J Mol Evol (2001) 53:180–190
DOI: 10.1007/s002390010208

JOURNAL OF MOLECULAR EVOLUTION

Special Evolutionary Properties of Genes Encoding a Protein with a Simple Amino Acid Repeat

Traci Meeds,[1] Erin Lockard,[1] Brian T. Livingston[1,2]

[1] Stowers Institute for Medical Research, Kansas City, MO 64110, USA
[2] Department of Cell Biology and Biophysics, School of Biological Sciences, University of Missouri—Kansas City, MO 64110, USA

Received: 31 October 2000 / Accepted: 20 March 2001

Abstract. We have examined the evolution of a gene, SM50, encoding a component of the spicule matrix, which plays an integral role in the formation of the echinoderm skeleton. This gene was originally characterized in *Strongylocentrotus purpuratus* and encodes an imperfect tandem repeat of six or seven amino acids. We have analyzed the sequence of this repeat in a number of sea urchin species and have determined that the repeat regions have undergone concerted evolution. There are differences in the repeat region between species, but the overall repeat structure is conserved, suggesting the repeat forms a structural domain important in biomineralization. The inherent conserved amino acid repeat structure promotes concerted evolution due to the high probability of misreplication and unequal crossing-over in the repeated segment of the gene. While there are constraints on the amino acids allowed in the repeat region, there are also variations, so that the sequences observed illustrate the balance between amino acid substitutions and concerted evolution. We have evidence that substitutions can alter the mechanisms of unequal crossing-over, altering the way concerted evolution occurs. The way in which concerted evolution occurred appears to be determined by the degree of sequence similarity between the repeats in a given gene, which influences how unequal crossing over may occur. We have mapped the differences in repeat regions on existing phylogenetic trees and indicate where concerted evolution has taken place. We also confirm an earlier report that *Hemicentrotus pulcherrimus* fits into the *Strongylocentrotus* genus and examine the evolution of the *H. pulcherrimus* SM50 repeat relative to other members of this genus.

Key words: Concerted evolution — Echinoderms — Spicule matrix proteins

Correspondence to: Brian T. Livingston; E-mail: livingstonbr@umkc.edu

Introduction

Genes encoding protein domains consisting of stretches of tandem repeats of amino acids are subject to two opposing mechanisms of change at the DNA level. One is concerted evolution, where unequal crossing-over and gene conversion can lead to a change in the number of repeats and homogenize the repeat sequences within a species (Smith 1976; Zimmer et al. 1980; Dover 1982). This leads to the repeat sequences within a species becoming more similar to each other than to the corresponding repeat in other species. The other types of changes are base substitutions, which can lead to changes in amino acids within individual repeats. Base substitutions tend to inhibit unequal crossing-over and gene conversion, reducing concerted evolution (Thomas et al. 1997). The balance of the rate of substitution versus concerted evolution determines the homogeneity of the repeat region. Repeat domains that have undergone concerted evolution relatively recently will have uniformity among repeat sequences that may differ substantially from closely related species (Swanson and Vacquier 1998). However, if concerted evolution occurs prior to

the divergence of species, closely related species may have very similar repeats which differ primarily due to single-amino acid changes, or small changes in repeat number due to local deletions, or additions of repeat units due to unequal crossing-over. We propose here that the accumulation of substitutions within a repeat region can restrict where unequal crossing-over can occur and change how the repeat region undergoes convergent evolution.

While the changes in the repeat region of a gene are carried out at the DNA level, when the repeat units encode a protein domain, the function of the protein imposes a constraint on the types of changes that can occur. The type of substitutions that can occur, as well as the allowed changes in the number of repeat units in a protein, is constrained by the degree of change that still results in a functional protein. The repeat nature of the protein also maintains the intrinsic simplicity of the repeat sequence by promoting concerted evolution. In this study we examine how homogenization of repeat sequences through concerted evolution and divergence of repeat sequences by base substitution and small deletions interact during the evolution of a repeat region located within the coding sequences of a gene encoding a skeletal matrix protein in sea urchins. While there are clearly constraints on the type of changes allowed in the repeat region of this protein, there is sufficient flexibility to allow us to observe the relative influences of base substitution and concerted evolution during the evolution of this protein in echinoderms. This system provides a good model to study the balance of these two processes, since the DNA sequences are not as tightly conserved as rRNA or microsattelites, but the critical level of structure required in the repeat region maintains the intrinsic repeat nature that leads to concerted evolution.

The organic matrix of the sea urchin skeletal spicules has been isolated (Benson et al. 1986), and four of the genes encoding spicule proteins have been cloned and their developmental expression characterized (Benson et al. 1987; Sucov et al. 1987; Livingston et al. 1991; George et al. 1991; Harkey et al. 1995; Lee and Britten 1998). Three of these genes encode proteins with tandem repeats of amino acids. One of these genes, the SM50 gene, has been extensively characterized. The amino acids in the SM50 repeats are largely nonpolar or uncharged, similar to what has been seen in a variety of mineralized tissue (Wilt and Benson 1988). The repeat motifs are relatively short, consisting of repeats of 3–10 amino acids. Analysis of the repeat sequences suggests that they may form a β-spiral structure found in elastic proteins such as elastin and silk proteins (Livingston et al. 1991). One of the spicule matrix proteins is encoded by the SM50 gene in *Strongylocentrotus purpuratus*. The SM50 gene has 29 repeat units encoding either six or seven amino acids, and these repeats are relatively diverse in sequence. SM50 orthologues have been cloned in sequenced in several species (Livingston et al. 1991, Table 1). All of these share a similar repeat region, although the number of repeat units varies, and in some species the repeats are homogeneous, while others show more sequence diversity.

Tandem repeats of amino acids similar to that found in SM50 repeats have been reported in genes encoding spider silk proteins, and these genes have undergone concerted evolution as well (Hayashi and Lewis, 2000). Similar repeats are also seen in sea urchin bindin genes; however, these repeats are subject to positive selection as well as concerted evolution (Biermann, 1998). Two other *S. purpuratus* spicule matrix genes have similar, albeit less extensive repeat regions (Harkey et al. 1995; Lee and Britten 1998). This suggests that these tandem repeats form an important structural motif that has been utilized in several different biological processes.

We have examined SM50 repeat sequences from members of this gene family in six different species and three genera of euechinoids. The repetitive region has undergone concerted evolution in all of the species examined. In some species, homogenization of the repeat region has occurred relatively recently, while in one group of strongylocentrotids, the appearance of truncated repeats and substitutions has made the repeat region relatively diverse. We see three variations in how concerted evolution occurred in the species examined. The way in which concerted evolution occurred appears to be determined by the degree of sequence similarity between the repeats in a given gene, which influences how unequal crossing over may occur. We have also mapped the differences in repeat regions on existing phylogenetic trees (Thomas et al. 1989; Smith et al. 1993; Turbeville et al. 1994; Littlewood and Smith 1995; Springer et al. 1995; Ferkowicz et al. 1998). One consequence of this analysis is that *Hemicentrotus pulcherrimus*, a Japanese sea urchin, has SM50 repeat sequences that clearly place it within the *Strongylocentrotus* clade. Analysis of mitochondrial DNA sequences confirms the placement of *Hemicentrotus* among the genus *Strongylocentrotus*, as first reported by Biermann (1998).

