The Myofibrillar Protein, Projectin, is Highly Conserved Across Insect Evolution Except for Its PEVK Domain

Agnes J. Ayme-Southgate · Richard J. Southgate · Richard A. Philipp · Erik E. Sotka · Catherine Kramp

Received: 30 April 2008 / Accepted: 8 October 2008 © Springer Science+Business Media, LLC 2008

Abstract All striated muscles respond to stretch by a delayed increase in tension. This physiological response, known as stretch activation, is, however, predominantly found in vertebrate cardiac muscle and insect asynchronous flight muscles. Stretch activation relies on an elastic third filament system composed of giant proteins known as titin in vertebrates or kettin and projectin in insects. The projectin insect protein functions jointly as a “scaffold and ruler” system during myofibril assembly and as an elastic protein during stretch activation. An evolutionary analysis of the projectin molecule could potentially provide insight into how distinct protein regions may have evolved in response to different evolutionary constraints. We mined candidate genes in representative insect species from Hemiptera to Diptera, from published and novel genome sequence data, and carried out a detailed molecular and phylogenetic analysis. The general domain organization of projectin is highly conserved, as are the protein sequences of its two repeated regions—the immunoglobulin type C and fibronectin type III domains. The conservation in structure and sequence is consistent with the proposed function of projectin as a scaffold and ruler. In contrast, the amino acid sequences of the elastic PEVK domains are noticeably divergent, although their length and overall unusual amino acid makeup are conserved. These patterns suggest that the PEVK region working as an unstructured domain can still maintain its dynamic, and even its three-dimensional, properties, without the need for strict amino acid conservation. Phylogenetic analysis of the projectin proteins also supports a reclassification of the Hymenoptera in relation to Diptera and Coleoptera.

Keywords Projectin · Titin · Elastic protein · Modular protein · Muscle · Flight · Insect evolution · PEVK

Introduction

Insect flight muscles can be classified as either synchronous or asynchronous based on differences in their physiology and ultrastructure (Pringle 1949, 1978, 1981). For example, synchronous muscles display a characteristic 1:1 ratio between nervous excitation and contraction, whereas asynchronous indirect flight muscles (IFM) exhibit multiple contractions for each nerve impulse. The asynchronous mode depends on a delayed increase in tension brought about by stretch (stretch activation), a physiological response associated with the high resting stiffness of muscle fibers (reviewed by Josephson et al. 2000; Moore 2006). Although observed in nearly all muscles, stretch activation is uniquely enhanced in insect asynchronous flight muscles and vertebrate cardiac muscles.

Several molecular models of stretch activation have implicated the insect C-filament system and its vertebrate counterpart (titin filaments) in providing the high resting stiffness necessary for sensing and transducing stretch to the actin-myosin lattice (Granzier and Wang 1993; Kulke et al. 2001; Vigoreaux et al. 2000; Fukuda and Granzier 2005; reviewed by Trombitas 2000; Moore 2006). The current model for C-filaments in Drosophila proposes
primarily two constituent proteins: kettin (and its longer isoform Sallimus [Sls]) and projectin (Bullard et al. 2000, 2005, 2006). Consistently, flies containing mutations in either projectin or Sls are, respectively, flight impaired and flightless (Moore et al. 1999; Hakeda et al. 2000; Kolmerer et al. 2000).

Projectin is a sizable (~1000-kDa) myofibrillar protein (Saide et al. 1989; Lakey et al. 1990; Trombitas 2000; Bullard et al. 2005) composed largely of two repeated motifs—fibronectin III (FnIII) and immunoglobulin (Ig) — arranged in a regular pattern [Fn-Fn-Ig] within its central core region (Ayme-Southgate et al. 1991, 1995; Fyberg et al. 1992; Daley et al. 1998). In contrast, the COOH and NH2 termini contain Ig motifs yet lack FnII domains. FnIII and Ig domains have been implicated in numerous examples of protein-protein interactions (reviewed by Gautel 1996), and the projectin protein is, therefore, proposed to serve as a scaffold for myofibril assembly through its interactions with the myosin filament. Projectin also includes several different nonrepeated amino acid sequences. One of these was initially characterized as analogous to titin PEVK because of its atypically high percentage of proline, glutamic acid, valine, and lysine (Southgate and Ayme-Southgate 2001). The PEVK domain is located within the NH2 terminus of the protein between two separate segments composed of eight and six Ig domains, respectively. In D. melanogaster IFM, projectin is aligned along the sarcomeric unit with its NH2 terminus embedded within the Z band, while its central core region is likely associated with the myosin filament (Ayme-Southgate et al. 2000, 2005). This orientation suggests that the projectin PEVK domain, together with some of the NH2-terminal Ig domains, can physically span the entire I band (Ayme-Southgate et al. 2005; Bullard et al. 2005). The presence of a PEVK region and the distribution of PEVK and Ig domains over the I band of IFM muscles are consistent with projectin serving as an elastic protein during stretch activation.

However, the molecular mechanism(s) underlying length-dependent stretch activation is(are) still not fully understood. In particular, several independent studies have been unable to correlate the protein composition of the myofibrils with the ability to generate stretch activation. For example, the protein arthrin was originally characterized only as an asynchronous muscle myofibril protein (Bullard et al. 1985), yet subsequently a more thorough investigation did not consistently find arthrin within all insect orders that are associated with asynchronous muscles (Schmitz et al. 2003). Troponin C is another such potential candidate where one isoform has been found to be specific to asynchronous flight muscle and sensitive to stretch in Lethocerus, Drosophila, and Anopheles (Qiu et al. 2003). However, a thorough investigation of more basal insect orders has not been performed for this protein.

