Evolution of thyroglobulin: an integrated view of its origin and complexity from a structural perspective.

Mauricio Gomes Pio[1,2], Wanderson Marques da Silva[1,3], Carina M. Rivolta[1,2] and Héctor M. Targovnik[1*]

[1] **Universidad de Buenos Aires. Facultad de Farmacia y Bioquímica. Departamento de Microbiología, Inmunología, Biotecnología y Genética/Cátedra de Genética. Buenos Aires, Argentina.**

[2] **CONICET-Universidad de Buenos Aires. Instituto de Inmunología, Genética y Metabolismo (INIGEM). Buenos Aires, Argentina.**

[3] **CONICET-Universidad de Buenos Aires. Instituto de Fisiología y Biofísica Bernardo Houssay (IFIBIO HOUSSAY). Buenos Aires, Argentina.**

Short title: Thyroglobulin evolution.

Key words: Evolution, Thyroglobulin, Nidogen 2, Sea lamprey, Bioinformatic tools

***Address correspondence and requests for reprints to:**

Dr. Héctor M. Targovnik

Universidad de Buenos Aires. Facultad de Farmacia y Bioquímica. Departamento de Microbiología, Inmunología, Biotecnología y Genética/Cátedra de Genética, Av. Córdoba 2351, Cuarto Piso, Sala 5, C1120AAR - Buenos Aires, Argentina.

Tel.: 54-11-5950-8805 E-mail:htargovn@ffyb.uba.ar

Abstract

This study provides an extensive bioinformatics evaluation of the origin and structural complexity of thyroglobulin (TG). We investigate the structural and evolutionary conservation of TG in *Petromyzon marinus* (sea lamprey) by reconstructing its complete TG sequence and comparing it to human TG. Using genomic data from NCBI, Uniprot, and Ensembl, we assembled a 2831-amino-acid sequence (TGPM), identifying key TG modules, including 11 TG type 1 modules, three TG type 2 modules, five TG type 3 modules, as well as the linker, hinge, and ChEL domains. Sequence alignment with human TG revealed strong conservation of hormonogenic tyrosines and cysteine residues, underscoring their structural significance in both species. We generated a PDB model for TGPM using the Swiss Model program for homology modeling. The full-length sea lamprey TG was subsequently aligned with full-length human TG through the Matchmaker command in UCSF ChimeraX, revealing a consistent structural overlap across all four regions (I, II, III, IV), highlighting their evolutionary conservation. Additionally, we identified a second sea lamprey TG transcript, $TGPM^{1746}$, which begins with the TG type 1-10 module and extends to the carboxyl terminus, exhibiting high homology with TGPM except in region III.

We also present an evolutionary model proposing that a nidogen-like precursor—characterized by the presence of TG type 1 modules—may have served as the ancestral form of thyroglobulin. Through genetic rearrangements and duplication events, a proto-TG likely emerged, influenced by environmental factors such as ionizing radiation. Structural modifications led to the loss of nidogen's globular regions, resulting in a secretory proto-TG containing several TG type 1 modules. Successive duplications expanded the TG structure, ultimately forming TG 11 type 1 modules, closely resembling the ancestral TG in *Petromyzon marinus*. The final evolutionary phase involved the incorporation of TG type 3 and TG type 2 modules, followed by the fusion of the ChEL domain, which enhanced TG secretion and

thyroid hormone biosynthesis. Our findings confirm that the TG complexation process is fully established in lampreys and has remained virtually unchanged throughout vertebrate evolution.

Introduction

Thyroid hormone (TH) biosynthesis played a crucial role in evolution, modulating gene expression across various tissues and regulating complex biological processes. These include the transition from larva to adult, cell differentiation, energy metabolism, seasonal physiological changes, and the development of the central nervous system. The significant role of TH biosynthesis in vertebrates contributed to their remarkable diversification and increasing complexity over time, enabling them to acquire unique traits. The emergence of an endocrine system capable of concentrating iodide from seawater was essential for TH production, providing metabolic support for the development of a more complex brain, the formation of a skeletal structure, and the regulation of body temperature.

