J Mol Evol (1994) 38:395–404

Journal of Molecular Evolution

The Evolution of Titin and Related Giant Muscle Proteins

Desmond G. Higgins, Siegfried Labeit, Mathias Gautel, Toby J. Gibson

European Molecular Biology Laboratory, Postfach 10.2209, Meyerhofstrasse 1, 69012 Heidelberg, Germany

Received: 15 October 1992 / Accepted: 29 May 1993

Abstract. Titin and twitchin are giant proteins expressed in muscle. They are mainly composed of domains belonging to the fibronectin class III and immunoglobulin c2 families, repeated many times. In addition, both proteins have a protein kinase domain near the C-terminus. This paper explores the evolution of these and related muscle proteins in an attempt to determine the order of events that gave rise to the different repeat patterns and the order of appearance of the proteins. Despite their great similarity at the level of sequence organization, titin and twitchin diverged from each other at least as early as the divergence between vertebrates and nematodes. Most of the repeating units in titin and twitchin were estimated to derive from three original domains. Chicken smooth-muscle myosin light-chain kinase (smMLCK) also has a kinase domain, several immunoglobulin domains, and a fibronectin domain. From a comparison of the kinase domains, titin is predicted to have appeared first during the evolution of the family, followed by twitchin and with the vertebrate MLCKs last to appear. The so-called C-protein from chicken is also a member of this family but has no kinase domain. Its origin remains unclear but it most probably pre-dates the titin/twitchin duplication.

Key words: Titin — Twitchin — Muscle — Myosin light-chain kinase — Immunoglobulin c2 domain — Fibronectin class III domain

Correspondence to: D.G. Higgins

Introduction

Immunoglobulin c2 and fibronectin class III domains mediate numerous protein/protein interactions in the extracellular compartment of multicellular animals. They are found, often in multiple copies, in a large number of different proteins (Bork and Doolittle 1992). Benian et al. (1989) demonstrated that homologous motifs, termed class I (immunoglobulin c2) and class II (fibronectin class III), are also present in the unusually large muscle protein twitchin, from *Caenorhabditis elegans*. This was the first example of the immunoglobulin c2 and fibronectin class III domains occurring in an intracellular context. Since then a growing family of muscle proteins has emerged that are mainly composed of these domains (Einheber and Fischman 1990; Labeit et al. 1990; Lakey et al. 1990; Olson et al. 1990; Shoemaker et al. 1990; Ayme-Southgate et al. 1991).

The almost-complete sequence of the *C. elegans* gene *unc*-22 was reported by Benian et al. (1989). It codes for a very large protein called twitchin with a relative molecular mass of approximately 700 kDa. Most of twitchin is composed of repeating class I and class II domains but the C-terminus also contains a domain homologous to the catalytic domains of protein kinases (Hanks et al. 1988). Partial sequences of a muscle protein from *Drosophila* called projectin show the same repeat pattern as sequences of twitchin (Ayme-Southgate et al. 1991; Fyrberg et al. 1992). Projectin appears to be the insect homologue of twitchin: The partial sequences can be exactly aligned with parts of twitchin.

Titin (also called connectin) is the largest-known protein, with a relative molecular weight of approxi-

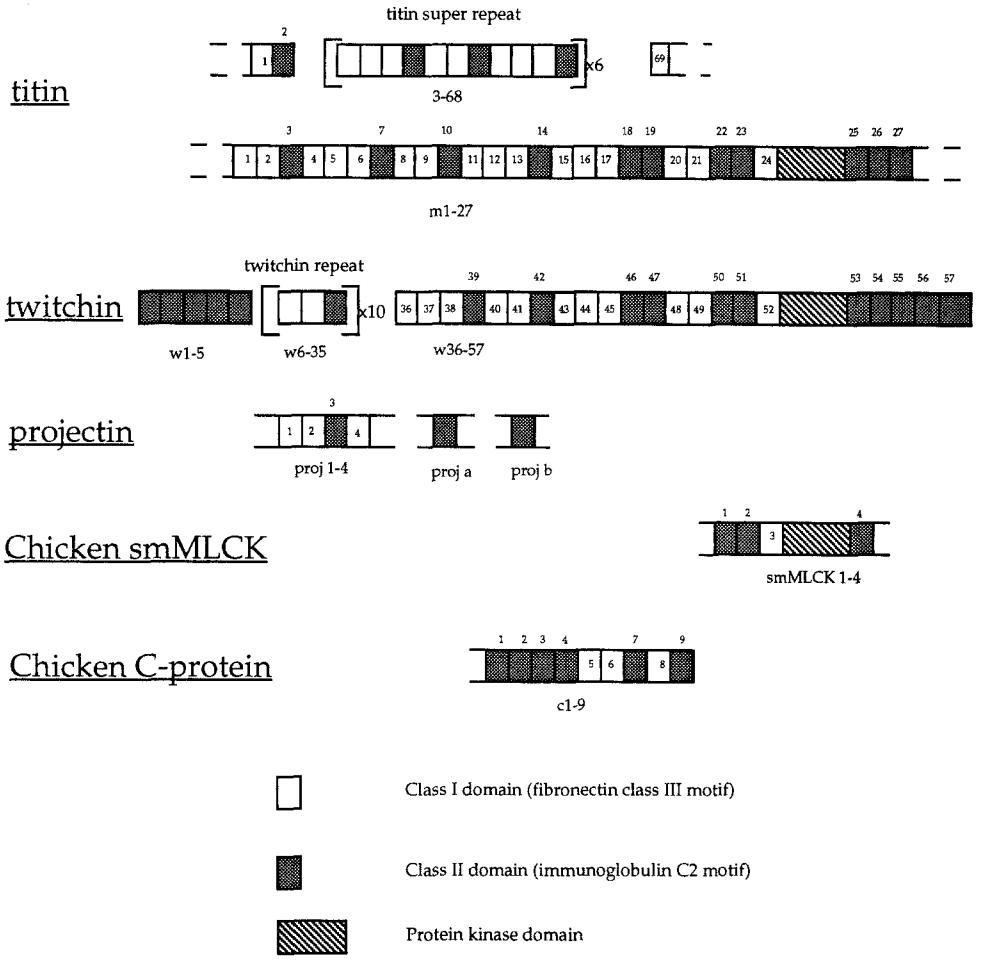

Fig. 1. The layout of the class I, class II, and kinase domains in each of the proteins. Two nonoverlapping titin sequences are used. The first titin sequence consists entirely of a series of 11-domain super repeats and is from the center of the protein while the second lies toward the C-terminus. The numbering system that is used in the figures and text is shown. The titin domains from the super repeat region are simply numbered from 1 to 69. The class I and class II domains in the second titin sequence are numbered from m1 to m27.

