GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene
ELSEVIER
Gene 232 (1999) 11–23

Molecular evolution of immunoglobulin and fibronectin domains in titin and related muscle proteins

Paraic A. Kenny *, [1], Eibhlis M. Liston, Desmond G. Higgins

Department of Biochemistry, University College, Cork, Ireland

Received 19 November 1998; received in revised form 3 March 1999; accepted 4 March 1999

Abstract

The family of regulatory and structural muscle proteins, which includes the giant kinases titin, twitchin and projectin, has sequences composed predominantly of serially linked immunoglobulin I set (Ig) and fibronectin type III (FN3) domains. This paper explores the evolutionary relationships between 16 members of this family. In titin, groups of Ig and FN3 domains are arranged in a regularly repeating pattern of seven and 11 domains. The 11-domain super-repeat has its origins in the seven-domain super-repeat and a model for the duplications which gave rise to this super-repeat is proposed. A super-repeat composed solely of immunoglobulin domains is found in the skeletal muscle isoform of titin. Twitchin and projectin, which are presumed to be orthologs, have undergone significant insertion/deletion of domains since their divergence. The common ancestry of myomesin, skelemin and M-protein is shown. The relationship between myosin binding proteins (MyBPs) C and H is confirmed, and MyBP–H is proposed to have given rise to MyBP–C by the acquisition of some titin domains. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: F12F3.2; Fibronectin type III domain; Immunoglobulin I set domain; Phylogenetic analysis; Projectin; Twitchin

1. Introduction

Immunoglobulin I set domains (Ig) and fibronectin type III domains (FN3) are independently folding protein modules of approximately 100 amino acids. These domains mediate protein–protein interactions in both the intracellular and extracellular compartments, and are found in hundreds of different proteins (Hunkapiller and Hood, 1989). They are found together in a family of large muscle proteins which includes titin and twitchin. These are the only intracellular proteins that contain these domains.

There has been one previous investigation into the evolution of these muscle proteins (Higgins et al., 1994) which examined 104 FN3 and 68 Ig domains from five proteins. Since this study was carried out, the sequence of titin has been completed (Labeit and Kolmerer, 1995) and a number of new proteins have been discovered and sequenced. This paper examines the evolutionary relationships between 333 Ig domains and 222 FN3 domains from 16 proteins.

The success of the immunoglobulin domain in evolution has been attributed to its potential to undergo diversification in the presence of a highly conserved structural framework, its protease resistance in the folded form, and its ability to readily form homo- and heterodimers through multiple interacting surfaces (Henikoff et al., 1997). The immunoglobulin domain can bind a wide variety of ligands by changing the nature of the variable polypeptide loops attached to the stable β-sheet core structure. It can also fulfil a purely structural function, acting as a modular 'spacer' to place an interacting module in the correct position to perform its function (Williams and Barclay, 1988). The latest release of the protein families database, Pfam (Version 3.1, August 1998) contains 4512 immunoglobulin and 2103 fibronectin type III domains.

Abbreviations: aa, amino acids; CaVPT, calcium vector protein target; FN3, fibronectin type III domain; Ig, immunoglobulin I set domain; kDa, kilodaltons; Mb, megabase pairs; MDa, megadaltons; MyBP–C, myosin binding protein C; MyBP–H, myosin binding protein H; NMR, nuclear magnetic resonance; PEVK, proline, glutamate, valine, lysine; smMLCK, smooth muscle myosin light chain kinase.

* Corresponding author. Tel: +44-171-352-8133;
fax: +44-171-352-3299.
E-mail address: paraic@icr.ac.uk (P.A. Kenny)

[1] Present address: Gene Function and Regulation, Chester Beatty Laboratories, Institute of Cancer Research, 237 Fulham Road, London SW3 6JB, UK.

PII: S0378-1119(99)00122-5

The tertiary structures of two immunoglobulin domains from titin have been determined by 2D NMR domains M5 (Pfuhl and Pastore, 1995) and I27 (Improta et al., 1996). The tertiary structure of these domains consists of two β-sheets packed against each other, each sheet containing four β-strands.

The immunoglobulin domains of these muscle proteins were originally classified as immunoglobulin C2 domains. They have been reclassified as immunoglobulin intermediate (I) set domains as they contain features which are intermediate between variable (V) and constant (C1 and C2) immunoglobulin frames (Harpaz and Chothia, 1994). Based on an analysis of the sequences, the other immunoglobulin domains in titin are also proposed to be members of the I set (Improta et al., 1996).

The 3D structure of one titin FN3 domain has been elucidated using 2D NMR (Muhle-Goll et al., 1998). FN3 domains have a similar 3D structure to immunoglobulin domains. The β-strands of Ig and FN3 domains can be closely superimposed in a 3D model with the exception of the C′ strand which is on different sheets in the two structures (Erickson, 1994).

The N- and C-termini of immunoglobulin and fibronectin domains are located at opposite ends of the structure. This facilitates the joining in series of many such independently folded domains. The muscle proteins under investigation in this study consist of these linear arrays of Ig and FN3 domains. The first muscle protein found to have these domains was twitchin, a *C. elegans* protein with a relative molecular mass of 750 kDa (Benian et al., 1989, 1993). Twitchin is composed of repeating Ig and FN3 domains and there is a serine/threonine kinase domain near the C-terminus. It is involved in the regulation of muscle contraction. The myosin light chain has been demonstrated to be the substrate for molluscan twitchin (Heierhorst et al., 1995).

Titin is the largest known protein. One skeletal muscle isoform of titin has 165 Ig and 132 FN3 domains, with a predicted molecular mass of 3.7 MDa. In vivo, the various titin isoforms have lengths of between 1 and 2 μm and span each half of the sarcomere from the Z-disk to the M-line (Trinick, 1994; Labeit et al., 1997). Like twitchin, titin has a serine/threonine kinase domain near the C-terminus. The mechanism by which titin phosphorylates its target, telethonin, in developing myocytes has recently been established (Mayans et al., 1998).

The I-band region, composed of immunoglobulin domains and a PEVK region (predominantly proline, glutamate, valine and lysine residues), has been implicated in the generation of passive tension and elasticity (Labeit and Kolmerer, 1995; Linke et al., 1996). As the sarcomere stretches (2.0–2.7 μm), the contracted tandem Ig domains straighten but the individual domains maintain their folded structure. Beyond 2.7 μm the Ig domains do not further extend and the PEVK extension dominates (Trombitas et al., 1998). The PEVK region acts as a compliant entropic spring at low stretch and as a stiffer enthalpic spring at higher extensions (Linke et al., 1998).

The A-band region consists of a series of seven- and 11-domain super-repeats of Ig and FN3 domains, although this repeat pattern breaks down close to the kinase domain. This region of titin is proposed to have a ruler/template function in myofibrillogenesis by regulating the assembly of the thick myosin filaments. The individual 11-domain titin super-repeats are 43 nm long and myosin binding protein C is found bound to titin and myosin at 43 nm intervals. Thereby titin is likely to be involved in controlling the assembly of the thick filament, although the mechanisms involved in the regulation are not yet understood (Whiting et al., 1989; Labeit et al., 1992; Trinick, 1994). A gene encoding a *Drosophila* chromosomal protein has recently been cloned which appears to be a homolog of vertebrate titin based on protein size, sequence similarity, developmental expression, subcellular localization and immunostaining with antibodies to multiple titin epitopes (Machado et al., 1998).