Materials and Methods

DNA Isolation

S. purpuratus, *S. franciscanus*, and *L. pictus* were purchased from Marinus (Orange County, CA). *S. droebachiensis* sperm was provided by Bruce Brandhorst (Simon Fraser University, Burnaby, B.C., Canada). *L. variegatus* DNA was provided by William Kinsey (Kansas University Medical Center). *H. pulcherrimus* DNA was provided by Koji Akasaka (Hiroshima University, Hiroshima, Japan). DNA was isolated from sperm using Qiagen Midi-columns (Qiagen). Five microliters of sperm was added to 100 μl Qiagen Buffer G2 (800 m*M* guanidine HCl, 30 m*M* Tris–Cl, pH 8.0, 30 m*M* EDTA, pH 8.0, 5% Tween-20, and 0.5% Triton X-100). The lysed sperm were loaded onto a Qiagen Genomic tip-100 and washed, and DNA was eluted according

Table 1. Nucleotide sequence of SM50 repeats

Repeat no.	*S. purpuratus*	*S. droebachiensis*	*H. pulcherrimus*
	GGC CAA	*GGC CAA*	*GGC CAA*
	CAA CCG GGC ATG GGA	*CAG CCG GGC ATG GGA*	*CAA CCG GGC ATG GGA*
	CAA GGC GGC TTT GGT AAT CAA	*CAA CCG GGC TTT GGT AAT CAA*	*CAA GGC −−− TTT GGC AAT CAA*
1	CAA CCA GGC ATG GGT GGG CGA	CAA CCA GGC ATG GGT GGG CGA	CAA CCA GGC TTT GGT AAT
2	CAA CCA GGC TTT GGT AAT	CAA CCA GGC TTT GGT AAT	CAA CCA GGC ATG GGT GGG CGA
3	CAA CCA GGA ATG GGT GGG CGA	CAA CCA GGC ATG GGT GGG CGA	CAA CCA GGC TTT GGC AAT
4	CAA CCA GGC TTT GGT AAT	CAA CCA GCC TGG GGT GGA CAA	CAA CCA GGT ATG GGT GGG CGA
5	CAA CCA GGA ATG GGA GGG CGA	CAA CCA GGT GTG GGA GGG CGA	CAA CCA GGC TTT GGC AAT
6	CAA CCA GGC TGG GGT AAT	CAA CCA GGC TGG GGT AAT	CAA CCA GGT ATG GGT GGG CGA
7	CAA CCC GGT GTG GGT GGG CGA	CAA CCC GGT GTG GGT GGG CGA	CAA CCA GGC TTT GGC AAT
8	CAA CCA GGC ATG GGT GGA CAA	CAA CCA GGC ATG GGT GGA CAA	CAA CCA GGT GTG GGT GGG CGA
9	CAA CCA GGC TGG GGT AAT	CAA CCA GGA GTG GTT GGG CGA	CAA CCA GGC TTT GGT AAT
10	CAA CCC GGT GTG GGT GGA CGA	CAA CCA GGC TTT GGT AAT	CAA CCC GGC ATG GGT GGG CGA
11	CAA CCA GGC ATG GGT GGA	CAA CCC GGC ATG GGG GGA CAA	CAA CCA GGC TTT GGC AAT
12	CAA CCA GGA GTG GGC GGG CGA	CAA CCA GGT GTG GGA GGG CAA	CAA CCA GGT GTG GGT GGG CGA
13	CAA CCA GGC TTT GGT AAT	CAA CCA GGC TGG GGT AAT	CAA CCA GGC TTT GGC AAT
14	CAA CCC GGC ATG GGT GGA CAA	CAA CCC GGT GTG GGT GGG CGA	CAA CCA GGC ATG GGT GGA CAA
15	CAA CCA GGC ATG GGT GGA CAA	CAA CCA GGC ATG GGT GGA CAA	CAA CCA GGT GTG GGT GGG CGA
16	CAA CCA GGC TGG GGT AAT	CAA CCA GGT GTG GGT GGA CGG	CAG CCA GGC TTT GGT AAT
17	CAA CCC GGT GTG GGT GGG CGA	CAA CCA GGC ATG GGT GGA CAA	CAA CCA GGT ATG GGT GGA AAC
18	CAA CCA GGC ATG GGT GGA	CAA CCA GGT GTG GGT GGA CGA	CAA CCC GGC ATG GGT GGA CAA
19	CAA CCA GGA GTG GGC GGG CGA	CAA CCA GGC TTT GGT AAT	CAA CCA GGC ATG GGC GGG CGA
20	CAA CCA GGT GTG GGT GGA CGA	CAG CCA GGT GTG GGT GGA CAA	CAA CCC GGC GTA GGT GGT CGA
21	CAA CCA GGC TTT GGT AAT	CAA CCA GGC ATG GGT GGA CAA	CAA CCA GGC ATG GGT GGG CAG
22	CAG CCA GGT GTG GGT GGA CGA	CAA CCA GGT GTG GGA GGG CGA	CAA CCA GGT ATG GGC GGG CGA
23	CAA CCA GGC ATG GGT GGA CAA	CAA CCA GGC TTT GGT AAT	CAA CCA GGC ATG GGT GGG CAG
24	CAA CCA GGT ATG GGT GGA	CAA CCA GGT GTG GGT GGG CGA	—
25	CAA CCA GGA GTG GGC GGG CGA	CAA CCA GGC ATG GGT GGC CAG	—
26	CAA CCA GGT ATG GGA GGG CGA	—	—
27	CAA CCA GGC TTT GGT AAT	—	—
28	CAA CCA GGT GTG GGT GGG CGA	—	—
29	CAA CCA GGC ATG GGT GGC CAG	—	—
	CAA CCG AAT AAC CCG AAT AAC	*CAA CCG AAT AAC CCG AAT AAC*	*CAA CCG AAT AAC CCG AAC AAC*

to the manufacturer's protocol for isolation of high molecular weight DNA from tissue.

Polymerase Chain Reaction (PCR) and Cloning

SM50 sequences were amplified from genomic DNA using primers 5′ and 3′ to the repeat region: 5′CGAAGCTTTGXAGCATDCGDGGYCKGTTRAA 3′ and 5′ ACGGATCCTTYTCXCARGAYAACCARATGGARATGGA 3′. Reactions were carried out using Taq polymerase (Perkin Elmer Applied Biosystems, Foster City, CA) using standard buffer containing 2 m*M* $MgCl_2$ and a 0.5 μ*M* concentration of primers. Cycling parameters were as follows: 95°C for 30 s; touchdown, 65 to 55°C for 30 s; 72°C for 30 s for 10 cycles; followed by 95°C for 30 s, 55°C for 30 s, and 72°C for 30 s for 30 cycles. PCR products were separated on a 1% agarose gel, isolated from the gel using Qiagen gel extraction kits (Qiagen, MA), and cloned into Bluescript (Stratagene). Ligation products were electroporated into *E. coli* strain XL1-blue (Stratagene) and plasmids containing inserts were identified using blue/white screening by standard methods. 12S rRNA sequences were amplified as described by Thomas et al. (1989) and sequenced as described above.

Sequencing

Sequencing was performed on an ABI Prism 377 automated sequencer using the dRhodamine Terminator Cycle Sequencing Kit (Perkin Elmer Applied Biosystems). Sequences were cleaned, corrected, and polished using Editview Version 1.01 from the ABI automated DNA sequence viewer. SM50 DNA sequences from GenBank were Nos. M16231, S48755, and X59616.