Projectin cannot be thought of as an asynchronous muscle-specific protein, as it is found in both synchronous and asynchronous insect muscles and crustacean muscles (Nave and Weber 1990; Vigoreaux et al. 1991; Oshino et al. 2003) and is considered an orthologue of the C. elegans myofibrillar protein, twitchin (Benian et al. 1989). However, there are several significant differences in the projectin isoforms present in asynchronous versus synchronous muscle types of D. melanogaster. The localization of projectin differs, as it is immunofluorescently localized exclusively over the A band in synchronous muscles but is found over the I band in asynchronous muscles (Vigoreaux et al. 1991). The IFM isoform is also notably shorter, most likely as a consequence of alternative splicing within specific regions of the transcript, in particular, the PEVK domain (Southgate and Ayme-Southgate 2001), and size variants of titin PEVK specifically correlate with differences in the resting tension of vertebrate muscle fibers (Linke et al. 1999; Freiburg et al. 2000; Granzier et al. 2000; Fukuda and Granzier 2005; reviewed by Granzier and Labeit 2004). Here we conducted an evolutionary analysis using both published and novel genomic sequence data to determine the changes in projectin sequences across several insect orders, to provide insight into how different regions of the protein may have changed under different evolutionary constraints related to the various functions attributed to projectin.

Materials and Methods

Gene Query and Manual Annotation

BLAST searches (Altschul et al. 1990, 1997) using sequences from the D. melanogaster core region (containing both Ig and Fn domains in a regular [Fn-Fn-Ig] pattern) were used to query available genome assemblies, contigs, or trace archives. Apis mellifera, Tribolium castaneum, and Anopheles gambiae genomes were at the stage of annotated genomes when the study began, whereas the Drosophila virilis, Drosophila ananassae, Drosophila pseudoobscura, Aedes aegypti, and Nasonia vitripennis genomes were assembled but not annotated. The following annotated genes/contigs were initially retrieved: A. mellifera, ENSAPMG00000008141 and ENSAPMP 00000014257; T. castaneum, CM000276.1/ GLEAN_04721; A. gambiae, ENSANGG00000014893; A. aegypti, LOCUS: AAGE02003896; D. virilis, scaffold_13052; D. pseudoobscura, group 8; and N. vitripennis, scaffold 113. In many cases, genomic data retrieved from annotated genomes include the predicted splicing pattern of the candidate gene and its derived translation data. Predicted cDNA and amino acid sequences were compared against
the \textit{D. melanogaster} amino acid sequences to identify missing domains or incorrect splice patterns. Regions not included in the initial annotation were manually annotated from available surrounding genomic sequences.

The \textit{Acyrthosiphon pisum} genome was completed, but not assembled, and only trace archives were available. The \textit{A. pisum} trace results from BLAST searches were assembled into contigs using the Vector NTI software (Invitrogen), and overlapping contigs were joined to generate a semicontiguous genomic sequence when possible. Assembled contig sequences were translated in all three forward frames and translation results were aligned against \textit{D. melanogaster} amino acid sequences using LaLign (Huang and Miller 1991). This allowed for manual annotation of the putative exon-intron splicing pattern over most of the gene. However, the PEVK domain could not be resolved using this approach.

**RNA Extraction and RT-PCR Sequencing**

Total RNA was purified from whole (live or frozen) animals, using Trizol (Invitrogen). \textit{Apis mellifera} was obtained from Dr. J. Evans (USDA, Beltsville, MD); \textit{Tribolium castaneum}, from Dr. S. Brown (Kansas State University, Manhattan, KS); \textit{Nasonia vitripennis}, from Dr. J. Werren (University of Rochester, Rochester, NY); \textit{Acyrthosiphon pisum}, from Dr. D. Stern (Princeton University, Princeton, NJ); and \textit{D. virilis} and \textit{D. pseudooobscura}, from the Tucson \textit{Drosophila} stock center. Isolated RNAs were used in RT-PCR reactions (Superscript One-Step RT-PCR mix; Invitrogen) with gene-specific primers designed from potential exons or predicted open reading frames (ORFs) as previously described (Southgate and Ayme-Southgate 2001). PCR-amplified cDNA products were separated by gel electrophoresis and either sequenced directly or subcloned into the pGEM-T-easy vector (Promega Inc.; DNA Core Facility, Medical University of South Carolina, Charleston, SC). The cDNA sequences were manually aligned by comparison to the genomic sequences to identify and verify putative splice site positions.

**Bioinformatics Analysis of Projectin Sequences**

Multiple sequence alignments (MSAs) were performed using the CLUSTALW algorithm (Thompson et al. 1994, 1997) and checked by eye within Jalview. Gaps originating from differences in the domain pattern, for example, the missing exon in mosquitoes, or gaps representing incomplete analysis, such as the PEVK domains in some species, were manually removed from the alignment when the entire sequences were compared. Two methods—maximum parsimony using ProtPars (Felsenstein 1989, 1996) and maximum likelihood using PhyML (Guindon and Gascuel 2003; Guindon et al. 2005)—were used to derive an evolutionary tree(s). ProtTest analysis determined that the most appropriate model of evolution was RtREV (Abascal et al. 2005, 2007), yet for phylogenies we expanded this to include additional models of evolution (JTT [Jones et al. 1992], RtREV [Dimmic et al. 2002], and WAG [Whelan and Goldman 2001]). The robustness of parsimony-based and likelihood-based phylogenies was assessed using 1000 to 2000 bootstrap replicates and summarized using Consense (Phylip package [Felsenstein 1989]).\textit{C. elegans} twitchin (Benian et al. 1989) or crayfish (\textit{Procambarus clarkii}) projectin (Oshino et al. 2003) sequences were used as outgroup sequences in the phylogenetic trees and the tree output was visualized using TreeView (Page 1996). Additionally, diagonal graphical alignments between pairs of sequences were generated using Dotlet (Junier and Pagni 2000) and manually evaluated.

