# The Evolutionary Relationship of the Domain Architectures in the RhoGEF-containing Proteins

Qing-Lan Sun, Hong-Jun Zhou, and Kui Lin*

*The Key Laboratory of the Ministry of Education for Biodiversity Science and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing 100875, China.*

**Domain insertions and deletions lead to variations in the domain architectures of the proteins from their common ancestor. In this work, we investigated four groups of the RhoGEF-containing proteins from different organisms with domain architectures RhoGEF-PH-SH3, SH3-RhoGEF-PH, RhoGEF-PH, and SH3-RhoGEF defined in the Pfam database. The phylogenetic trees were constructed using each individual domain and/or the combinations of all the domains. The phylogenetic analysis suggests that RhoGEF-PH-SH3 and SH3-RhoGEF-PH might have evolved from RhoGEF-PH through the insertion of SH3 independently, while SH3-RhoGEF of proteins in fruit fly might have evolved from SH3-RhoGEF-PH by the degeneration of PH domain.**

**Key words: protein domain, domain architecture, phylogenetic analysis, domain loss, RhoGEF-containing proteins**

## Introduction

Protein domains are the structural, functional, and evolutionary units of proteins (*1*, *2*). In structural biology, a domain is defined as a spatially distinct, compact, and stable protein structural unit that could conceivably fold and function in isolation. SCOP (Structural Classification of Protein; ref. *3*) and SUPERFAMILY (*4*, *5*) are two of the most useful and important structural domain databases. On the other hand, domains are often delineated as distinct regions of protein sequences that are highly conserved in evolution. Indeed, sequence-based domain definitions, central to the methods of domain discovery and assignment, represent one of the most convenient and practically important levels at which to understand the evolution and function of either proteins or domains. Many methods are available for detecting remote homologous domains in sequences, including those using position specific score matrices (*6*) and hidden Markov models (HMMs; ref. *7*, *8*). This leads to different domain databases such as CDD (Conserved Domain Database; ref. *9*), Pfam (Protein family database; ref. *10*), and SMART (Simple Modular Architecture Research Tool; ref. *11*). CDD is a curated Entrez database of conserved domain alignments that imports alignments from SMART, Pfam, and COGs (Clusters of Orthologous Groups; ref. *12*). The SMART web tool provides the annotation of the mobile eukaryotic domains and the analysis of domain architectures, on the basis of sensitive database searches and multiple alignments. Pfam is constructed from seed alignments by searching against all proteins from SP-TrEMBL and SWISS-PROT (*13*). For each obtained protein family, a profile HMM is computed. The library of HMMs is used to identify all family members, including remote ones in public protein databases, and to search a query sequence via tools such as HMMER packages (*7*).

Most proteins often consist of multiple domains (*14*). Evolutionary changes of these proteins in the domain architectures often have obvious functional implications. Variations in the domain organization of multidomain proteins have largely been attributed to domain shuffling, intramolecular duplications, domain loss, and novel domain acquisition by the fusion of distinct proteins from multifunctional polypeptides and pathways during evolution (*15*). Differences in domain architectures among multidomain proteins would often raise the problem that how the domain architectures have evolved in proteins and what relationship the domain architectures in the orthologous proteins or paralogous proteins have. In theory, the closer the protein phylogenetic relationship is, the more similar their domain architectures are. In most

*** Corresponding author.**
**E-mail: linkui@bnu.edu.cn**

cases, proteins in similar domain architectures may have close evolutionary relationships except for convergent evolution in the domain architectures.

The evolutionary relationships of domain architectures and protein sequences are often analyzed widely using the phylogenetic methods. Multidomain proteins often contain different numbers or types of domains in different orders. The different domain arrangements in protein sequences cause considerable difficulties in sequence alignment analyses. Therefore, to infer the evolutionary relationships among the different regions of multidomain protein sequences requires careful analysis of each domain that possesses a distinct evolutionary history (*2*). Traditional phylogenetic analysis is often based on the entire protein sequences. Whereas, for multidomain proteins, it is better to isolate the separate domains and carry out the phylogenetic analysis based on each domain respectively. Thus, as we can see, the domain architecture evolution can be revealed by comparing all the phylogenetic trees of individual domains and their combinations. Understanding how a given domain architecture has evolved from simpler modules is not only important to understand what the functional implications of this evolution are, but also especially meaningful for analyzing orthologous relationships between proteins (*16*).

From the Pfam database, four different domain architectures of the RhoGEF-containing proteins are extracted and analyzed (Figure 1). The RhoGEF domain is a novel module in the Guanine nucleotide exchange factors (GEFs) in the Dbl family for Rho/Rac/Cdc42-like GTPases or the Dbl-homologous (DH) domain. It encodes a GEF activity specific for some members in the Rho protein family and is about 200 amino acid residues long. The RhoGEFs are regulators of Rho proteins and control the activation state of small Rho proteins (*17*), which undergoes interconversion between active (GTP-bound) and inactive (GDP-bound) forms (GTP-Rho and GDP-Rho, respectively) (*18*). The Rho family GTPases Rho, Rac, and Cdc42 regulate a diverse array of cellular processes, including cell proliferation, apoptosis, differentiation, cytoskeletal reorganization, and membrane trafficking. The pleckstrin homology (PH) domain occurs in many proteins that are involved in intracellular signaling or as constitutes of the cytoskeleton; it is about 100 amino acid residues long. All proteins in the Dbl family share a RhoGEF domain and a PH domain. The PH domain is always located at the C-terminal of the RhoGEF domain (*19*). Biochemical

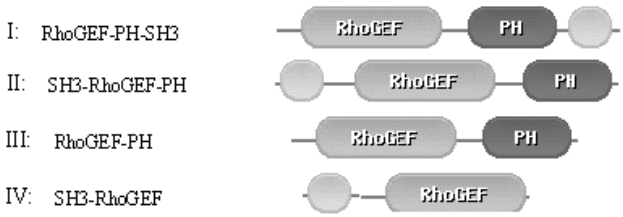


**Fig. 1** The four different domain architectures in the RhoGEF-containing proteins.

data have established the role of the conserved DH domain in the Rho GTPase interaction and activation. The DH domain of Dbl has been shown to mediate the oligomerization that is mostly homophilic in nature. The SH3 (Src homology 3) domains are often indicative of a protein involved in the signal transduction related to the cytoskeletal organization, first described in the Src cytoplasmic tyrosine kinase. They are small protein modules containing approximately 50 amino acid residues. The RhoGEFs of the Dbl family also contain the SH3 domains. There are four groups of the RhoGEF-containing proteins with similar domain architectures in the Pfam database 11.0 that we used. Three of them are found in lots of eukaryotic species, the fourth is only observed in fruit fly. Our analyses indicate that the two domain architectures, RhoGEF-PH-SH3 and SH3-RhoGEF-PH, may have evolved from the domain architecture RhoGEF-PH by the insertion of domain SH3 independently, and the domain architecture SH3-RhoGEF in fruit fly proteins may have evolved from their ancestor architecture SH3-RhoGEF-PH by the loss of domain PH.