Projectin is believed to be the *Drosophila* homolog of twitchin as it has a very similar domain organization (Ayme-Southgate et al., 1991). There are specific isoforms in different muscle types. In the asynchronous indirect flight muscle, projectin is found in the I-band where it is believed to be involved in stretch activation, while in synchronous muscle, projectin is localized over the A-band where it is proposed to regulate myosin. In vitro, projectin has been found to phosphorylate a thin filament associated 30 kDa protein, tentatively assumed to be troponin I (Weitkamp et al., 1998). Unc-89 is a structural component of the *C. elegans* M-band. Unc-89 mutants have a disorganized muscle structure in which there are no M-lines and the thick filaments are not organized into A-bands. The sequence consists of signal transduction domains (SH3, dbl/CDC24 and PH domains) and 53 immunoglobulin domains (Benian et al., 1996). A new member of this family, F12F3.2, was sequenced as part of a 2.2 Mb contig from *C. elegans* (Wilson et al., 1994). It differs from other members of this protein family, as the kinase domain is located at the N-terminus and it is followed by 18 immunoglobulin and one fibronectin domains. In the other kinases, the kinase domain is followed by between one and 10 immunoglobulin domains. No information has yet been published on F12F3.2.

Myomesin, skelemin and M-protein are structural proteins, each with the same domain organization, which are found in the M-line. Their function is to stabilize the 3D arrangement of the thick filament. Myomesin is a ubiquitous protein of vertebrate sarcomeric M-bands (Vinkemeyer et al., 1993). It binds myosin through a unique sequence at the N-terminus, and the region spanning the first three FN3 domains of myomesin binds to the M4 domain of titin (Obermann et al., 1997). M-protein was localized to the M-band of striated muscle by immunoelectron microscopy (Vinkemeyer et al., 1993). Its expression is tissue-specifically and developmentally regulated and is found only in fast muscle fibers (Carlsson et al., 1990). Skelemin is concentrated at the M-band periphery and contains intermediate filament core-like motifs. It is postulated to be a link between myofibrils and the intermediate filament cytoskeleton (Price and Gomer, 1993).

Myosin binding protein C (MyBP–C) is found exclusively in the C-zone of the A-band of skeletal muscle where it binds both myosin and titin at intervals of 43 nm. Roles in filament assembly during myofibrillogenesis and in regulation of contraction have been proposed (Offer et al., 1973; Weber et al., 1993; Okagaki et al., 1993). The C-terminal immunoglobulin domain has been implicated in myosin binding (Okagaki et al., 1993; Gilbert et al., 1996). The C-terminal 40 kDa of myosin binding protein H (MyBP–H) shares 49.6% sequence identity with the C-terminal four domains of MyBP–C (Vaughan et al., 1993b), the region required for myosin binding. The subcellular localization of MyBP–H is species specific. In rabbit fast muscle fibers it is found in the P-zone and the last stripe of the C-zone nearest the M-band (Bennett et al., 1986), while chicken MyBP–H is localized to the C-zone (Vaughan et al., 1993b).

Smooth muscle myosin light chain kinase (smMLCK) is a Ca^{2+}/calmodulin-dependent protein kinase (Gallagher et al., 1991). It regulates contraction by phosphorylating myosin light chains. Telokin is the independently expressed C-terminal immunoglobulin domain of smMLCK (Gallagher and Herring, 1991). Calcium vector protein target (CaVPT) is a small amphioxus muscle protein containing a calmodulin-binding domain and two immunoglobulin domains. It is believed to act on the thick filament, either directly or via an interaction with other immunoglobulin domain-containing proteins (Takagi and Cox, 1990). Kettin is a 500–700 kDa modular protein composed solely of Ig domains alternating with less conserved 35 amino acid linkers. It is necessary for stable cross-linking of actin by α-actinin in the Z-disk of insect muscle. It binds F-actin with high affinity and a stochiometry of one Ig domain per actin protomer (Lakey et al., 1993; van Straaten et al., 1999).

2. Materials and methods

2.1. Sequences

The protein sequences were obtained from EMBL (Release 55) and PIR (Release 57.05). Proteins containing immunoglobulin and fibronectin domains were identified using the BLAST search algorithm running on the server of the Bork group at EMBL (http://www.bork.embl-heidelberg.de/Blast2/) with the BLOSUM-62

Table 1
Proteins included in the study

Protein	Abbreviations	Accession[a]	Species	Immunoglobulin domains	Fibronectin domains	References
Titin – cardiac isoform	Z, I, A and M	X90568	*Homo sapiens*	112	132	Labeit and Kolmerer, 1995
Titin – skeletal isoform (I-band segment)	TS	X90569	*Homo sapiens*	53	0	Labeit and Kolmerer, 1995
Drosophila titin[b]	DT	AF045775	*Drosophila melanogaster*	6	0	Machado et al., 1998
Twitchin	W	X15423	*Caenorhabditis elegans*	27	31	Benian et al., 1989
Projectin[b]	PJ	AFO47475	*Drosophila melanogaster*	25	37	Daley et al., 1998
M-protein	MP	X69089	*Homo sapiens*	7	5	Vinkemeyer et al., 1993
Myomesin	MY	X69090	*Homo sapiens*	7	5	Vinkemeyer et al., 1993
Skelemin	SK	Z22866	*Mus musculus*	7	5	Price and Gomer, 1993
Myosin binding protein C	CP	X73114	*Homo sapiens*	7	3	Weber et al., 1993
Myosin binding protein H	HP	L05606	*Homo sapiens*	2	1	Vaughan et al., 1993a
Unc-89	U	U33058	*Caenorhabditis elegans*	53	0	Benian et al., 1996
Telokin	TLK	M76233	*Oryctolagus cuniculus*	1	0	Gallagher and Herring, 1991
Calcium vector protein target	CAVPT	PIR:A37982	*Branchiostoma lanceolatum*	2	0	Takagi and Cox, 1990
Muscle localized kinase	MLK	U37708	*Mus musculus*	3	0	
Smooth muscle myosin light chain kinase	smK	M76233	*Oryctolagus cuniculus*	3	1	Gallagher et al., 1991
F12F3.2	F	U80022	*Caenorhabditis elegans*	18	2	Wilson et al., 1994
Kettin[b]	KT	X72709	*Drosophila melanogaster*	4	0	Lakey et al., 1993

[a] Accession numbers are EMBL unless stated.

[b] Incomplete sequences.