**Results**

**Characterization of the Projectin Gene in Different Insect Species**

The projectin gene was identified in the genomes of the following insects for which genome projects were either under way or completed: Diptera—\textit{Drosophila virilis}, \textit{Drosophila ananassae}, \textit{Drosophila pseudoobscura}, \textit{Anopheles gambiae} (mosquito/malaria), and \textit{Aedes aegyptii} (mosquito/yellow fever); Hymenoptera—\textit{Apis mellifera} (honeybee) and \textit{Nasonia vitripennis} (jewel wasp); Coleoptera—\textit{Tribolium castaneum} (red flour beetle); and Hemiptera—\textit{Acyrthosiphon pisum} (pea aphid). All of these insects possess asynchronous flight muscles.

Typically, BLAST searches were done with a sequence from the projectin core region of \textit{D. melanogaster} (containing both Ig and FnIII domains) to query available genome assemblies, contigs, or trace archives (see Materials and Methods). We determined whether candidate sequences identified using this approach were for projectin orthologues rather than for other Ig domain-containing proteins (such as stretchin or kettin/Sls [Champagne et al. 2000; Kolmerer et al. 2000]) by verifying that they contain the [Fn-Fn-Ig] characteristic pattern of the core region of projectin. For example, kettin does not possess any FnIII domains and its longer Sls isoform contains only a few FnIII domains near its COOH terminus.

As a result of this initial search we retrieved the \textit{A. mellifera} and \textit{A. gambiae} projectin orthologues, but determined that the sequences equivalent to the \textit{D. melanogaster} NH$_2$-terminus and PEVK domains were not
included in the initial gene builds. We retrieved adjacent 5’ genomic sequences, and the exon-intron pattern of the NH₂
terminus was manually annotated by alignment of the
translation data with D. melanogaster sequences, together
with matches to Expressed Sequence Tag (EST) sequences
when available. If necessary, the predicted exon-intron
pattern was confirmed by RT-PCR analysis (see Materials
and Methods). The D. virilis, D. pseudoobscura, D. anan-
assae, A. aegyptii, and N. vitripennis genomes were
assembled, but not annotated, so we manually annotated
the exon-intron splicing of all regions except for the PEVK
domain using alignment to the Drosophila orApis (for
Nasonia) amino acid sequences. Any ambiguities were
further resolved by RT-PCR analysis (see Materials and
Methods). For the A. pisum genome, only trace archives
were available, and BLAST searches returned a total of
315 trace archives, which were assembled to generate most
of the A. pisum projectin gene (see Materials and Methods).

There are still gaps for which we could not retrieve any
traces, but all of these missing sequences are found within
introns. Since the start of this project, the first assembly of
the A. pisum genome has been released, and the projectin
genomic sequence we generated is identical to the genomic
sequence found in the assembly (some of the gaps still
exist). As before, the exon-intron splicing of all regions
except for the PEVK domain was established using align-
ment of EST matches to the genomic sequence and
comparison of the translation products for homology to
other projectin sequences. Ambiguities were further
resolved by RT-PCR analysis (see Materials and Methods).

Following this initial annotation, genomic sequences
found between Ig8 and Ig9 in each of these genes were
considered potential PEVK regions. Because of low
homology (see below) they could not be manually anno-
tated, and the coding regions were identified by RT-PCR as
previously performed in D. melanogaster (Southgate and
Ayme-Southgate 2001).

Characterization of the projectin genes and their exon-
intron splicing is complete in five insect species repre-
senting four orders: Acyrtosiphon pisum (pea aphid;
Hemiptera), Tribolium castaneum (red flour beetle; Cole-
optera), Apis mellifera (honeybee; Hymenoptera), Nasonia
vitripennis (jewel wasp; Hymenoptera), and Drosophila
virilis (Diptera). We also have almost-complete charac-
terization in Drosophila pseudoobscura, Drosophila
ananassae, Anopheles gambiae, and Aedes aegyptii, where
the entire gene except for parts of the PEVK domain has
been annotated.

General Motif Pattern Within the Projectin Proteins

All characterized projectin genes, apart from the two
mosquito genes, contain 39 copies of both the Ig and the
FnIII domains, as previously identified in D. melanogaster.
Multiple alignments with D. melanogaster sequences
demonstrate that the Ig18-Fn9-Fn10 block is missing from
the core domain of A. gambiae and A. aegyptii genes (data
not shown). In Drosophila species and in T. castaneum this
module is in fact a one-exon entity, allowing for the loss of
these three domains without actually affecting the ORF.

The loss of this specific exon must have occurred at some
point after the Drosophila and Culicidae lineages
separated.

