Diversity and Evolution of the Thyroglobulin Type-1 Domain Superfamily

Marko Novinec, Dušan Kordiš,* Vito Turk,* and Brigita Lenarčič*†*

*Department of Biochemistry and Molecular Biology, Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia; and †Department of Chemistry and Biochemistry, Faculty of Chemistry and Chemical Technology, University of Ljubljana, Aškerčeva 5, Ljubljana, Slovenia

Multidomain proteins are gaining increasing consideration for their puzzling, flexible utilization in nature. The presence of the characteristic thyroglobulin type-1 (Tg1) domain as a protein module in a variety of multicellular organisms suggests pivotal roles for this building block. To gain insight into the evolution of Tg1 domains, we performed searches of protein, expressed sequence tag, and genome databases. Tg1 domains were found to be Metazoa specific, and we retrieved a total of 170 Tg1 domain–containing protein sequences. Their architectures revealed a wide taxonomic distribution of proteins containing Tg1 domains followed or preceded by secreted protein, acidic, rich in cysteines (SPARC)-type extracellular calcium-binding domains. Other proteins contained lineage-specific domain combinations of peptidase inhibitory modules or domains with different biological functions. Phylogenetic analysis showed that Tg1 domains are highly conserved within protein structures, whereas insertion into novel proteins is followed by rapid diversification. Seven different basic types of protein architecture containing the Tg1 domain were identified in vertebrates. We examined the evolution of these protein groups by combining Tg1 domain phylogeny with additional analyses based on other characteristic domains. Testicans and secreted modular calcium binding protein (SMOCs) evolved from invertebrate homologs by introduction of vertebrate-specific domains, nidogen evolved by insertion of a Tg1 domain into a preexisting architecture, and the remaining four have unique architectures. Thyroglobulin, Trops, and the major histocompatibility complex class II–associated invariant chain are vertebrate specific, while an insulin-like growth factor–binding protein and nidogen were also identified in urochordates. Among vertebrates, we observed differences in protein repertoires, which result from gene duplication and domain duplication. Members of five groups have been characterized at the molecular level. All exhibit subtle differences in their specificities and function either as peptidase inhibitors (thyropins), substrates, or both. As far as the sequence is concerned, only a few conserved residues were identified. In combination with structural data, our analysis shows that the Tg1 domain fold is highly adaptive and comprises a relatively well-conserved core surrounded by highly variable loops that account for its multipurpose function in the animal kingdom.

Introduction

Many proteins in multicellular organisms are made from combinations of several autonomously folding domains or modules. During evolution, these units have been highly mobile and have spread into previously existing proteins or gave birth to new architectures with novel biological functions, principally by mechanisms of exon shuffling and duplication (Patthy 2003; Vogel et al. 2004*a*). By participating in interactions with multiple partners, multidomain proteins contribute greatly to organism complexity (Patthy 2003). Although the repertoire of possible domain combinations is, in principle, virtually unlimited, only a constrained number is actually present as a consequence of strong evolutionary selection. Some domains have been shown to be highly versatile, whereas most have only a limited set of neighboring domains (Vogel et al. 2004*a*). With the huge increase in available genomic and expressed sequence tag (EST) data over the past few years, our knowledge of the presence and versatility of many different modules has been greatly enlarged.

Thyroglobulin type-1 (Tg1) domains are classified as a superfamily belonging to the class of small proteins in the Structural Classification of Proteins database (Murzin et al. 1995). They were originally identified as cysteine-rich motifs (Mercken et al. 1985), organized into 10 tandem repeated units in the N-terminal third of thyroglobulin (Parma et al. 1987). Molina et al. (1996) identified an eleventh repeat and characterized the repeats as being modules found in proteins from different families. Based on the number of disulfide bonds, this group also divided Tg1 domains into two subtypes, subtype 1A (Tg1A) containing three disulfide bonds and subtype 1B (Tg1B) without the second disulfide bond.

Since their original identification in thyroglobulin, Tg1 domains have been found as parts of many proteins with different origins and functions. Some Tg1 domain–containing proteins inhibit certain peptidases, mostly cysteine cathepsins. These proteins were named thyropins (Lenarčič and Bevec 1998) and classified as peptidase inhibitor family I31 (Rawlings, Tolle, and Barrett 2004). For some thyropins, it has been shown that their inhibitory properties reside on Tg1 domains (Bevec et al. 1996; Lenarčič and Turk 1999; Lenarčič et al. 2000; Meh et al. 2005). The best studied thyropin is the p41 splice variant of the major histocompatibility complex class II–associated invariant chain, which is involved in antigen processing (Stumptner-Cuvelette and Benaroch 2002). The crystal structure of the p41 fragment in complex with cathepsin L (Gunčar et al. 1999) is the foundation of our understanding of thyropin action. Other well-studied thyropins include equistatin from sea anemone *Actinia equina* (Lenarčič and Turk 1999; Galeša et al. 2003), saxiphilin from the bullfrog *Rana catesbeiana* (Lenarčič et al. 2000), the chum salmon egg peptidase inhibitor (Yamashita and Konagaya 1996), and human testican-1 (Bocock et al. 2003). The latter is unique among the thyropins, being not only an inhibitor but also a substrate for certain cysteine cathepsins (Meh et al. 2005). The evidence for Tg1 domain functions other than peptidase inhibition comes from the insulin-like growth factor–binding protein (IGFBP)-6. The Tg1 domains of the IGFBPs have long been proposed to be involved in the binding of insulin-like growth factors (IGFs), and these interactions have recently been

Key words: thyroglobulin type-1 domain, protein module, evolution, Metazoa.

E-mail: brigita.lenarcic@ijs.si.

Mol. Biol. Evol. 23(4):744–755. 2006
doi:10.1093/molbev/msj082
Advance Access publication December 20, 2005

confirmed and characterized on the molecular level (Headey et al. 2004).

Most of the identified Tg1 domains remain functionally uncharacterized. However, they are among the most versatile domains found in metazoans. In this study, we searched protein, EST, and genome databases for proteins containing Tg1 domains and examined their domain architectures. We have analyzed the phylogenetic relationships between Tg1 domains and evaluated the correlation between their phylogeny and protein architecture. Further, we focused on the origins and evolution of vertebrate Tg1 domain–containing proteins in terms of domain architecture, Tg1 domain phylogeny, and additionally performed analyses. Finally, we looked into the functional and structural diversity within the Tg1 domain superfamily in the light of our study and other available data.

Methods

Data Collection

Sequences of Tg1 domain–containing proteins were obtained by Blast (www.ncbi.nlm.nih.gov/Blast/) searches of the GenBank nonredundant protein database, EST databases, and available genome sequences (see Supplement 1 of the Supplementary Material online). Coding sequences were extracted from genomic data using GENSCAN (Burge and Karlin 1997). Whenever possible complete protein coding sequences were obtained.

The set of vertebrate proteins includes sequences obtained from human (*Homo sapiens*), mouse (*Mus musculus*), cow (*Bos taurus*), chicken (*Gallus gallus*), frog (*Xenopus laevis*), zebrafish (*Danio rerio*), and pufferfish (*Tetraodon nigroviridis*). Invertebrate sequences were obtained from available EST databases and the following genome databases: sea urchin (*Strongylocentrotus purpuratus*), sea squirt (*Ciona intestinalis*), amphioxus (*Branchiostoma floridae*), *Schmidtea mediterranea*, nematode (*Caenorhabditis elegans*), fruit fly (*Drosophila melanogaster*), honeybee (*Apis mellifera*), flour beetle (*Tribolium castaneum*), starlet sea anemone (*Nematostella vectensis*), and marine sponge (*Reniera* sp. JGI-2005).

Determination of Domain Architecture

For characterized proteins, domain architectures were taken from annotations. For uncharacterized proteins, architectures were determined by querying the Conserved Domain Database v2.02 (www.ncbi.nlm.nih.gov/structure/cdd/) and the InterPro release 8.1 database (www.ebi.ac.uk/interpro/). Additionally, we performed BlastP searches of the GenBank nonredundant protein database. Partial protein sequences were used as queries, and homologous regions of Blast hits were examined for known structural elements.

Sequence Alignment and Analysis

Amino acid sequences were aligned with ClustalW version 1.83 (Thompson, Higgins, and Gibson 1994) using the BLOSUM series matrices. The created alignments were manually checked and refined with SEAVIEW (Galtier, Gouy, and Gautier 1996). Residue conservation was determined manually.