mately 3,000 kDa (Maruyama et al. 1984; Kurzban and Wang 1988). It is an abundant protein of vertebrate striated muscle; it spans from the M to the Z lines and is therefore over 1 μm in length (Fürst et al. 1989; Whiting et al. 1989). Recently, a total of 30 kb of titin cDNA sequence was reported (Labeit et al. 1992). Like twitchin, titin is mainly composed of repeated class I and class II domains and also has a kinase domain near the C-terminus. Thus there is extensive similarity between titin and twitchin organization at the sequence level. Notwithstanding the sequence similarity, different functions have been proposed for the two proteins. For twitchin, analysis of genetic defects implies an involvement in regulation of muscle contraction. For titin, a structural function for thick filament assembly appears to have been established (Labeit et al. 1992).

The layout of the different domain types in two fragments of titin, in twitchin, and in three other sequenced muscle proteins with class I and class II domains—projectin, chicken smooth-muscle myosin light-chain kinase (smMLCK), and C-protein—is shown in Fig. 1. Chicken smMLCK also has a kinase domain. Titin and twitchin are mostly made up of repeating class I and class II domains. In titin, the repeats form an 11-domain super repeat with the following pattern: I-I-I-II-I-I-II-I-I-I-II (Labeit et al. 1990, 1992). Dot-matrix-plot self-comparisons show that the domains at the same position in each super repeat are more similar to each other than to homologous domains in the same repeat. This presumably reflects the original duplication events that gave rise to the repeated pattern. In twitchin, the super repeat is much shorter and has the pattern I-I-II, repeated 10 times (Benian et al. 1989). The C-terminal parts of titin and twitchin, flanking the kinase domain, are almost identical in the arrangement of domains. Chicken smMLCK is too short to have a super repeat but its three class II and one class I domain are arranged in the same pattern around the kinase domain as in titin and twitchin.

This paper reports an analysis of these proteins from an evolutionary perspective in an attempt to establish the order of events that gave rise to the different proteins. The strategy adopted was to analyze separately the three

types of domain: class I, class II, and kinase. In each case, all of the available domains from titin, twitchin, projectin, chicken smMLCK, and C-protein were extracted and aligned with one another. Phylogenetic trees were constructed from the three alignments. The original events that gave rise to the different proteins must have occurred at least as early as the common ancestor of vertebrates and nematodes, approximately 700 million years ago (Dayhoff 1978). The class I and class II domains are short (approximately 100 amino acids [aa] each) and it is not likely that a statistically reliable, detailed picture of the evolution of the domains can be derived from such an analysis. Nonetheless, the very large number of domains available makes it possible to infer the major events involved.

Materials and Methods

Sequences. Sequences were either extracted from the EMBL nucleotide sequence database, release 31 (Higgins et al. 1992b), or from the Swissprot protein sequence database, release 22 (Bairoch and Boeckmann 1992). The sequence entry names and accession numbers are listed below.

1. *Titin.* Two titin A-band cDNA sequences were published by Labeit et al. (1992). These are a 20-kb-long contig from rabbit (EMBL: X64696, OCTITINR) from the central portion of the protein and a 9-kb-long contig from human (EMBL: X64697, HSTITINC3) located near the C-terminus. The contigs do not overlap. OCTITINR contains 69 class I and class II domains labeled 1 to 69 in Fig. 1. It encodes the C-protein-binding region of the thick filament. HSTITINC3 contains 27 class I and class II domains labeled m1 to m27. It also contains a protein kinase domain near the C-terminus. In vivo, this fragment is located 100 nm from the M-line. Labeit et al. (1992) compared the aa sequences of human and rabbit titin over a 3.3-kb stretch (1,100 aa) and found them to be 95% identical, or 99% similar when conservative aa replacements were considered. Therefore we do not distinguish between rabbit and human sequences for the purposes of the present study.
2. *Twitchin.* The genomic DNA sequence of twitchin from *C. elegans* was reported by Benian et al. (1989) (EMBL: X15423, CEUNC22). The sequence is nearly complete but lacks the promoter, RNA 5′ end, and possibly a small number of N-terminal class II domains. It contains a total of 57 class I and class II domains labeled w1 to w57 in Fig. 1. Close to the C-terminus, a protein kinase domain is encoded.
3. *Projectin.* Ayme-Southgate et al. (1991) reported the nucleotide sequences of three fragments of projectin from *Drosophila melanogaster* (EMBL: DMPROJA, M73433; DMPROJB, M73434; DMPROJC, M73435). DMPROJA and DMPROJB each contain one complete class II domain (labeled "proj a" and "proj b" in Fig. 1), and DMPROJC contains three class I domains and one class II domain (labeled as "proj 1" to "proj 4" in Fig. 1).
4. *C protein.* The partial sequence of C protein from chicken was published by Einheber and Fischman (1990) (Swissprot: P16419, CPSF_CHICK). It contains three class I and six class II domains (labeled c1–c9 in Fig. 1) but is not complete at the 5′ end.
5. *Myosin light-chain kinase.* A cDNA sequence of smooth-muscle myosin light-chain kinase from chicken was reported by Olson et al. (1990) (EMBL: M31048, GGSMMLCK). It contains one class I domain (labeled smMLCK3), three class II domains (smMLCK1, smMLCK2, and smMLCK4), and a protein kinase domain. It is likely to be congenic with the nonmuscle (nmMLCK) myosin light-chain kinase sequence (Shoemaker et al. 1990), in which case the sequence is incomplete at the 5′ end and should encode several additional class II domains (Gibson and Higgins 1992). Skeletal-muscle myosin light-chain kinase sequences from rat (Roush et al. 1988; Swissprot: P20689, KMLC_RAT) and rabbit (Herring et al. 1990; Swissprot: P07313, KMLC_RABIT) as well as myosin light-chain kinase from *Dictyostelium discoideum* (Tan and Spudich 1991; EMBL: M64176, DDMLCK) each contain a protein kinase domain but no class I or class II domains.