# Results

## Phylogenetic relationships for proteins with domain architectures RhoGEF-PH-SH3 (I), SH3-RhoGEF-PH (II), and RhoGEF-PH (III)

All members of the Dbl family possess a RhoGEF domain and a PH domain. The PH domain is invariably located at the C-terminal of the RhoGEF domain. This invariant topology suggests a functional interdependence between these two structural modules (*20*), implying that the two domains might have evolved together. In order to validate this observation, two phylogenies of the proteins in all of the three domain architectures (I, II, and III) were created based on

the RhoGEF domain sequences and the PH domain sequences respectively. The phylogeny of the proteins based on the SH3-RhoGEF domain combination was also constructed according to the coexistence of the two domains. By careful examination, we found that these three phylogenetic trees display similar overall topologies. Figure 2 shows the phylogenetic tree for the PH domain. The other two trees for the RhoGEF domain and the combination of domains RhoGEF and PH are shown in Figures 3 and 4. Therefore, we can determine the phylogeny of these proteins and infer the phylogenetic relationships among the three domain architectures (I, II, and III).

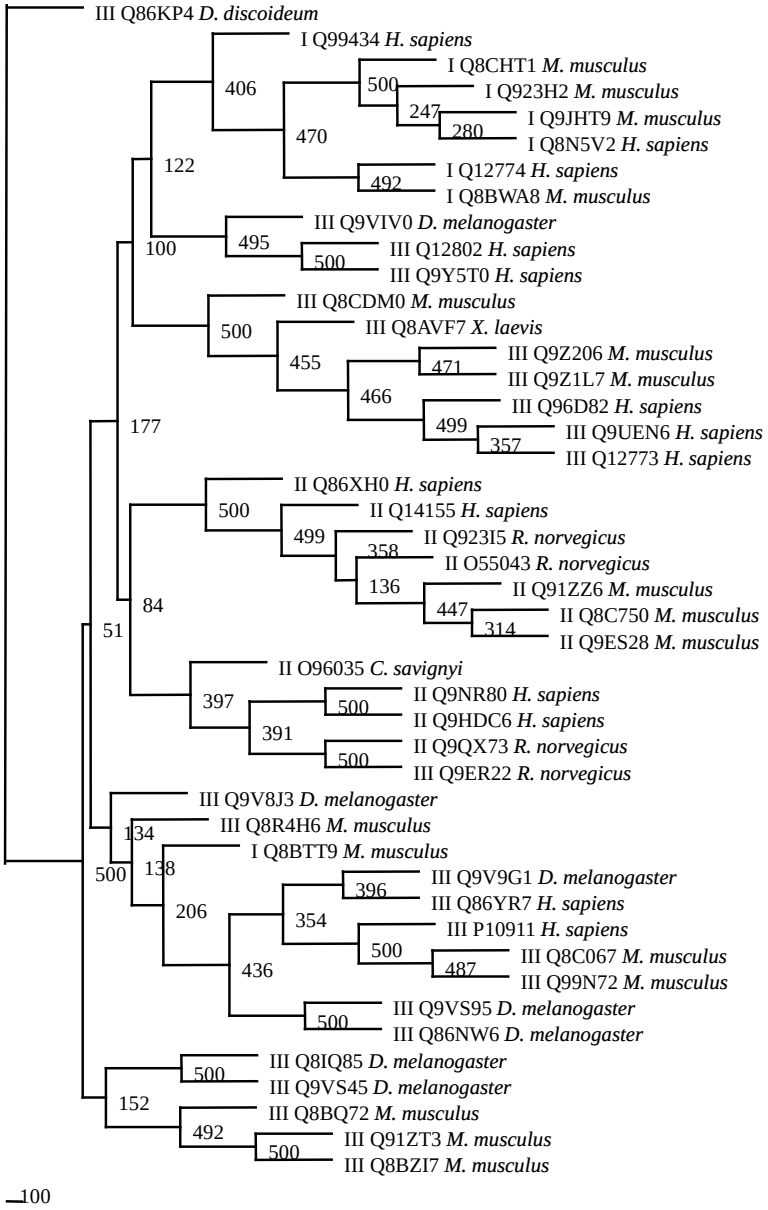


**Fig. 2** Phylogenetic tree of the proteins for the domain architectures RhoGEF-PH-SH3 (I), SH3-RhoGEF-PH (II), and RhoGEF-PH (III) based on the PH domain sequences. The bootstrapping numbers are marked at the center nodes of the clades. The first word of each OTU label (one protein) indicates to which domain architecture it belongs.