Phylogenetic Analysis

Phylogenetic analyses were performed with the PHYLIP package version 3.63 (Felsenstein 1989). Multiple sequence alignments were used to create 100 replicates with the SEQBOOT module. Then, PROTDIST was used to calculate protein distances with the Probability Matrix from Blocks model (Veerassamy, Smith, and Tillier 2003). From the generated distance matrices, trees were created using the NEIGHBOR module with the neighbor-joining (NJ) method (Saitou and Nei 1987). Trees were visualized and manipulated, and consensus trees were created with the TreeExplorer program of Koichiro Tamura.

Results and Discussion

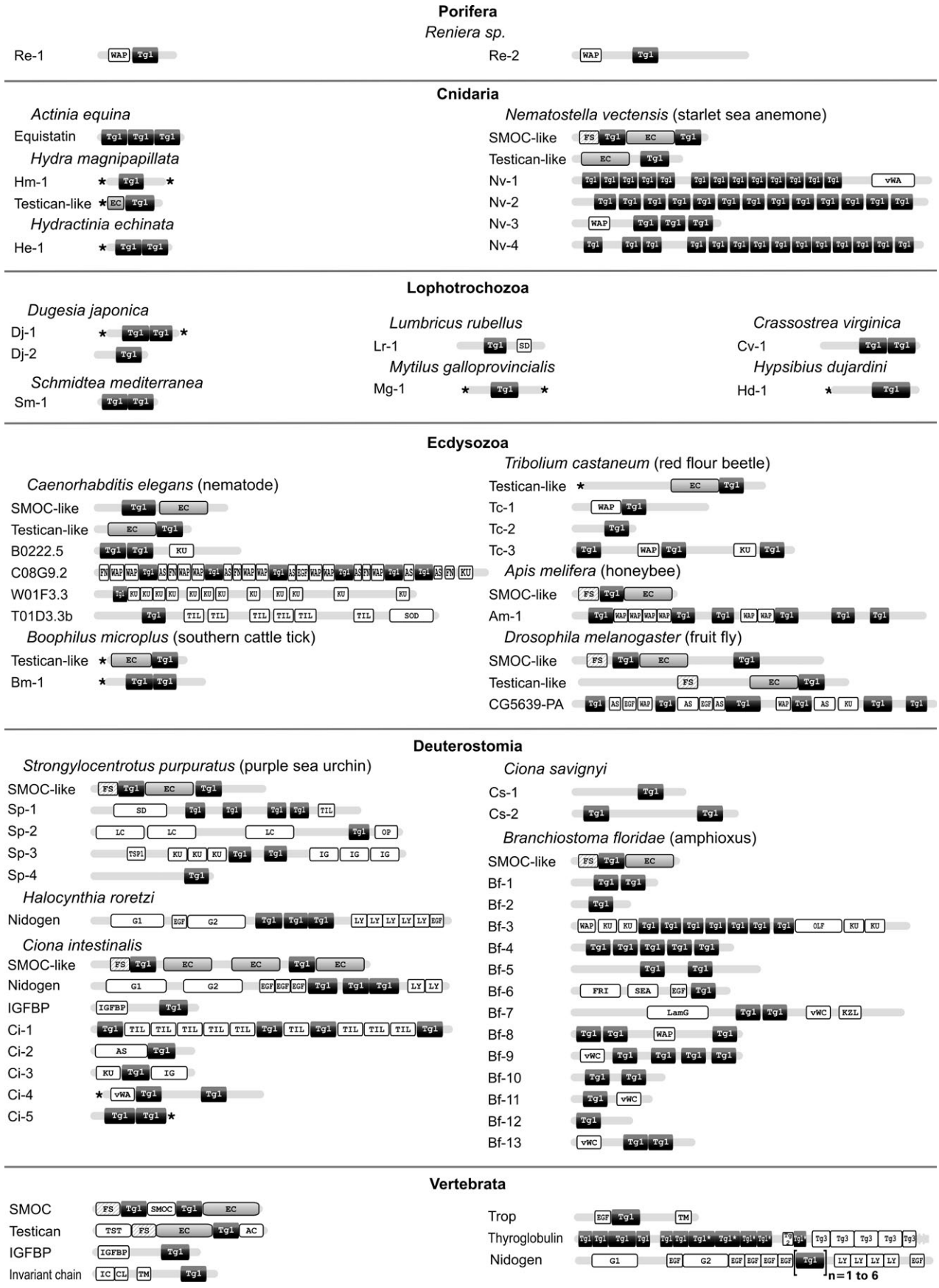
Domain Architecture and Taxonomic Distribution of Tg1 Domain–Containing Proteins

We retrieved 170 protein sequences containing 333 Tg1 domains. Some were already available, but most were reconstructed from EST and genomic data. Complete coding sequences were retrieved for the majority of proteins, while for some only partial sequences were obtained. Domain architectures were determined for previously unidentified proteins, and the complete inventory is available as Supplement 1 of the Supplementary Material online.

The recognized domain architectures show that Tg1 domains are highly versatile modules, which were mobilized into proteins with greatly different configurations during metazoan evolution (fig. 1). Tg1 domain–containing proteins are unequally distributed among diverse metazoan lineages, and distinct domain compositions predominantly occur in lineage-specific patterns. The repertoire of domain combinations is broad and includes peptidase inhibitory modules of different classes, such as BPTI/Kunitz domains (Laskowski 1986), whey acidic protein (WAP) domains (Hennighausen and Sippel 1982), and trypsin inhibitor–like cysteine-rich domains (Grasberger, Clore, and Gronenborn 1994). Of modules with distinct functions frequently found in other extracellular proteins, von Willebrand factor–like domains, immunoglobulin-like domains, and epidermal growth factor–like domains are most frequent, while others occur to a more limited extent. Tg1 domains can also be incorporated into proteins with otherwise nonmodular nature, for example the invariant chain.

In the demosponge *Reniera*, the representative of the most basal animal lineage, both identified Tg1 domain–containing proteins contain combinations of a Tg1 domain and a WAP domain (fig. 1). This ancient domain combination that is conserved and further expanded by additional domains throughout the animal kingdom was however lost during the evolution of most deuterostome lineages.

In cnidarians, proteins are predominantly composed of tandem Tg1 domain repeats. The number of repeated units varies from 2 to as many as 14 in *N. vectensis* (fig. 1). Equistatin, a well-characterized protein from the sea anemone *A. equina*, comprises three Tg1 domains which exhibit a remarkable degree of inhibitory diversity. The first domain inhibits cysteine cathepsins, the second one the aspartic peptidase cathepsin D, while no activity toward peptidases has been identified for the third one (Galeša et al. 2003).



Two proteins containing Tg1 domains and SPARC-type extracellular calcium-binding (EC) domains were identified in Cnidaria (fig. 1). Similar proteins were identified throughout the animal kingdom, all sharing a Tg1-EC domain pair. This element resembles a supradomain, as defined by Vogel et al. (2004*b*), and may be organized in two different orientations. In vertebrates, one orientation is found in testicans (domain order EC-Tg1) and the other in SMOCs (domain order Tg1-EC); therefore, we term them testican-type and SMOC-type supradomain, respectively. Accordingly, proteins containing one of these supradomains were termed testican-like or SMOC-like proteins.

Lophotrochozoa are the only metazoan group besides sponges in which none of the supradomains were found, and the small number of retrieved sequences indicates secondary gene loss of Tg1 domains in this lineage. However, with the availability of new genome data, further Tg1 domain sequences might be identified. In Ecdysozoa, Tg1 domains are more versatile than in cnidarians and proteins comprise unique arrangements of different types of peptidase-inhibiting modules, but in general Tg1 domains appear to have been infrequently utilized during protostome evolution.

Tg1 domains, however, have been highly mobile during deuterostome evolution, and each taxonomic group possesses a characteristic set of Tg1 domain–containing proteins with some unique domain combinations. The vertebrate repertoire consists of seven architecturally distinct groups. Three of these, the invariant chain, Trops, and thyroglobulin are found exclusively in vertebrates. Nidogen and IGFBP orthologs were also found in the urochordate *C. intestinalis* but surprisingly not in the cephalochordate *B. floridae*. Homologs of the remaining two groups, testicans and SMOCs, are present throughout metazoans; however, vertebrate proteins are characterized by addition of unique vertebrate-specific domains (the testican-unique domain in testicans and the SMOC-unique domain in SMOCs).