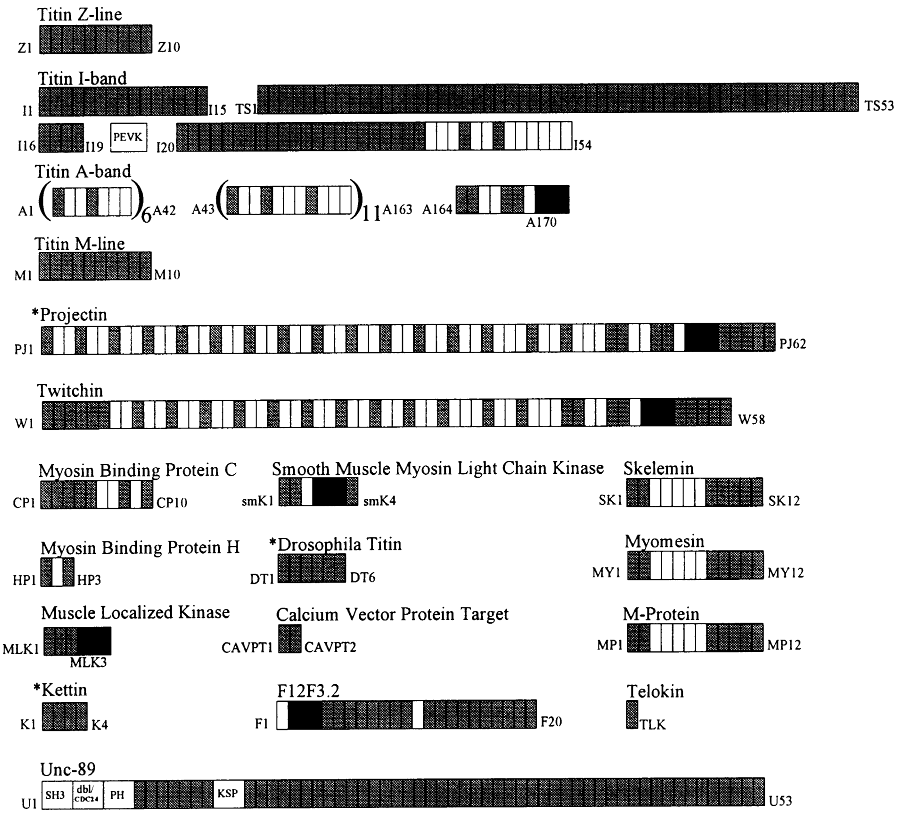

Fig. 1. Domain organization of the proteins: immunoglobulin I set – gray; fibronectin type III – white; kinase domains – black. Note that only the relative positions of these three domain types are shown and all intervening unique sequences have been omitted. Incomplete sequences are marked *. The titin I-band domains follow the numbering system for the cardiac isoform (I1 to I54), and the additional domains in the titin skeletal isoform are numbered TS1 to TS53. The skeletal muscle isoform of myosin binding protein C is shown.

weight matrix. Sequence retrieval searches were carried out at the European Bioinformatics Institute (http://srs.ebi.ac.uk). The sequences used are listed in Table 1 and a schematic diagram of their domain organization is given in Fig. 1. The domains are numbered sequentially from the N-terminus.

2.2. Multiple alignments

Domain boundaries were determined based on alignments of the 112 immunoglobulin domains and 132 fibronectin domains found in the human cardiac titin sequence (Labeit and Kolmerer, 1995) annotated by Labeit and coworkers (http://www.embl-heidelberg.de/ExternalInfo/Titin/annotation.html). Extra information on domain boundaries within the super-repeats was supplied by Siegfried Labeit (personal communication). These alignments were used to generate consensus sequences for immunoglobulin and fibronectin domains. These consensus sequences were used to locate the domain boundaries in the other proteins in this family. The individual immunoglobulin and fibronectin domain sequences were aligned using ClustalX (Thompson et al., 1997) with the default parameters. The alignments were checked for errors and manually refined where appropriate. The alignments are available by e-mail from paraic@icr.ac.uk. The alignments were checked against known tertiary structures of immunoglobulin I set domains [Titin M5 (Pfuhl and Pastore, 1995) and I27 (Improta et al., 1996)] to ensure that the gaps introduced during the alignment process were not located within the β-strands of the domains.

2.3. Phylogenetic trees

Phylogenetic trees were calculated using the neighbor-joining method. Pairwise distances were determined with PROTDIST using the Dayhoff PAM matrix and neighbor-joining trees were calculated using NEIGHBOR. Both programs are from PHYLIP 3.5.

3. Results and discussion

3.1. Phylogenetic trees

The phylogenetic trees, which were constructed from the two multiple alignments, show the approximate evolutionary relationships between 333 immunoglobulin domains (Fig. 2) and 222 fibronectin domains (Fig. 3).

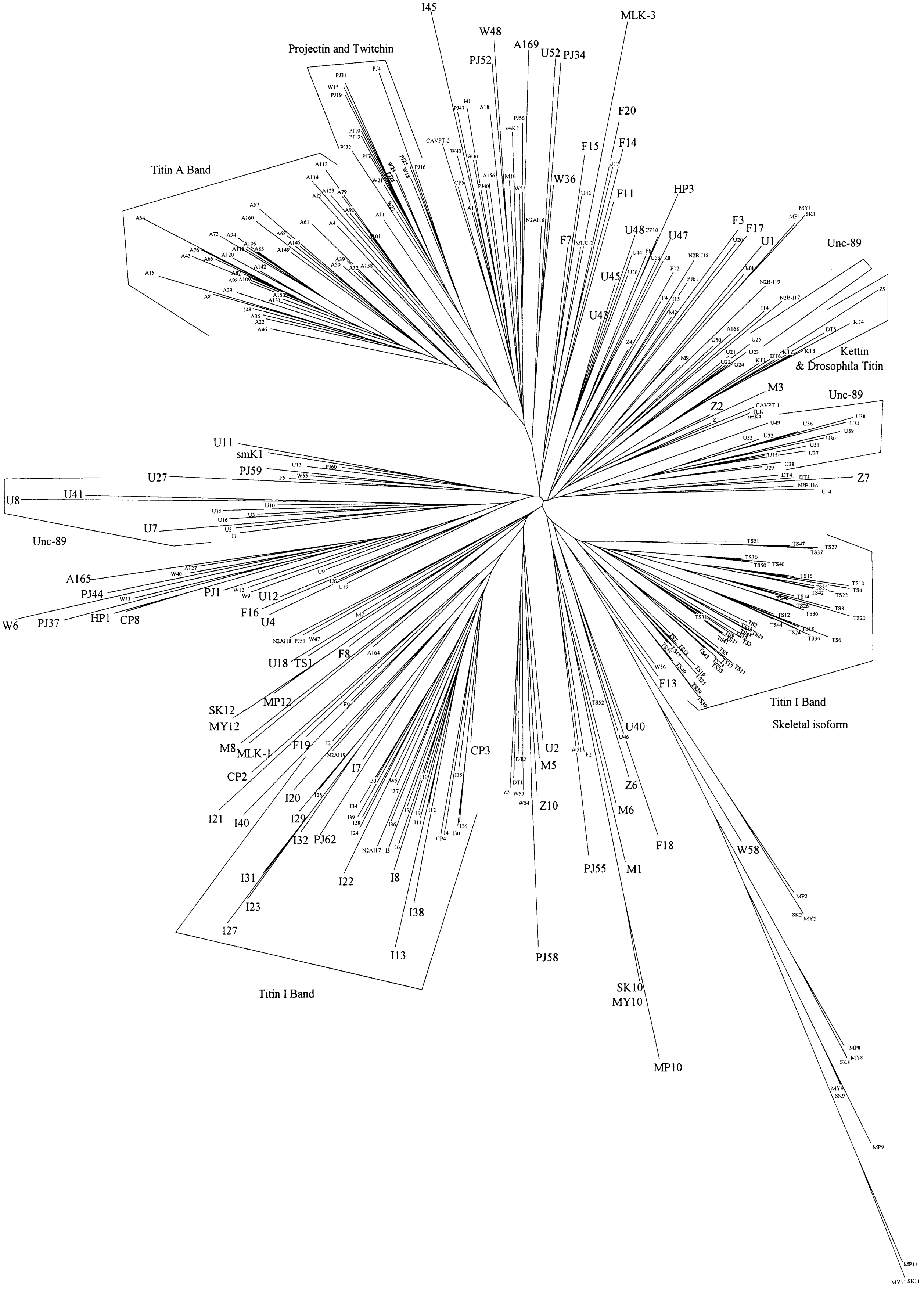

Fig. 2. Unrooted neighbor-joining tree of 333 immunoglobulin I set domains.