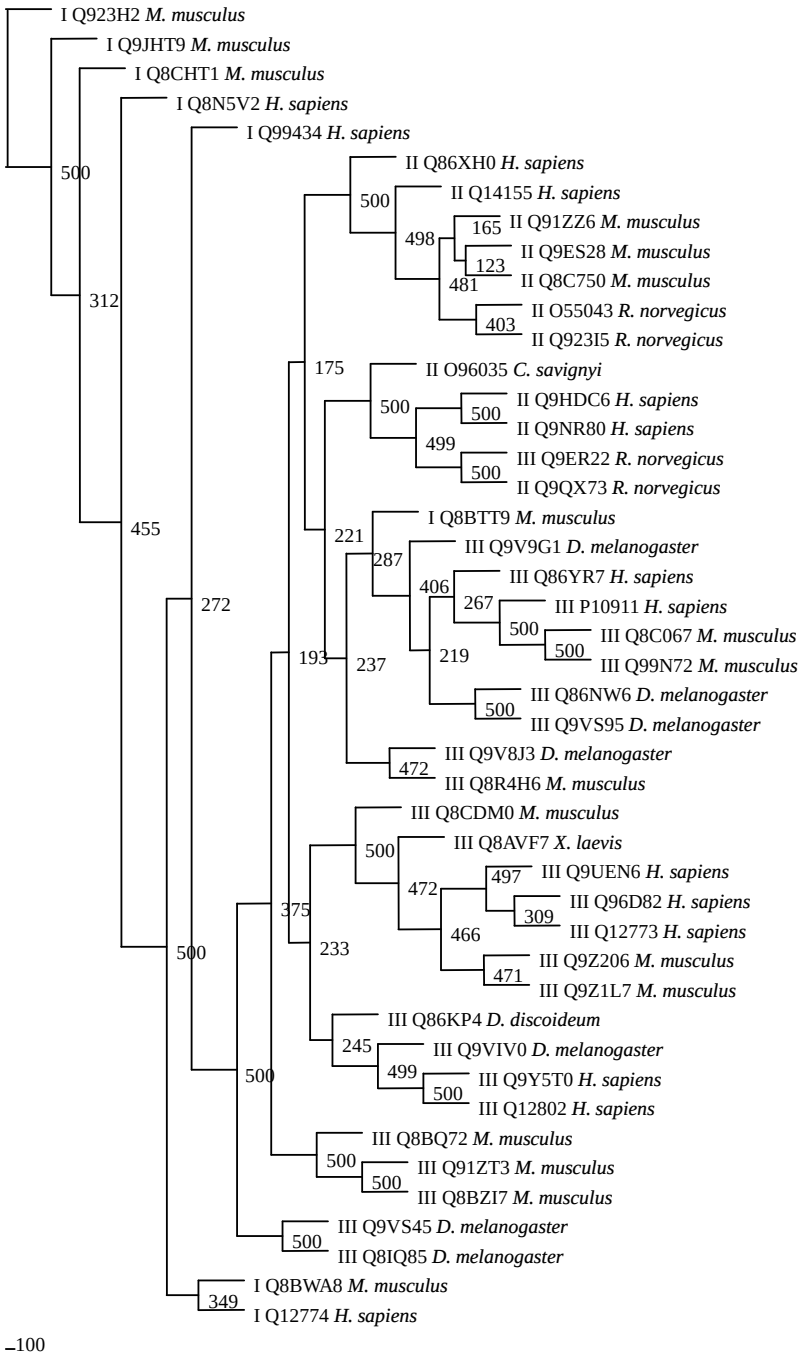


**Fig. 3** Phylogenetic tree of the proteins in the domain architectures I, II, and III based on the RhoGEF domain sequences.

In Figure 2, the proteins in the domain architectures I and II are grouped in different clades respectively, suggesting the different ancestors. Most proteins, if not all, in the same domain architecture therefore have evolved from a latest common ancestor. Interestingly, both groups of proteins in the domain architectures I and II share an SH3 domain but in different locations, one at the N-terminal and the other at the C-terminal (Figure 1). Further phylogenetic analysis based on the sequences of SH3 domains also shows that these two groups of proteins have evolved from two different latest common ancestors (Figure 5). The very similar overall topologies of the phylogenetic trees suggest that the proteins in the same domain architecture have closer evolutionary relationships. Thus, these three domains in each group have evolved in a congruent way.

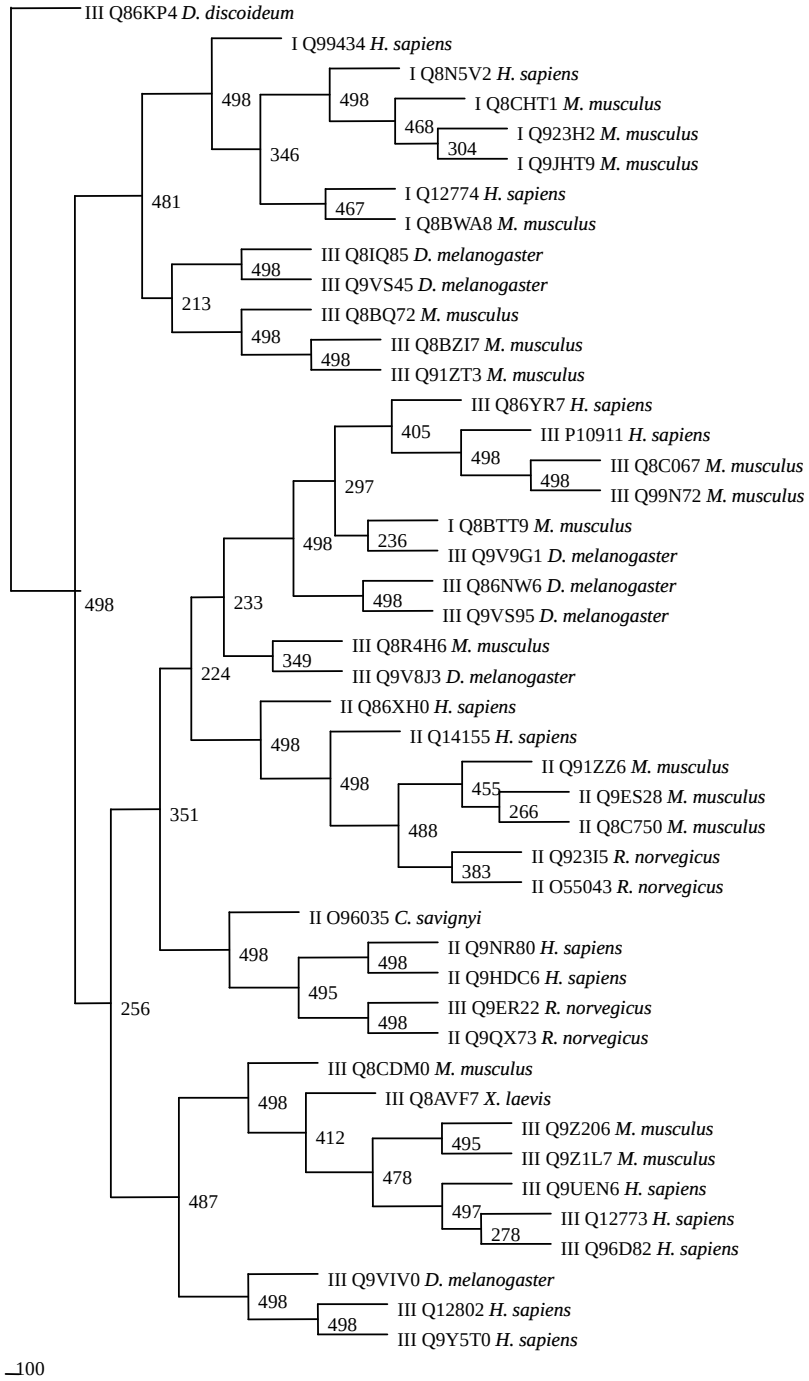


**Fig. 4** Phylogenetic tree of the proteins in the domain architectures I, II, and III based on the combination of RhoGEF and PH domain sequences.