Alignments. All 104 class I and 68 class II domains were excised and manually aligned. The most conserved regions were identified and aligned first, followed by the more weakly conserved sections. Care was taken to minimize the number of gaps. This was done by always aligning gaps to gaps whenever adjacent conserved segments were fitted together, allowing only a single gap in any unconserved sequence segment. Conserved segments match the known secondary structures (beta-strands) in the solved structures of the domains. The protein kinase catalytic domains were excised and aligned automatically using the CLUSTAL V program (Higgins et al. 1992).

Data Analysis. Phylogenetic trees were mainly calculated using the neighbor-joining method of Saitou and Nei (1987). Distance matrices were calculated from the three multiple alignments by calculating percentage identity values between all pairs of sequence, i.e., the number of differences divided by the number of sites compared, ignoring positions with a gap in either sequence. For the class I and class II trees, distances were not corrected for multiple substitutions ("multiple hits") as some of the distances involved were greater than 85%. The usual formula from Kimura (1983), used to correct protein distances only, applies in the range 0 to 80%. For the kinase domain tree, Kimura's correction (Kimura 1983) was used. Confidence intervals were calculated using a bootstrap procedure (Felsenstein 1985). All of the above calculations were carried out using the CLUSTAL V program. The kinase domains were also analyzed using maximum parsimony with the PROTPARS, SEQBOOT, and CONSENSE programs of the PHYLIP package (Felsenstein 1989). The trees were drawn using the DRAWTREE program of the same package (Felsenstein 1989).

Results

The results of the phylogenetic analysis of the three domain types are shown in Figs. 2–4. The two large trees are shown unrooted. Bootstrap confidence levels are given for the kinase domain trees (Fig. 4) but not for the class I and class II domain trees. In all cases the figures were very low, reflecting the short sizes of the domains used. Therefore the exact details of the branching orders in each case are not stable. This was also reflected in the way the details of the trees varied depending on subtle differences in the distance calculations such as whether or not to remove all sites in the alignments where any sequence had a gap. Despite these reservations, however, it is shown below that the various domains mainly group into classes corresponding to their positions in the super repeats in the different proteins. This is to be expected if the super repeats arose by duplication and indicates that the branching orders in the class I and