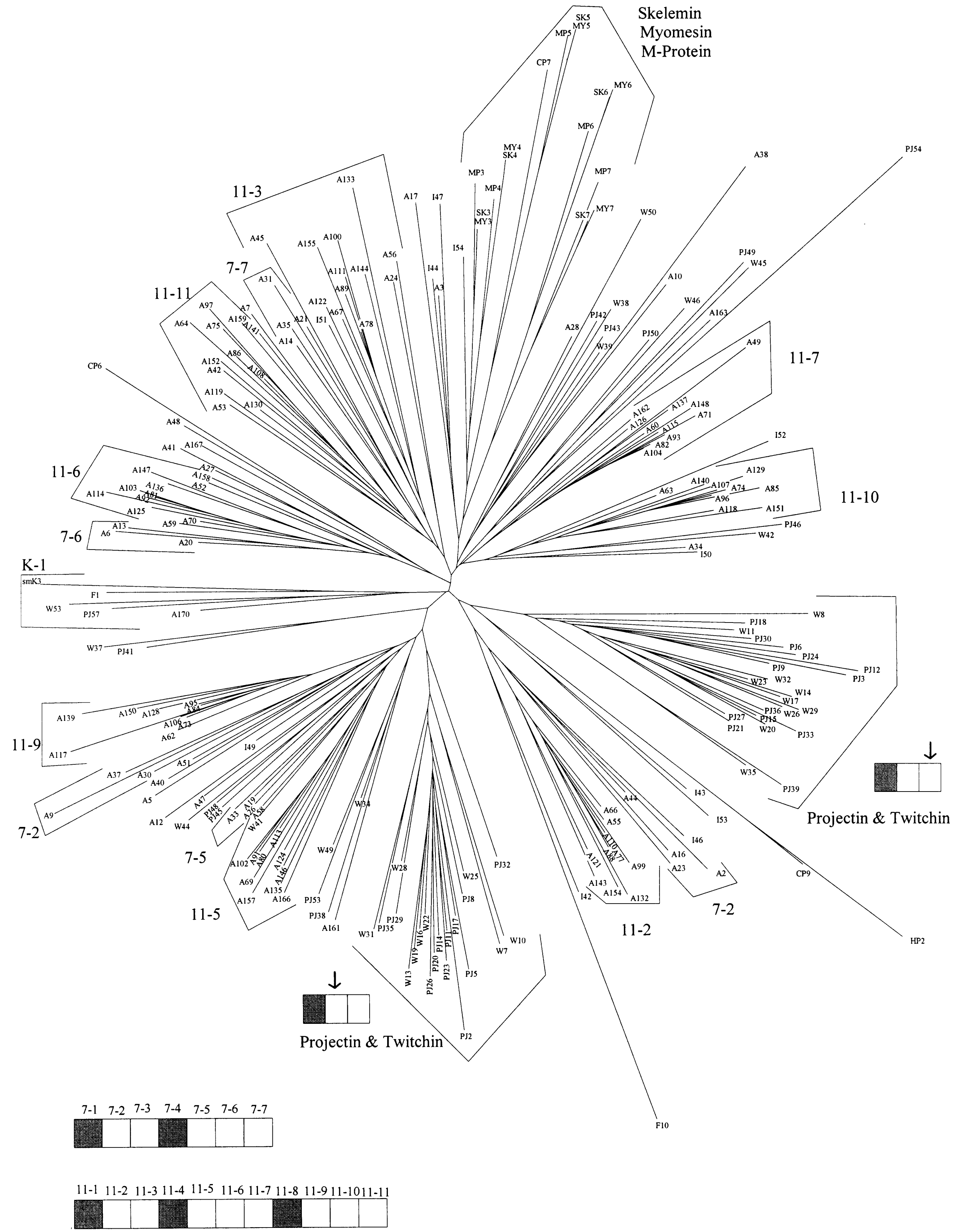

Fig. 3. Unrooted neighbor-joining tree of 222 fibronectin type III domains. Groups of domains from the titin A-band super-repeat are annotated on the tree.

The branches are packed tightly at the center of the trees. This reflects the early and rapid duplication of the immunoglobulin and fibronectin domains in these proteins. The roots of the trees are impossible to locate accurately, but they are inferred to lie approximately at the centers. Because the domain sequences are short (~100 amino acids), it is not likely that a statistically reliable, detailed picture of their evolution can be derived from this analysis. However, the large number of domains makes it possible to infer the major events involved.

3.2. *Titin A-band super-repeats*

The A-band domains of titin, located in the thick filament region of the sarcomere, form two super-repeats. There are six copies of a seven-domain super-repeat and 11 copies of an 11-domain super-repeat (Fig. 1). The majority of the domains are found in groups on the trees that correspond to their positions in the super-repeat (Fig. 4B and C). These groups have been annotated on the trees. The groups 7–2, 11–1, 11–4, 11–6, and 11–9 contain all the domains from these positions in the super-repeat. The 11–2, 11–5, 11–7 and 11–10 groups each contains 10 of their 11 domains. The groups 7–4, 7–6, and 11–3 are each missing only two domains. The 11–8 and 11–10 groups each contain eight of their 11 domains. The 7–3 and 7–7 domains are found intermingled on the fibronectin tree and the 7–1 domains are spread around the immunoglobulin tree. These results show that the domains at corresponding positions in the seven- and 11-domain super-repeats are generally more closely related to each other than to other domains, of the same type, within the same super-repeat. This shows that the super-repeats of A-band titin arose by duplication of domains at the level of the whole super-repeat. This agrees with the results obtained using a partial sequence of the 11-domain super-repeat region (Labeit et al., 1992; Higgins et al., 1994). A more detailed analysis of the branching pattern on the trees (Fig. 4B and C) reveals some clues about the evolutionary origin of the A-band super-repeats. Large groups are formed on the trees composed of domains from a number of super-repeats. These large groups represent the positions of the super-repeat that share a common evolutionary ancestor. The groups 7–1, 7–2, 7–5 and 7–6 branch with the 11–1, 11–2, 11–5, and 11–6 groups respectively. Considering this, and the similar domain arrangement, we assume that the seven-domain super-repeat is directly related to the first seven domains of the 11-domain super-repeat. By examining the remaining four domains from the 11-domain super-repeat, it is possible to determine the evolutionary origin of this conserved repeat pattern. The 7–4 domains branch with the 11–8 domains, 7–5 with the 11–9 domains, 7–6 with the 11–10 domains, and 7–7 with the 11–11 domains. Hence, it can be deduced that the 11-domain super-repeat arose by a duplication of the C-terminal four domains (7–4 to 7–7) of one original seven-membered super-repeat. Based on a tree calculated from an alignment of whole 11-domain super-repeat segments (Fig. 4D), the scheme of duplications in Fig. 4E is proposed.

3.3. *Titin I-band immunoglobulin super-repeat*

Of the several skeletal muscle isoforms of titin, the largest contains 53 more Ig domains in the elastic I-band region than the cardiac isoform. The PEVK region of this skeletal muscle isoform has 2174 amino acids and the cardiac isoform has 163 (Labeit and Kolmerer, 1995). Biophysical studies have shown that the skeletal isoform is more elastic than the cardiac isoform because of the additional immunoglobulin domains and longer PEVK region in the I-band (Linke et al., 1996, 1998; Trombitas et al., 1998). These domains from the skeletal isoform group together on the tree, separately from the other I-band immunoglobulin domains. This implies an independent evolutionary origin, from a single immunoglobulin domain. Analysis of these domains reveals the existence of a super-repeat composed solely of immunoglobulin domains. These domains appear to have developed by a series of duplications resulting in three copies of a six-domain repeat followed by three copies of a 10-domain repeat. The periodicity of this repeat was established using a dotplot (data not shown) in conjunction with a phylogenetic tree of these domains (Fig. 5). Clearly, each of the domains 1*–6* of the six-membered super-repeat is related to a corresponding domain (1–6) in the 10-membered super-repeat, as they branch together on the tree. While the groups of domains at each position of the super-repeat are stable, the order of branching of these super-repeat groups is unstable and hence it is impossible to determine precisely the origin of domains 7–10 of the 10-domain super-repeat with respect to the other domains. The functional significance of this repeat pattern, if any, is unknown. Increasing the number of Ig domains in the I-band is known to increase the elasticity of muscle, but a conserved pattern such as the one described would not be a prerequisite for this. This pattern may be only an artefact of the way the duplication has occurred. However, the strong conservation of a repeat pattern, through a large number of duplications, suggests that this regularity could have functional significance. In the A-band of titin, the repeating pattern of immunoglobulin and fibronectin domains is likely to be involved in the organization of the thick filament during myofibrillogenesis (Whiting et al., 1989; Trinick, 1994). It is possible that the regular repeating pattern of immunoglobulin domains in the I-band may be responsible for binding other proteins to allow the construction of an