Therefore, the domain architecture III may be ancestral and can be regarded as a supra-domain, which is an evolutionary unit larger than single protein domains. The proteins in the architecture III are divided into three sectors (Figure 2): 15 proteins form an individual clade, 10 proteins are clustered together with those proteins in the domain architecture I, and a singleton is grouped into the domain architecture II. It implies that, by the parsimonious explanation, the domain architectures I and II have evolved from the domain architecture III by the SH3 domain insertions at different ends independently. If it is the case, the proteins in the domain architectures I and II must have evolved from their ancestors in the domain architecture III. However, it may be that some proteins with the architecture III might have evolved from an

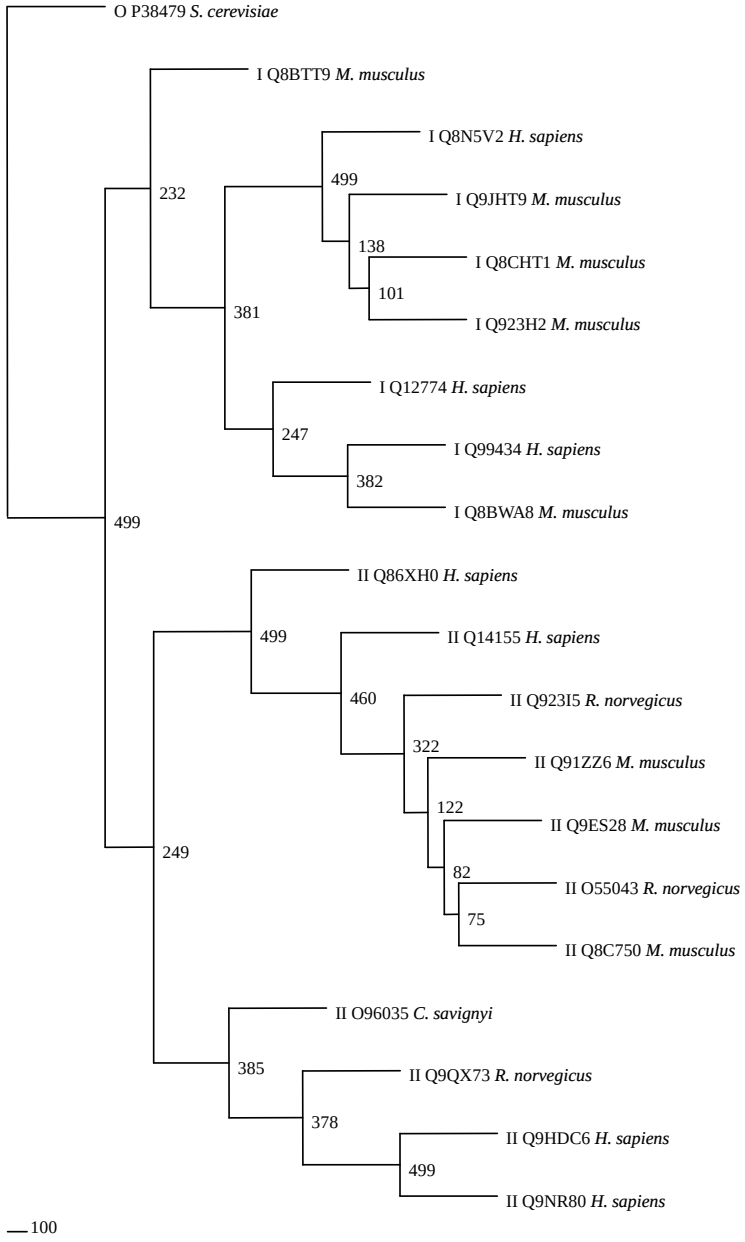


**Fig. 5** Phylogenetic tree of the proteins in the domain architectures I and II based on the SH3 domain sequences. The SH3 domain sequence of the *Saccharomyces exiguus* protein (P38479) is used as the outgroup (O).

ancestor with either architecture I or II by loss of the SH3 domain. To clarify this point, we used the profile HMMs to analyze the proteins in the architecture III. However, none of the sequence evidence of the SH3 domain degeneration has been found, which indicates that the proteins in the domain architecture III might not have evolved from those in the domain architectures I and II by the degeneration of the SH3 domain. Therefore, we conclude that the domain architectures I and II have evolved from a simpler module, the domain architecture III, with high likelihood.

## Phylogenetic relationships for proteins with domain architectures RhoGEF-PH-SH3 (I), SH3-RhoGEF-PH (II), and SH3-RhoGEF (IV)

There is another group of proteins having the domain architecture SH3-RhoGEF (IV), which is different from the patterns of the Dbl family mentioned above. The combination of domains SH3 and RhoGEF is inconsistent with the supra-domain idea of the RhoGEF-PH. This group of proteins is very interesting because all the proteins were only found