Sequence Analysis

Sequences were aligned by eye using ESEE (Cabot and Beckenback 1989). Regions of the ambiguous alignment were removed from the analysis (Sidow and Thomas 1994; Swofford et al. 1996). Maximum-parsimony (MP), neighbor-joining (NJ), and maximum-likelihood (ML) methods [as implemented in PAUP* (Swofford 1998), MEGA (Kumar et al. 1993), and PHYLIP (Felsenstein, 1993)] were employed to infer phylogenies. MP searches used 100 random input orders, nearest-neighbor interchanges, and branch swapping to increase the probability of recovering the best tree. Weighting schemes considered the work of others (Hollar and Springer 1996; Krajewski et al. 1997; Lavergne et al. 1996; Springer et al. 1995; Stanhope et al. 1996). NJ method sequence divergence was estimated using a variety of models [e.g., the Jukes–Cantor (1969), Tamura–Nei (1993), ML (Swofford et al. 1996), and logdet (Lockhart et al. 1994) methods]. ML employed empirical base frequencies and, where possible, ML estimates of ts/tv ratio. We tested phylogenetic hypotheses using winning sites (Prager and Wilson 1988), Templeton (1983), and Kishino–Hasegawa (1989) tests, as well as bootstrapping (Felsenstein 1985; Hillis and Bull 1993; Hillis et al. 1996; Huelsenbeck et al. 1995; Sanderson 1989, 1995).

Table 1. Extended

S. franciscanus	*L. pictus*	*L. variegatus*
GGC CAA		
CAA CCA GGC ATG GGA		
CCG GGC ATG GGC ATG GGA		
CCG GGC GGT GGT GGT CGA	*GGT CAA*	*GGT CAA*
CAA CCA GGC TTT GGG CAA	CAA CCT GGC TTC GGT GGG CAA	CAA CCT GGC ATC GGC GGG CAA
CAA CCA GGC ATG GGA GGG CGA	CAA CCT GGC TTC GGC GGA CGA	CAA CCT GGC TTC GGC GGG CAA
CAA CAA GGC ACG GGT GGG TGG	CAA CCT GGC TTC GGT GGG CAA	CAA CCT GGC GTC GGC GGA CGA
CAA CCA GGC ATG GGT GGA CAA	CAA CCT GGC TTC GGG CAA	CAA CCT GGC TTC GGT GGG CAA
CAA CCA GGC ATG GGA GGG CGA	CAA CCT GGC TTC GGG GGG CGA	CAA CCT GGC TTC GGC GGG CAA
CAA CCA GGC ATG GGT GGA CAA	CAA CCT GGC TTC GGG GGG CGA	CAA CCT GGC TTC GGC GGG CGA
CAA CCA GGC ATG GGA GGG CGA	CAA CCT GGC TTC GGC GGG CAA	CAA CCT GGC TTC GGT GGG CAA
CAA CCA GGC ATG GGT GGA CAA	CAA CCT GGC TTT GGG GGA CAA	CAA CCT GGC TTC GGC GGG CAG
CAA CCA GGC ATG GGA GGG CGA	CAA CCT GGC TTC GGC GGG CAA	CAA CCT GGC TTC GGT GGG CAA
CAA CCA GGC ATG GGT GGA CAA	CAA CCT GGC TTT GGC GGG CAG	CAA CCT GGC TTC GGC GGG CAA
CAA CCA GGT ATG GGA GGG CGA	CAG CCT GGC TTT GGG GGA CAA	CAA CCT GGC TTC GGT GGG CAA
CAA CCA GGC ATG AGT GGA CAG	CAA CCT GGC TTT GGC GGG CAG	CAA CCT GGC TTT GGG GGA CAA
CAA CCA GGC ATG GGT GGA CGA	CAG CCT GGC TTT GGG GGA CAA	CAA CCG GGT TTT GGT GGG GGA CCA
CAA CCG GGC ATG GGT GGA CAG	CAA CCG GGT TTT GGT GGG GGA CCA	CAA CGA CCT GGC ATG GGG GGA
CAA CCA GAC ATG GGT GGA CGA	CAA CGA CCT GGC ATG GGG GGA	—
CAA CCA GGC ATG GGT GGG CAG	—	—
—	—	—
—	—	—
—	—	—
—	—	—
—	—	—
—	—	—
—	—	—
—	—	—
—	—	—
—	—	—
—	—	—
—	—	—
—	—	—
CAA CCA GAT AAC CCA AAC CCC	*CAG CCA AAC*	*CAG CCA AAT*

Results

Genomic DNA was isolated from three species of the genus *Strongylocentrotus*, two species of *Lytechinus*, and *Hemicentrotus pulcherrimus.* PCR was then performed using primers corresponding to conserved amino acid sequences that flank the SM50 and PN repeat domains in *S. purpuratus* and *L. pictus.* The PCR products varied in size from 1100 bp in *S. purpuratus* to 650 bp in *L. variegatus.* Sequence analysis of the PCR products indicates that the variation in size of the PCR products is due entirely to variation in the number of repeats in the two repeat domains and is due primarily to differences in the number of SM50 repeats (Fig. 3). The positions and sizes of the repeat regions, as well as the location of the PCR primers, in the genes from the various species are shown in Fig. 1.

To examine the evolution of the repeat regions in these genes, we needed to be able to map changes onto an existing phylogenetic tree. Phylogenetic trees for most of these species have been published (Thomas et al. 1989; Smith et al. 1993; Turbeville et al. 1994; Littlewood and Smith 1995; Springer et al. 1995; Ferkowicz et al. 1998), and recently the relationship of *Hemicentrotus pulcherrimus* to other sea urchins has been examined using mitochondrial DNA and bindin sequences (Biermann, 1998). We confirmed this analysis by using PCR to amplify mitochondrial rRNA sequences from *H. pulcherrimus* and determined their sequence. We also amplified and sequenced mitochondrial rRNA genes from *Eucidaris tribuloides* to use as an outgroup. These sequences were then used, along with published sequences for the other species, to construct the phylogenetic tree shown in Fig. 2. The information relevant to our analysis of the SM50 repeats is that *H. pulcherrimus, S. purpuratus,* and *S. droebachiensis* fall into one phylogenetic group, *S. franciscanus* is in a related but separate group, and the two *Lytechinus* species form a third, more distantly related group. We have also used the sequence of the nonrepeat coding region between our 5′ PCR primer