In all studied genes, the Ig and Fn domains are orga-
nized in the same basic pattern and order, as in
D. melanogaster (see Fig. 1). This arrangement defines five
distinct regions within the protein. In the NH₂-terminal the
PEVK region separates two tracts of Ig domains containing
eight (N-8Ig) and six (N-6Ig) Ig domains, respectively. The
core is composed of 14 repeats of the [Fn-Fn-Ig] module.
Finally, the intermediate region has a conserved, but non-
modular arrangement of FnIII and Ig domains, while the
COOH terminal contains five Ig domains (C-Ig). A similar
pattern has also been found in groups more ancestral than
the insect lineage, including (i) the crayfish (P. clarkii)
sequence, which is similar except for the presence of seven
rather than eight Ig domains in the initial NH₂-terminal
region (Oshino et al. 2003; see below), and (ii) the
C. elegans twitchin protein, which has a similar arrange-
ment, but with a lower number of the two domains in all its
regions (Benian et al. 1989, 1993). The projectin protein
contains two other nonrepeated regions, the kinase and the
PEVK domains (Fig. 1; see below), which are not com-
poved of either Ig or FnIII motifs.

Exon-Intron Pattern of the Projectin Genes

Because of the modular nature of the projectin protein, the
relationship between exons and the boundaries of the Ig
and FnIII domains was examined. Even though there are
elements of a single domain encoded by a single exon in
all species studied, such instances are more frequent in A.
pisum (23 examples) and A. mellifera (24 examples) than in
T. castaneum (14 examples) or Drosophila sp. (only 3
examples). In A. pisum, single domains are frequently split
between 2 exons (25 domains split between two exons),
with seven such occurrences in A. mellifera and only 2
domains split between two exons in Drosophila sp. Con-
versely, no exons in A. pisum contain more than one
complete domain, whereas six such exons are present in A.
mellifera containing 2–4 complete domains, and 32
domains (both Ig and FnIII) are included within the largest
exon of Drosophila sp. (Supplemental Materials 1).

Because of its remarkably conserved protein organiza-
tion, the complete projectin cDNA sequences of the insects
characterized in this study are all similar in size, ranging
from 26 to 27 kb in length (Table 1). In contrast, the size of the genomic sequences differs considerably (from 51 to > 70 kb), a difference attributable to both the number and the size of the introns (Supplemental Materials 1). As summarized in Table 1, the number of exons is variable, ranging from 41 in D. virilis to 144 in A. pisum, reflecting a reduction in the number of introns as one progresses from more basal (A. pisum) to more derived (Drosophila sp.) insect lineages.

Most of the documented intron losses occurred within the core and intermediate regions of the gene. For insects that are more closely related, such as the different Drosophila species and the A. mellifera/N. vitripennis pair, the number of exons and the position of the splice sites are more similar, even when the size of their corresponding introns differs significantly (Supplemental Materials 1).

A thorough comparison of the position of splice sites throughout the entire projectin genes was carried out. As an example, we present an analysis of the genomic sequence encoding the N-6Ig region and one scenario (of several alternatives) by which the observed pattern of intron losses could have occurred (Fig. 2). In this genomic region there are 10 introns in the A. pisum gene; 8 of them are still found in the hymenopteran genes but only 5 are still present in both T. castaneum and the Diptera (with only 4 in A. aegyptii). As a possible alternative, for example, intermediate 1 in Fig. 2 could be (i) more ancestral to A. pisum, and (ii) A. pisum could have acquired two additional introns within the Ig12- and 13-encoding exons. Regardless of the true evolutionary scenario, the splicing pattern of the two hymenopteran genes indicates that the positions of certain splice sites (in particular, between exons encoding Ig10 and 11 and between exons encoding Ig12) are identical to the splice sites found in A. pisum, suggesting that additional intron removal events would have occurred to yield the intron patterns in T. castaneum and in Diptera (Fig. 2). This observation holds true when other regions of the projectin genes are examined (data not shown).

Until recently, accepted lineages of insect orders indicated that the Coleoptera was ancestral to the lineage that led to Hymenoptera and Diptera (e.g., Wheeler et al. 2001; reviewed by Beutel and Pohl 2006) (Fig. 3b). Based on our observations, if this phylogeny was correct, the exact same exon fusion events would have had to occur independently in Coleoptera and in Diptera after the divergence from Hymenoptera. Alternatively, persistence of these “ancestral” introns in the available Hymenoptera projectin genes could indicate that the relative position of the Hymenoptera

### Table 1 Gene data: summary of the main information on gene and cDNA sizes, as well as the number and size of exons, for the insect projectin genes described in this study together with the Drosophila melanogaster gene

<table>
<thead>
<tr>
<th></th>
<th>Dmel</th>
<th>Dvir</th>
<th>Dpse</th>
<th>Agam</th>
<th>Aaeg</th>
<th>Nvi</th>
<th>Amel</th>
<th>Tcas</th>
<th>Api</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene size (bp)</td>
<td>51,073</td>
<td>51,490</td>
<td>&gt;39,641</td>
<td>&gt;31,000</td>
<td>&gt;58,353</td>
<td>48,634</td>
<td>66,952</td>
<td>37,040</td>
<td>&gt;70,000</td>
</tr>
<tr>
<td>cDNA size (bp)</td>
<td>27,154</td>
<td>26,867</td>
<td>24,338</td>
<td>24,389</td>
<td>23,809</td>
<td>24,215</td>
<td>26,547</td>
<td>26,012</td>
<td>26,184</td>
</tr>
<tr>
<td>No. of exons</td>
<td>46</td>
<td>45</td>
<td>&gt;40</td>
<td>&gt;27</td>
<td>&gt;27</td>
<td>&gt;27</td>
<td>&gt;27</td>
<td>&gt;27</td>
<td>&gt;27</td>
</tr>
<tr>
<td>Largest exon (bp)</td>
<td>9,462</td>
<td>9,472</td>
<td>9,474</td>
<td>10,046</td>
<td>11,964</td>
<td>4,371</td>
<td>2,325</td>
<td>2,670</td>
<td>1,054</td>
</tr>
<tr>
<td>Smallest exon (bp)</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>No. of amino acids</td>
<td>8,923</td>
<td>8,847</td>
<td>&gt;8,258</td>
<td>&gt;7,957</td>
<td>&gt;8,068</td>
<td>8,912</td>
<td>8,730</td>
<td>8,379</td>
<td>8,607</td>
</tr>
<tr>
<td>Calculated MW</td>
<td>998,292</td>
<td>990,025</td>
<td>&gt;922,695</td>
<td>&gt;892,182</td>
<td>&gt;907,283</td>
<td>995,868</td>
<td>995,346</td>
<td>933,354</td>
<td>960,331</td>
</tr>
</tbody>
</table>