in *Drosophila melanogaster* while no protein with the domain architecture II was found in *D. melanogaster*. It is thought that the domain architecture IV may be related with the domain architecture II by loss of the PH domain in the course of evolution. To this end, we constructed the protein phylogenetic trees for the domain architectures I, II, and IV based on the RhoGEF domain and the SH3 domain respectively. In order to get a reliable phylogeny, an outgroup was used to determine the position of the root of evolution in the group. The phylogenetic tree based on the RhoGEF domain is shown in Figure 6, and Figure 7 presents the tree for the SH3 domain. The phylogenetic tree based on the combination of domains RhoGEF and SH3 was also constructed, which shows a very similar topology with individual domain trees (data not shown). In all phylogenetic trees, the proteins in the domain architectures I and II are clustered in two separate clades. However, the proteins in the domain architecture IV are grouped with those in the domain architecture II, and in turn, they are divided into two subgroups (Figures 6 and 7). The closer relationship between the architectures II and IV indicates that the proteins in the architecture IV may have evolved from an ancestral protein with the architecture II. In addition, the ancestor of the proteins in the architecture IV might also contain a PH domain, which has been lost or degenerated during the evolutionary history. To validate this hypothesis, we analyzed these ten proteins using the profile HMMs with a high E-value cutoff (50.0, default is 10.0) to detect whether there are any information of the degenerated PH domains. Interestingly, one insignificant PH domain at the C-terminal was found for each protein, with the E-value ranging from 3.8 to 41. With this discovery, we constructed a phylogenetic tree based on all the PH domain sequences, including these ten insignificant ones, from the proteins in all of the three domain architectures I, II, and IV. The result is shown in Figure 8. It is interesting that the topology of the tree is similar with those trees based on individual domains (SH3, RhoGEF) and their combination (SH3-RhoGEF). This phenomenon consolidates the correct phylogeny among these proteins. Therefore, the architecture IV have evolved from the architecture II by the degeneration of the PH domain due to loss of its function or lesser functional constraints in *D. melanogaster*. The observation that the degeneration only likely occurs in *D. melanogaster* implies that the evolution of the architecture IV may be lineage-specific.

# Discussion

We have constructed the phylogenies of four groups of the RhoGEF-containing proteins with similar domain architectures, RhoGEF-PH-SH3 (I), SH3-RhoGEF-PH (II), RhoGEF-PH (III), and SH3-RhoGEF (IV). The phylogenetic relationships of these proteins were examined carefully by comparing these trees. The proteins in the same domain architecture are suggested having evolved from the same latest common ancestors. Domain insertions or deletions make differences among related domain architectures. The analyses on the evolution of the four domain architectures may share light on the complication that how a given domain architecture has evolved. Within the proteins we investigated, the architectures RhoGEF-PH-SH3 (I) and SH3-RhoGEF-PH (II) have evolved from the architecture RhoGEF-PH (III) by the SH3 domain insertion at different polypeptide ends, although it is possible that some proteins in the architecture RhoGEF-PH might contain the SH3 domain(s) undetected by the profile HMM analysis. Interestingly, by carefully examining the results from the phylogenetic analyses based on sequences of each individual domain, it is very confident that the domain architecture SH3-RhoGEF (IV) have evolved from the architecture SH3-RhoGEF-PH (II) by degenerating the PH domain in respective proteins in fruit fly.

The analysis of the RhoGEF domain-containing proteins shows that the phylogeny construction of the proteins, which is based on each individual domain or their combinations rather than based on the whole protein sequences, can give insights into the constraints and mechanisms of the protein evolution. Any approach of analyzing the phylogenetic relationships of the proteins based on the whole sequences may fail because of variations in domain architectures. In our analyses, the overall topology of all phylogenetic trees based on individual domains is very similar except for minor incongruence in deep branches. This may be caused by too many variations within these proteins from different species, including different evolution rates of these proteins. On the other hand, as we know, the maximum parsimony (MP) does not use all the sequence information and does not correct for multiple mutations (no such model of evolution). Meanwhile it does not provide information on branch lengths and is sensitive to codon bias. Although it is of high quality, the Pfam database may contain a small proportion of false positives and false negatives. The protein Q8BTT9 from *Mus musculus*

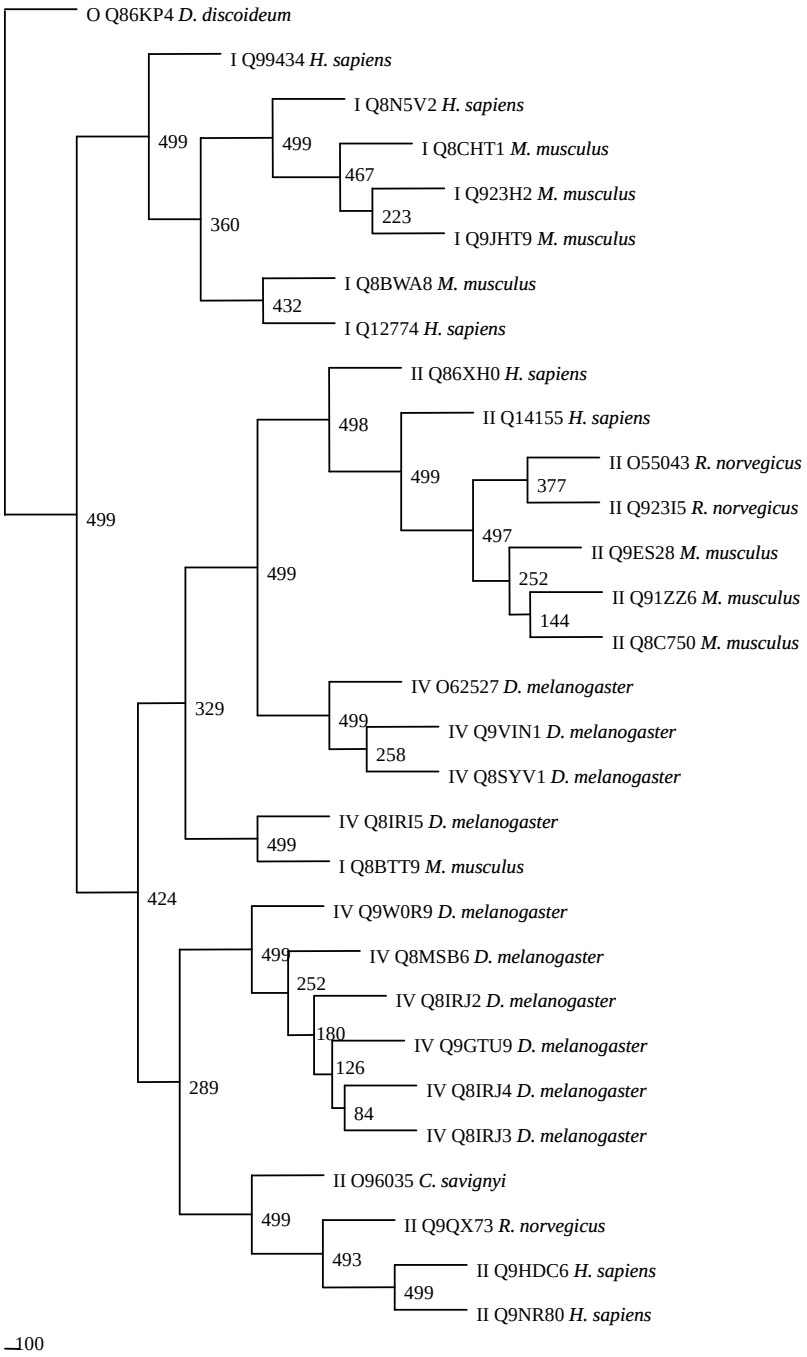


**Fig. 6** Phylogenetic tree of the proteins for the domain architectures I, II, and IV based on the RhoGEF domain sequences. The RhoGEF domain sequence of the *Dictyostelium discoideum* protein (Q86KP4) is used as the outgroup.