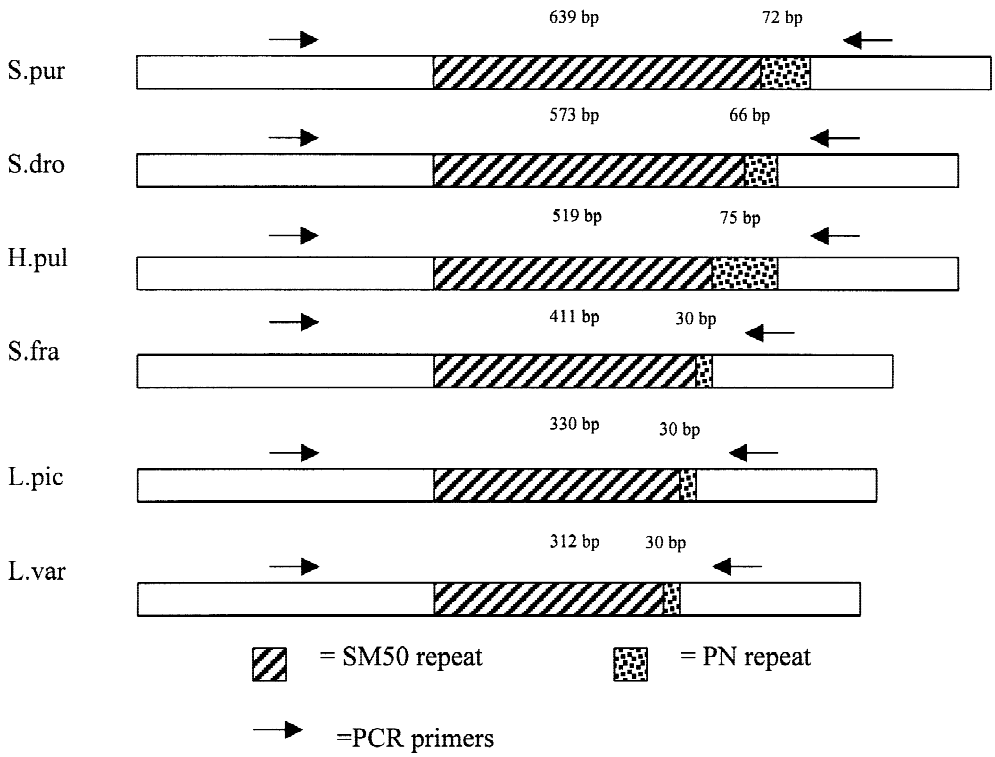


Fig. 1. Diagram of SM50 genes. *Hatched areas* indicate the SM50 repeat. The size of the repeat in base pairs is indicated above the hatched area. The *dotted area* indicates the PN repeat, with the sizes of the repeat in base pairs indicated above the dotted area.

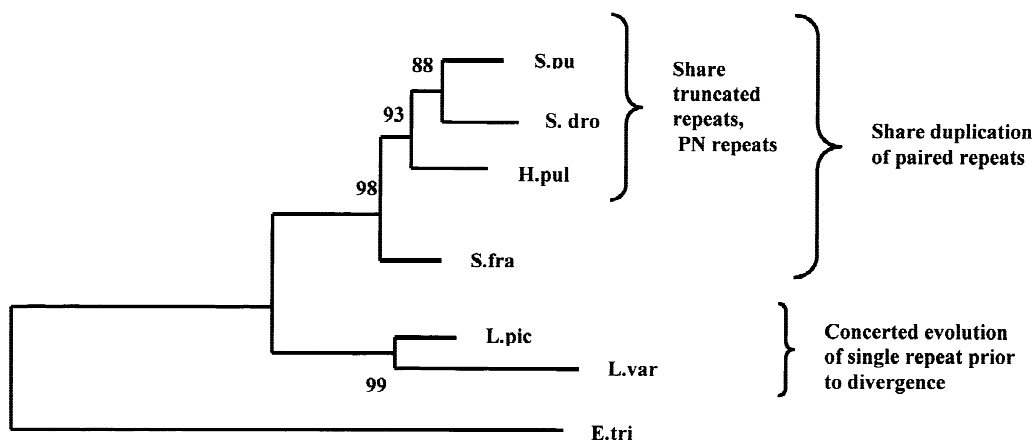


Fig. 2. Phylogenetic trees based on sequence comparisons of the 16S small ribosomal RNA gene (srRNA). Evolutionary changes in the SM50 gene are indicated on the *right.* GenBank sequences used for A: M27523, M27672, M27673, M27674, and M27675. The branching pattern was derived using neighbor joining and was implemented using MEGA. The same branching pattern was seen using maximum likelihood and maximum parsimony (not shown). The tree was rooted with *Eucidaris tribuloides* as the outgroup. The numbers represent bootstrap values with 1000 replicates.

and the SM50 repeat from the different species to construct a phylogenetic tree (not shown). Again, *H. pulcherrimus* falls into a group with the members of the Strongylocentrotus genus. *S. purpuratus* and *S. droebachiensis* appear more closely related to each other than to either *H. pulcherrimus* or *S. franciscanus.* Analysis of the SM50 repeat region shows that *H. pulcherrimus* is more closely related to *S. purpuratus* and *S. droebachiensis* than *S. franciscanus* and that *S. purpuratus* and *S. droebachiensis* are closely related.

Table 1 shows the nucleotide sequence of the SM50 repeats in the six species examined. The basic repeat unit shared by all species is a 21-bp sequence which we depict starting with a CAA and ending with a CAA or CGA. Twelve of 21 bases in the 21-bp repeats are absolutely conserved across all species. There is some variation in the wobble positions of some, but not all, codons. The greatest variation occurs in the fourth and seventh codons. This corresponds to bases 10–12 and 20–21. *S. purpuratus, S. droebachiensis,* and *H. pulcher-*

Table 2. Amino acid sequence of SM50 repeats

S. purpuratus	*S. droebachiensis*	*H. pulcherrimus*	*S. franciscanus*	*L. pictus*	*L. variegatus*
QPGMGGR	QPGMGGR		QPGFGQ	QPGFGGQ	QPGIGGQ
QPGFGN	QPGFGN	QPGFGN	QPGMGGR	QPGFGGR	QPGFGGQ
QPGMGGR	QPGMGGR	QPGMGGR	QPGTGGW	QPGFGGQ	QPGVGGR
QPGFGN	QPGWGGQ	QPGFGN	QPGMGGQ	PQGFGQ	QPGFGGQ
QPGMGGR	QPGVGGR	QPGMGGR	QPGMGGR	QPGFGGR	QPGFGGQ
QPGWGN	QPGWGN	QPGFGN	QPGMGGQ	QPGFGGR	QPGFGGR
QPGVGGR	QPGVGGR	QPGMGGR	QPGMGGR	QPGFGGQ	QPGFGGQ
QPGMGGQ	QPGMGGQ	QPGFGN	QPGMGGQ	QPGFGGQ	QPGFGGQ
QPGWGN	QPGVVGR	QPGVGGR	QPGMGGR	QPGFGGQ	QPGFGGQ
QPGVGGR	QPGFGN	QPGFGN	QPGMGGQ	QPGFGGQ	QPGFGGQ
QPGMGG	QPGMGGQ	QPGMGGR	QPGMGGR	QPGFGGQ	QPGFGGQ
QPGVGGR	QPGVGGQ	QPGFGN	QPGMSGQ	QPGFGGQ	QPGFGGQ
QPGFGN	QPGWGN	QPGVGGR	QPGMGGR	QPGFGGQ	QPGFGGGQ
QPGMGGQ	QPGVGGR	QPGFGN	QPGMGGQ	QPGFGGGP	QPPGMGG
QPGMGGQ	QPGMGGQ	QPGMGGQ	QPDMGGR	QRPGMGG	
QPGWGN	QPGVGGR	QPGVGGR	QPGMGGQ		
QPGVGGR	QPGMGGR	QPGFGN			
QPGMGG	QPGVGGR	QPGMGGN			
QPGVGGR	QPGFGN	QPGMGGQ			
QPGVGGR	QPGVGGQ	QPGMGGR			
QPGFGN	QPGMGGQ	QPGVGGR			
QPGVGGR	QPGVGGR	QPGMGGQ			
QPGMGGQ	QPGFGN	QPGMGGR			
QPGMGG	QPGVGGR	QPGMGGQ			
QPGVGGR	QPGMGGR				
QPGMGGR					
QPGFGN					
QPGVGGR					
QPGMGGR					