**Note:** MW, molecular weight. cDNA sizes are calculated from the initiation to the termination codons and do not include possible 5’ and 3’ UTR. NA and “>” indicate that the information for that gene is incomplete due to the lack of characterization of the PEVK domain in that species.
in the reconstruction of insect evolutionary history should be reconsidered, consistent with more recent studies (Savard et al. 2006; Krauss et al. 2008).

Phylogenetic Relation of the Projectin Proteins

To address these alternatives, we investigated the phylogenetic relationships among insect orders for which we had projectin sequences. The analysis was carried out using (i) the whole amino acid sequence for insects with completed PEVK data or (ii) individual protein regions such as the core, kinase, and C-terminal regions for all insects studied. We also conducted phylogenetic analysis using either *Apis mellifera* or *Nasonia vitripennis* sequences. These various alignments were evaluated under multiple models of evolution using maximum likelihood or
maximum parsimony analyses (See Materials and Methods for details; aligned sequences are presented in Supplemental Materials 2). Irrespective of the computational method, evolutionary model, or sequence combinations used, a concordant tree with nodal support of at least 80% was consistently generated (Fig. 3a). This phylogeny supports a more basal position of the Hymenoptera in relation to the Coleoptera that is consistent with recent reports by Savard et al. (2006) and Krauss et al. (2008) (Fig. 3c).

Duplication Within the N-8Ig Region

The presence of eight Ig domains within the first part of the NH$_2$ terminus of projectin has been reported only within insects. Projectin has been characterized in only one other arthropod, the crayfish ($P$. clarkii), which was completely derived from cDNA sequencing. Crayfish projectin contains only seven Ig domains (a number that can be further reduced to six by alternative splicing [Oshino et al. 2003]). This increase to eight Ig domains appears to have occurred very early in the insect lineage based on preliminary data from analysis of the projectin gene in the silverfish (Apterygote; order Zygoptera, Thysanura) (R. Southgate, unpublished observation). Maximum parsimony analysis indicates that crayfish Ig1 to Ig7 correspond to insect Ig1 to Ig7, making the Ig8 in insects the probable result of a duplication event (Fig. 4a). The six Ig domains of the N-6Ig region are present in both crayfish and insects, and maximum parsimony analysis indicates that crayfish Ig8 to Ig13 correspond to insect Ig9 to Ig14 (data not shown). The more ancestral protein, twitchin, in $C$. elegans contains a total of only nine Ig domains at its NH$_2$ terminus. Both maximum parsimony and maximum likelihood analyses indicate that twitchin Ig4 to Ig9 correspond to insect Ig9 to Ig14, whereas $C$. elegans Ig1 clusters with arthropod Ig4, $C$. elegans Ig2 with arthropod Ig5/6, and $C$. elegans Ig3 with arthropod Ig7/8 (Fig. 4a). This clustering leads to a possible pathway for domain duplications to explain the increase in the first Ig stretch from three domains in $C$. elegans twitchin to eight in insect projectin (Fig. 4b). In such a model, Ig5 and Ig6 would be the product of duplication and divergence from twitchin Ig2. Similarly, Ig7 and Ig8 would be the product of duplication from twitchin Ig3. Interestingly the tree seems to indicate that twitchin Ig3 is more closely related to insect Ig8 than arthropod Ig7. This may suggest that Ig7 diverged more than Ig8 after the duplication event that created the eighth domain in basal insects. The tree also suggests that arthropod Ig1 to Ig3 are the products of two successive duplications from either $C$. elegans Ig3 or Ig1, the split creating arthropod Ig1 and Ig2 probably being the last one to occur (Fig. 4).

Crayfish projectin is known to undergo alternative splicing within the N-7Ig region, where the first half of Ig5 can be spliced directly to the second half of Ig6, removing one exon as part of a larger intron and creating a new Ig domain hybrid between Ig5 and Ig6 (Oshino et al. 2003). Comparison of the splice site positions within the various insect genes reveals that this alternative splicing is still possible in the A.
pisum gene, as the exon-intron pattern is consistent with the formation of an in-frame hybrid Ig domain by alternative splicing. This possibility is lost, however, in more derived insects, even though the potential for other alternative splicing events still exists within this region.