in the domain architecture I and the protein Q8IRI5 from *D. melanogaster* in the domain architecture IV are always at the anomalous positions. We have checked Q8BTT9 and found that the annotation of the protein in Pfam version 16.0 has been changed to possess a different architecture RhoGEF-PH-SH3 from RhoGEF-PH-SH3_2 in Pfam 11.0. So we argue that some anomalous positions of proteins in the phylogenetic trees may be resulted from the misannotations in the databases we used.

The commonly used methods for reconstructing phylogenic relationships are the maximum likelihood (ML), MP, and various distance-based methods. The distance-based methods are able to predict evolutionary relationship when variation among the sequences is present and the amount of variation is intermediate.

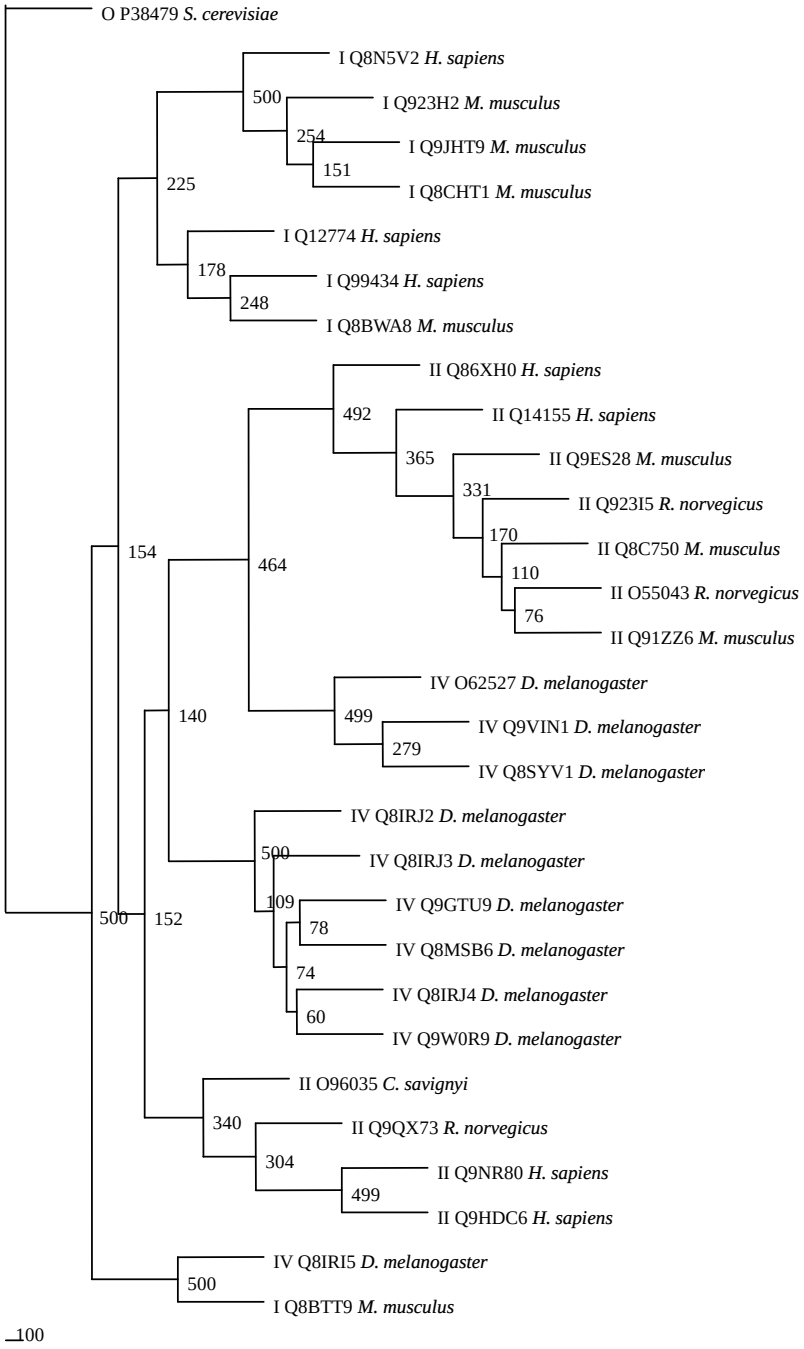


**Fig. 7** Phylogenetic tree of the proteins for the domain architectures I, II, and IV based on the SH3 domain sequences. The SH3 domain sequence of the *Saccharomyces exiguus* protein P38479 is used as the outgroup.

The ML method is particularly useful for more variable sequences, and it has lower variance than other methods but is very slow and depends on the model of evolution (*21*). The MP method searches all possible tree topologies for the optimal tree and evaluates different trees; it only uses the informative sites and tries to provide information on the ancestral sequences; it is cogent when the amount of variation among sequences is small. Besides MP, we have also used the distance-based method (protdist + neighbor in the Phylip toolkit) to infer potential phylogenies of these proteins. However, the outgroup is misplaced in the respective trees (data not shown). As it is considered that the sequences of the members in the domain family detected by HMMs in Pfam are highly conserved, as mentioned above, the distance-based method might be not suitable for the dataset we collected. In addition, we think that 500 bootstrap