During the evolution of chordates—including the subphyla Cephalochordata, Urochordata (or Tunicata), and Vertebrata (or Craniata)—thyroid hormone formation preceded the morphological differentiation of thyroid cells and their organization into follicles [Suzuki and Kondo, 1971; Suzuki et al., 1975]. The storage of thyroid hormones bound to thyroglobulin (TG) within the follicular lumen is a characteristic feature of vertebrates. Cephalochordates, urochordates, and vertebrates share a common ancestor that likely lived around 550 million years ago [Putnam et al., 2008]. Vertebrates are classified into gnathostomes and cyclostomes (also known as agnathans or jawless fish), which diverged early in vertebrate evolution before the emergence of hinged jaws, a defining trait of gnathostomes [Schimeld et al., 2012].

The only surviving lineages of cyclostomes today fall into two categories: lampreys (also known as Hyperoartia) and hagfishes (also known as Myxini) [Xu et al., 2016]. Cyclostomes are believed to have first appeared during the Ordovician, reached their peak in the Silurian and Devonian, and then gradually declined toward extinction in the late Devonian [Xu et al., 2016]. Recently, using fossil calibrations, Marlétaz et al. [2024] estimated that the divergence between lampreys and hagfishes occurred in the late Ordovician, approximately 449 million years ago,

while the split between cyclostomes and gnathostomes took place near the Cambrian-Ordovician boundary, around 493 million years ago.

The oldest known lamprey fossils date back to the late Devonian period, approximately 360 million years ago, and their morphology has remained virtually unchanged over hundreds of millions of years of evolution [Xu et al., 2016]. The most extensively studied modern lamprey species include *Petromyzon marinus* (sea lamprey), *Lethenteron japonicum*, *Lampetra fluviatilis*, and *Lampetra planeri*. Lampreys produce TG 18-19S before larval metamorphosis, meaning before the thyroid follicles observed in adults are fully organized [Monaco et al., 1978]. In contrast, hagfishes contain iodinated glycoproteins of a non-TG nature, distinguished by differences in molecular size, solubility, amino acid composition, and carbohydrate and iodoamino acid content and structure [Ohmiya et al., 1989]. However, the possibility that the amino acid sequence of the hagfish iodoprotein may correspond to or be related to a TG domain cannot be ruled out.

The monomeric human TG preprotein consists of a polypeptide comprising 2,768 amino acids (NCBI: NP_003226.4; Uniprot: P01266) [- Holzer et al., 2016; Malthiéry and Lissitzky, 1987; van de Graaf et al., 2001]. The classical model of TG's primary structure is organized into four distinct regions (I, II, III, and IV) [Citterio et al., 2021; Holzer et al., 2016; Malthiéry and Lissitzky, 1987; Mercken et al., 1985; Molina et al., 1996; Parma et al., 1987; van de Graaf et al., 2001]. Region I contains 10 of the 11 TG type 1 repeats, along with two modules known as the linker and hinge, plus an amino-terminal T_4-forming site. Region II includes three TG type 2 repeats and the 11th TG type 1 repeat, while Region III encompasses all five TG type 3 repeats. The fourth region consists of a domain homologous to esterases, known as the cholinesterase-like (ChEL) domain [Park and Arvan, 2004; Swillens et al., 1986], along with a carboxi-terminal T_3-forming site.

The TG type 1 domain first appeared at the base of metazoan evolution [Novinec et al., 2006; Takagi et al., 2022; Yurchenco and Wadsworth, 2004]. Several proteins containing TG type 1 domains emerged during the early stages of life development, including nidogen-1 (NID-1),

nidogen (NID-2), SMOC and SMOC-like, testican and testican-like, the invariant chain, IGFBP, and trop. Homology between esterases and TG was first identified by comparing the primary structures of *Torpedo californica* acetylcholinesterase and bovine TG [Schumacher et al., 1986], suggesting that both evolved from a common ancestral gene. Tandem duplication events in the amino-terminal and amino-central regions, followed by fusion with an acetylcholinesterase-like (ChEL) domain, enabled TG to acquire new folding, secretion, and TH biosynthesis capabilities [Holzer et al., 2016].