The Phylogeny of Tg1 Domains in Correlation with Protein Architectures

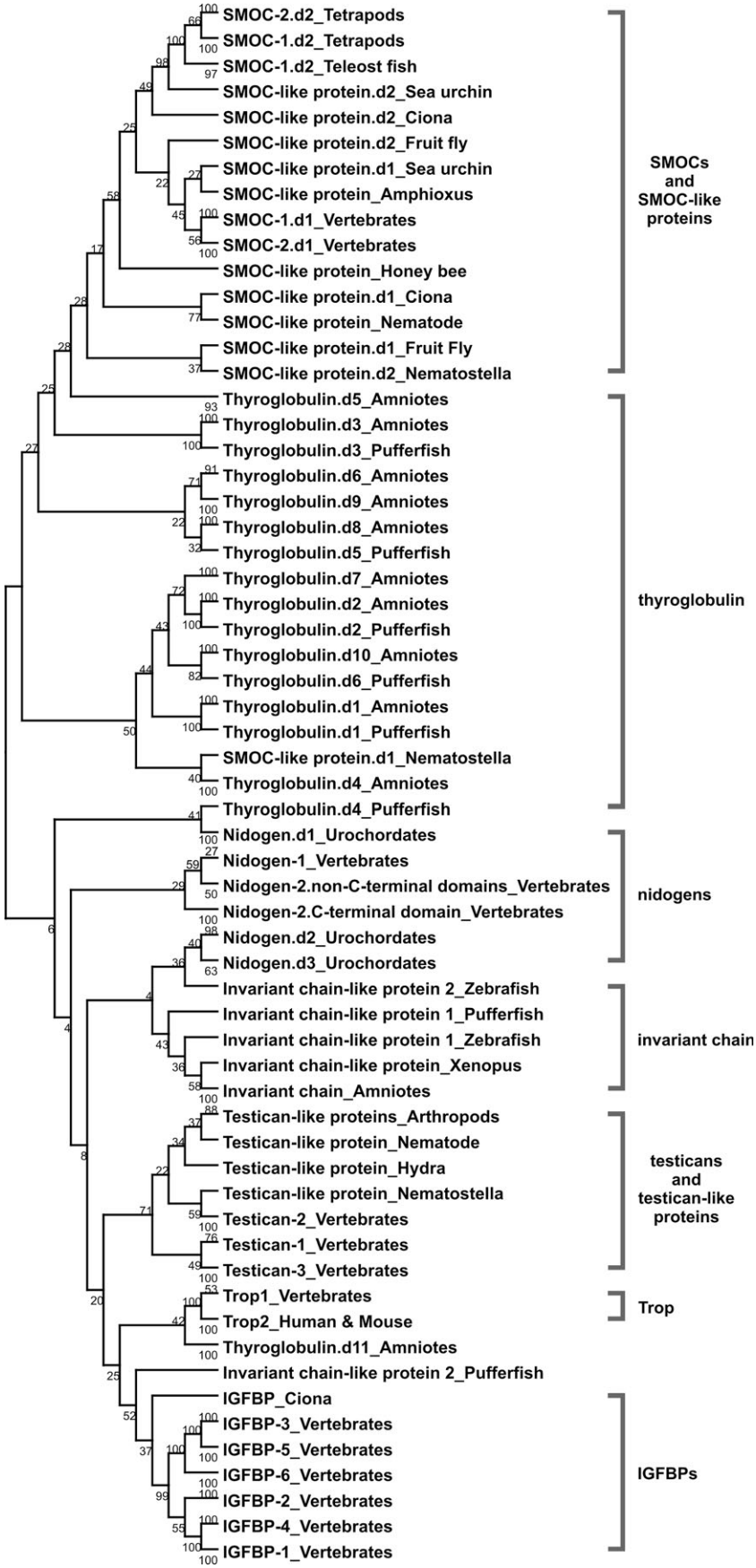
The domain architecture of Tg1 domain–containing proteins differs markedly between diverse metazoans. To investigate a possible correlation between domain architecture and Tg1 domain phylogeny, we conducted two separate phylogenetic analyses, one including all retrieved Tg1 domain sequences and the second including only domains of vertebrate proteins and their invertebrate homologs. The obtained NJ trees are presented as Supplement 2 of the Supplementary Material online, and a condensed NJ tree of the second analysis is shown in figure 2.

Tg1 domains of vertebrates and their invertebrate homologs are divided into seven well-defined groups which correspond to protein architectures identified in vertebrates (fig. 2 and Supplement 2A of the Supplementary Material online). Within groups further subdivision is observed. Groups containing multiple paralogs may be subdivided according to protein paralogy, as is seen for IGFBPs and testicans. The IGFBPs are divided into two subgroups, each comprising three paralogs, while the *C. intestinalis* IGFBP has basal position as the oldest representative. Testican Tg1 domains form two groups, one containing domains of invertebrate testican-like proteins and vertebrate testican-2 and the other vertebrate testican-1 and -3 domains. In contrast to the former, proteins with more than one Tg1 domain exhibit more complex division patterns. In the SMOC group, Tg1 domains of vertebrate homologs are divided into two subgroups. One comprises domains near the N-terminus and the other domains near the C-terminus. Within each subgroup, further division corresponding to SMOC paralogy is observed. The Tg1 domains of invertebrate SMOC-like proteins are not clustered, rather they are dispersed throughout the group. Vertebrate nidogen Tg1 domains are subdivided into three branches. Nidogen-2 domains form two subgroups, one comprising domains nearest the C-terminus in individual proteins and the other comprising the rest, while the third is formed by domains of nidogen-1. Similarly, two subgroups are formed by the Tg1 domains of ascidian nidogens. In the case of thyroglobulin a diffuse pattern is observed. The majority form a loosely related group, whereas domain 11 does not cluster with the rest.

The described seven groups are only partly observed in the tree constructed from all retrieved sequences (Supplement 2B of the Supplementary Material online), as expected due to low bootstrap support for the branches. This is most clearly observed for Tg1 domains of thyroglobulin, which are separated into four groups, and for those of ascidian nidogens, which do not cluster with their vertebrate paralogs. Further differences are observed for individual domains and for branching patterns within groups, as well as for relatedness between groups. Nevertheless, the existence of these groups shows that, overall, Tg1 domain conservation is tightly linked to the conservation of protein architecture, and this effect is observed over large evolutionary distances. In contrast, introduction of Tg1 domains into novel protein architectures appears to be followed by rapid diversification because clear relationships between Tg1 domains from different architectures cannot be determined.

←

FIG. 1.—Domain architectures of selected Tg1 domain–containing proteins. Proteins are represented as horizontal lines, and rectangles indicate protein modules. Tg1 domains are colored black, EC domains gray, follistatin-like domains are striped, and the remainder are white. Proteins are identified with protein name or abbreviation. Larger images are not to scale. Asterisks (*) indicate missing protein termini. Domain designations: AC, testican acidic region; AS, antistasin-like domain; CL, invariant chain class II-associated invariant chain peptide (CLIP) fragment; EGF, epidermal growth factor–like domain; FN, FOLN domain; FRI, FRI domain; FS, follistatin-like domain; G1, nidogen G1 domain; G2, nidogen G2 domain; IC, invariant chain intracellular domain; IG, immunoglobulin-like domain; KU, BPTI/Kunitz domain; LamG, laminin G domain; LC, lipocalin domain; LY, low-density lipoprotein receptor–like YWTD repeat; OLF, olfactomedin-like domain; OP, osteopontin domain; SD, syndecan domain; SEA, SEA domain; SMOC, SMOC-unique domain; Tg2, thyroglobulin type-2 repeat; Tg3, thyroglobulin type-3 repeat; TIL, trypsin inhibitor–like cysteine-rich domain; TM, transmembrane region; TSP1, thrombospondin type-1 repeat; TST, testican-unique domain; vWA, von Willebrand factor type-A domain; vWC, von Willebrand factor type C domain; and WAP, whey acidic protein–type four-disulfide core domain.



In addition to vertebrate groups, several smaller groups appear in the tree constructed from all available Tg1 domains (Supplement 2 of the Supplementary Material online). These mostly comprise domains from closely related species or domains from a single species, which probably evolved by domain duplication or recombination. Interestingly, close relationships are also observed for *N. vectensis* and *B. floridae* Tg1 domains, indicating that several proteins of the eumetazoan last common ancestor (LCA) persisted in the deuterostome lineage at least until the divergence of the cephalochordates were however modified or lost in protostomes. This observation is in agreement with other studies that show a strong resemblance between cnidarians and higher deuterostomes (Kortschak et al. 2003; Raible and Arendt 2004).