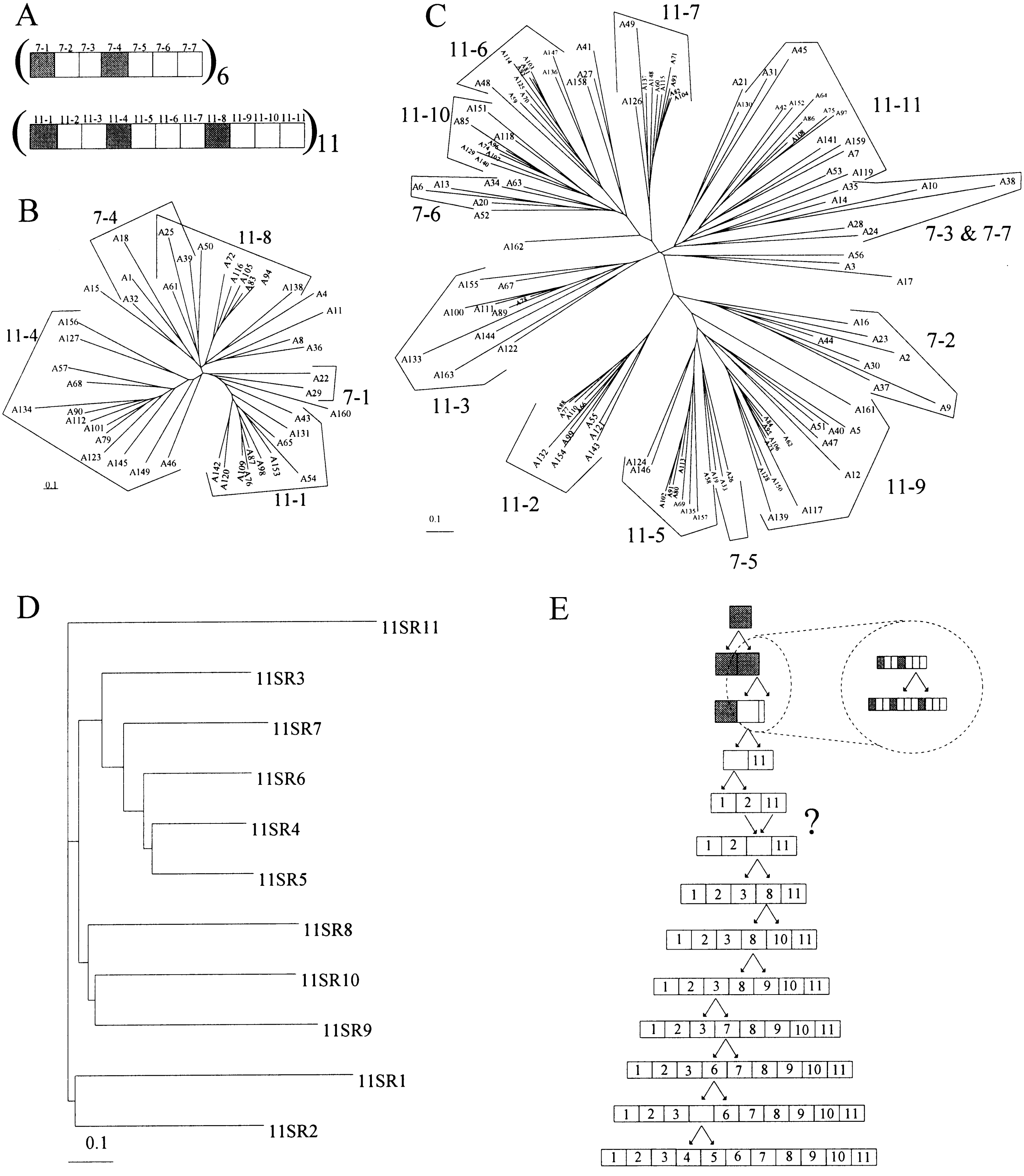

Fig. 4. Titin A-band evolution. Domain nomenclature of the A-band super-repeats (A). Neighbor-joining tree of the immunoglobulin domains from the A-band (B). Neighbor-joining tree of fibronectin type III domains from the A-band (C). Unrooted neighbor-joining tree of whole 11-domain super-repeat sequences, shown as rooted for clarity (D). Proposed order of duplications beginning with a seven-domain super-repeat (gray) which gave rise to the 11-domain super-repeat (E).

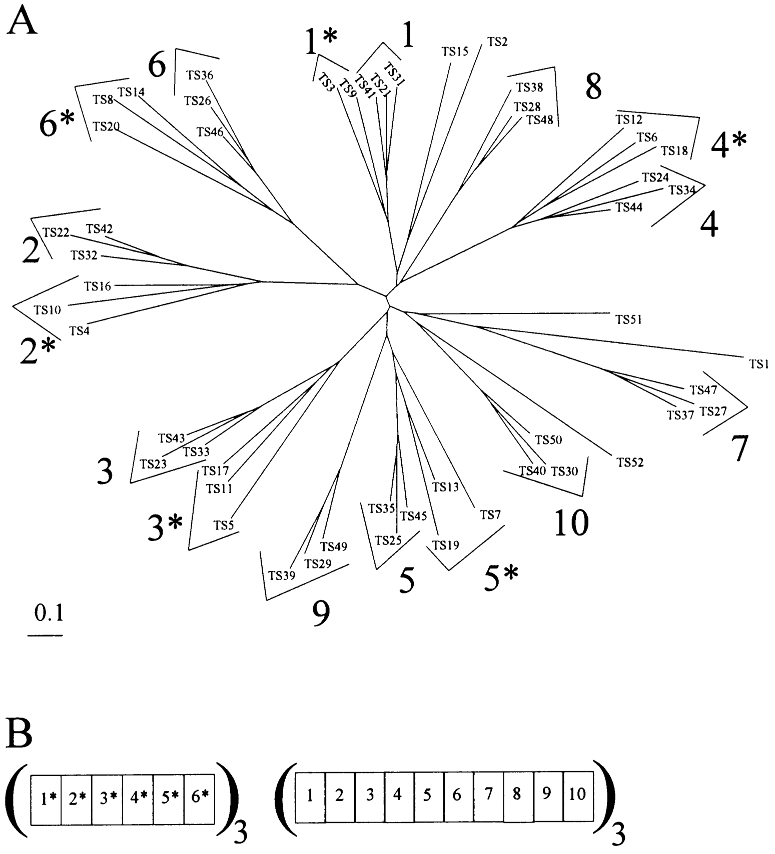

Fig. 5. Unrooted neighbor-joining tree of the immunoglobulin domains from the I-band of the skeletal isoform of titin (A) and the periodicity and nomenclature of the repeat (B).

ordered framework or lattice. Further details on this super-repeat have recently been published by Witt et al. (1998).

