain (pdb code 1TIT21) is shown in Figure 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^{12,13,42}.

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^{12,13}. 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 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 2B). We show this example to illustrate how the recordings are analyzed. However, all the data shown in figure 2C was taken from Li *et. al.* 2002, and 2003. Seven I-band Ig domains have been characterized in similar fashion^{13,49} (Figure 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^{9,14,50}.

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 domains⁵¹. 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 coaggregation^{51,52}. The sequence variation between these different Ig domains is illustrated in figure 3. Immediately visible is a short region of parallel diagonal lines of high sequence identity indicative of possible gene duplication^{52,53}. 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 proteins²⁴. This forms an interesting pattern of 9 domains in which the first, fifth, and ninth domain are 40-60% identical and the 2nd-4th domains are similar to the 6th-8th domains.

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. 4A) and the four domains I27, I28, I32, and I34 of the ‘*strong*’ family (Fig. 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.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.

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 illustrates the distribution of motifs (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 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 unfolding³⁷⁻³⁹.

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 size⁴⁸. Table 2 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, 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 7a displays significant differences in hydrophilicity, and figure 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 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 7a that no overall trend toward hydrophobic or hydrophilic side chains is present in either inward or outwardly facing residues. Figure 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 sidechains 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 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.

DISCUSSION

Titin Ig domains of different species but at identical positions show a remarkable conservation across vertebrates (Fig.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^{2,54}. 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 titin⁵⁵. 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 fibronectin⁵⁰ and the titin-like proteins projectin and kettin from *Drosophila* flight muscle¹⁴.

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^{2,50}. 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.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 I2843⁵⁶. 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^{38,39}. Proline mutagenesis experiments aimed at breaking these key bonds showed that there are other important interactions³⁷. Recently Sharma et al⁵⁷ 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 alone⁵⁸. Consistent with these findings we found other regions in titin Ig domains which may contribute to determining its mechanical stability.

CONCLUSIONS

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.

Acknowledgments

This work is supported by grants from the National Institutes of Health (R01 AI064913 to WB and in part by R01 DK073394 to AFO) and a training fellowship from the Keck Center for Computational and Structural Biology of the Gulf Coast Consortia (National Library of Medicine grant 5T15LM07093 to T.G.). In addition grant support from the John Sealy Memorial Endowment Fund to AFO is gratefully acknowledged.

REFERENCES

1. Lange S, Ehler E, Gautel M. From A to Z and back? Multicompartiment proteins in the sarcomere. *Trends Cell Biol.* 2006; 16(1):11–18. [PubMed: 16337382]
2. Linke WA. Sense and stretchability: the role of titin and titin-associated proteins in myocardial stress-sensing and mechanical dysfunction. *Cardiovasc Res.* 2008; 77(4):637–648. [PubMed: 17475230]
3. Linke WA, Grutzner A. Pulling single molecules of titin by AFM-recent advances and physiological implications. *Pflugers Arch.* 2007
4. Tskhovrebova L, Trinick J. Properties of titin immunoglobulin and fibronectin-3 domains. *J Biol Chem.* 2004; 279(45):46351–46354. [PubMed: 15322090]
5. Granzier HL, Irving TC. Passive tension in cardiac muscle: contribution of collagen, titin, microtubules, and intermediate filaments. *Biophys J.* 1995; 68(3):1027–1044. [PubMed: 7756523]
6. White DC. The elasticity of relaxed insect fibrillar flight muscle. *J Physiol.* 1983; 343:31–57. [PubMed: 6557139]
7. Bang ML, Centner T, Fornoff F, Geach AJ, Gotthardt M, McNabb M, Witt CC, Labeit D, Gregorio CC, Granzier H, Labeit S. 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.* 2001; 89(11):1065–1072. [PubMed: 11717165]
8. Linke WA, Rudy DE, Centner T, Gautel M, Witt C, Labeit S, Gregorio CC. I-band titin in cardiac muscle is a three-element molecular spring and is critical for maintaining thin filament structure. *J Cell Biol.* 1999; 146(3):631–644. [PubMed: 10444071]
9. Oberhauser AF, Carrion-Vazquez M. Mechanical biochemistry of proteins one molecule at a time. *J Biol Chem.* 2008
10. Linke WA, Grutzner A. Pulling single molecules of titin by AFM-recent advances and physiological implications. *Pflugers Arch.* 2008; 456(1):101–115. [PubMed: 18058125]
11. Kellermayer MS, Smith SB, Granzier HL, Bustamante C. Folding-unfolding transitions in single titin molecules characterized with laser tweezers. *Science.* 1997; 276(5315):1112–1116. [PubMed: 9148805]
12. Rief M, Gautel M, Oesterhelt F, Fernandez JM, Gaub HE. Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science.* 1997; 276(5315):1109–1112. [PubMed: 9148804]
13. Li H, Linke WA, Oberhauser AF, Carrion-Vazquez M, Kerkvliet JG, Lu H, Marszalek PE, Fernandez JM. Reverse engineering of the giant muscle protein titin. *Nature.* 2002; 418(6901):998–1002. [PubMed: 12198551]
14. 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; 103(12):4451–4456. [PubMed: 16537423]
15. Granzier HL, Labeit S. The giant protein titin: a major player in myocardial mechanics, signaling, and disease. *Circ Res.* 2004; 94(3):284–295. [PubMed: 14976139]
16. Labeit S, Kolmerer B, Linke WA. The giant protein titin. Emerging roles in physiology and pathophysiology. *Circ Res.* 1997; 80(2):290–294. [PubMed: 9012751]
17. Linke WA. Titin elasticity in the context of the sarcomere: force and extensibility measurements on single myofibrils. *Adv Exp Med Biol.* 2000; 481:179–202. discussion 203–176. [PubMed: 10987073]
18. Makarenko I, Opitz CA, Leake MC, Neagoe C, Kulke M, Gwathmey JK, del Monte F, Hajjar RJ, Linke WA. Passive stiffness changes caused by upregulation of compliant titin isoforms in human dilated cardiomyopathy hearts. *Circ Res.* 2004; 95(7):708–716. [PubMed: 15345656]
19. Labeit S, Barlow DP, Gautel M, Gibson T, Holt J, Hsieh CL, Francke U, Leonard K, Wardale J, Whiting A, et al. A regular pattern of two types of 100-residue motif in the sequence of titin. *Nature.* 1990; 345(6272):273–276. [PubMed: 2129545]
20. Amodeo P, Fraternali F, Lesk AM, Pastore A. Modularity and homology: modelling of the titin type I modules and their interfaces. *J Mol Biol.* 2001; 311(2):283–296. [PubMed: 11478861]
21. Improta S, Politou AS, Pastore A. Immunoglobulin-like modules from titin I-band: extensible components of muscle elasticity. *Structure.* 1996; 4(3):323–337. [PubMed: 8805538]