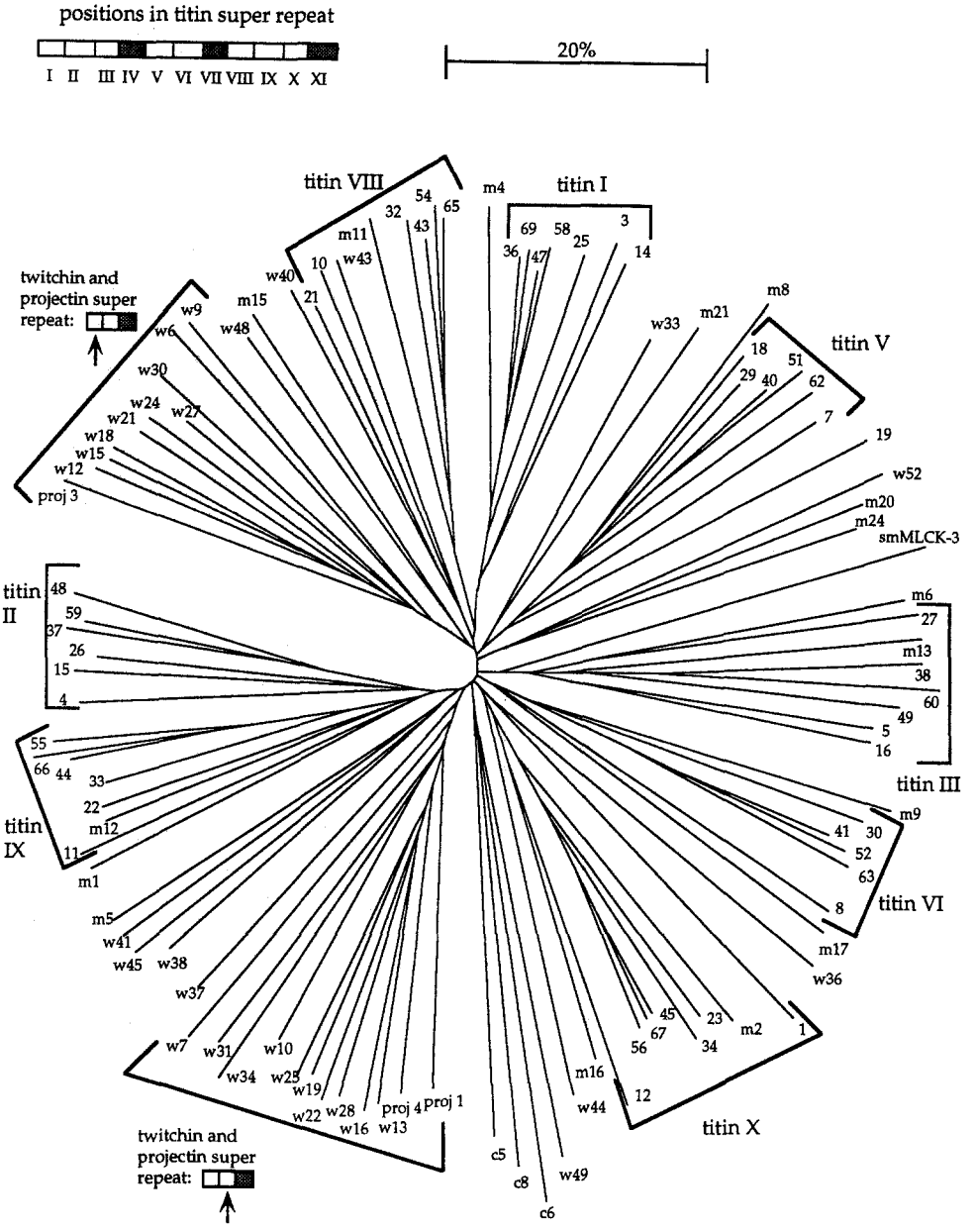

Fig. 2. Neighbor-joining tree of uncorrected distances for 104 class I domains from titin (domains 1–69 and m1–m27), twitchin (domains w6–w52), projectin (proj 1–4), smMLCK-3, and C-protein (c5, c6, and c8). The *scale bar* shows 20% distance. A *diagram* of one titin super repeat is shown with *open boxes* for class I domains and *shaded boxes* for class II domains.

class II trees do reflect the main events in the evolution of these proteins.

The Class I (Fibronectin Class 3) Domains

Figure 2 shows the class I domain tree. This is a very complicated diagram with 104 domains. Most of the domains branch together very tightly at the center of the tree. Efforts to root the tree using different fibronectin class 3 domains from human fibronectin (Kornblihtt et al. 1985) showed the exact position of the root of the tree to be difficult to estimate accurately; there are too many closely packed branches at the center. The approximate position of the root is inferred to be at the center of the diagram.

In order to help navigate the tree, the domains were divided into several classes and these groupings are indicated on the figure. First, the domains from the super repeat part of twitchin (w6–w34) and projectin group in two classes, defined by the two positions in the super repeat. The only exception is domain w33. The rest of the twitchin domains, from the C-terminal part of the protein (domain w36–w52), are spread around the tree.

Second, the domains from the super repeat of titin

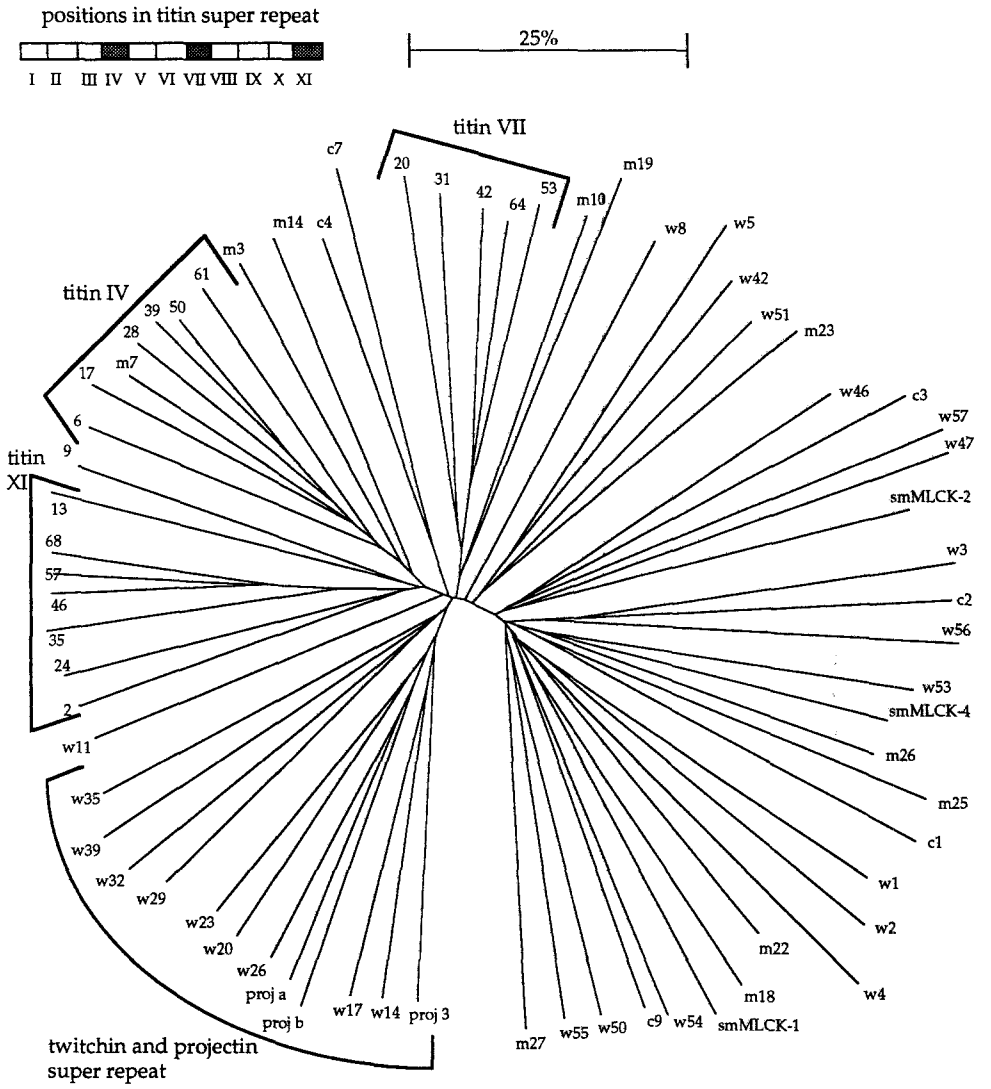

Fig. 3. Neighbor-joining tree based on uncorrected distances for 68 class II domains from titin (domains 2–68 and m3–m27), twitchin (domains w1–w57), projectin (proj 3, proj a, and proj b), smMLCK-3 (smMLCK-1–4) and C-protein (c1–c8). The *scale bar* shows 25% distance. A *diagram* of one titin super repeat is shown with *open boxes* for class I domains and *shaded boxes* for class II domains.