3.4. Twitchin and projectin

The domains from the super-repeat regions of twitchin and projectin are found together in groups on the large trees. This is expected as they share a very similar domain organization (Fig. 1). The N-terminus is composed of a series of copies of a three-domain Ig–FN3–FN3 repeat — 13 copies in projectin and 10 copies in twitchin. The domain organization of the C-termini is the same in both proteins. The domains from the super-repeat regions of twitchin and projectin are found in a single group on the immunoglobulin tree (Fig. 2) and in two groups on the fibronectin tree (Fig. 3). The domains from the C-termini of the proteins are spread throughout the trees, where they mostly group in pairs of one twitchin and one projectin domain. The projectin sequence used is incomplete at the N-terminus. The twitchin and projectin fibronectin domains group by their position in the super-repeat up to domains W32 in twitchin and PJ36 in projectin, three domains N-terminal of the point where the periodicity of the super-repeat in the primary structure of the proteins breaks down. Twitchin and projectin have been proposed to be orthologs (Ayme-Southgate et al., 1991). From the trees, however, it appears that significant insertion or deletion of domains has taken place since the divergence of the common ancestor of these proteins. If the super-repeat regions of twitchin and projectin were directly related then one would expect each projectin domain to branch with its corresponding twitchin domain. Instead, several projectin domains from the super-repeat region group together and not with the corresponding twitchin domains. In the group on the tree corresponding to the first fibronectin domain in the super-repeat, nine projectin domains are found together, branching separately from a group of four twitchin domains. Similarly in the group corresponding to the second fibronectin domain in the super-repeat, six projectin domains branch separately from a group of six twitchin domains (Fig. 3). A tree was constructed to show the relationship between the super-repeats of twitchin and projectin (Fig. 6A). The super-repeats 2–9 of projectin group together on the tree. This indicates that they are more closely related to each other than to any domains in twitchin. The same is true for the super-repeats 4–6 of twitchin. Super-repeat 3 of twitchin appears also to be unique to twitchin

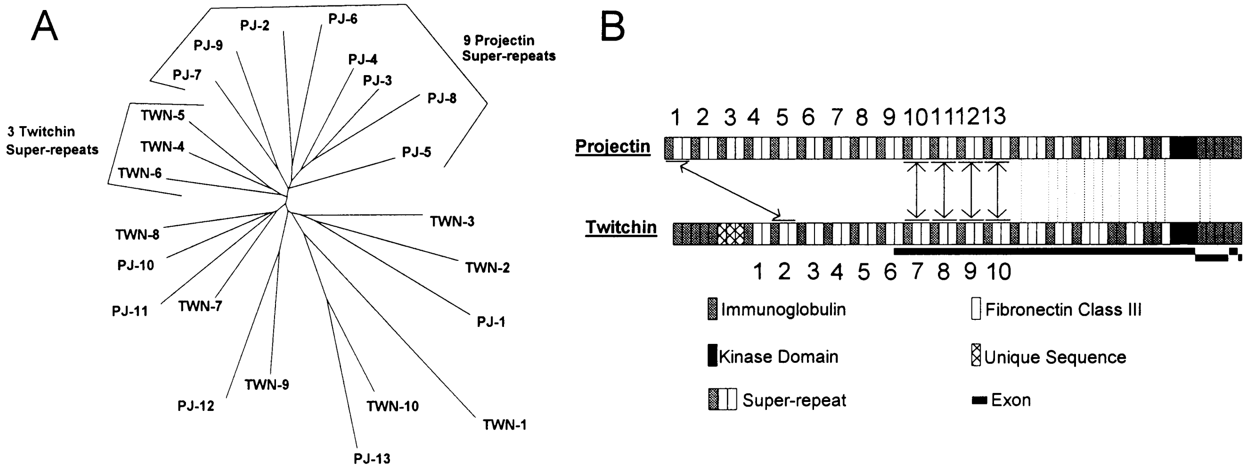

Fig. 6. Unrooted neighbor-joining tree of the three-domain Ig–FN3–FN3 super-repeats of twitchin and projectin (A). Domain organization of twitchin and projectin showing the super-repeats which branch together in A (arrows) and the individual domains which branch together on the large trees in Figs. 2 and 3 (dotted lines) (B).

as none of its domains branch with projectin domains in the large trees. It can be concluded that these particular super-repeats arose from duplications that occurred subsequent to the divergence of insects and nematodes. Conversely, the domains from the C-termini of the molecules are mostly found in pairs on both the immunoglobulin and fibronectin trees. This implies that these domains were present in the putative common ancestor of twitchin and projectin. This hypothesis is supported by the genomic DNA sequence of twitchin (Benian et al., 1989). The 31-domain region common to both proteins is encoded by a single exon which also codes for the kinase domain. The five immunoglobulin domains C-terminal to the kinase domain, which are also common to both proteins, are encoded by three exons. It is possible that the ancestral gene consisted of these four exons, which includes the kinase domain. These four exons are illustrated by black bars in Fig. 6B. The region upstream of this, which differs between the two molecules, is encoded by smaller exons in twitchin. Similarly, in projectin, this super-repeat region is encoded by smaller exons with intron/exon boundaries corresponding to the Ig–FN3–FN3 super-repeat structure. Alternative splicing of these exons leads to a number of projectin isoforms (Daley et al., 1998).

3.5. Skelemin, myomesin and M-protein

The 15 fibronectin domains of these three proteins all group in the same part of the tree (Fig. 3). The domains at each of the five corresponding positions in the three proteins are more closely related to each other than they are to the other fibronectin domains in the same protein. Therefore the duplications that led to the five fibronectin domains in the middle of each molecule did not occur independently in each of the proteins. Like the fibronectin domains, the immunoglobulin domains at each position in these proteins group together on the tree. Domains 1, 2 and 12 branch at separate points on the tree; and domains 8, 9, 10 and 11 branch together from a single node. This branching pattern reflects the common ancestry of skelemin, myomesin and M-protein, and suggests that they arose by the duplication of an ancestral gene which encoded a protein with the same domain organization as the extant proteins. From the tree, it is clear that skelemin and myomesin are the most closely related proteins. In each case the M-protein domain branches off before the node which gives rise to the corresponding skelemin and myomesin domains. Alignment of the amino acid sequences shows that human myomesin and mouse skelemin are 88.5% identical (data not shown).

3.6. Myosin binding proteins C and H

The fibronectin domains HP2 and CP9 branch together on the tree and domains CP6 and CP7 branch separately elsewhere (Fig. 3). On the immunoglobulin tree (Fig. 2), the domains HP1 and CP8, and HP3 and CP10 branch together. The domains CP2, CP3 and CP4 branch with the immunoglobulin domains from the I-band region of titin. From the branching pattern it can be deduced that MyBP–H is related to the C-terminus of MyBP–C. They have the same domain organization Ig–FN3–Ig. The C-terminal 40 kDa of MyBP–H has been shown to share 49.6% sequence identity with the C-terminus of MyBP–C (Vaughan et al., 1993b). This is the region of the MyBP–C implicated in myosin binding (Okagaki et al., 1993; Gilbert et al., 1996). The similarity of the N-terminus of MyBP–C to the titin I band immunoglobulin domains, and of its C-terminus to MyBP–H, suggests that MyBP–H could have given rise to MyBP–C by acquisition of two or three titin domains.

3.7. Other proteins

There are three main groups of Unc-89 immunoglobulin domains annotated on the tree: groups containing domains U21 to U25, U28 to U39, and a mixed group of nine domains. The remaining domains are spread around the tree individually and in small groups. This is evidence that internal duplications occurred in *unc-89* during its evolution. The 12 domains in the group U28–U39 arose by duplications from a common ancestral immunoglobulin domain. The same can be said of the five domains in the U21–U25 group.