rimus all have a truncated, 18-bp repeat interspersed among the longer 21-bp repeat. The appearance of the truncated repeat is mapped onto the phylogenetic tree in Fig. 2. These truncated repeats are missing the GGX codon at bases 16 to 18 of the 21-bp repeat and end with AAT. There is even less variation in the sequence of these truncated repeats: 17 of 25 truncated repeats found in these three species are identical. The terminal sequence of 3 of 25 is GGA, and in 5 of 25 there is a TGG instead of a TTT in base positions 10–12. There are single truncated repeats in *S. franciscanus* and *L. pictus* that appear to have lost bases 16–18 in a single repeat unit. There are sequences before and after the repeating sequences in all species that resemble partial or imperfect repeats. These sequences are identical within genera and are shown in italics in Table 1. The SM50 sequences from *S. purpuratus* represent a correction of a minor error in the published sequence (Katoh-Fukui et al. 1991).

The amino acids encoded by the SM50 repeat domains are shown in Table 2. Five of the seven amino acids in the long repeats are identical both within the repeats of a single species and between all the species examined. The amino acid in the fourth position is the most variable but is always nonpolar. In *S. purpuratus, S. droebachiensis,* and *H. pulcherrimus,* there is considerable variation in the amino acid found in the fourth position in the SM50 repeat of each species. In *S. franciscanus, L. pictus,* and *L. variegatus,* the fourth position is largely homogeneous, with a methionine found in all but one repeat in *S. franciscanus* and phenylalanine predominant in the two *Lytechinus* species. The amino acid in the seventh position of the long repeat can be either an arginine or a glutamine. The truncated six amino acid repeats in *S. purpuratus, S. droebachiensis,* and *H. pulcherrimus* are identical to the seven amino acid repeats in the first five positions but have an arginine in the sixth position. All seven of the truncated repeats in *H. pulcherrimus* have the amino acid sequence Q P G F G N. In *S. droebachiensis,* four of six are Q P G F G N, while the remaining two have a tryptophan in place of the phenylalanine. In *S. purpuratus,* the fourth position varies among phenylalanine (5 of 11), tryptophan (3 of 11), and methionine (3 of 11).

Immediately following the SM50 repeat region there is a repetitive sequence of proline and arginine residues (PN repeat) in *S. purpuratus, S. droebachiensis,* and *H. pulcherrimus* (Table 3). The number of repeats varies between these three species. The PN repeat is not present in *S. franciscanus, L. pictus,* or *L. variegatus,* although the amino acid sequences on either side of the PN repeat are highly conserved. The nucleotide sequence encoding the PN repeat is 100% identical between *S. purpuratus* and *S. droebachiensis* with the exception of two regions

Table 3. Comparison of C-terminal PN repeats

S. purpuratus																									
	Q	P	N	N	P	N	N	P	N		P	N	N	P	N	N	P	N	N	P	N	P	R	F	N
CAA	CCG	AAT	AAC	CCG	AAT	AAC	CCG	AAC	---	CCG	AAC	AAC	CCG	AAC	AAC	CCG	AAT	AAC	CCA	AAC	CCC	AGG	TTC	AAC	
S. droebachiensis																									
	Q	P	N	N	P	N	N	P	N	N	P	N	N	P				N	N	P	N	P	R	F	N
CAA	CCG	AAT	AAC	CCG	AAT	AAC	CCG	AAC	AAC	CCG	AAC	AAC	CCG	---	---	---	AAT	AAC	CCA	AAC	CCC	AGG	TTC	AAC	
H. pulcherrimus																									
	Q	P	N	N	P	N	N	P	N	N	P	N	N	P	N	N	P	N	N	P	N	P	R	F	N
CAA	CCG	AAT	AAC	CCG	AAC	AAC	CCG	AAT	AAC	CCG	AAT	AAC	CCG	AAT	AAC	CCG	AAT	AAC	CCC	AAA	CCC	AGG	TTC	AAC	
S. franciscanus																									
	Q	P	D	N																P	N	P	R	F	N
CAA	CCA	GAT	AAC	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	CCA	AAC	CCC	AGG	TTC	AAC
L. pictus																									
	Q	P	N	S																P	N	P	R	F	N
CAG	CCA	AAC	AGT	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	CCA	AAC	CCG	AGG	TTC	AAC
L. variegatus																									
	Q	P	N	S																P	N	P	R	F	N
CAG	CCA	AAT	AGT	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	CCA	AAC	CCG	AGG	TTC	AAC

where repeats have been lost. *H. pulcherrimus* is 91.7% identical to *S. purpuratus* and 92.5% identical to *S. droebachiensis.* The most-parsimonious explanation is that the common ancestor to these three species had a PN repeat number equal to that of *H. pulcherrimus,* since *H. pulcherrimus* contains the PN repeat units missing in both *S. purpuratus* and *S. droebachiensis.*

To analyze the mode of convergent evolution that led to the current arrangement of SM50 repeats in the different species examined, we treated each repeat as a separate sequence, and performed a phylogenetic analysis of the repeats within a species. Only the first two nucleotide positions of codons were used in this analysis. The repeats of *L. pictus* (not shown) and *L. variegatus* (Fig. 3A) both fell largely into a single group. Each group of repeats was given a number, and the order of the group number of each repeat was placed in order from 5′ to 3′ in the coding sequence (Fig. 3B). The simplest explanation for the pattern seen is that expansion of the repeat region occurred by duplication of a single 21-nucleotide repeat unit, with a group 1 repeat serving as template. Substitutions in repeats following expansion likely resulted in the repeats that fall into different groups.

When the repeats of *S. franciscanus* were analyzed, they fell into two groups (Fig. 3C). Repeats from the two groups alternate when they are aligned from 5′ to 3′ (Fig. 3D). This suggests that this repeat region expanded by duplication of pairs of slightly dissimilar repeats, giving rise to a repeating 42-nucleotide sequence.

S. purpuratus, S. droebachiensis, and *H. pulcherrimus* share the presence of a truncated 18-nucleotide repeat interspersed between 21-nucleotide repeats. The long (21-nucleotide) repeats of all three of these species were analyzed together and fell into six groups, labeled L1 to L6 in Fig. 4. All but four of the repeats from all three species fell into three groups. The short (18-nucleotide) repeats from all three species were also analyzed together, and these fell into three groups, S1, S2, and S3 (Fig. 3). Figure 4 shows the order from 5′ to 3′ of the repeats labeled according to their homologous group for each of the three species. *H. pulcherrimus* shows an alternating pattern of seven short and long repeats, followed by six long repeats in a row. *S. purpuratus* and *S. droebachiensis* also have an overall pattern of alternating long and short repeats. However, there has been duplication of individual long repeats, resulting in from one to four long repeats between some of the short repeats. There are no short repeats adjacent to one another in any of the three genes examined. The array of SM50 repeat sequences from *S. purpuratus* and *S. droebachiensis* are similar enough to be aligned. The repeats are aligned in Fig. 4B, with repeats that can be aligned in boldface. The *H. pulcherrimus* gene can be aligned to *S. purpuratus* and *S. droebachiensis* in the nonrepeat sequences on either side of the SM50 repeat region. This homology extends into the first two repeats at the 5′ end of the SM50 repeats and picks up again in the last two repeats prior to the PN repeat. The internal repeats of *H. pulcherrimus,* however, cannot be aligned to *S. purpuratus* or *S. droebachiensis* (not shown).