(domains 1–69) almost all fall into eight groups, defined by the eight class I domains in each titin super repeat. Within each group, the index numbers of the domains are in multiples of 11, reflecting the spacing of each domain along the protein. Not all of the domains fall into neat groups: Domain 19 is distant from the rest of its class (position VI) and there is some mixing of the members between the groups at positions I and VIII. Nonetheless, the grouping of the domains at repeat positions is very clear. Interestingly, domains m1–m13 from the C-terminal region mainly group with various super repeat groupings from the rest of titin. Out of 10 class I domains between m1 and m13 inclusive, all except two group with corresponding super repeat groupings. The two exceptions are domain m5, which should group at super repeat position II, and domain m13, which should group at position X. This indicates that the super repeat extends as far as domain m12 or m13. From domain m14 to the end of the titin sequence, the pattern of repeats along the protein is identical between titin and twitchin: These domains are spread around the tree and do not group together.

The rest of the domains in the tree are from smMLCK (smMLCK-3) and the chicken C-protein (c5, c6, and c8). The three C-protein domains group with one twitchin domain (w49). The single smMLCK domain groups with the titin super repeat domains at position III.

The Class II (Immunoglobulin c2) Domains

The class II tree is shown in Fig. 3. As with the class I tree, the branches are tightly packed at the center of the figure. Again, the root was impossible to locate accurately but is inferred to lie approximately at the center

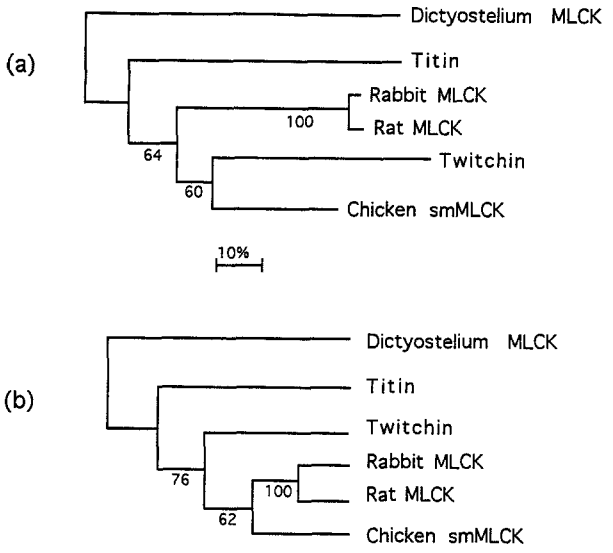

Fig. 4. Neighbor-joining tree (**a**) and maximum parsimony tree (**b**) calculated using the PROTPARS program for the kinaselike domains. In **a,** branch lengths are proportional to corrected distances along each branch. In **b,** one most parsimonious tree was found requiring 608 steps. In this case, branch lengths are not drawn to scale. Bootstrap confidence levels are shown as percentages (based on 1,000 bootstrap samples in the neighbor-joining tree and 100 samples in the PROTPARS tree). Both trees were rooted using the *Dictyostelium* sequence.

of the tree. The class II domains have diverged more than the class I domains, and this is reflected in the way most of the domains branch very close to the center. The exact branching order at the center cannot be inferred with any accuracy, and in general, the tree is less stable than the one for the class I domains. Despite this reservation, the domains from the super repeat regions of titin and twitchin group into their respective classes.

All of the twitchin and projectin domains from the super repeat region group together, with the exception of domains w8 and w11. Mixed in with these is one other twitchin domain (w39). All except one of the titin domains from the super repeat lie in three groups, corresponding to the three positions in each 11-domain repeat. The only exception is domain 9, which is expected to group with the domains at position VII. Two of the first three class II domains from the C-terminal sequence (m7 and m10) group with the rest of the super repeat domains. The rest of the twitchin and titin domains are spread around the tree along with the three smMLCK domains (smMLCK-1, -2, and -4) and the six domains from the chicken C-protein (c1–c4, c7, and c9).

The Kinase Catalytic Domains

Figure 4 shows two trees for the kinase domains from titin, twitchin, and chicken smMLCK along with those from two mammalian skeletal muscle MLCK proteins and one from *Dictyostelium.* These domains are very similar; only four gaps were needed to align them (data not shown). The trees are rooted using the *Dictyostelium* sequence. This is required on taxonomic and evolutionary grounds, while, in addition, it places the root along the longest branch in the neighbor-joining tree. Bootstrap confidence levels are shown for each internal branch. In all cases except for the grouping of the two mammalian skeletal muscle MLCK domains, the figures are low. Therefore, the branching orders must be treated with great caution. If the branching order in either tree is accepted uncritically, then it implies that the common ancestor of the twitchin and smMLCK domains is more recent than the common ancestor of the twitchin and titin domains, a rather unexpected result. This implies that the vertebrate MLCKs derived from titin/twitchin by one or two truncation events (each requiring two deletions: One each from the 5′ and 3′ ends of the kinase domain). In the neighbor-joining tree, two separate series of truncation events are required to explain the tree topology: One leading to chicken smMLCK and one to the skeletal muscle MLCKs. In the maximum parsimony tree, the vertebrate MLCKs form one group, and only one truncation event (of two deletions) is required to explain the topology.