The generally low bootstrap support for the obtained relationships prevents the classification of all Tg1 domains into distinct families. Nevertheless, both analyses on Tg1 domain phylogeny do provide substantial insight into certain aspects of Tg1 domain evolution. To reinforce these data, we performed additional phylogenetic analyses of vertebrate Tg1 domain–containing proteins. These analyses are available as Supplement 2C of the Supplementary Material online and were based on modules, which are abundant in individual protein groups. For SMOCs and testicans the analyses were performed on the respective supradomains, for nidogens on their G1 domains, and IGFBPs were examined as whole proteins. Together with Tg1 domain phylogeny, we used the data obtained to deduce the evolution of these proteins.

The Birth of the Tg1 Domain

Tg1 domains are Metazoa-specific protein modules, which evolved very early in the evolution of this kingdom. They are predominantly found in extracellular proteins, and their occurrence is probably related to the appearance of multicellular animals, which was accompanied by rapid evolution of modules and proteins involved in cell-cell interactions as well as in interactions of cells with their environment. A comparison between yeast, fruit fly, and nematode proteins shows a 10-fold increase in the number of proteins involved in these interactions in Metazoa (Hazkani-Covo et al. 2004). Based on the broad phyletic distribution of the thyropins (Lenarčič and Bevec 1998; Lenarčič et al. 2000), we can assume that the primordial Tg1 domain originated as a regulator of peptidase activity and that most, if not all, Tg1 domains have retained this function throughout metazoan evolution.

The Evolution of Testicans and SMOCs

The eumetazoan LCA appears to have contained a well-developed set of Tg1 domain–containing proteins, including both testican-like and SMOC-like proteins, as is demonstrated by their presence in Cnidaria (fig. 1 and Supplement 1 of the Supplementary Material online). The Tg1 domains of both are well conserved in most metazoan groups (fig. 2 and Supplement 2 of the Supplementary Material online), indicating that they were among the crucial innovations of the first Eumetazoa. Their existence in basal metazoans further suggests that two distinct Tg1 domain lineages originated early in the evolution of the Tg1 domain superfamily and that all members might derive from one of them.

Unfortunately, the physiological roles of invertebrate testican-like and SMOC-like proteins are unknown. Testicans are involved in the regulation of cell attachment (Marr and Edgell 2003; Schnepp et al. 2005), as well as regulation of cysteine cathepsin (Bocock et al. 2003; Meh et al. 2005) and metalloprotease activity (Nakada et al. 2001, 2003). SMOCs, on the other hand, are little known glycoproteins. SMOC-1 was found to be localized predominantly to basement membranes (Vannahme et al. 2002), whereas the precise localization of SMOC-2, which exhibits a somewhat broader expression pattern (Vannahme et al. 2003), remains to be determined.

Both SMOC-like and testicans-like proteins are widely distributed throughout the animal kingdom and have a rich evolutionary history of lineage-specific modifications, which are summarized in figure 3. To evaluate the relationships within each group, we performed phylogenetic analyses on the respective supradomains (Supplement 2C of the Supplementary Material online). In both, supradomain phylogeny correlates well with species phylogeny, although a few exceptions are seen for SMOC-type supradomains.

The domain architecture of the *N. vectensis* SMOC-like protein is remarkably similar to those identified in higher metazoans (fig. 3). On the other hand, cnidarian testican-like proteins are comprised solely of a testican-type supradomain. A follistatin-like domain was introduced into this architecture before the deuterostome-protostome split, as demonstrated by its presence in both vertebrates and fruit fly. Interestingly, both nematode proteins are comprised solely of the respective supradomains, probably as a result of lineage-specific domain losses.

In deuterostomes, SMOC-like proteins are ubiquitous, whereas testican-like proteins appear only in vertebrates. It is not exactly clear whether the absence of testican-like proteins is accounted for by independent secondary gene losses in these lineages or is simply due to incompleteness of data.

Comparison of vertebrate SMOCs and testicans with invertebrate homologs shows the presence of additional vertebrate-specific domains, the testican-unique domain in testicans, and the SMOC-unique domain in SMOCs. These domain architectures were conserved during vertebrate evolution, and further functional diversification was achieved by gene duplication. Two SMOC paralogs are already present in pufferfish, indicating that the gene duplication producing these paralogs occurred before the teleost

←

FIG. 2.—Condensed representation of the NJ tree showing Tg1 domains from vertebrate proteins and their invertebrate orthologs. Seven major Tg1 domain groups are evident. The tree was constructed using the PHYLIP package with the probability matrix from Blocks (PMB) distance matrix, and bootstrap values were calculated using 100 replicates. The tree was visualized with TreeExplorer.

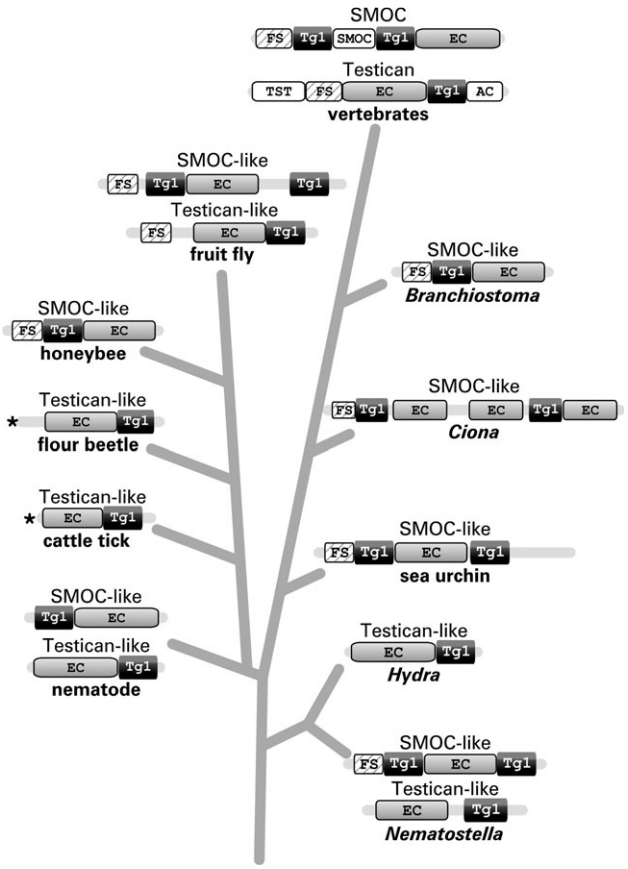


FIG. 3.—Summary of SMOC-like and testican-like protein architectures in different metazoan lineages. Branch lengths do not represent evolutionary distances between species. Proteins are represented as horizontal lines, and domains are shown as rectangles. Tg1 domains are colored black, EC domains gray, follistatin-like (FS) domains are striped, and the SMOC-unique (SMOC) domain, the testican-unique (TST) domain, and the testican acidic region (AC) are white.

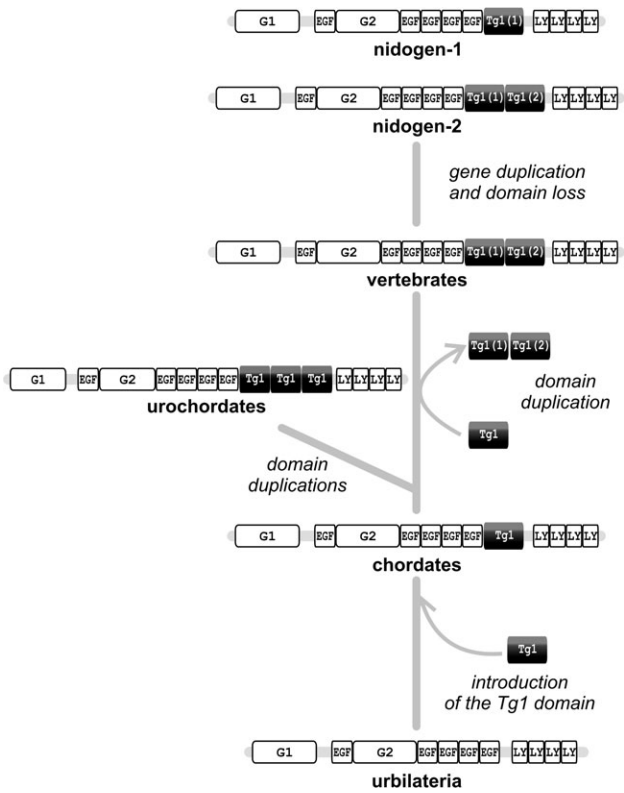


FIG. 4.—The proposed evolutionary pathway of nidogens in the deuterostome lineage. Proposed evolutionary events are in italic. Domain abbreviations: G1, nidogen G1 domain; G2, nidogen G2 domain; EGF, epidermal growth factor–like domain; and LY, low-density lipoprotein receptor–like YWTD repeat.

fish-tetrapod split 450 MYA. Similarly, three testican paralogs were produced before the divergence of these two lineages by two subsequent gene duplications. As indicated by both Tg1 domain (fig. 2) and testican-type supradomain phylogeny (Supplement 2C of the Supplementary Material online), the first duplication gave rise to testican-2 and the common ancestor of testican-1 and -3, and a following duplication of the second gave rise to the other two paralogs.