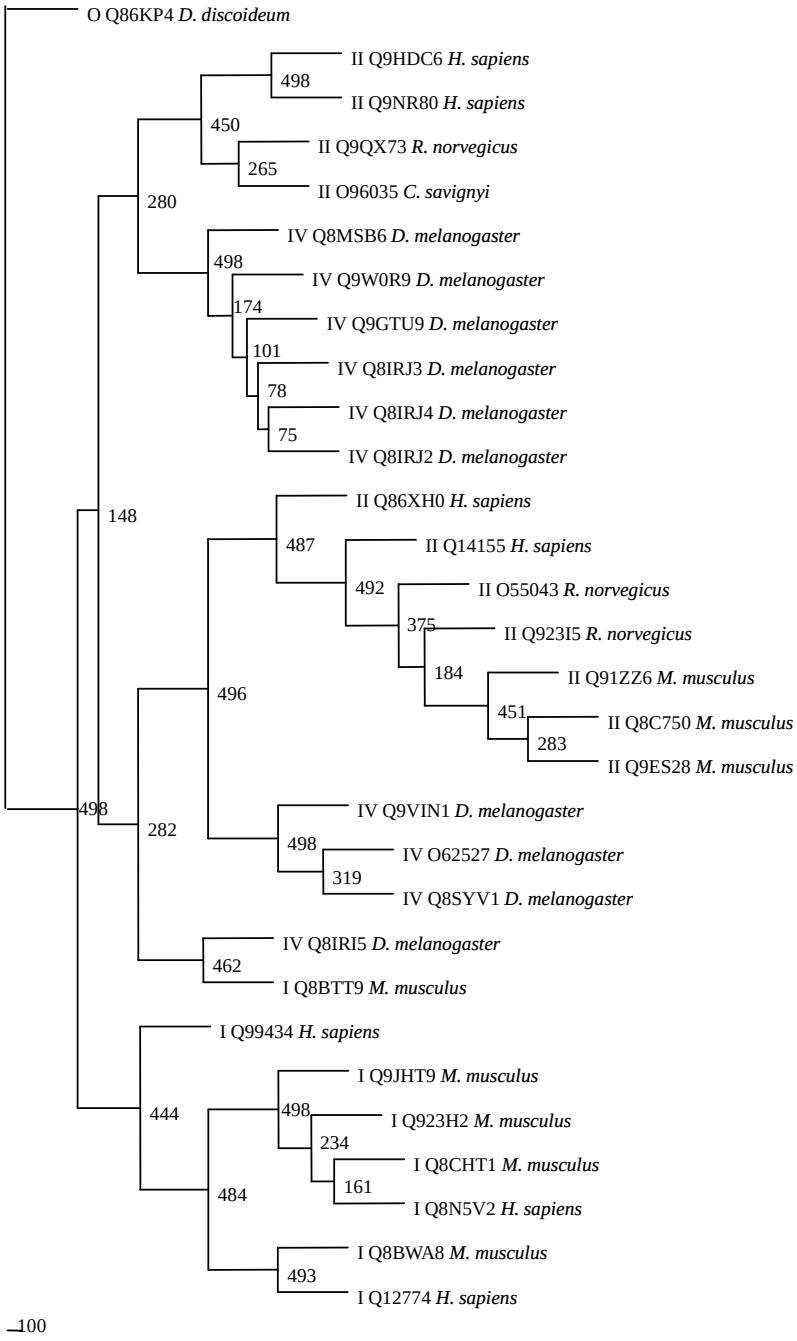


**Fig. 8** Phylogenetic tree of the proteins for the domain architectures I, II, and IV based on the PH domain sequences. The PH domain sequence of the *Dictyostelium discoideum* protein (Q86KP4) is used as the outgroup.

replicates in general are enough to get a reliable tree statistically.

Differences in domain architectures among genes that have clearly arisen, at least in part, from a common ancestor, raise the question of whether these genes are orthologous. Hence, tracing the history of a certain domain architecture is important for functional annotation of multidomain proteins, and for understanding the function of individual domains (*16*). Meanwhile, it is worth investigating the mechanisms of the variations in the domain architectures of multidomain proteins and analyzing the relative contributions of the domain shuffling, domain duplication, domain loss, and polypeptide fusion leading to domain rearrangement. We must note that descriptions of orthology are most appropriately applied to domains, rather than proteins, except when proteins contain identical domain architectures (*2*).

# Materials and Methods

## Domain datasets

The protein domains and evolutionary families were extracted from the Protein family database (Pfam 11.0 as of November 2003). Pfam is an accurate and comprehensive collection of protein domains and families. It is composed of two parts; the first part, Pfam-A, is the manually curated collection of protein families and is believed in high quality. The second part is Pfam-B, in which sequence segments that are not included in Pfam-A are clustered automatically. We used the Pfam version 11.0 and only the Pfam-A part. The files of the Pfam MySQL database were downloaded from the Pfam ftp server (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/database_files/) and stored in MySQL DBMS locally. We only focused on the RhoGEF (PF00621), PH (PF00169), and SH3 (PF00018, in the version 17.0 named SH3_1) domains and extracted all proteins that contained at least these three domains. We excluded all of those proteins containing the RhoGEF domain that are not annotated in good quality, such as the putative and hypothetical proteins. Two groups of proteins that contain all three domains of interest were identified and designated by RhoGEF-PH-SH3 and SH3-RhoGEF-PH, respectively. For all of the other proteins containing both the RhoGEF domain and either the PH or the SH3 domain, we grouped them into the other two groups according to their domain architectures and assigned these two groups with RhoGEF-PH and SH3-RhoGEF, respectively. Thus we have four groups of proteins with different domain architectures. There are 8 proteins in RhoGEF-PH-SH3, 11 in SH3-RhoGEF-PH, 26 in RhoGEF-PH, and 10 in SH3-RhoGEF. These proteins are from *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Ciona Savignyi*, *Xenopus laevis*, *Dictyostelium discoideum*, and *Drosophila melanogaster*, respectively. The four domain architectures are shown in Figure 1 and denoted as the domain architecture I, II, III, and IV, respectively. The list of protein accession numbers defined in SWISS-PROT is given in Table 1. The corresponding sequences and annotations of the proteins were extracted from the SWISS-PROT database release 41.25 and the SP-TrEMBL release 24.14.