In the fibronectin tree (Fig. 3) there is a group of five domains (smK3, F1, W53, PJ57 and A170; annotated K-1), each of which is immediately N-terminal to the kinase domain in these proteins. This suggests that the kinase region of these proteins predates their divergence and it is evidence for the inclusion of F12F3.2 in this family of muscle proteins. The inclusion of F12F3.2 is further supported by the branching pattern of immunoglobulin domains on the tree. F3 and the titin domain M2 branch together — these domains are both two domains C-terminus to the kinase domain. Similarly, F2 and M1, and F8 and M7 are found in the same groups on the tree. Each pair is at the same position with respect to the kinase domain in the two proteins. The close relationship between these three pairs of domains from F12F3.2 and titin suggests that the kinase regions of these proteins share a common ancestor in evolution. The immunoglobulin domains from F12F3.2 are very divergent and are found spread around the tree (Fig. 2). Many of them are associated with Unc-89 or with twitchin and projectin domains although it was not possible to establish a coherent evolutionary relationship between these proteins. Muscle localized kinase (MLK) is a tyrosine kinase. Its domains branch separately from those of myosin light chain kinase (a serine/threonine kinase). Therefore it appears to be from a different part of this muscle protein family to the group of serine/threonine kinases. The immunoglobulin domains MLK-2 and MLK-3 branch together on the tree and are deduced to have arisen by duplication.

The smooth muscle myosin light chain kinase smK4 domain and telokin occupy the same branch on the tree. This is consistent with the observation that telokin is the independently expressed C-terminus domain of smMLCK (Gallagher and Herring, 1991). The two immunoglobulin domains of calcium vector protein target, CaVPT1 and CaVPT2, branch separately on the tree — with CP5 and smK4/telokin respectively. It did not prove possible to make any evolutionary inferences about these domains.

The six immunoglobulin domains from the putative *Drosophila* titin fragment (Machado et al., 1998) mostly branch with the human titin Z-disk domains (DT1 and DT2 with Z5, DT3 and DT4 with Z7, and DT5 and DT6 with Z9). This, together with the data presented by Machado et al., suggests that the *Drosophila* molecule is, in fact, a homolog of titin. The four immunoglobulin domains from kettin fragment (KT) branch together on the tree with the human titin domain Z9 and two domains from *Drosophila* titin (DT5 and DT6). The size of kettin (~600 kDa) and its domain architecture (immunoglobulin domains separated by conserved 35 aa spacers) suggests that it is not a titin homolog.

4. Conclusions

(1) The ability of both immunoglobulin I set and fibronectin class III domains to adopt similar tertiary structures, while varying their amino acid sequences and hence their binding properties, has facilitated the widespread distribution of these domains in this family of structural and regulatory muscle proteins. In total 555 domains were examined in this study. This presents a more complete view of the evolution of this protein family than the previous study which involved 172 domains (Higgins et al., 1994). All of these proteins are related evolutionarily but fall into three groups which, for convenience, we have termed kinase, myosin binding and structural M-line.

(2) The kinase group contains all of the serine/threonine kinase proteins in the family — titin, twitchin, projectin, F12F3.2, smMLCK and its independently expressed C-terminal domain telokin. We also provisionally include the putative *Drosophila* titin in the kinase group even though the small published sequence fragment does not contain a kinase domain. Experimental evidence shows that this protein is similar to titin in terms of size, sequence similarity, developmental expression and subcellular localization (Machado et al., 1998) and in our study each of the six immunoglobulin domains in the fragment branch with domains from the titin Z-disk. Further sequence data and/or functional studies are necessary to properly classify this protein.

(3) The myosin binding group contains two proteins — myosin binding proteins H and C — both of which bind myosin through their C-terminal immunoglobulin domains. The sequence similarity between their C-terminal regions justifies their inclusion in a separate group. Data from the phylogenetic tree suggests that MyBP–C may have evolved via the acquisition by MyBP–H of a small number of titin I-band domains and so this group may overlap the kinase group.

(4) The structural M-line group contains three proteins with the same domain organization — myomesin, skelemin and M-protein. The domains from these pro-

teins all branch together on the trees. It is clear that they have evolved by duplication from a common ancestor with the same domain organization.

(5) Four proteins could not be classified by these criteria and these must be examined in isolation. The two-domain CaVPT protein does not contain sufficient domains to establish an unambiguous relationship to any of the above groups. The small number of domains does not make it possible to establish accurate evolutionary relationships due to the short length of the domain sequences.

(6) Muscle localized kinase (MLK) is a tyrosine kinase and this sets it apart from the proteins in this family that contain serine/threonine kinase domains. None of its three immunoglobulin domains branches with domains in close proximity to a kinase domain in the other proteins.

(7) Unc-89 is the only muscle architectural protein which contains SH3, dbl/CDC24 and PH signal transduction domains (Benian et al., 1996). There are a number of groups on the tree consisting solely of immunoglobulin domains from Unc-89. In each case, it can be deduced that these domains developed by duplication from a common ancestral domain in the evolutionary precursor of Unc-89. The remaining domains are spread around the tree where they branch with a variety of proteins. This wide variety of proteins suggests that many of these pairings are spurious and consequently it did not prove possible to determine the protein to which Unc-89 is most closely related. This study gives no information as to the origin or method of acquisition of the three signal transduction domains.

(8) The four immunoglobulin domains of kettin branch with two *Drosophila* titin domains and a domain from the titin Z-disk. As such, it is tempting to provisionally include kettin in the kinase group. However, unlike the *Drosophila* titin case, we feel that the available experimental and sequence data are insufficient to merit its inclusion.

(9) As more sequences become available it will become easier to establish coherent evolutionary relationships between these proteins. In particular, sequences of new kinases from this family (if any more exist) might allow the unambiguous order of evolution to be established for this group of proteins. The sequence of only one new kinase (F12F3.2) has been determined in the four years which have elapsed since the study of Higgins et al., and despite its inclusion it has not yet been possible to establish unequivocally the order of evolution. In addition, further progress in experimental studies of these proteins should present a better model of the structure of the sarcomere and the functions, binding properties and localization of its constituent proteins.

(10) Whether immunoglobulin and fibronectin domains have themselves diverged from a common ancestor remains an open question. Despite the striking structural similarity, it has not been possible to demonstrate an evolutionary relationship using sequence homology. However, given the length of time which would have elapsed since their divergence from a putative common ancestor and the relatively low homology between members of the same family, this lack of observed homology cannot be taken as evidence that these domains are unrelated. We cannot determine conclusively whether these domains have diverged from a common ancestor or have convergently evolved on a common fold.

Acknowledgements

The authors are extremely grateful to Siegfried Labeit for his encouragement and for providing data before publication. We also thank Donald Fischman for his advice, Carol Gregorio for providing an advance copy of her review, and Toby Gibson for getting us interested in these proteins in the first place. The work was partially supported by a grant from the EU Biotechnology Programme (BIO4-CT95-0130).

References

Ayme-Southgate, A., Vigoreaux, J., Benian, G., Pardue, M.L., 1991. *Drosophila* has a twitchin/titin related gene that appears to encode projectin. Proc. Natl. Acad. Sci. USA 88, 7973–7977.

Benian, G.M., Kiff, J.E., Necklemann, N., Moerman, D.G., Waterson, R.H., 1989. Sequence of an unusually large protein implicated in regulation of myosin activity in *C. elegans*. Nature 342, 45–50.

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

Benian, G.M., Tinley, T.L., Tang, X., Borodovsky, M., 1996. The *Caenorhabditis elegans* gene *unc-89*, required for muscle M-line assembly, encodes a giant modular protein composed of Ig and signal transduction domains. J. Cell. Biol. 132, 835–848.

Bennett, P., Craig, R., Starr, R., Offer, G., 1986. The ultrastructural localization of C-protein, X-protein and H-protein in rabbit muscle. J. Muscle Res. Cell. Motil. 7, 550–567.

Carlsson, E., Grove, B.K, Wallimann, T., Eppenberger, H.M., Thornell, L.E., 1990. Myofibrillar M-band proteins in rat skeletal muscles during development. Histochemistry 95, 27–35.

Daley, J., Southgate, R., Ayme-Southgate, A., 1998. Structure of the *Drosophila* projectin protein: isoforms and implication for projectin filament assembly. J. Mol. Biol. 279 (1), 201–210.