In this study, we present a comprehensive bioinformatics analysis of the potential origin and structural complexity of TG. The primary objectives were to examine the organization and architecture of full-length TG and its alternative transcripts in a primitive jawless vertebrate fish, the sea lamprey, to hypothesize the origin of TG from a protein present at the base of vertebrate evolution contanining TG type 1 domains—such as an ancestral nidogen-like protein; and to propose a hypothetical model outlining the evolutionary events that shaped the distinct domains of TG.

Results

Assembly of full-length thyroglobulin from sea lamprey and comparative domain analysis.

To obtain the full-length TG from *Petromyzon marinus* and analyze its structural organization, we retrieved three peptides from the NCBI, Uniprot, and Ensembl databases, each containing sequences characteristic of TG architecture**.** A 2475 amino-acids sequence (Supplementary Figure S1a), referred to here as $TGPM^{2475}$, is indexed in the Uniprot database (S4R814_PETMA) and extracted from the *Petromyzon marinus* genome assembly available on ENSEMBL (ENSPMAG 00000001187, ENSPMAT00000001350.1), located on scaffold GL476337: 707,134-808,438 forward strand [Smith et al., 2013]. Alignment with human TG (NP_003226.4) using EMBOSS Needle reveals that $TGPM^{2475}$ contains eleven TG type 1 modules, three TG type 2 modules and five TG type 3 modules, linker and hinge domains, and spacers 1, 2 and 3, along with only the first 256 amino acids of the ChEL domain. However, $TGPM^{2475}$ lacks the signal peptide, the amino-terminal hormonogenic acceptor tyrosine required for T_4 formation, the remaining amino acids of the ChEL domain, and the carboxy-terminal hormonogenic site necessary for T_3 formation. Our upstream analysis to the first amino acid in $TGPM^{2475}$, corresponding to the first cysteine of the TG type 1-1 module in scaffold GL476337, allowed us to identify the first 11 amino acids of exon 2 (YSDSKTSLASA). Among these, three amino acids correspond to the first three amino acids of the TG type 1-1 module (ASA) and the amino-terminal hormongenic acceptor site (Y) (Supplementary Figure S2a,b). The glutamic acid residue (gag codon) overlaps the splice site between exons 1 and 2. In upstream exon 1, following the 687-nt intron, the first 25 amino-terminal amino acids of sea lamprey TG were identified (Supplementary Figure S2a,b). The SignalP 6.0 program predicts that this sequence contains a signal peptide with a cleavage site between positions 24 and 25 (MRTSPLLPATTTLYLVLWIGTISA) with a probability of 0.981653 (Supplementary Figure S2c).

A second sequence of 1746 amino acids, corresponding to *Petromyzon marinus* TG was identified in the NCBI database and is referred to here as TGPM^{1746} (*Petromyzon marinus* isolate kPetMar1 chromosome 27, kPetMar1.pri, whole genome shotgun sequence, NCBI reference sequence: NC_046095, XP_032817730.1, Supplementary Figure S3a). Pairwise sequence alignment of TGPM^{2475} and TGPM^{1746} using the EMBOSS Needle program revealed that TGPM^{1746} begins at TG type 1-10 and, notably, contains the complete ChEL domain and the carboxy-terminal hormogenic site required for T_3 formation—which are partially and entirely absent in TGPM^{2475}, respectively.

The pairwise sequence alignment of TGPM^{1746} and the scaffold GL476337 using the Blat Search Genome, with assembly WUGSC 7.0/petMar2, revealed alignement with the initial 258 amino acids of the ChEL domain (positions 1172-1429 of TGPM^{1746}) and the last 115 carboxy-terminal amino acids (positions 1632-1746 of TGPM^{1746}), including the final 68 amino acids (positions 1632-1699 TGPM^{1746}) of the ChEL domain and the carboxy-terminal hormonogenic site required for T_3 formation (tyrosine^{1745}). However, no alignments was observed between residues 1430-1631 of TGPM^{1746} (202 amino acids) with scaffold GL476337, as this region is absent from the genome data (positions 811252-833817). This absence creates a gap in the ChEL domain of the *Petromyzon marinus* TG sequence previously published by Holzer et al. [2017]. A 26-amino-acids was identified immediately preceding the TG type 1-10 of TGPM^{1746}. This sequence begins with methionine and contains 3 alanines at positions 13, 14 and 15, as well as two cysteines at positions 8 and 19. (MPSRVRACVRAPAAATFLCSALPSPG). The SignalP 6.0 program predicted that TGPM^{1746} contains a signal peptide, with a probability of 0.4996 (Supplementary Figure S3b).