Discussion

The SM50 repeat encodes a conserved protein domain found within an integral matrix protein of the sea urchin skeleton. Each repeat consists of six or seven amino acids with four or five invariant amino acids, respectively. Each repeat has one nonpolar amino acid that can vary and a terminal amino acid that is either positively charged or uncharged. The variation seen between re-

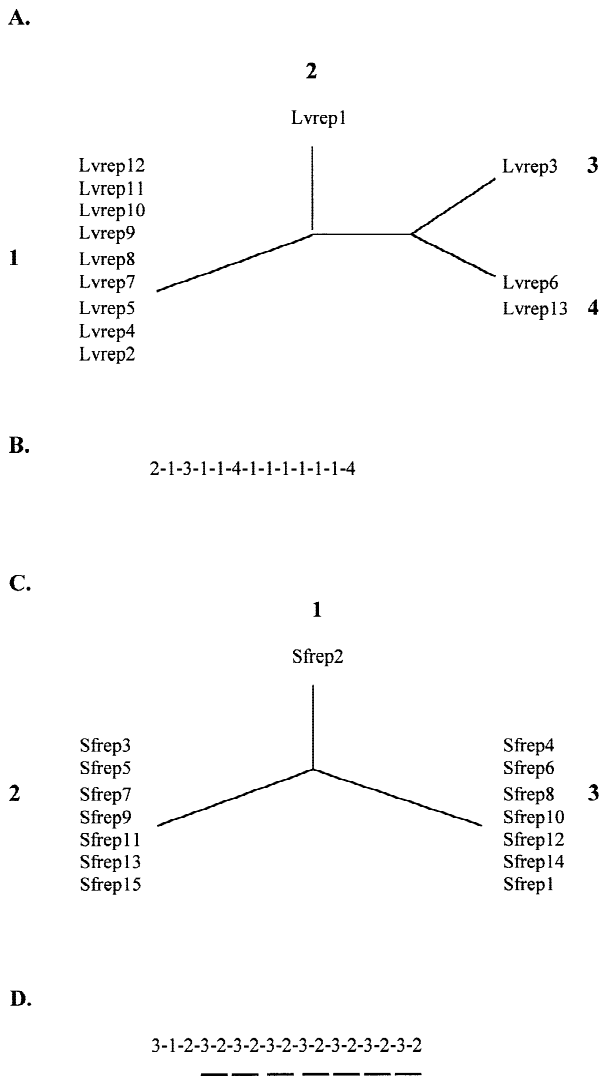


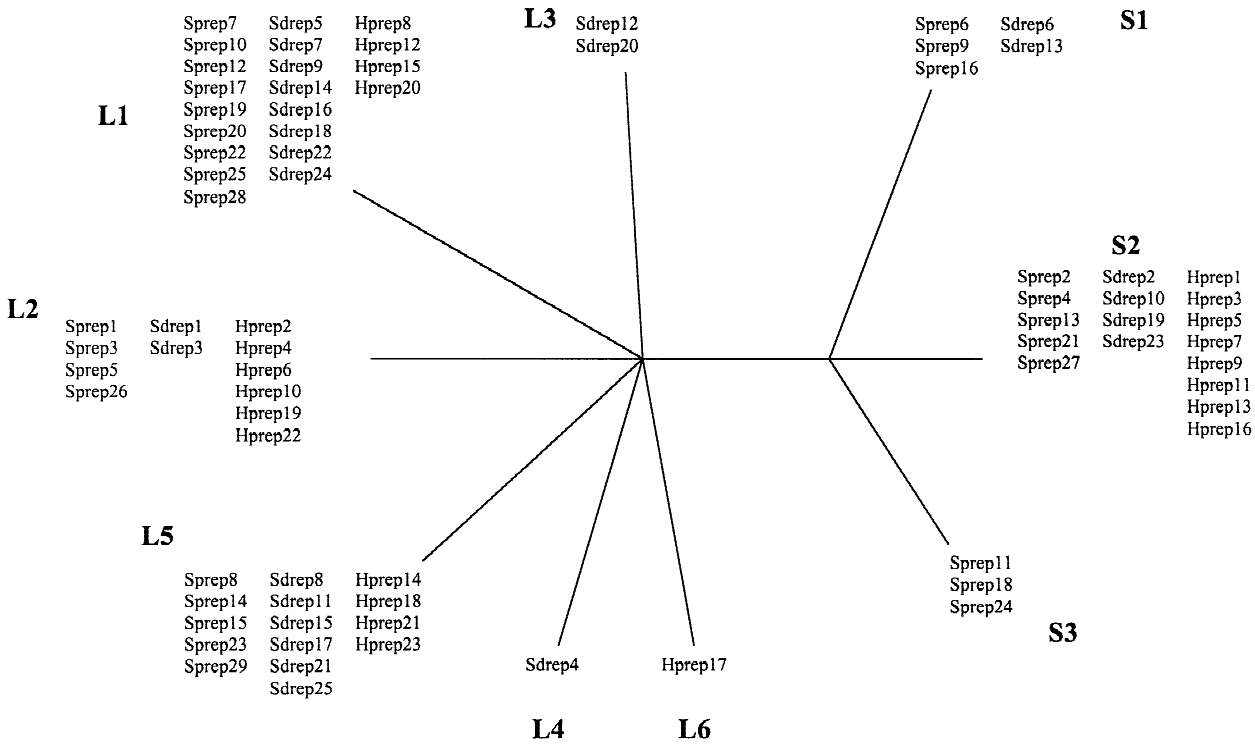
Fig. 3. Analysis of the similarity between SM50 repeats within the same gene for *Lytechinus variegatus* (**A** and **B**) and *Strongylocentrotus franciscanus* (**C** and **D**). The relationship between the repeats is shown in A and C. The linear order of repeats of each group in the SM50 genes is shown in B and D. The duplication of pairs of repeats in *Lytechinus pictus* is indicated by bars in D. Each repeat was treated as a separate species, and the branching pattern was derived using neighbor joining and implemented using MEGA. The same branching pattern was seen using maximum likelihood and maximum parsimony (not shown). The tree was unrooted.

peats both within a gene and between species suggest that the repeat region represents a protein structural domain with a three-dimensional shape that is of more importance than its primary sequence. The presence of similar repeat domains in proteins encoded by other spicule matrix genes suggests that this structural motif is important in skeleton formation and biomineralization in echinoderms. A similar repeat domain is also seen in bindin (Gao et al. 1986; Minor et al. 1991; Biermann 1998), however, suggesting the possibility that this repeat motif is utilized in urchins for a more general function, such as protein–protein interactions or cell adhesion. Similar repeats have been found in several other proteins with disparate functions, such as elastin (Urry 1982), wheat glutenin (Miles et al. 1991), and spider silk proteins (Hayashi and Lewis 2000). Repeat regions composed of the amino acids seen in SM50 proteins and other proteins have been predicted to form a β-spiral structure (Livingston et al. 1991). It could be that such repeats form a general structural motif, and that proteins containing these repeats share this structure, but function is conferred by other protein domains present on the proteins.