Origins and Evolution of Other Vertebrate Tg1 Domain–Containing Proteins

In contrast to the former, other vertebrate Tg1 domain–containing proteins evolved by different evolutionary scenarios. Nidogens, ubiquitous basement membrane components, have been thoroughly studied and have diverse functions (Erickson and Couchman 2000 for review). The domain architecture found in vertebrates evolved from a preexisting protein lacking Tg1 domains. Such proteins were found in Ecdysozoa (Lee and Cheung 1996), showing that an ancestral nidogen existed in Urbilateria. The introduction of the Tg1 domain occurred specifically in deuterostomes, and the proposed evolutionary pathway of nidogen in this lineage is shown in figure 4. The oldest nidogen containing Tg1 domains was identified in the ascidian *Halocynthia roretzi* (Nakae et al. 1993), and we have also identified a *C. intestinalis* ortholog. The three Tg1 domains of ascidian nidogen form a separate group not closely related to those of vertebrate nidogens (fig. 2 and Supplement 2A and B of the Supplementary Material online) and probably arose by lineage-specific domain duplications. This indicates that nidogen Tg1 domains were exposed to further evolutionary pressure after the divergence of urochordates from the chordates and argues that the insertion of a Tg1 domain occurred not long before the split of both lineages 550 MYA. In vertebrates, two distinct Tg1 domains of nidogen originated from domain duplication. From this ancestral protein two nidogens evolved by gene duplication, and the birth of nidogen-1 was accompanied by the loss of the Tg1 domain nearer the C-terminus (Tg1(2) in fig. 4), as suggested by Tg1 domain phylogeny (fig. 2).

Another group of vertebrate proteins, which can be traced back to urochordates, are the IGFBPs. These were already shown to be widely present in vertebrates (Upton et al. 1993) and were classified as a family within the IGFBP superfamily (Hwa, Oh, and Rosenfeld 1999*a*). The presence of this protein in *C. intestinalis* demonstrates that the origins of the IGF system, which is involved in the development of the central nervous system (Hwa, Oh, and Rosenfeld 1999*b* for review), by far predate the emergence of vertebrates.

The remaining three protein groups appear to be restricted to vertebrates because we could not find

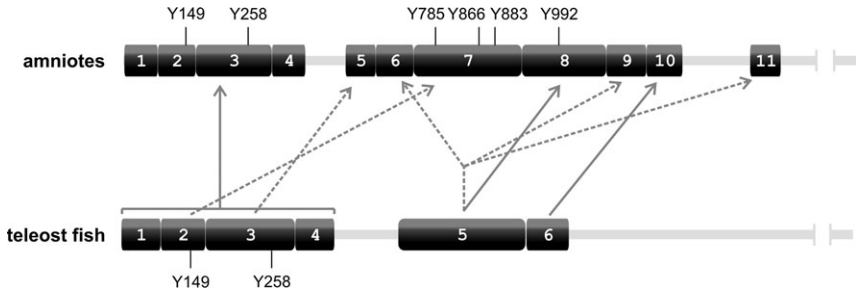


FIG. 5.—Proposed evolutionary pathway of thyroglobulin in vertebrates. Tg1 domains are shown as black rectangles. Solid arrows represent conserved domains, and dotted arrows represent closely related domains. The positions of conserved iodinated Tyr residues within Tg1 domains are marked.

invertebrate homologs. All are highly specialized proteins with a restricted expression pattern, which appear to have been formed during vertebrate evolution. The invariant chain is a critical component of the immune system (Stumptner-Cuvelette and Benaroch 2002 for review), Trops are involved in cell adhesion (Litvinov et al. 1997) and epithelia development (Guillemot et al. 2001), while thyroglobulin is involved in hormone signaling.

The evolution of thyroglobulin is perhaps the most intriguing of all vertebrate Tg1 domain–containing proteins. The Tg1 domains occupy the N-terminal portion of the chain, and domain architectures of orthologs from diverse taxonomic groups (Supplement 1 of the Supplementary Material online) show that this part of the molecule experienced large rearrangements during vertebrate evolution. The phylogeny of amniote and pufferfish Tg1 domains (Supplement 2C of the Supplementary Material online) mostly agrees with analyses made on larger samples (fig. 2 and Supplement 2A and B of the Supplementary Material online) and identifies the relationships between individual domains. As shown in figure 5, the N-terminal cluster is conserved between pufferfish and amniotes, whereas additional domains were introduced into the second cluster after the teleost fish-tetrapod split 450 MYA by internal domain duplications.

A unique feature of thyroglobulin is the presence of multiple iodinated Tyr residues, the precursors of thyroid hormones. Lamas et al. (1989) identified numerous iodination sites in human thyroglobulin, and six of these are located within Tg1 domains. Two of them, Tyr149 and Tyr258 (numbering applies to human thyroglobulin), are located in domains 2 and 3, respectively, and are conserved from teleost fishes to mammals. Three are located in domain 7 and are conserved in amniotes, are however missing in teleost fish, which lack this domain. The sixth is located in domain 8 and is also conserved among amniotes. However, it is not found in domain 5 of pufferfish thyroglobulin, which is most closely related to amniote domain 8.

Lineage-specific Variability in the Repertoires of Tg1 Domain–Containing Proteins in Vertebrates

Although the seven Tg1 domain–containing protein groups are ubiquitous in vertebrates, there is substantial variability in the repertoires between different vertebrate lineages. The observed differences are largely due to complete or partial gene duplications and domain duplications within proteins, while novel protein architectures are rarely observed and are derived from one of the seven major architectures. Rare examples of such proteins are egg chum salmon inhibitor (ECI) and saxiphilin, two well-characterized thyropins. The first is an isoform of the cysteine peptidase inhibitor from chum salmon oocytes (Yamashita and Konagaya 1996) and is composed of a single Tg1 domain, while the second is a transferrin homolog (Li and Moczydlowski 1991) with high binding affinity for saxitoxin and its derivatives (Mahar et al. 1991). Interestingly, Tg1 domain phylogeny suggests that Tg1 domains of both derive from non–C-terminal domains of nidogen-2 (fig. 2 and Supplement 2A of the Supplementary Material online).

Several lineage-specific products resulting from gene duplications were identified in teleost fishes (Supplement 1 of the Supplementary Material online). To distinguish between these paralogs, we termed one of them the same as its orthologs in higher vertebrates, whereas the other was denoted with a prime suffix. In pufferfish, two orthologs of testican-2 of higher vertebrates were identified and were therefore termed testican-2 and testican-2′, respectively. Similarly, three distinct nidogens were identified in both examined species. The pufferfish orthologs of nidogen-1 were termed nidogen-1 and nidogen-1′, whereas the zebrafish orthologs of nidogen-2 were termed nidogen-2 and nidogen-2′. Additionally, lineage-specific gene duplications of some IGFBPs were identified and appropriately termed in pufferfish and also in xenopus. Furthermore, teleost fish contain two proteins which resemble the invariant chain of higher vertebrates and probably result from gene duplication of an ancestral invariant chain–like protein. Because only one invariant chain protein is known in higher vertebrates, these proteins were termed invariant chain–like protein 1 and 2, respectively.