Discussion

There are two questions that one can attempt to answer using the data just presented. First, what was the series of events that gave rise to the repeat patterns found in each protein? Second, what was the order of appearance of the different proteins? One can use the results from the three trees and also use the pattern of the repeats in the different proteins to obtain possible answers. Unfortunately, the within- and between-protein duplication events happened early and rapidly in the evolution of the family. The speed of the duplication can be seen from the densely packed branches at the center of each of the two large trees. The species distribution of the various members shows that the earliest events must have occurred at least as early as the common ancestor of vertebrates and nematodes—presumably during the original development of organized muscle tissue. Therefore, it is not yet possible to give very detailed answers to these questions but the main events can be discerned.

Evolution of the Super Repeats

The most spectacular events in the evolution of the family gave rise to the super repeats in titin and twitchin/projectin. The limited data from projectin agree

with the proposal (Ayme-Southgate et al. 1991) that it is the *Drosophila* homologue of twitchin. The results from the class I and class II trees show that the duplication events happened independently in the two proteins. In each case, the individual domains group into sets corresponding to position in the super repeats. The super repeat duplicated from one original repeat unit, three domains long in twitchin and projectin and 11 domains long in titin. It is intriguing that such a massive internal duplication occurred in such a similar way in the two proteins. Further, as will be shown below, it appears that some of the domains involved in the duplication were the same in the two proteins. That the two duplications happened independently can also be inferred from the repeat patterns. The alternative hypothesis, that the super repeat in one protein was derived from that in the other protein by consistent insertion or deletion of domains along the protein, is very unlikely.

Where did the super repeats in titin and twitchin come from? The titin super repeat extends toward the C-terminus as far as domain m12 or m13. Both the repeat pattern and the observed grouping of most of the individual domains from m1 to m13 with the other super repeat domains confirm this. Within the main super repeat region (domains 1–69), the most diverged domains in each super repeat group are the most N-terminal ones while the most recently duplicated are toward the C-terminus. This can be seen by examining the titin super repeat groupings in Fig. 2. The first domain to branch off from the root in seven out of eight of the groups is the most N-terminal (lowest number) while the rest of the domains appear in roughly rising sequence, with the last three usually the closest. This shows that the duplication of super repeat units happened mainly by successive addition of extra units to the C-terminal end of the super repeat region. The twitchin super repeat extends as far as domains w34 or w25 but cannot reliably be traced any further toward the C-terminus. The N-terminal domains of the twitchin super repeat are the oldest (domains w6, w7, and w9) as these are the first to branch away from the root of their repeat groups on the tree, followed by the C-terminal ones (w31, w30, and w33; w33 branches in another part of the tree). The middle domains (w13–28) have the shortest branch lengths and therefore, in contrast to titin, duplication of super repeats occurred mainly in the center of the molecule. This result confirms the separate development of the titin and twitchin super repeats.

The individual domains in the original super repeats also arose by duplication, possibly of domains from the C-termini of the proteins. It is more difficult to tell where these original domains came from. This was an older event than the duplication of the super repeats and is confounded with the duplications that gave rise to the C-terminal pattern of repeats in the first place. One clue to the origin of the first super repeats comes from an examination of the groupings of the major repeat classes on the class I tree in Fig. 2. There are two groups of twitchin class I super repeat domains corresponding to the first and second domains in each repeat. There are three domains in each titin super repeat, immediately C-terminal to a class II domain: Positions I, V, and VIII. These three domain classes group with the domains at the first position in the twitchin super repeat. The non-super repeat domains from twitchin that are to the right of a class II domain are also all in this part of the tree (domains w40, w43, w48, and w52) except for domain w36. Further, all of the non-super repeat titin domains that are to the right of a class II domain are grouped here (domains m15, m20, and m24). The titin repeats at positions II and IX group with the domains at the second position in each twitchin super repeat. Both of these titin positions are two domains to the right of a class II domain. Domains w37 and w41 from twitchin are also two domains to the right of a class II domain and group here, as do domains w38 and w45, which lie three domains to the right. Domain w49 groups just outside the above cluster along with the three C-protein domains. The other three titin super repeat classes (positions III, VI, and X) are all immediately to the left of a class II domain and join the center of the tree separately, although at adjacent branch points.

The above grouping of super repeat domains from titin and twitchin is strong evidence that all of the twitchin and five of the titin class I repeats from each super repeat derive from just two original domains. Most of the C-terminal class I domains also derive from the same two original repeats. Possible candidate domains for this origin are the last two class I domains in each protein. Domains w52 and m24 are the last class I domains in twitchin and titin, respectively. They form a group with domain m20 on the tree. This group lies immediately outside of one of the large groupings of super repeat domains: Those immediately to the right of a class II domain. Why do these class I domains not only group together but also have the same orientation with respect to class II domains? The only reasonable explanation is that three domains consisting of two class I and one class II duplicated originally to give rise to this part of the super repeat. The class II tree confirms the possibility that domains m29/w51 and m24/w52 duplicated originally. Domains w51 and m23 are at exactly the same positions in twitchin and titin and these lie just outside of the large grouping of the class II super repeat domains.

Figure 5 shows a possible scheme for the initial duplications that gave rise to the original titin super repeat. Starting with a twitchinlike repeat unit of two class I and

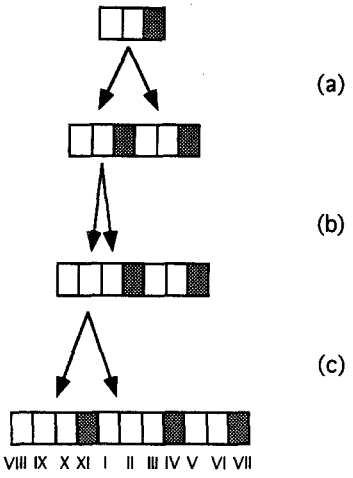

Fig. 5. Scheme for the evolution of the original titin super repeat by a series of three duplication events. Starting with an initial twitchinlike unit of two class I domains and one class II domain, the duplication at **a** gives a unit six domains long. One of the class I domains then duplicated in **b** and finally, a fragment with three class I domains and one class II domain duplicated in **c** to give the final super repeat.

one class II domains, it is possible to derive an 11-long repeat by a series of just three duplication events. The data from the class I and class II trees provide support for parts of this scheme but the exact details remain vague as they depend on knowing the exact branching order between the super repeat groupings and the roots of the trees. Evidence for this scheme comes from the way the closest relationships between titin super repeat domains are between class II domains IV and XI and between class I domains I and VIII and then II and IX. These groupings reflect the last duplication event in the scheme. The scheme also requires domain positions III and X to group. While these groups are near each other on the class I tree, they may be separated by the root of the tree.