22. Witt CC, Olivieri N, Centner T, Kolmerer B, Millevoi S, Morell J, Labeit D, Labeit S, Jockusch H, Pastore A. A survey of the primary structure and the interspecies conservation of I-band titin's elastic elements in vertebrates. *J Struct Biol.* 1998; 122(1-2):206–215. [PubMed: 9724622]
23. Gautel M. The super-repeats of titin/connectin and their interactions: glimpses at sarcomeric assembly. *Adv Biophys.* 1996; 33:27–37. [PubMed: 8922100]
24. Kenny PA, Liston EM, Higgins DG. Molecular evolution of immunoglobulin and fibronectin domains in titin and related muscle proteins. *Gene.* 1999; 232(1):11–23. [PubMed: 1033517]
25. Marino M, Svergun DI, Kreplak L, Konarev PV, Maco B, Labeit D, Mayans O. 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.* 2005; 26(6-8):355–365. [PubMed: 16341830]
26. Mrosek M, Labeit D, Witt S, Heerklotz H, von Castelmur E, Labeit S, Mayans O. Molecular determinants for the recruitment of the ubiquitin-ligase MuRF-1 onto M-line titin. *FASEB J.* 2007; 21(7):1383–1392. [PubMed: 17215480]
27. Muhle-Goll C, Habeck M, Cazorla O, Nilges M, Labeit S, Granzier H. 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.* 2001; 313(2):431–447. [PubMed: 11800567]
28. Ma K, Kan L, Wang K. Polyproline II helix is a key structural motif of the elastic PEVK segment of titin. *Biochemistry.* 2001; 40(12):3427–3438. [PubMed: 11297408]
29. Greaser M. Identification of new repeating motifs in titin. *Proteins.* 2001; 43(2):145–149. [PubMed: 11276084]
30. von Castelmur E, Marino M, Svergun DI, Kreplak L, Ucurum-Fotiadis Z, Konarev PV, Urzhumtsev A, Labeit D, Labeit S, Mayans O. 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.* 2008; 105(4):1186–1191. [PubMed: 18212128]
31. Mathura VS, Schein CH, Braun W. Identifying property based sequence motifs in protein families and superfamilies: application to DNase-1 related endonucleases. *Bioinformatics.* 2003; 19(11):1381–1390. [PubMed: 12874050]
32. Schein CH, Zhou B, Oezguen N, Mathura VS, Braun W. Molego-based definition of the architecture and specificity of metal-binding sites. *Proteins.* 2005; 58(1):200–210. [PubMed: 15505785]
33. Izumi T, Schein CH, 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/aprimidinic endonuclease (APE1). *Biochemistry.* 2004; 43(3):684–689. [PubMed: 14730972]
34. Ivanciuc O, Oezguen N, Mathura VS, Schein CH, Xu Y, Braun W. Using property based sequence motifs and 3D modeling to determine structure and functional regions of proteins. *Curr Med Chem.* 2004; 11(5):583–593. [PubMed: 15032606]
35. Negi SS, Kolokoltsov AA, Schein CH, Davey RA, Braun W. Determining functionally important amino acid residues of the E1 protein of Venezuelan equine encephalitis virus. *J Mol Model.* 2006; 12(6):921–929. [PubMed: 16607494]
36. Schein CH, Zhou B, Braun W. Stereophyicochemical variability plots highlight conserved antigenic areas in Flaviviruses. *Virol J.* 2005; 2:40. [PubMed: 15845145]
37. Li H, Carrion-Vazquez M, Oberhauser AF, Marszalek PE, Fernandez JM. Point mutations alter the mechanical stability of immunoglobulin modules. *Nat Struct Biol.* 2000; 7(12):1117–1120. [PubMed: 11101892]
38. Lu H, Isralewitz B, Krammer A, Vogel V, Schulten K. Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophys J.* 1998; 75(2):662–671. [PubMed: 9675168]
39. Lu H, Schulten K. The key event in force-induced unfolding of Titin's immunoglobulin domains. *Biophys J.* 2000; 79(1):51–65. [PubMed: 10866937]
40. Dietz H, Bertz M, Schlierf M, Berkemeier F, Bornschlogl T, Junker JP, Rief M. Cysteine engineering of polypeptides for single-molecule force spectroscopy. *Nat Protoc.* 2006; 1(1):80–84. [PubMed: 17406215]