Conservation Within the Ig, FnIII, and Kinase Domains

Multiple alignment analysis of FnIII and Ig domains from all the insect species included in this study indicates a very high degree of amino acid conservation across different species between Ig or FnIII domains found at identical positions within the protein, e.g., Ig1 in A. pisum is more similar to Ig1 in D. melanogaster than it is to Ig5 in A. pisum (see Figs. 5 and 6). The conserved residues between domains within one species tend to be the ones corresponding to the consensus positions as originally defined for the Ig and FnIII domains of C. elegans twitchin (Benian et al. 1989). Figure 5a represents the alignment of the first eight Ig domains within the N-8Ig region of A. pisum. It shows that, after excluding conserved consensus positions (26; gray positions), only 17 positions of 98 amino acids are conserved (including conservative substitutions) in at least five of the eight Ig domains (ocher positions in Fig. 5a). In Fig. 5b, the alignment for Ig1 in all studied insects indicates that, after excluding the consensus positions (23; gray positions), 50 positions of the total 98 amino acids are conserved (including conservative substitutions) across all 10 insect species (blue positions), and another 16 are shared by 6 of 10 insect species (red positions).

A similar analysis was performed for all 39 FnIII domains within all 10 studied insect species and Fig. 6a presents the comparison for T. castaneum as an example. The projectin FnIII domains were modeled manually on the representation of the titin FnIII fold, and the A-G β-sheets are indicated above the alignments in Fig. 6 (Amodeo et al. 2001). All 39 FnIII domains in all 10 insects were aligned against both the twitchin consensus (con-tw) and the titin consensus (con-ti) sequences. This alignment generates a slightly different consensus for insect projectin FnIII domains (new con). As described above for the Ig domains, there is an overall higher conservation across species for domains at similar positions (red and blue residues in Fig. 6b) than between FnIII domains within one species.

The repeated pattern in the core region is less complex in projectin than in titin, consisting of 14 simple [Fn-Fn-Ig] modules representing Fn1-28 (domains in the blue brackets...
The alignment presented in Fig. 6a reveals another interesting pattern of conservation specific for that region of the protein. All the even-numbered FnIII domains are missing a residue at position 7. Position 9 is usually a highly conserved proline, but only in odd-numbered domains. Position 10 is a proline in even-numbered, but a non-conserved residue in odd-numbered FnIII domains. The highly variable C strand and C'-E loop are highlighted in light blue and light pink, respectively. Blue brackets on the side of the alignment are for the odd- and even-numbered FnIII domains of the core region. Fig. 6b shows the alignment of Fn5 and Fn17 from all available projectin sequences in this study. Conserved amino acids are highlighted with different colors depending on whether or not these amino acids coincide with the conserved positions in the consensus sequence.
leucine in odd-numbered domains. Several other residues (e.g., cysteine at position 20) follow a similar pattern, where the conservation is higher between either odd- or even-numbered domains. This pattern does not extend to domains Fn29–Fn39, probably because they belong to the intermediate region of the protein (Fig. 1). This analysis holds true in all the analyzed insect projectins.

Two amino acid stretches at the center of the FnIII domains vary in both their length and their amino acid sequence among the 39 Fn domains except for the highly conserved tryptophan residue at position 50 (equivalent to W54 in the titin Fn3 model [Muhle-Goll et al. 2001]) and a hydrophobic residue at position 53. When the projectin fibronectin domains are modeled on the representation of titin Fn3 fold, these two variable clusters correspond to strand C′ (boxed in light blue in Fig. 6) and loop C′E (boxed in light pink in Fig. 6) (Amodeo et al. 2001; Muhle-Goll et al. 2001). Even though the residues in the C′ strand are very different from domain to domain within a species, when the FnIII domains at the same position in the 10 different insects are compared, there is a higher overall conservation of the residues (e.g., residues highlighted in blue and red in Fig. 6b). The C′E loop shows greater length variation (three to nine amino acids) and is less conserved even between domains at the same position within all 10 insects (Fig. 6b).

The kinase domain of all 10 insects is extremely conserved as shown in Fig. 7, with the conservation including amino acids other than the ones implicated in the different loops and pockets, such as the ATP and substrate binding sites, as well as the catalytic and activation loop (Conserved Domain Database [CDD] reference CD00180 [Marchler-Bauer et al. 2007]). In all species except for the two mosquitoes, the kinase region is encoded by three or more exons (see Supplemental Materials), leaving no possibility for alternative splicing to either remove or inactivate the kinase activity without actually changing the reading frame for the downstream COOH terminus of the protein.

Linker Unique Sequences

In the N-8Ig-encoding region, the Ig domains are linked by short unique sequences (between 6 and 11 amino acids long) that can be considered as “extensions” of the Ig domains. In a given species, the amino acid sequences of these various linkers are very different from one another, but they are well conserved between all characterized projectin proteins at a specific location, reminiscent of the pattern described above for the Ig and FnIII domains: higher conservation between species at a given position rather than between domains within one protein. In D. melanogaster and D. viridis, the linker between Ig1 and Ig2 is actually encoded by two alternatively spliced exons, as demonstrated by RT-PCR analysis (Southgate and Ayme-Southgate 2001; data not shown). The presence of two similar alternative exons has also been predicted from the sequence in D. ananassae and D. pseudoobscura, but not verified by RT-PCR. In D. melanogaster the alternative splicing is known to be muscle type specific, one exon (and therefore one linker) being IMF specific, whereas the other exon is used in all other synchronous muscles (Southgate and Ayme-Southgate 2001). The muscle type specificity of the alternative exons has not been confirmed by RT-PCR analysis in the other three Drosophila species. The presence of two alternatively spliced exons between Ig1 and Ig2 may well be specific to Drosophila sp., as examination of the genomic sequences between Ig1 and Ig2 in non-Drosophila insects has revealed the presence of only one of these two possible small exons.