Order of Appearance of the Proteins

The results from the kinase domain trees in Fig. 4 were disappointing with regard to the accuracy with which the topology could be predicted. This was potentially the most direct way of estimating the order of appearance of the proteins. If the trees do reflect the true history of the group, then the vertebrate MLCKs are derived from titin/twitchin by one or two sets of truncation events. An examination of the arrangement of repeats from smMLCK in the two large trees supports this hypothesis. The three class II domains from smMLCK are not grouped together on the class II tree. If smMLCK is similar to the common ancestor of the group, then its domains should group together on a long branch, reflecting their original duplication from each other, before the origin of titin/twitchin. This branch would in fact provide a root for the tree.

In Fig. 6, a possible scheme for the evolution of all of the proteins with a kinase domain is shown. Hypothetical intermediate proteins are shown at the internal nodes. The scheme starts with an MLCK which contains a kinase domain but no class I or class II domains. This is still found in *Dictyostelium.* This protein acquired a class I and a class II domain which duplicated internally to give rise to a protein with the same pattern of domains found around the kinase domain in titin and twitchin, i.e., a protein with approximately six class I and eight class II domains. This hypothetical protein is at positions (a) and (b) in Fig. 6. It duplicated internally to produce the super repeat in titin. Later, the super repeat in twitchin (and projectin) arose by a similar but independent process. Finally, the vertebrate MLCKs arose from the common ancestor of titin and twitchin by a series of truncation events.

The origin of the C-protein is not clear. The available data cannot be used to distinguish between models involving C-protein as the ancestor of the entire group and models in which it is derived from parts of titin/twitchin. Its class I domains do group together on a long branch in Fig. 2, with domain w49 from twitchin. This is consistent with these domains having arisen from a duplication event in the C-protein, separate from those in titin/twitchin. Further, when domains from human fibronectin were used to try to provide a root for the class I tree, these three C-protein domains branched off first from the root, suggesting that they are the oldest in the tree. Its class II domains, however, are spread around the tree in Fig. 3 which suggests that they may be derived from several different titin/twitchin class II domains. Again, it is quite possible that the domains from C-protein have diverged too much to be placed correctly.

Conclusions

The main conclusion from this discussion is that the extraordinary series of duplication events which gave rise to most of titin and twitchin (the two largest protein sequences to date) happened independently in the two proteins but apparently from the same ancestral class I and class II domains. The kinase domain trees suggest that titin diverged first in the family, followed by twitchin, and finally the vertebrate MLCKs. This scheme appears counterintuitive as it has the longest and most complicated protein appearing first. It is also interesting that the vertebrate skeletal muscle MLCKs and the *Dictyostelium* sequence are not directly related. Therefore, the *Dictyostelium* MLCK may not be an ap-

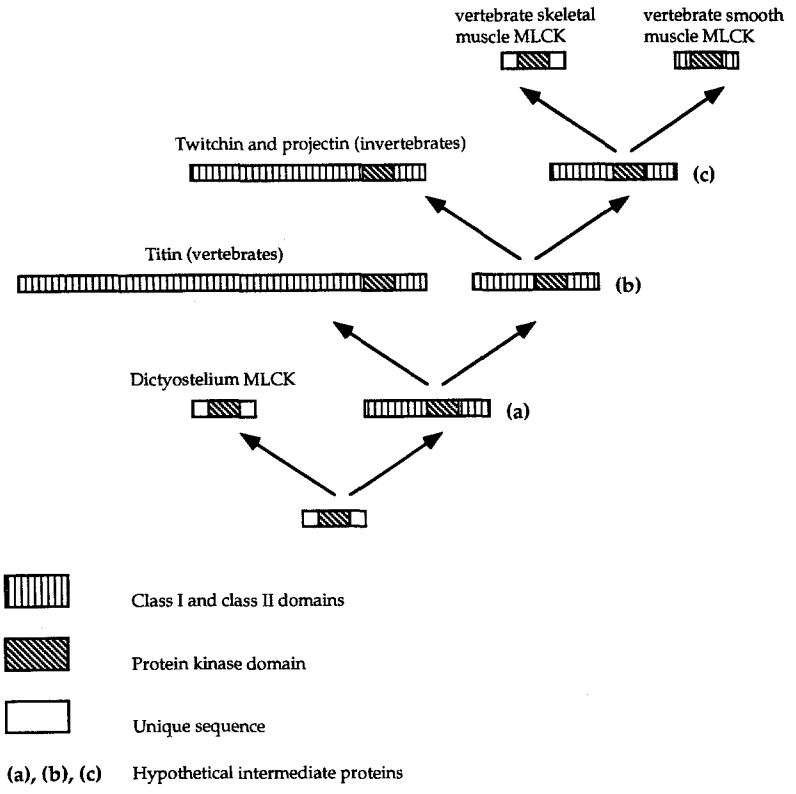

Fig. 6. Scheme for the evolution of the complete proteins. Three types of *shading* are used to indicate the presence of a kinase domain, repeated class I and class II domains, or unique sequence. The proteins are not drawn to scale. Hypothetical ancestral proteins are labeled **(a)**, **(b)**, and **(c)**.

propriate model for the vertebrate MLCKs, with respect to their function in muscle regulation.