41. Fisher TE, Oberhauser AF, Carrion-Vazquez M, Marszalek PE, Fernandez JM. The study of protein mechanics with the atomic force microscope. *Trends Biochem Sci.* 1999; 24(10):379–384. [PubMed: 10500301]
42. Leake MC, Grutzner A, Kruger M, Linke WA. Mechanical properties of cardiac titin's N2B-region by single-molecule atomic force spectroscopy. *J Struct Biol.* 2006; 155(2):263–272. [PubMed: 16682230]
43. Rounsevell R, Forman JR, Clarke J. Atomic force microscopy: mechanical unfolding of proteins. *Methods.* 2004; 34(1):100–111. [PubMed: 15283919]
44. Forman JR, Clarke J. Mechanical unfolding of proteins: insights into biology, structure and folding. *Curr Opin Struct Biol.* 2007; 17(1):58–66. [PubMed: 17251000]
45. Marko JF, Siggia ED. Stretching DNA. *Macromolecules.* 1995; 28(26):8759–8770.
46. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* 1998; 11(9):739–747. [PubMed: 9796821]
47. 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.* 2001; 9(4):331–340. [PubMed: 11525170]
48. Venkatarajan MS, Braun W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *Journal of Molecular Modeling.* 2001; 7(12):445–453.
49. Li H, Fernandez JM. Mechanical design of the first proximal Ig domain of human cardiac titin revealed by single molecule force spectroscopy. *J Mol Biol.* 2003; 334(1):75–86. [PubMed: 14596801]
50. Oberhauser AF, Badilla-Fernandez C, Carrion-Vazquez M, Fernandez JM. The mechanical hierarchies of fibronectin observed with single-molecule AFM. *J Mol Biol.* 2002; 319(2):433–447. [PubMed: 12051919]
51. Wright CF, Teichmann SA, Clarke J, Dobson CM. The importance of sequence diversity in the aggregation and evolution of proteins. *Nature.* 2005; 438(7069):878–881. [PubMed: 16341018]
52. Han JH, Batey S, Nickson AA, Teichmann SA, Clarke J. The folding and evolution of multidomain proteins. *Nat Rev Mol Cell Biol.* 2007; 8(4):319–330. [PubMed: 17356578]
53. Bjorklund AK, Ekman D, Elofsson A. Expansion of protein domain repeats. *PLoS Comput Biol.* 2006; 2(8):e114. [PubMed: 16933986]
54. Grater F, Shen J, Jiang H, Gautel M, Grubmuller H. Mechanically induced titin kinase activation studied by force-probe molecular dynamics simulations. *Biophys J.* 2005; 88(2):790–804. [PubMed: 15531631]
55. Watanabe K, Muhle-Goll C, Kellermayer MS, Labeit S, Granzier H. Different molecular mechanics displayed by titin's constitutively and differentially expressed tandem Ig segments. *J Struct Biol.* 2002; 137(1-2):248–258. [PubMed: 12064950]
56. Li H, Oberhauser AF, Fowler SB, Clarke J, Fernandez JM. Atomic force microscopy reveals the mechanical design of a modular protein. *Proc Natl Acad Sci U S A.* 2000; 97(12):6527–6531. [PubMed: 10823913]
57. Sharma D, Cao Y, Li H. Engineering proteins with novel mechanical properties by recombination of protein fragments. *Angew Chem Int Ed Engl.* 2006; 45(34):5633–5638. [PubMed: 16856192]
58. Best RB, Fowler SB, Herrera JL, Steward A, Paci E, Clarke J. 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.* 2003; 330(4):867–877. [PubMed: 12850153]