In the N-6Ig-encoding region, there is only one such “linker” sequence present between Ig9 and Ig10 (Fig. 2). This linker is much longer, with 46 amino acids, and well conserved across insect species at both the DNA and the amino acid levels (average, 72% and 75.8%, respectively). We referred to this new unique sequence as the “FRAM” domain based on the 100% conservation of these four specific amino acids across all the species studied. This sequence is also present at the same relative position within crayfish projectin and even in C. elegans twitchin. It is not, however, found in titin, or other proteins as tested using tBLASTn, nor does the FRAM sequence match any consensus domain from the CDD (v2.13) at NCBI (data not shown).

PEVK Sequences

Identification of the PEVK region by RT-PCR analysis was completed for T. castaneum, A. mellifera, N. vitripennis, D. viridis, and A. pisum. The data indicate that the length of the region is relatively conserved, from 448 (A. pisum) to 655 amino acids (D. melanogaster) (Table 2). Markedly different from the rest of the gene, the PEVK region is assembled from a comparable number of exons in all available insects (between 14 and 17; Table 2), yet the actual sizes of the exons and introns are not conserved, sometimes even between related species (Fig. 8a; Supplemental Materials 1). Pairwise graphical alignments using the Dotlet software (Junier and Pagni 2000), as well as multiple sequence alignments using CLUSTALW, indicate that the amino acid sequences present between Ig8 and Ig9 are highly variable through most of their length, except for a stretch of 140–150 highly conserved amino acids just before Ig9 (Fig. 8b). For example, LALIGN alignment between the A. mellifera and the D. melanogaster PEVK indicates only a 32.5% identity (score: 263 E [10,000];
We would like to redefine, therefore, the amino acid sequence located between Ig8 and Ig9 as two distinct regions: the PEVK domain per se (hereafter referred to as the PEVK region) and a new unique sequence just before the Ig9 domain, which we refer to as the “YERP” sequence. Part of the YERP sequence is also included at the same position within the previously described crayfish EK domain and in unique sequence 3 of *C. elegans* twitchin. This YERP sequence is not found in any other proteins, including vertebrate titin.

Even though there is no substantial sequence conservation between the PEVK regions of the different insects studied, all revealed an elevated frequency of the amino acids P, E, V, and K ranging from 44% (*D. melanogaster*) to 63% (*A. mellifera*). In contrast, the percentage of P, E, V, and K within the YERP sequence decreases from 39% in *A. pisum* to a low of 28% in *D. melanogaster* (Table 2).

The PEVK domains found in vertebrate titin and in *D. melanogaster* Sallimus (D-titin) have, on average, a higher PEVK content (≥70%) and, also, are characterized by the presence of repeats, such as the PPAK and poly(E) repeats in titin (Greaser 2001; Nagy et al. 2005). In contrast, no PPAK-type repeated pattern has been identified in any of the projectin PEVK domains characterized so far by either LaLign, Dotlet analysis, or visual observation (data not shown). There is, however, a poly(E) stretch present at the very beginning of the projectin PEVK region in both *T. castaneum* and the two hymenopteran projectins.

### Discussion

Projectin is unique among muscle proteins in several aspects, including its dual location within the sarcomere of different muscle types (i.e., synchronous and asynchronous) and its proposed functions as both a scaffold for myofibril assembly and an elastic protein in stretch activation. Our evolutionary analysis of projectin provided...
important insights into how separate regions of the protein may have been modified under different evolutionary constraints.

The observed amplification in the number of NH2-Ig domains from three in nematode twitchin to seven (or eight) in arthropods could account for the apparent extension of the NH2-terminal part of the molecule into the I-band region. This would allow for its anchoring at the Z band while still maintaining its association with myosin within the A band. Both nematode twitchin and crayfish projectin have been shown to be localized to the A bands of oblique sarcomeres and giant sarcomeres of the claw and flexor muscles, respectively (Hu et al. 1990; Manabe et al. 1993; Oshino et al. 2003). The study by Oshino et al. (2003) revealed, however, that a part of the NH2-terminal region of crayfish projectin does extrude into the I band but does not physically reach the Z band in either the closer or the flexor sarcomeres. In many insects that use asynchronous muscles, the current model for C-filament structure proposes that projectin and its companion protein kettin/Sls are physically anchored to the Z bands through their NH2-terminal regions and overlap with at least part of the A band (Ayme-Southgate et al. 2005; Bullard et al. 2006). The combination of the shorter I band of insect indirect flight muscles and the longer NH2-terminal Ig regions would conceivably allow projectin to be long enough both to be anchored at the Z band and to maintain its association with the myosin filaments.

It is of interest that all the *Drosophila* species included in this study, and *D. melanogaster*, possess a small alternatively spliced exon corresponding to a short N-terminal extension of Ig2, with one of these alternative extensions being asynchronous specific in *D. melanogaster*. Numerous studies using several titin Ig domains have described the direct effect of terminal extensions on the stability of the Ig fold (Politou et al. 1994; Pfuhl et al. 1997). In the case of one titin domain, two constructs differing only in the length of their N-terminal extensions show noticeable differences, as the longer NH2 terminus is typically more stable (Politou et al. 1994). The significance of an alternative N-terminal extension for projectin Ig2 is intriguing in relation to any interactions with other proteins during its anchoring to the Z band and/or a change in the stability between different muscle isoforms. This hypothesis will require further studies.