In addition to lineage-specific protein products, diversity in the number of Tg1 domains in nidogens is observed over the entire vertebrate subphylum. Nidogen-1 orthologs always contain a single Tg1 domain, whereas the number of Tg1 domains in nidogen-2 varies from two to six (Supplement 1 of the Supplementary Material online) as a result of internal domain duplications. Interestingly, these always involve only the Tg1 domain nearest the N-terminus (Tg1(1) in fig. 4, see also fig. 2 and Supplement 2A and B of the Supplementary Material online).

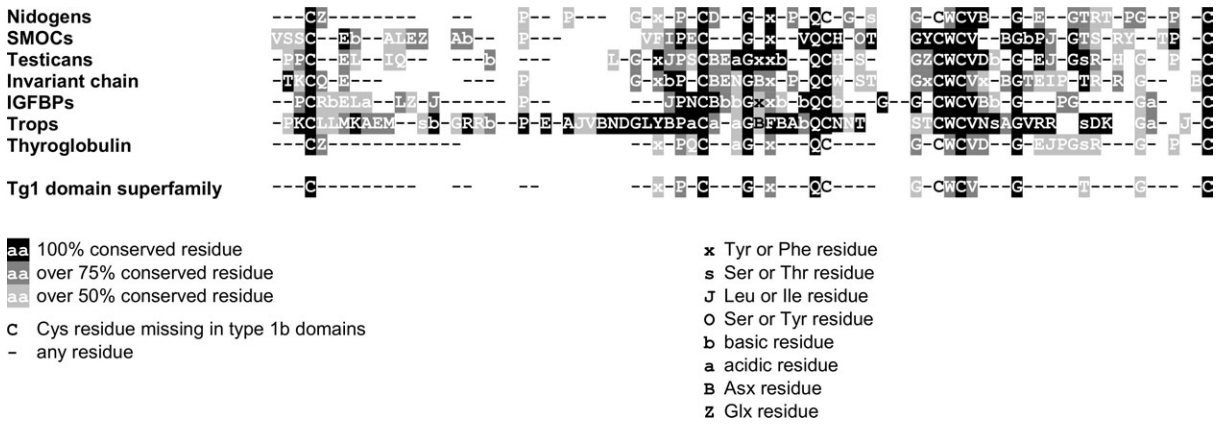


FIG. 6.—Residue conservation within vertebrate Tg1 domain groups and within the whole Tg1 domain superfamily. Positions correspond to average distances between conserved residues in the sequence alignment (see Supplement 3 of the Supplementary Material online). Alignments were created using ClustalX and manually edited with SEAVIEW. Residue conservation was determined manually.

Sequence Conservation Within the Tg1 Domain Superfamily

On average, Tg1 domains comprise 65 residues but vary from 50 to over 150 residues. From the alignment of all Tg1 domain sequences included in our work (Supplement 3 of the Supplementary Material online), we determined consensus sequences for the seven groups of vertebrate domains and for the whole superfamily, which are shown in figure 6. As expected, substantial residue conservation is observed between domains derived from protein homologs. On the superfamily level, Tg1 domains exhibit high primary structure variability, which is reflected in a very low degree of residue conservation. Besides the characteristic pattern of disulphide-bonded Cys residues, the only absolutely conserved residue is Gly49, whereas residues Gly28 and Gln34 are missing only in a few cases. The characteristic CWCV motif, defined by Molina et al. (1996), is only conserved in about 80% of all sequences. From the residues comprising this motif, Val45 is over 90% conserved, while Trp43 is conserved only in about 80% of all sequences and is usually replaced by a Phe or a Tyr residue. Somewhat less conserved residues include a residue with an aromatic side chain, usually a Tyr or Phe, at position 30, and Pro22, which occur in over 75% of sequences. Also, a few residues were found to occur at specific positions in over 50% of all Tg1 domains.

Additionally, we searched the sequences for putative N-glycosylation signals of the Asn-Xaa-Ser/Thr type. The search revealed the presence of such signals in over 25% of all sequences, indicating that some Tg1 domains are glycosylated.

Structural Diversity Within the Tg1 Domain Superfamily

The Tg1 domain fold was first identified in the crystal structure of the p41 fragment-cathepsin L complex (Gunčar et al. 1999). The mode of interactions revealed by this structure is assumed to reflect the general mode of thyropin action, and homology modeling based on this template has already been successfully used to elucidate interactions between other thyropins and their targets (Meh et al. 2005). Considering this, adaptation to structures of target peptidases appears to be a major driving force in Tg1 domain evolution.

Comparison of the p41 fragment with the recently determined structure of the C-terminal portion of IGFBP-6 revealed the major regions of structural deviation between the Tg1 domains, as has been thoroughly discussed (Headey et al. 2004). A superposition of both structures is shown in figure 7*A*, and the data obtained from it can be projected onto the whole Tg1 domain superfamily. The major region of variability is the N-terminal α-helix followed by the first loop. In the IGFBP-6 domain, the α-helix is longer than in the p41 fragment. This might also be the case in numerous other Tg1 domains, which contain insertions near their N-termini, based on the p41 fragment (see Supplement 3 of the Supplementary Material online). The prolongation of these α-helices has also been predicted using various secondary structure prediction algorithms (data not shown). The large insertion in the second loop is restricted to the IGFBP family of Tg1 domains and probably evolved by adaptive evolution.

The greatest difference between the two structures is observed in the C-terminal loop, the conformation of which is restricted by the disulfide bond. In the p41 fragment, the *cis*-disulfide bond keeps the loop oriented toward the bottom of the molecule, whereas in IGFBP-6 the *trans* configuration orients the loop toward the top (fig. 7*A*). In principle, any Tg1 domain may exist in either of these conformations, which makes tertiary structure predictions more difficult. Both solution structures show this loop to be flexible, which is more pronounced in IGFBP-6 (Headey et al. 2004) than in the p41 fragment (Chiva et al. 2003). Upon complex formation, the loop becomes fixed but interacts only weakly with the peptidase (Gunčar et al. 1999). This indicates that it is not absolutely necessary for peptidase inhibition.

The low number of conserved residues indicates the Tg1 domain fold to be highly adaptive with few structural constraints. As shown in figure 7*B*, the highly conserved Gly28 and Gly49 residues are located in loops on the top of the domain. These two residues are probably essential for the conformation of these loops and hence for the conservation of the Tg1 domain fold. Gln34, the third