**Table 1 Proteins in Four Groups of Domain Architectures**

| Domain architecture | Protein accession number in SWISS-PROT and TrEMBL | |
|---|---|---|
| RhoGEF-PH-SH3 (I) | *H. sapiens* | Q12774, Q8N5V2, Q99434 |
| (8 proteins) | *M. musculus* | Q8CHT1, Q8BTT9, Q8BWA8, Q923H2, Q9JHT9 |
| SH3-RhoGEF-PH (II) | *H. sapiens* | Q14155, Q86XH0, Q9H0C6, Q9NR80 |
| (11 proteins) | *M. musculus* | Q8C750, Q91Z26, Q9ES28 |
| | *R. norvegicus* | O55043, Q923I5, Q9QX73 |
| | *C. Savignyi* | O96035 |
| RhoGEF-PH (III) (26 proteins) | *H. sapiens* | P10911, Q12773, Q12802, Q86YR7, Q96D82, Q9UEN6, Q9Y5T0 |
| | *M. musculus* | Q8BQ72, Q8BZI7, Q8C067, Q8CDM0, Q8R4H6, Q91ZT3, Q99N72, Q9Z1L7, Q9Z206 |
| | *R. norvegicus* | Q9ER22 |
| | *X. laevis* | Q8AVF7 |
| | *D. discoideum* | Q86KP4 |
| | *D. melanogaster* | Q86NW6, Q8IQ85, Q9V8J3, Q9V9G1, Q9VIV0, Q9VS45, Q9VS95 |
| SH3-RhoGEF (IV) (10 proteins) | *D. melanogaster* | Q8IRJ2, Q8IRJ3, Q9GTU9, Q8IRI5, O62527, Q8IRJ4, Q9VIN1, Q9W0R9, Q8MSB6, Q8SYV1 |

### Phylogenetic analysis

According to the domain organizations of the proteins, we first analyzed the phylogenetic relationship between the domain architectures RhoGEF-PH-SH3 (I), SH3-RhoGEF-PH (II), and RhoGEF-PH (III), and then discussed the phylogenies among the domain architectures RhoGEF-PH-SH3 (I), SH3-RhoGEF-PH (II), and SH3-RhoGEF (IV).

The domain sequences were extracted from their protein sequences according to the respective annotations by Pfam. The individual domain sequences were aligned using CLUSTAL W version 1.83 (*22*) installed on local Linux platform. All alignments were manually checked up for good results prepared for further analysis. The alignments are given in Figures S1-S7 (Supporting Online Material).

Phylogenetic analyses using amino acid sequences were conducted using the MP method. It is a character-based cladistic method that infers a phylogenetic tree by minimizing the total number of evolutionary steps required to explain a given set of data, or in other words by minimizing the total tree length. When applied to protein sequence data, the MP method either considers each site of the sequence as a multistate unordered character with 20 possible states (the amino acids), or may take into account the genetic code and the number of mutations, 1, 2, or 3, which is required to explain an observed amino-acid substitution. The latter method is implemented in the PROTPARS program included in the PHYLIP package (*23*). All phylogenetic analyses were performed using the PHYLIP package version 3.6b locally. Each protein sequence dataset was analyzed under the optimality criteria of MP. Tree space was searched using the branch-and-bound algorithm, which guarantees to find the optimal tree(s). Tree reliability under both optimality criteria was assessed using non-parametric bootstrap re-sampling of 500 replicates. The trees are displayed with the TreeView package version 1.6.6 (*24*). The datasets and trees are available from the authors on request.

### Degeneration of domain detection using HMMs

In order to test whether the proteins of interest may contain information of the degeneration of any of these three domains, the profile HMM analysis was performed. HMMs, which are usually more sensitive for detecting remote homologous domains than pairwise approaches, are widely used to search the protein sequences for remote homologues. The HMM library file Pfam_ls (version 11.0) was downloaded from the Pfam ftp site and the profile HMMs of RhoGEF, PH, and SH3 were retrieved. These three profiles were searched against all of the sequences for significantly similar domain matches with the E-value cutoff = 50.0 using the HMMER software run locally. Such high E-value cutoff used is to detect the insignificantly potential domains that are excluded in the Pfam annotations.

## Acknowledgements

We thank anonymous reviewers for their helpful comments. This work was supported by the National High-Tech Research and Development Program of China (Grant No. 2003AA231030), the Excellent Young Teachers Program of the Ministry of Education of China (2003), and Beijing Normal University.

## References

1. Murzin, A.G., *et al.* 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247: 536-540.
2. Ponting, C.P. and Russell, R.R. 2002. The natural history of protein domains. *Annu. Rev. Biophys. Biomol. Struct.* 31: 45-71.
3. Andreeva, A., *et al.* 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* 32: D226-229.
4. Gough, J. and Chothia, C. 2002. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* 30: 268-272.
5. Madera, M., *et al.* 2004. The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.* 32: D235-239.
6. Altschul, S.F., *et al.* 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
7. Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* 14: 755-763.
8. Karplus, K., *et al.* 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14: 846-856.
9. Marchler-Bauer, A., *et al.* 2003. CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* 31: 383-387.

10. Bateman, A., *et al.* 2004. The Pfam protein families database. *Nucleic Acids Res.* 32: D138-141.
11. Letunic, I., *et al.* 2004. SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* 32: D142-144.
12. Tatusov, R.L., *et al.* 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4: 41.
13. Boeckmann, B., *et al.* 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31: 365-370.
14. Henikoff, S., *et al.* 1997. Gene families: the taxonomy of protein paralogs and chimeras. *Science* 278: 609-614.
15. Braun, E.L. and Grotewold, E. 2001. Fungal Zuotin proteins evolved from MIDA1-like factors by lineage-specific loss of MYB domains. *Mol. Biol. Evol.* 18: 1401-1412.
16. Storm, C.E. and Sonnhammer, E.L. 2001. NIFAS: visual analysis of domain evolution in proteins. *Bioinformatics* 17: 343-348.
17. Schwartz, M. 2004. Rho signalling at a glance. *J. Cell Sci.* 117: 5457-5458.
18. Takai, Y., *et al.* 1995. Rho as a regulator of the cytoskeleton. *Trends Biochem. Sci.* 20: 227-231.
19. Soisson, S.M., *et al.* 1998. Crystal structure of the Dbl and pleckstrin homology domains from the human Son of sevenless protein. *Cell* 95: 259-268.
20. Cerione, R.A. and Zheng, Y. 1996. The Dbl family of oncogenes. *Curr. Opin. Cell Biol.* 8: 216-222.
21. Thornton, J.W. and DeSalle, R. 2000. Gene family evolution and homology: genomics meets phylogenetics. *Annu. Rev. Genomics Hum. Genet.* 1: 41-73.
22. Thompson, J.D., *et al.* 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.
23. Felsenstein, J. 1989. PHYLIP–phylogeny inference package (Version 3.2). *Cladistics* 5: 164-166.
24. Page, R.D. 1996. TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* 12: 357-358.

## Supporting Online Material

http://www.gpbjournal.org/journal/pdf/GPB3(2)-05.pdf
Figures S1–S7