A further unexpected result from the kinase trees is that twitchin, despite being known only from invertebrates, is on the branch leading to the vertebrate MLCKs. At one stage in the common ancestor of vertebrates and invertebrates, titin and twitchin must have coexisted, and indeed they could still do so in extant species. Specifically, titin could be found in invertebrates, co-existing with twitchin/projectin. Locker and Wild (1986) surveyed muscle tissue from a range of phyla for the presence of large proteins. They found proteins of the same molecular weight as titin in annelids, arthropods, and molluscs.

Independent roles have been proposed for titin and twitchin. Twitchin is implicated in the regulation of muscle contraction while titin has been proposed to be involved in sarcomere thick-filament assembly (Whiting et al. 1989), acting as a "molecular ruler." Twitchin may also have a limited "ruler" capacity. Titin kinase substrate is probably an M-line protein if it regulates sarcomere assembly, as predicted. Twitchin kinase substrate is probably myosin light-chain, given its role in muscle regulation and its apparent ancestry to vertebrate MLCKs. Therefore it is attractive to propose that titin and twitchin did arise by gene duplication and subsequently acquired new functions since this event. Finally, if smMLCK did arise from the large proteins by a series of truncation events, then this suggests that smooth muscle appeared later than skeletal muscle in the development of vertebrate muscle tissues, despite its simpler ultrastructure.

Acknowledgments. The authors thank Dan Graur, Belinda Bullard, Kevin Leonard, and John Trinick for commenting on an early version of the manuscript.

References

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

Bairoch A, Boeckmann B (1992) The SWISS-PROT protein sequence data bank. Nucleic Acids Res 20:2019–2022

Benian GM, Kiff JE, Neckelmann N, Moerman DG, Waterston RH (1989) Sequence of an unusually large protein implicated in regulation of myosin activity in *C. elegans*. Nature 342:45–50

Bork P, Doolittle RF (1992) Proposed acquisition of an animal protein domain by bacteria. Proc Natl Acad Sci USA 89:8990–8994

Dayhoff MO (1978) Survey of new data and computer methods of analysis. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5, suppl 3. National Biomedical Research Foundation, Washington, DC, pp 1–8

Einheber S, Fischman DA (1990) Isolation and characterisation of a cDNA clone encoding avian skeletal muscle C-protein: an intracellular member of the immunoglobulin superfamily. Proc Natl Acad Sci USA 87:2157–2161

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Felsenstein J (1989) PHYLIP: phylogeny inference package (version 3.2). Cladistics 5:164–166

Fyrberg CC, Labeit S, Bullard B, Leonard K, Fyrberg E (1992) *Drosophila* projectin: relatedness to titin and twitchin and correlation with lethal (4)102 CDa and bent-dominant. Proc R Soc Lond B 249:33–40

Fürst DO, Osborn M, Weber K (1989) Myogenesis in mouse embryo: differential onset of myogenic proteins and the involvement of titin in myofibril assembly. J Cell Biol 109:517–527

Gibson TJ, Higgins DG (1992) Myosin light-chain kinase: no end in sight. Submitted to J DNA Seq Map

Hanks SK, Quinn AM, Hunter T (1988) The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science 241:42–52

Herring BP, Stull JT, Gallagher PJ (1990) Domain characterisation of rabbit skeletal muscle myosin light-chain kinase. J Biol Chem 265: 1724–1730

Higgins DG, Bleasby AJ, Fuchs R (1992) Clustal V: improved software for multiple sequence alignment. Comput Appl Biosci 8:189–191

Higgins DG, Fuchs R, Stoehr PJ, Cameron GN (1992) The EMBL data library. Nucleic Acids Res 20:2071–2074

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kornblihtt AR, Umezawa K, Vibe-Pedersen K, Baralle FE (1985) Primary structure of human fibronectin: differential splicing may generate at least 10 polypeptides from a single gene. EMBO J 4:1755–1759

Kurzban GP, Wang K (1988) Giant polypeptides of skeletal muscle titin: sedimentation equilibrium in guanidine hydrochloride. Biochem Biophys Res Commun 155:1155–1161

Labeit S, Barlow DP, Gautel M, Gibson T, Holt J, Hsieh CL, Francke U, Leonard K, Wardale J, Whiting A, Trinick J (1990) A regular pattern of two types of 100-residue motif in the sequence of titin. Nature 345:273–276

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

Lakey A, Ferguson C, Labeit S, Reedy M, Larkins A, Butcher G, Leonard K, Bullard B (1990) Identification and localisation of high molecular weight proteins in insect flight and leg muscles. EMBO J 9:3459–3467

Locker RH, Wild DJC (1986) A comparative study of high molecular weight proteins in various types of muscle across the animal kingdom. J Biochem 99:1473–1484

Maruyama K, Kimura S, Yoshidomi H, Sawada H, Kikuchi M (1984) Molecular size and shape of beta-connectin, an elastic protein of striated muscle. J Biochem 95:1423–1493

Olson NJ, Pearson RB, Needleman DS, Hurwitz MY, Kemp BE, Means AR (1990) Regulatory and structural motifs of chicken gizzard myosin light-chain kinase. Proc Natl Acad Sci USA 87:2284–2288

Roush CL, Kennelly PJ, Glaccum MB, Helfman DM, Scott JD, Krebs EG (1988) Isolation of the cDNA encoding rat skeletal muscle myosin light-chain kinase. J Biol Chem 263:10510–10516

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Shoemaker MO, Lau W, Shattuck RL, Kwiatkowski AP, Matrisian PE, Guerra-Santos L, Wilson E, Lukas TJ, van Eldik LJ, Watterson DM (1990) Use of DNA sequence and mutant analyses and antisense deoxynucleotides to examine the molecular basis of nonmuscle myosin light-chain kinase autoinhibition, calmodulin recognition, and activity. J Cell Biol 111:1107–1125

Tan JL, Spudich JA (1991) Characterisation and bacterial expression of the *Dictyostelium* myosin light-chain kinase cDNA: identification of an autoinhibitory domain. J Biol Chem 266:16044–16049

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