The SM50 repeat region has undergone concerted evolution in all species examined. In the two *Lytechinus* species, the repeats are more similar within each species and between homologous repeats within their genus than to any of the *Strongylocentrotus* species (Table 1; analysis not shown). Analysis of the pattern of repeats indicates that the most likely mechanism for concerted evolution was mispairing of repeat units with similar, adjacent repeats, followed by unequal crossing-over. This resulted in the duplication of single repeat units. Most of these duplications occurred prior to divergence of these two species, although some duplications near the 5′ end of the repeat sequences may have occurred subsequent to their divergence. There have been a number of silent substitutions in the repeat regions between the genes of these two species, but relatively few replacement substitutions.

The species examined from the genus *Strongylocentrotus* fall into two distinct groups. The *S. franciscanus* repeats are relatively homogeneous at both the amino acid and the nucleotide level, indicating that this species underwent concerted evolution relatively recently. The variable fourth amino acid is a methionine in 14 of 16 repeats, while the variable seventh amino acid is either glutamine or arginine in approximately equal numbers. Phylogenetic analysis of the *S. franciscanus* repeats indicates that they expanded as pairs of repeats. This is likely explained by the difference sequences at the terminal seventh codon position in alternating repeats. This would favor mispairing of repeats with similar sequences, resulting in the looping-out and duplication of a pair or pairs of repeats. It appears that the divergence of sequences at the ends of a repeat can alter the mechanism of convergent evolution.

The other members of the genus *Strongylocentrotus* (*S. purpuratus, S. droebachiensis,* and *H. pulcherrimus*) share a more recent common ancestor. These species all have a highly conserved, truncated, six-amino acid repeat interspersed among the seven-amino acid SM50 repeats, they all have a repeated region encoding a PN amino acid repeat following the SM50 repeats, and phylogenetic analysis of the sequence of the SM50 gene provides evidence for a recent common ancestor. The truncated repeat always ends in an asparagine, and primarily has a phenylalanine in the variable fourth position, although some tryptophan residues are present in



S.purp: **L2** **S2 L2** S2 L2 **S1 L1** **L5** S1 **L1** S3 L1 **S2 L5** L5 **S1 L1** S3 **L1** -- **L1** **S2** L1 **L5** S3 **L1** L2 **S2 L1** **L5**

S.dro: **L2** **S2 L2** L4 L1 **S1 L1** **L5** -- **L1** -- -- **S2 L5** L3 **S1 L1** L5 **L1** L5 **L1** **S2** L3 **L5** -- **L1** -- **S2 L1** **L5**

H. pul: **S2 L2** S2 L2 S2 L2 S2 L1 S2 L2 S2 L1 S2 L5 L1 S2 L6 L5 L2 L1 L5 **L2** **L5**

Fig. 4. Analysis of the similarity between SM50 repeats in the genus *Strongylocentrotus.* Long repeats (21 nucleotides) were analyzed together and placed into six groups, L1 to L6, based on amino acid sequence similarity (shown on the *left*). Short repeats (18 nucleotides) were analyzed together and placed into three groups, S1 to S3 (shown on the *right*). The linear order of repeats classified by group for each species is shown at the *bottom.* Duplications of short-long repeat pairs are *underlined.* Each repeat was treated as a separate species, and short and long repeats were analyzed separately. The branching pattern was derived using neighbor joining and was implemented using MEGA. The same branching pattern was seen using maximum likelihood and maximum parsimony (not shown). The tree was unrooted.

this position in *S. purpuratus* and *S. droebachiensis.* In the long repeats, the fourth amino acid position of these species shows more variation than in the other species examined. Three different nonpolar amino acids are present at the fourth position in repeats from *H. pulcherrimus,* while *S. purpuratus* and *S. droebachiensis* have four nonpolar amino acids present, including tryptophan, which is not present in the repeats of any of the other species.

All three of the species that have a truncated repeat show an underlying pattern of alternating long then short repeat. The truncation would favor this type of mispairing during unequal crossing-over, since like repeats are more likely to pair with one another. This is most obvious in the *H. pulcherrimus* repeat. Two other types of mispairing and crossing-over events also seem to be at work in the *S. purpuratus* and *S. droebachiensis* repeat regions. One is pairing of close, but not immediately adjacent repeats, which causes duplication/loss of larger segments (multiple repeat units). This could explain the sequence similarities between the *S. purpuratus* repeats 9–15 and repeats 16–23. The second is mispairing of dissimilar repeats that leads to a gain/loss of a single long or short repeat. This would result in a loss of the clear alternating pattern of long and short repeats.

Overall it is clear that evolution of DNA repeats within the coding region of a gene is constrained by the fact that the function of the protein must be maintained. In the case of the SM50 family of proteins, the repeat region likely encodes a structural domain, and while some variation in amino acid sequence is allowed between repeats and between species, these changes are conservative. Without an analysis of amino acid substitution rates, we cannot rule out that compensatory changes have occurred elsewhere in the genome; however, it appears that differences in numbers of repeats

and some variation in amino acid sequence are not detrimental. However, even small variations in the nucleotide sequences between repeats can alter how these repetitive regions undergo concerted evolution. Variations in the nucleotide sequence, especially at the ends of a repeat, can alter the ability of repeats to misalign, favoring pairing of repeats whose sequences are most alike. Pairing with like repeats that are close by would be most favored. This would lead to expansion or loss of pairs of repeats, as in *S. franciscanus* and *H. pulcherrimus.* Pairing with like repeats that are farther away is less likely, though possible, and would lead to duplication of larger groups of repeats, as implied in *S. purpuratus.* Mispairing of repeats that are more divergent in sequence should be less frequent and would lead to a divergence from an alternating pattern of a pair of repeats with dissimilar sequence. A great deal of sequence divergence between repeats within a gene would limit the types of mispairing that is possible even further and would inhibit the large-scale homogenization that leads to identical repeats within a gene. One prediction of this hypothesis is that *S. purpuratus* and *S. droebachiensis* will not undergo concerted evolution to the extent that the repeat region will become homogeneous, as seen in some of the other species.

Acknowledgments. We would like to thank Kelley Thomas, Linda Frisse, Krystalynne Morris, and Dee Denver for their help in phylogenetic analysis and critical review of the manuscript.