Erickson, H.P., 1994. Reversible unfolding of fibronectin type III and immunoglobulin domains provides the structural basis for stretch and elasticity of titin and fibronectin. Proc. Natl. Acad. Sci. USA 91, 10114–10118.

Gallagher, P.J., Herring, B.P., 1991. The carboxyl terminus of the smooth muscle myosin light chain kinase is expressed as an independent protein, telokin. J. Biol. Chem. 266, 23945–23952.

Gallagher, P.J., Herring, B.P., Griffin, S.A., Stull, J.T., 1991. Molecular characterization of a mammalian smooth muscle myosin light chain kinase. J. Biol. Chem. 266, 23936–23944.

Gilbert, R., Kelly, M.G., Mikawa, T., Fischman, D.A., 1996. The carboxyl terminus of myosin binding protein C (MyBP–C, C-protein) specifies incorporation into the A-band of striated muscle. J. Cell. Sci. 109, 101–111.

Harpaz, Y., Chothia, C., 1994. Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. J. Mol. Biol. 238, 528–539.

Heierhorst, J., Probst, W.C., Kohanski, R.A., Buku, A., Weiss, K.R., 1995. Phosphorylation of myosin regulatory light chains by the molluscan twitchin kinase. Eur. J. Biochem. 233, 426–461.

Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K., Hood, L., 1997. Gene families: the taxonomy of protein paralogs and chimeras. Science 278, 609–614.

Higgins, D.G., Labeit, S., Gautel, M., Gibson, T.J., 1994. The evolution of titin and related muscle proteins. J. Mol. Evol. 38, 395–404.

Hunkapiller, T., Hood, L., 1989. Diversity of the immunoglobulin supergene family. Adv. Immunol. 44, 1–63.

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

Labeit, S., Gautel, M., Lakey, A., Trinick, J., 1992. Towards a molecular understanding of titin. EMBO J. 11, 1711–1716.

Labeit, S., Kolmerer, B., 1995. Titins: giant proteins in charge of muscle ultrastructure and elasticity. Science 270, 293–296.

Labeit, S., Kolmerer, B., Linke, W.A., 1997. The giant protein titin. Emerging roles in physiology and pathophysiology. Circ. Res. 80 (2), 290–294.

Lakey, A., Labeit, S., Gautel, M., Ferguson, C., Barlow, D.P., Leonard, K., Bullard, B., 1993. Kettin, a large modular protein in the Z-disc of insect muscles. EMBO J. 12, 2863–2871.

Linke, W.A., Ivemeyer, M., Olivieri, N., Kolmerer, B., Ruegg, J.C., Labeit, S., 1996. Towards a molecular understanding of the elasticity of titin. J. Mol. Biol. 261, 62–71.

Linke, W.A., Ivemeyer, M., Mundel, P., Stockmeier, M.R., Kolmerer, B., 1998. Nature of PEVK–titin elasticity in skeletal muscle. Proc. Natl. Acad. Sci. USA 95, 8052–8057.

Machado, C., Sunkel, C.E., Andrew, D.J., 1998. Human autoantibodies reveal titin as a chromosomal protein. J. Cell. Biol. 141, 321–333.

Mayans, O., van der Ven, P.F.M., Wilm, M., Mues, A., Young, P., Fürst, D.O., Wilmanns, M., Gautel, M., 1998. Structural basis for the activation of the titin kinase domain during myofibrillogenesis. Nature 395, 863–869.

Muhle-Goll, C.M., Pastore, A., Nilges, M., 1998. The three-dimensional structure of a type I module from titin: a prototype of intracellular fibronectin type III domains. Structure 6, 1291–1302.

Obermann, W.M.J., Gautel, M., Weber, K., Fürst, D.O., 1997. Molecular structure of the sarcomeric M band: mapping of titin and myosin binding domains in myomesin and the identification of a potential regulatory phosphorylation site in myomesin. EMBO J. 16, 211–220.

Offer, G., Moos, C., Starr, R., 1973. A new protein of the thick filaments of vertebrate skeletal myofibrils. Extractions, purification and characterization. J. Mol. Biol. 15, 653–676.

Okagaki, T., Weber, F.E., Fischman, D.A., Vaughan, K.T., Mikawa, T., Reinach, F.C., 1993. The major myosin-binding domain of skeletal muscle MyBP–C (C-protein) resides in the COOH-terminal, immunoglobulin C2 motif. J. Cell. Biol. 123, 619–626.

Pfuhl, M., Pastore, A., 1995. Tertiary structure of an immunoglobulin-like domain from the giant muscle protein titin: a new member of the I set. Structure 3, 391–401.

Price, M.G., Gomer, R.H., 1993. Skelemin, a cytoskeletal M-disc periphery protein, contains motifs of adhesion/recognition and intermediate filament proteins. J. Biol. Chem. 268, 21800–21810.

Takagi, T., Cox, J.A., 1990. Primary structure of the target of calcium vector protein of amphioxus. J. Biol. Chem. 265 (32), 19721–19727.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25 (24), 4876–4882.

Trinick, J., 1994. Titin and nebulin: protein rulers in muscle? Trends Biochem. Sci. 19, 405–409.

Trombitas, K., Greaser, M., Labeit, S., Jin, J.P., Kellermayer, M., Helmes, M., Granzier, K., 1998. Titin extensibility in situ: entropic elasticity of permanently folded and permanently unfolded molecular segments. J. Cell. Biol. 140, 853–859.

van Straaten, M., Goulding, D., Kolmerer, B., Labeit, S., Clayton, J., Leonard, K., Bullard, B., 1999. Association of kettin with actin in the Z-disk of insect flight muscle. J. Mol. Biol. 285, 1549–1562.

Vaughan, K.T., Weber, F.E., Reid, T., Ward, D.C., Reinach, F.C., Fischman, D.A., 1993a. Human myosin-binding protein H (MyBP–H): complete primary sequence, genomic organization and chromosomal localization. Genomics 16, 34–40.

Vaughan, K.T., Weber, F.E., Einheber, S., Fischman, D.A., 1993b. Molecular cloning of chicken myosin-binding protein MyBP–H (86 kDa protein) reveals extensive homology with MyBP–C (C-protein) with conserved immunoglobulin C2 and fibronectin type III motifs. J. Biol. Chem. 268, 3670–3676.

Vinkemeyer, U., Obermann, W., Weber, K., Fürst, D.O., 1993. The globular head of titin extends into the center of the sarcomeric M band. cDNA cloning, epitope mapping and immunoelectron microscopy of two titin associated proteins. J. Cell. Sci. 106, 319–330.

Weber, F.E., Vaughan, K.T., Reinach, F.C., Fischman, D.A., 1993. Complete sequence of human fast-type and slow-type muscle myosin-binding protein C (MyBP–C). Differential expression, conserved domain structure and chromosome assignment. Eur. J. Biochem. 216, 661–669.

Weitkamp, B., Jurk, K., Beinbrech, G., 1998. Projectin–thin filament interactions and modulation of the sensitivity of the actomyosin ATPase to calcium by projectin kinase. J. Biol. Chem. 273, 19802–19808.

Whiting, A., Wardale, J., Trinick, J., 1989. Does titin regulate the length of muscle thick filaments? J. Mol. Biol. 205, 263–268.

Williams, A.F., Barclay, A.N., 1988. The immunoglobulin superfamily — domains for cell surface recognition. Ann. Rev. Immunol. 6, 381–405.

Wilson, R., et al., 1994. 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. Nature 368, 32–38.

Witt, C.C., Olivieri, N., Kolmerer, B., Millevoi, S., Morell, J., Labeit, D., Labeit, S., Jockusch, H., Pastore, A., 1998. 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.