When the 39 copies of either the Ig or the FnIII domain from one species are compared among themselves, the conserved residues tend to correspond almost exclusively to the consensus positions. This is consistent with maintaining amino acids essential for the fold of the domain (Bork et al. 1994; Pfuhl and Pastore 1995; Politou et al. 1995; Fraternali and Pastore 1999; Amodeo et al. 2001; Muhle-Goll et al. 2001; Lee et al. 2007). Assuming that projectin Ig and FnIII domains found at identical positions within the projectin proteins in different insect species, mainly within the central portion of both domains. NMR and crystallography studies of both domains from twitchin and titin indicate that these residues are found on the surface of both the Ig and the FnIII folds and are more likely to participate in position-specific protein interactions (Fong et al. 1996; Pfuhl and Pastore 1995; Politou et al. 1995; Fong et al. 1996; Kenny et al. 1999). On the other hand, there is a higher level of conservation among Ig and FnIII domains at equivalent positions may actually reflect the likely involvement of specific Ig or FnIII domains in different protein-protein interactions.

The analysis of the gene structure presented in this study supports the observed trend of extensive intron losses that
occurred during insect evolution. Within that general trend, the *Apis* gene is unusual, with 95 introns, compared to only 70 in the other hymenopteran *Nasonia* gene. Most of these additional *Apis* introns are found within the core region of the gene, where most of the intron losses have otherwise occurred in the other projectin genes. Because the positions of these additional introns in the *Apis* sequence are conserved from the more basal aphid gene, it is likely that this represents a lack of intron losses rather than the gain of new introns. A similar observation reporting the lack of intron loss has been described for the family of odorant receptor genes in the honeybee (Foret and Maleszka 2006).
Even with only 70 introns, the *Nasonia* projectin gene is still more ancestral in its gene structure than the coleopteran and dipteran projectin genes. Our phylogenetic analysis provides further evidence that the Hymenoptera are the most basal group of the Holometabola, and that they diverged from the Diptera significantly before the Coleoptera. Our study is consistent with findings from morphological data (Kukalová-Peck and Lawrence 2004), as well as sequence analysis of ESTs (Savard et al. 2006), genomic sequences (Zdobnov and Bork 2007), and intron evolution (Krauss et al. 2008).

The region between insect Ig8 and insect Ig9 was described in *D. melanogaster* as the equivalent of titin PEVK domain because of its 50% content in the four relevant amino acids. The current study indicates that the projectin PEVK domains are highly variable within most of their sequence, except for a conserved C-terminal segment (termed the YERP sequence). The projectin PEVK, however, shows conservation in length and unusual amino acid composition. Multiple studies of various titin PEVK regions indicate that no conventional secondary structures such as α-helix or β-sheet are evident in titin PEVK sequences (Greaser et al. 2000). Instead, the titin PEVK domain is thought to be in an open, flexible conformation, with stable structural folds consisting of polyproline (PPII) left-handed helices (Gutierez-Cruz et al. 2001; Li et al. 2001; Ma and Wang 2003). Electrostatic interactions between positively and negatively charged residues could also contribute to different coexisting configurations arising upon stretching and release (Forbes et al. 2005). Other authors, however, did not detect PPII helices in other regions of titin PEVK and have classified the PEVK modules as intrinsically disordered proteins (IDPs) (Duana et al. 2006). Whether various examples of the projectin PEVK domains contain PPII helices and whether they resemble IDPs remain to be investigated. It is a tentative idea that comparable mechanical properties could be accomplished by PEVK domains with very divergent amino acid sequences, by maintaining features such as an unusually high P, E, V, and K content, potential charge interactions, and the possibility to generate PPII-like helices. A recent study by Daughdrill et al. (2007) of RPA70 (replication protein A) and an “intrinsically unstructured linker domain” found between two of its globular domains supports the possibility of conservation of dynamic behavior (and potentially molecular function) in the apparent absence of amino acid conservation. As further support for this idea, a study by Granzier et al. (2007) has shown that the titin PEVK domain of chicken is significantly different in length, sequence, and PEVK content compared to mammalian PEVK titin, yet is expected to accomplish essentially similar molecular functions.

The reported analysis of projectin provides new insight into how different regions of the protein may have evolved over time. For the Ig and FnIII domains, the evolutionary constraints on the domain sequences are linked to the maintenance of their specific three-dimensional conformation in order to preserve both the functional folds and the potential protein-protein interactions. On the other hand, the PEVK region must maintain features such as length and unusual amino acid composition, but without any strict amino acid sequence conservation. Continued analysis to expand our projectin data to other insect groups, including more basal lineages, may help refine this evolutionary analysis to further understand the conservation as well as the changes within the projectin protein in relation to its dual functions, dual localizations, and modifications of insect flight physiology during evolution.

Acknowledgments This research was supported in part by National Institutes of Health (NIH) Grant 1R15AR053137-01 to AAS. This publication was also made possible by NIH Grant P20 RR-016461 from the National Center for Research Resources. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH. We thank Dr. J. Evans (USDA, Beltsville, MD), Dr. S. Brown (Kansas State University, Manhattan, KS), Dr. J. Werren (University of Rochester, Rochester, NY), Dr. D. Stern (Princeton University, Princeton, NJ), and the Tucson Drosophila stock center for providing the live insects. We also acknowledge the Human Genome Sequencing Center at Baylor College of Medicine for making accessible the genome sequencing data for *Tribolium castaneum*, *Nasonia vitripennis*, and *Acyrthosiphon pisum* before publication.

References