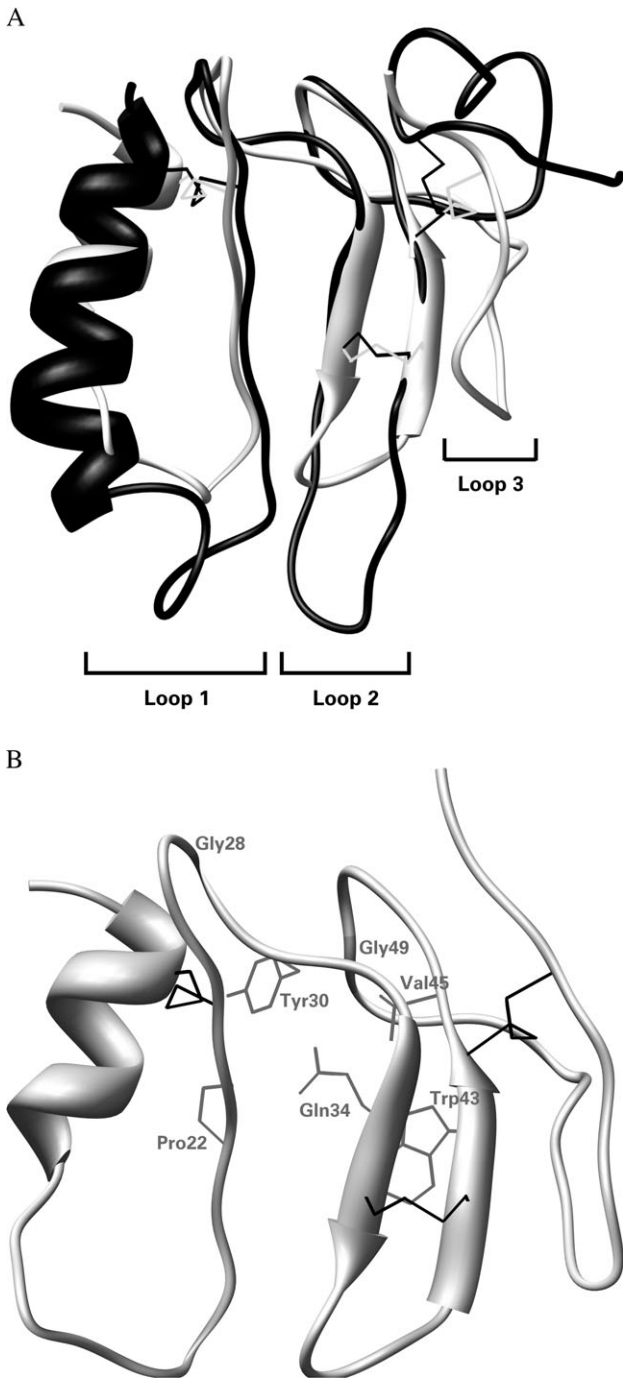


FIG. 7.—(*A*) Superposition of three-dimensional structures of IGFBP-6 Tg1 domain (Headey et al. 2004) and p41 fragment (Chiva et al. 2003). IGFBP-6 domain is colored black and p41 fragment white. Disulfide bonds are shown as sticks. Structures were superposed with DeepView Swiss-PdbViewer 3.7. (*B*) Positions of highly conserved residues (fig. 6) in the three-dimensional structure of p41 fragment (Gunčar et al. 1999). Disulfide bonds are colored black. Both images were created using UCSF Chimera.

highly conserved residue, as well as residues which are conserved in over 75% of all domains are located in the core of the domain. All together, this data shows that Tg1 domains are comprised of a relatively well-conserved core surrounded by variable loop regions.

Although the three disulfide bonds would be predicted as being essential for the stabilization of the Tg1 domain fold, Tg1B domains, which contain only two disulfide bonds, were identified in all three major metazoan lineages. They appear to evolve from Tg1A domains by replacement of one of the Cys residues forming the second disulfide bond. Once the initial mutation has occurred, the remaining Cys residue is apparently also quickly substituted. The only Tg1B domain found to contain one of the involved Cys residues is pufferfish thyroglobulin domain 5, which is most closely related to amniote thyroglobulin domain 8 (fig. 5), a Tg1A domain. This indicates that, in the LCA of teleost fishes and mammals, this domain was of Tg1A and that a lineage-specific subtype switching occurred in some teleost fishes.

Whether the initial subtype-switching mutation is random or driven by an evolutionary force can only be speculated. However, the fixation of Tg1B domains in the population indicates that the second disulfide bond is not essential for the stability of the Tg1 domain fold. Indeed, this bond is located within the antiparallel β-sheet (fig. 7*A* and *B*), which is by itself stabilized by hydrogen bonding. The disulfide bond establishes the orientation of the second loop, and its abolition probably increases the flexibility of this loop. This might lead to broadened specificity in regards to the domain's peptidase-binding properties by means of increasing the ability of the second loop to adapt to the active sites of structurally different peptidases. Although the experimental data to support this prediction is lacking, this indicates the evolution of Tg1B domains by gain-of-function mutations and explains their fixation in the population.

Functional Diversity Within the Tg1 Domain Superfamily

The conservation of Tg1 domains in homologous proteins indicates that these domains retain their functions within protein architectures. In contrast, their insertion into novel proteins is followed by rapid diversification. Therefore, domains in different protein architectures are likely to perform distinct biological functions.

Thus far, members of five vertebrate Tg1 domain groups have been characterized on the molecular level, and all were found to be involved in regulation of proteolysis. However, a high degree of specificity was observed in all cases, indicating that the physiological functions of Tg1 domains are precisely defined and regulated. Members of nidogen and invariant chain groups were shown to be highly selective inhibitors of individual cysteine cathepsins (Bevec et al. 1996; Yamashita and Konagaya 1996; Lenarčič et al. 2000). In contrast, domains of thyroglobulin were suggested to be substrates for these enzymes (Pungerčič et al. 2002). Recently, the testican-1 Tg1 domain was shown to combine both features (Meh et al. 2005). Although most frequent, the regulatory effects of Tg1 domains are not confined to cysteine peptidases, and some were shown to inhibit peptidases of the aspartate and metalloprotease classes (Fowlkes et al. 1997; Lenarčič and Turk 1999). The high specificities of all Tg1 domains studied thus far indicate that adaptation and coevolution with their targets are among the major driving forces in Tg1 domain evolution.

Additionally, some Tg1 domains evolved to perform secondary functions. The Tg1 domains in IGFBPs are involved in the binding of IGFs, and these interactions have been explained on molecular level for IGFBP-6 (Headey et al. 2004). In thyroglobulin, several highly conserved iodinated Tyr residues were identified within some Tg1 domains, indicating direct involvement of these domains in prohormone storage in the thyroid gland. Most iodination sites are located within large insertions, which are unique to these Tg1 domains (see Supplement 3 of the Supplementary Material online), indicating that these additional sequences were introduced specifically to expand the functional range of Tg1 domains.

Conclusion

Apart from the studies of thyropins and the IGFBPs, Tg1 domains have been largely overlooked, although they have been identified in a variety of proteins with different domain architectures, functions, and phyletic distributions. The studies on thyropins revealed Tg1 domains to be a unique group of proteolysis regulators with respect to their high specificity as well as their diversity. In this work, we have shown them to be a superfamily of versatile modules with a rich evolutionary history. The first Tg1 domain originated early in the evolution of Metazoa, presumably as a peptidase inhibitor, and has quickly spread and diversified. During metazoan evolution, Tg1 domains were incorporated into proteins with greatly different domain architectures, and two supradomain organizations, the testican-type and the SMOC-type supradomains, that comprise Tg1 domains and EC domains, occur in most major metazoan branches. Unfortunately, many identified Tg1 domains remain structurally and functionally uncharacterized, and future work will certainly reveal more interesting, as yet unknown properties of these domains.

Supplementary Material

Supplementary figures and tables are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Prof. R. H. Pain for critical reading of the manuscript and Gregor Gunčar for technical support. This work was supported by the Slovenian Research Agency by research program P1-0140. This work was done entirely at the Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia.

Literature Cited

Bevec, T., V. Stoka, G. Pungerčič, I. Dolenc, and V. Turk. 1996. Major histocompatibility complex class II-associated p41 invariant chain fragment is a strong inhibitor of lysosomal cathepsin L. J. Exp. Med. **183**:1331–1338.

Bocock, J. P., C. J. Edgell, H. S. Marr, and A. H. Erickson. 2003. Human proteoglycan testican-1 inhibits the lysosomal cysteine protease cathepsin L. Eur. J. Biochem. **270**:4008–4015.

Burge, C., and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268**:78–94.

Chiva, C., P. Barthe, A. Codina et al. (14 co-authors). 2003. Synthesis and NMR structure of p41icf, a potent inhibitor of human cathepsin L. J. Am. Chem. Soc. **125**:1508–1517.

Erickson, A. C., and J. R. Couchman. 2000. Still more complexity in mammalian basement membranes. J. Histochem. Cytochem. **48**:1291–1306.

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

Fowlkes, J. L., K. M. Thrailkill, C. George-Nascimento, C. K. Rosenberg, and D. M. Serra. 1997. Heparin-binding, highly basic regions within the thyroglobulin type-1 repeat of insulin-like growth factor (IGF)-binding proteins (IGFBPs) -3, -5, and -6 inhibit IGFBP-4 degradation. Endocrinology **138**:2280–2285.

Galeša, K., R. Pain, M. A. Jongsma, V. Turk, and B. Lenarčič. 2003. Structural characterization of thyroglobulin type-1 domains of equistatin. FEBS Lett. **539**:120–124.

Galtier, N., M. Gouy, and C. Gautier. 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. **12**:543–548.

Grasberger, B. L., G. M. Clore, and A. M. Gronenborn. 1994. High-resolution structure of Ascaris trypsin inhibitor in solution: direct evidence for a pH-induced conformational transition in the reactive site. Structure **2**:669–678.

Guillemot, J. C., M. Naspetti, F. Malergue, P. Montcourrier, F. Galland, and P. Naquet. 2001. Ep-CAM transfection in thymic epithelial cell lines triggers the formation of dynamic actin-rich protrusions involved in the organization of epithelial cell layers. Histochem. Cell Biol. **116**:371–378.

Gunčar, G., G. Pungerčič, I. Klemenčič, V. Turk, and D. Turk. 1999. Crystal structure of MHC class II-associated p41 Ii fragment bound to cathepsin L reveals the structural basis for differentiation between cathepsins L and S. EMBO J. **18**: 793–803.