References

Benson SC, Benson NC, Wilt F (1986) The organic matrix of the skeletal spicule of sea urchin embryos. J Cell Biol 102:1878–1886

Benson S, Sucov H, Stephens L, Davidson E, Wilt F (1987) A lineage-specific gene encoding a major matrix protein of the sea urchin embryo spicule I. Authentication of the cloned gene and its developmental expression. Dev Biol 120:499–506

Biermann CH (1998) The molecular evolution of sperm bindin in six species of sea urchins (Echinoida: Strongylocentrotidae). Mol Biol Evol 15(12):1761–1771

Cabot EL, Beckenbach AT (1989) Simultaneous editing of multiple nucleic acid and protein sequences with ESEE. Comput Appl Biosci 5:233–234

Dover G (1982) Molecular drive: A cohesive mode of species evolution. Nature 299:111–117

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

Felsenstein J (1993) PHYLIP (phylogeny inference package), version 3.5. University of Washington, Seattle

Ferkowicz MJ, Stander MC, Raff RA (1998) Phylogenetic relationships and developmental expression of three sea urchin Wnt genes. Mol Biol Evol 15:809–819

Gao B, Klein LE, Britten RJ, Davidson EH (1986) Sequence of mRNA coding for bindin, a species-specific sea urchin sperm protein required for fertilization. Proc Natl Acad Sci USA 83:8634–8638

George NC, Killian CE, Wilt FH (1991) Characterization and expression of a gene encoding a 30.6-kDa Strongylocentrotus purpuratus spicule matrix protein. Dev Biol 147(2):334–342

Harkey MA, Klueg K, Sheppard P, Raff RA (1995) Structure, expression, and extracellular targeting of PM27, a skeletal protein associated specifically with growth of the sea urchin larval spicule. Dev Biol 168:549–566

Hayashi CY, Lewis RV (2000) Molecular architecture and evolution of a modular spider silk protein gene. Science 287:1477–1479

Hillis DM, Bull JJ (1993) An empirical test of bootstrapping as a method for assessing confidence on phylogenetic analysis. Syst Biol 42:182–192

Hillis DM, Moritz C, and Mable BK (eds) (1996) Molecular systematics. Sinauer Associates, Sunderland, MA

Hollar LJ, Springer MS (1996) Old world fruit bat phylogeny; Evidence for convergent evolution and an endemic African clade. Proc Natl Acad Sci USA 94:5716–5721

Huelsenbeck JP, Hillis DM, Jones R (1995) Parametric bootstrapping in molecular phylogenetics: Applications and performance. In: Ferraris J, Palumbi S (eds) Molecular zoology: Strategies and protocols. Wiley, New York

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN, (ed) Mammalian protein metabolism. Academic Press, New York

Katoh-Fukui Y, Noce T, Ueda T, Fujiwara Y, Hashimoto N, Higashinakagawa T, Killian CE, Livingston BT, Wilt FH, Benson SC (1991) The corrected structure of the SM50 spicule matrix protein of Strongylocentrotus purpuratus. Dev Biol 145:201–202

Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol 29:170–179

Krajewski CM, Blacket M, Buckley L, Westerman (1997) A multigene assessment of phylogenetic relationships within the *dasyurid* marsupial subfamily *Sminthopsinae.* Mol Phylogenet Evol 8:236–248

Kumar S, Tamaura K, Nei M (1993) MEGA: Molecular evolutionary genetics analysis, version 1.01. Pennsylvania State University, University Park

Lavergne AE, Douzery E, Stichler FM, Catzeflis FM, Springer MS (1996) Interordinal mammalian relationships: Evidence for paenungulate monophyly is provided by complete mitochondrial 12S rRNA sequences. Mol Phylogenet Evol 6:245–258

Lee Y-H, Britten RJ (1998) SM37, a new skeletogenetic gene of the sea urchin embryo isolated by regulatory target site screening. GenBank Submission AF068737

Littlewood DT, Smith AB (1995) A combined morphological and molecular phylogeny for sea urchins (Echinoidea: Echinodermata). Philos Trans R Soc Lond B Biol Sci 347:213–234

Livingston BT, Shaw R, Bailey A, Wilt F (1991) Characterization of a cDNA encoding a protein involved in formation of the skeleton during development of the sea urchin *Lytechinus pictus.* Dev Biol 148:473–480

Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605–612

Miles MJ, Carr HJ, McMaster TC, I'Anson KJ, Belton PS, Morris VJ, Field JM, Shewry PR, Tatham AS (1991) Scanning tunneling microscopy of a wheat storage protein reveals details of an unusual supersecondary structure. Proc Natl Acad Sci USA 88:68–71

Minor JE, Fromson DR, Britten RJ, Davidson EH (1991) Comparison of the bindin proteins of *Strongylocentrotus franciscanus, S. purpuratus,* and *Lytechinus variegatus:* Sequences involved in the species specificity of fertilization. Mol Biol Evol 8:781–795

Prager EM, Wilson AC (1988) Ancient origin of lactalbumin from lysozyme: analysis of DNA and amino acid sequences. J Mol Evol 27:326–335

Sanderson MJ (1989) Confidence limits on phylogenies: The bootstrap revisited. Cladistics 5:113–129

Sanderson MJ (1995) Objections to bootstrapping phylogenies: A critique. Syst Biol 44:299–320

Sidow A, Thomas WK (1994) A molecular evolutionary framework for eukaryotic model organisms. Curr Biol 4:596–603

Smith GP (1976) Evolution of repeated Dna sequences by unequal crossover. Science 191(4227):528–535

Smith MJ, Arndt A, Gorski S, Fajber E (1993) The phylogeny of echinoderm classes based on mitochondrial gene arrangements. J Mol Evol 36:545–554

Springer MS, Tusneem NA, Davidson EH, Britten RJ (1995) Phylogeny, rates of evolution, and patterns of codon usage among sea urchin retroviral-like elements, with implications for the recognition of horizontal transfer. Mol Biol Evol 12:219–230

Stanhope MJ, Smith VG, Waddell VG, Porter CA, Shivji MS, Goodman M (1996) Mammalian evolution and the interphotoreceptor retinoid binding protein (IRBP) gene: Convincing evidence for several superordinal clades. J Mol Evol 43:83–92

Sucov HM, Benson S, Robinson JJ, Britten RJ, Wilt F, Davidson EH (1987) A lineage-specific gene encoding a major matrix protein of the sea urchin embryo spicule II. Structure of the gene and derived sequence of the protein. Dev Biol 120:507–519

Swanson WJ, Vacquier VD (1998) Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein. Science 281: 710–712

Swofford DL (1998) PAUP*, phylogenetic analysis using parsimony (* and other methods), version 4.0. Sinauer Associates, Sunderland, MA

Swofford DL, Olsen GP, Waddell PJ, Hillis (1996) Phylogenetic inference. In: HIllis DM, Moritz C, Mable BK (eds) Molecular systematics. Sinauer Associates, Sunderland MA, pp 407–514

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Templeton AR (1983) Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and apes. Evolution 37:221–224

Thomas WK, Maa J, Wilson AC (1989). Shifting constraints on tRNA genes during mitochondrial DNA evolution in animals. New Biol 1(1):93–100

Thomas GH, Newbern EC, Korte CC, Bales MA, Muse SV, Clark AG, Kiehardt DP (1997). Intragenic duplication and divergence in the spectrin superfamily of proteins. Mol Biol Evol 14(12):1285–1295

Turbeville JM, Schulz JR, Raff RA (1994) Deuterostome phylogeny and the sister group of the chordates: Evidence from molecules and morphology. Mol Biol Evol 11:648–655

Urry, DW (1982) Characterization of soluble peptides of elastin by physical techniques. In: Cunningham LW, Fredericksen DW (eds) Methods of enzymology: Structural and contractile proteins, Part A. Extracellular matrix. Academic Press, New York, Vol 82, pp 673–716

Wilt FH, Benson SC (1988) Development of the endoskeletal spicule of the sea urchin embryo. In: Varner J (ed) Self-assembling architecture. Alan R. Liss, New York, pp 203–227

Zimmer EA, Martin SL, Beverley SM, Kan YW, Wilson AC (1980) Rapid duplication and loss of genes coding for the alpha chains of hemoglobin 21. Proc Natl Acad Sci USA 77:2158–2162