To construct the full-length *Petromyzon marinus* TG sequence, we integrated: 36 upstream amino acids of the TGPM^{2475} identified in the scaffold GL476337, 204 downstream amino acids of the TGPM^{2475} identified in NC_046095 (TGPM^{1746}), 202 of which correspond to the missing

region in scaffold GL476337 and the 115 carboxy-terminal amino acids identified in the scaffold GL476337. This assembly results in the full-length sea lamprey TG (TGPM), comprising 2831 amino acids. Figure 1 presents the alignment between TGPM and *Human sapiens* TG (TGHS) using EMBOSS Needle, highlighting the location of the different domains, cysteines and hormononegic tyrosines. Although the primary structures of sea lamprey TG and human TG diverge (identity: 35.1%, similarity: 50.5%), the hormonogenic tyrosines and over 90% of the cysteine residues are highly conserved at identical positions. (Figure 1).

In summary, according to our results the sea lamprey TG gene is a single-copy gene, approximately 121 kb long, divided into 53 exons. Its mRNA encodes 8493 nt, corresponding to coding sequences (Supplementary Figure S4). The sea lamprey monomeric TG preprotein consists of 24-amino-acid leader peptide followed by a 2807-amino-acid polypeptide.

Remarkably, the comparison of TG sequences between *Petromyzon marinus* (TGPM) and two other lamprey species, *Lampetra fluviatilis* (TGLF) and *Lampetra planeri* (TGLP), using Clustal Omega with its visualization tool Mview, reveals a high degree of sequence conservation among the three species. The sequence identity is 86.9% between TGPM and TGLF and 86.7% between TGPM and TGPLP (Supplementary Figure S5). This finding confirms that our strategy for obtaining the full-length TG of *Petromyzon marinus* was accurate. Additionally, it suggests that the divergence between these three lamprey lineages occurred relatively recently. A comparative analysis of TG organization in TGPM, TGLF and TGLP demonstrates that TG in each species contains all regions found in mammalian TG (Supplementary Figure S5). The four TG regions, along with their domains and modules, are similar in size to human TG, with strict conservation of cysteines residues. Furthermore, the amino-terminal and carboxyl-terminal hormonogenic sites are highly conserved across all three species (Figure 1, Supplementary Figure S5).

A comparative sequence analysis between positions 1189 and 2831 of TGPM and the full length of TGPM^{1746}, conducted using the EMBOSS Needle program, revealed five major differences at

the following positions: 1388 (hinge domain), 1673 to 1710 (TG type 3a1), 1831 to 1850 (TG type 3b1), 1881 (TG type 3b1) and 2233 to 2235 (Spacer 3) of the TGPM (see Supplementary Figure S6). Interestingly, in the latter region, $TGPM^{1746}$ exhibits a 39 amino acid sequence containing 8 cysteines, resembling the TG type 3b modules (SLGTSCAARCGLYSPGQPCQCNRECQRFGDCCPDAAVCL).

In addition to TG^{1746} a second sequence of 567 amino acids, related to *Petromyzon marinus* TG, was identified in the NCBI database and is referred to here as $TGPM^{567}$ (XP_032817835.1, Supplementary Figure S7a). Pairwise sequence alignment of TGPM and $TGPM^{567}$ using the EMBOSS Needle program revealed that $TGPM^{567}$, in addition to linker domain, contains the TG type 1-4, TG type 1-5, TG type 1-6 modules and the first 42 amino acids of TG type 7. 51 amino acids at the amino-terminal end and 22 amino acids at the carboxy-terminal end of $TGPM^{567}$ are not represented in the TGPM (Supplementary Figure S8). The amino-terminal 51-amino-acid sequence begins with methionine and includes three alanines, at positions 13, 14 and 15, as well as one cysteine at position 51 (MELLPLGEKDDDAAALTAPALAPSVQEAPVFRLLRRRFLALRLLLSGSFRC). However, the SignalP 6.0 program does not predict this sequence to contain a signal peptide (Supplementary Figure S7b).