Hazkani-Covo, E., E. Y. Levanon, G. Rotman, D. Graur, and A. Novik. 2004. Evolution of multicellularity in Metazoa: comparative analysis of the subcellular localization of proteins in Saccharomyces, Drosophila and Caenorhabditis. Cell Biol. Int. **28**:171–178.

Headey, S. J., D. W. Keizer, S. Yao, G. Brasier, P. Kantharidis, L. A. Bach, and R. S. Norton. 2004. C-terminal domain of insulin-like growth factor (IGF) binding protein-6: structure and interaction with IGF-II. Mol. Endocrinol. **18**:2740–2750.

Hennighausen, L. G., and A. E. Sippel. 1982. Mouse whey acidic protein is a novel member of the family of 'four-disulfide core' proteins. Nucleic Acids Res. **10**:2677–2684.

Hwa, V., Y. Oh, and R. G. Rosenfeld. 1999*a*. Insulin-like growth factor binding proteins: a proposed superfamily. Acta Paediatr. Suppl. **88**:37–45.

———. 1999*b*. The insulin-like growth factor-binding protein (IGFBP) superfamily. Endocr. Rev. **20**:761–787.

Kortschak, R. D., G. Samuel, R. Saint, and D. J. Miller. 2003. EST analysis of the cnidarian Acropora millepora reveals extensive gene loss and rapid sequence divergence in the model invertebrates. Curr. Biol. **13**:2190–2195.

Lamas, L., P. C. Anderson, J. W. Fox, and J. T. Dunn. 1989. Consensus sequences for early iodination and hormonogenesis in human thyroglobulin. J. Biol. Chem. **264**:13541–13545.

Laskowski, M. Jr. 1986. Protein inhibitors of serine proteinases—-mechanism and classification. Adv. Exp. Med. Biol. **199**:1–17.

Lee, M., and H. T. Cheung. 1996. Isolation and characterization of Caenorhabditis elegans extracellular matrix. Biochem. Biophys. Res. Commun. **221**:503–509.

Lenarčič, B., and T. Bevec. 1998. Thyropins—new structurally related proteinase inhibitors. Biol. Chem. **379**:105–111.

Lenarčič, B., G. Krishnan, R. Borukhovich, B. Ruck, V. Turk, and E. Moczydlowski. 2000. Saxiphilin, a saxitoxin-binding protein

with two thyroglobulin type 1 domains, is an inhibitor of papain-like cysteine proteinases. J. Biol. Chem. **275**:15572–15577.

Lenarčič, B., and V. Turk. 1999. Thyroglobulin type-1 domains in equistatin inhibit both papain-like cysteine proteinases and cathepsin D. J. Biol. Chem. **274**:563–566.

Li, Y., and E. Moczydlowski. 1991. Purification and partial sequencing of saxiphilin, a saxitoxin-binding protein from the bullfrog, reveals homology to transferrin. J. Biol. Chem. **266**:15481–15487.

Litvinov, S. V., M. Balzar, M. J. Winter, H. A. Bakker, I. H. Briaire-de Bruijn, F. Prins, G. J. Fleuren, and S. O. Warnaar. 1997. Epithelial cell adhesion molecule (Ep-CAM) modulates cell-cell interactions mediated by classic cadherins. J. Cell. Biol. **139**:1337–1348.

Mahar, J., G. L. Lukacs, Y. Li, S. Hall, and E. Moczydlowski. 1991. Pharmacological and biochemical properties of saxiphilin, a soluble saxitoxin-binding protein from the bullfrog (Rana catesbeiana). Toxicon **29**:53–71.

Marr, H. S., and C. J. Edgell. 2003. Testican-1 inhibits attachment of Neuro-2a cells. Matrix Biol. **22**:259–266.

Meh, P., M. Pavšič, V. Turk, A. Baici, and B. Lenarčič. 2005. Dual concentration-dependent activity of thyroglobulin type-1 domain of testican: specific inhibitor and substrate of cathepsin L. Biol. Chem. **386**:75–83.

Mercken, L., M. J. Simons, S. Swillens, M. Massaer, and G. Vassart. 1985. Primary structure of bovine thyroglobulin deduced from the sequence of its 8,431-base complementary DNA. Nature **316**:647–651.

Molina, F., M. Bouanani, B. Pau, and C. Granier. 1996. Characterization of the type-1 repeat from thyroglobulin, a cysteine-rich module found in proteins from different families. Eur. J. Biochem. **240**:125–133.

Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247**: 536–540.

Nakada, M., H. Miyamori, J. Yamashita, and H. Sato. 2003. Testican 2 abrogates inhibition of membrane-type matrix metalloproteinases by other testican family proteins. Cancer Res. **63**:3364–3369.

Nakada, M., A. Yamada, T. Takino, H. Miyamori, T. Takahashi, J. Yamashita, and H. Sato. 2001. Suppression of membrane-type 1 matrix metalloproteinase (MMP)-mediated MMP-2 activation and tumor invasion by testican 3 and its splicing variant gene product, N-Tes. Cancer Res. **61**:8896–8902.

Nakae, H., M. Sugano, Y. Ishimori, T. Endo, and T. Obinata. 1993. Ascidian entactin/nidogen. Implication of evolution by shuffling two kinds of cysteine-rich motifs. Eur. J. Biochem. **213**:11–19.

Parma, J., D. Christophe, V. Pohl, and G. Vassart. 1987. Structural organization of the 5′ region of the thyroglobulin gene. Evidence for intron loss and "exonization" during evolution. J. Mol. Biol. **196**:769–779.

Patthy, L. 2003. Modular assembly of genes and the evolution of new functions. Genetica **118**:217–231.

Pungerčič, G., I. Dolenc, M. Dolinar, T. Bevec, S. Jenko, S. Kolarič, and V. Turk. 2002. Individual recombinant thyroglobulin type-1 domains are substrates for lysosomal cysteine proteinases. Biol. Chem. **383**:1809–1812.

Raible, F., and D. Arendt. 2004. Metazoan evolution: some animals are more equal than others. Curr. Biol. **14**:R106–R108.

Rawlings, N. D., D. P. Tolle, and A. J. Barrett. 2004. Evolutionary families of peptidase inhibitors. Biochem. J. **378**:705–716.

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

Schnepp, A., P. Komp Lindgren, H. Hulsmann, S. Kroger, M. Paulsson, and U. Hartmann. 2005. Mouse testican-2. Expression, glycosylation, and effects on neurite outgrowth. J. Biol. Chem. **280**:11274–11280.

Stumptner-Cuvelette, P., and P. Benaroch. 2002. Multiple roles of the invariant chain in MHC class II function. Biochim. Biophys. Acta **1542**:1–13.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 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.

Upton, Z., S. J. Chan, D. F. Steiner, J. C. Wallace, and F. J. Ballard. 1993. Evolution of insulin-like growth factor binding proteins. Growth Regul. **3**:29–32.

Vannahme, C., S. Gosling, M. Paulsson, P. Maurer, and U. Hartmann. 2003. Characterization of SMOC-2, a modular extracellular calcium-binding protein. Biochem. J. **373**:805–814.

Vannahme, C., N. Smyth, N. Miosge, S. Gosling, C. Frie, M. Paulsson, P. Maurer, and U. Hartmann. 2002. Characterization of SMOC-1, a novel modular calcium-binding protein in basement membranes. J. Biol. Chem. **277**:37977–37986.

Veerassamy, S., A. Smith, and E. R. Tillier. 2003. A transition probability model for amino acid substitutions from blocks. J. Comput. Biol. **10**:997–1010.

Vogel, C., M. Bashton, N. D. Kerrison, C. Chothia, and S. A. Teichmann. 2004*a*. Structure, function and evolution of multidomain proteins. Curr. Opin. Struct. Biol. **14**:208–216.

Vogel, C., C. Berzuini, M. Bashton, J. Gough, and S. A. Teichmann. 2004*b*. Supra-domains: evolutionary units larger than single protein domains. J. Mol. Biol. **336**:809–823.

Yamashita, M., and S. Konagaya. 1996. A novel cysteine protease inhibitor of the egg of chum salmon, containing a cysteine-rich thyroglobulin-like motif. J. Biol. Chem. **271**: 1282–1284.

Claudia Kappen, Associate Editor

Accepted December 14, 2005