Figure 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 1TIT21) 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.

Figure 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 ± 20 pN ($n=115$). **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.

Figure 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.

Figure 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.

A)

I1 APKIFERISQSTVGQGSDAHFRVRVVGKPDPECEWYKNGVKIERSDRIYWYWPEDNVCELVIRDVTGEDSASIMVKAINIAGETSSHAFLLVQAK
 I4 PIAILQGLSDQKVCEGDIVQLEVKVSL-ESVEGVWMKDGQEVQPSDRVHIVI-DKQSHMLLIEDMTKEDAGNYSFTIPA--LGLSTSGRVSVYSV
 I5 SVDVITPLKDVNVIEGTKAVLECKVSVDPVTSVKWYLNDEQIKPDDRQAIIV-KGTKQRLVINRTHASDEGPYKLIVG----RVETNCNLSVEKI

I27 LIEVEKPLYGVEVFVGETAHFEIELSE-PDVHGQWKLKGQPLTASPDCEIIE-DGKKHILILHNCQLGMTGEVSFQAA----NAKSAANLKVKEK
 I28 PLIFITPLSDVKVFEKDEAKFECEVSR-EPKTFRWLKGTEITGDDRFELIK-DGTHSMVIKSAAFEDEAKYMFEEA----DKHTSGKLIIEGI
 I32 VIGLLRPLKDVTVTAGETATFDCELSY-EDIPVEWYLKGKKLEPSDKVVPVS-EGKVHTLTLRDVKLEDAGEVQLTAK----DFKTHANLFVKEP
 I34 HVEFLRPLTDLQVREKEMARFECELSR-ENAKVKWFKDGAIEKKGKKYDIIS-KGAVRILVINKCLLDDEAEYSCEVR----TARTSGMLTVLEE

B)

I1 APKIFERISQSTVGQGSDAHFRVRVVGKPDPECEWYKNGVKIERSDRIYWYWPEDNVCELVIRDVTGEDSASIMVKAINIAGETSSHAFLLVQAK
 I4 PIAILQGLSDQKVCEGDIVQLEVKVSL-ESVEGVWMKDGQEVQPSDRVHIVI-DKQSHMLLIEDMTKEDAGNYSFTIPA--LGLSTSGRVSVYSV
 I5 SVDVITPLKDVNVIEGTKAVLECKVSVDPVTSVKWYLNDEQIKPDDRQAIIV-KGTKQRLVINRTHASDEGPYKLIVG----RVETNCNLSVEKI

I27 LIEVEKPLYGVEVFVGETAHFEIELSE-PDVHGQWKLKGQPLTASPDCEIIE-DGKKHILILHNCQLGMTGEVSFQAA----NAKSAANLKVKEK
 I28 PLIFITPLSDVKVFEKDEAKFECEVSR-EPKTFRWLKGTEITGDDRFELIK-DGTHSMVIKSAAFEDEAKYMFEEA----DKHTSGKLIIEGI
 I32 VIGLLRPLKDVTVTAGETATFDCELSY-EDIPVEWYLKGKKLEPSDKVVPVS-EGKVHTLTLRDVKLEDAGEVQLTAK----DFKTHANLFVKEP
 I34 HVEFLRPLTDLQVREKEMARFECELSR-ENAKVKWFKDGAIEKKGKKYDIIS-KGAVRILVINKCLLDDEAEYSCEVR----TARTSGMLTVLEE

Figure 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 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.

Figure 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 1TIT21) with a small C_{α} sphere and large C_{β} 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.

Table 1

Positions with significant t-values*

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

* Positions in which significant differences were detected in conserved properties between *strong* and *weak* families. The first and second columns give the positions in the alignment and the residue name at that position in the human I1 domain as a reference. The third and fourth columns give the t-values in the E1 and E2 vectors respectively with values above the 3.55 cutoff for a 0.001 confidence interval in bold.