To analyze the homology profile among the 11 TG type 1 modules from TGPM and TGHS we performed an alignment using the T-Coffee program. The analysis revealed a high total consistency value of 75 points, indicating strong sequence conservation (Supplementary Figure S9). Among the 11 of TG type 1 modules, the six cysteine residues are highly conserved across both species. Figure 2 presents the classical alignment of the 11 TG type 1 modules, highlighting the six conserved cysteines. Most TG type 1 repeats exhibit similar sizes, except for TG Type 1-3, TG Type 1-7, and TG Type 1-8, which contain long sequence insertions. Remarkably, these insertions are highly conserved in both species. In TGPM, TG Type 1-3 exhibits a 40-amino-acids

insertion in the middle of the module and an additional 46-amino-acids insertion at the end of the module (40 and 45 amino acids, respectively, in human TG). Similarly, TG Type 1-7 contains a 128-amino-acids insertion in the middle of the module (123 in human TG), while TG Type 1-8 features a 100-amino-acids insertion at the beginning of the module, which is identical in human TG (Figure 2). A pairwise comparison of the individual modules between TGHS and TGPM, conducted using the EMBOSS Needle program, revealed an average identity of 37.94% and an average similarity of 51.24% **(**Table 1, Supplementary Figure S10**).**

The alignment of the three TG type 2 modules from TGPM and TGHS, performed using the T-Coffee program, yielded a total consistency value of 94, indicating strong sequence conservation (Supplementary Figure 11). This value is higher than that obtained in the the TG type 1 module comparison, suggesting greater homology. Both cysteines residues are fully conserved between the TG type 2 modules of TGPM and TGHS (Figure 3). A pairwise comparison of the individual modules, conducted using the EMBOSS Needle program, revealed an average identity of 45.70% and an average similarity of 55.86% (Table 1, Supplementary Figure S12).

The alignment of the five TG type 3 modules from TGPM and TGHS performed using T-Coffee program yielded a total consistency value of 69, indicating moderate sequence conservation (Supplementary Figure S13). This value is lower than that obtained in the TG type 1 modules and TG type 2 modules comparisons, suggesting lower homology. All six cysteine residues are fully conserved in the TG type 3a modules (Figure 4), as are all eight cysteines in the TG type 3b modules of TGPM and TGHS, except for the eighth cysteine in TG type 3b1 and TG type 3b2 which the T-Coffee program failed to align (Supplementary Figure S13). A pairwise comparison of the individual modules, conducted using the EMBOSS Needle program**,** revealed an average identity of 30.26% and an average similarity of 48.58% **(**Table 1, Supplementary Figure S14). In all cases, the six cysteines of the TG type 3a modules and the eight cysteines of the TG type 3b modules are conserved, except for the eighth cysteine in the TG type 3b1 module, which EMBOSS

Needle also failed to align, consistent with the T-Coffee results**.** Additionally, an extra cysteine was identified at position 1682 in the TG type 3a1 module of TGPM (Supplementary Figure S14).

Finally, the comparative analysis of the ChEL domain between TGPM and TGHS conducted using the EMBOSS Needle program revealed an average identity of 41.70% and an average similarity of 57.40% between the two species (Table 1). Six of the seven cysteines located in the ChEL domain of TGHS are conserved in TGPM (Figure 1).

Comparative 3D structural analysis of sea lamprey thyroglobulins and human thyroglobulin using the Matchmaker tool.

To compare the 3D structure of each of sea lamprey TG isoforms with human TG, as well as among the sea lamprey TG isoforms between themselves, we generated PDB for $TGPM^{567}$, $TGPM^{1746}$ and $TGPM^{2475}$, since PDBs were not available. We used the Swiss Model program to perform homology modeling for the sequence of these three proteins ($TGPM^{567}$: NCBI, XP_032817835.1, Supplementary Figure S7c; NCBI, TG^{1746}: XP_032817730.1, Supplementary Figure S3c; TG^{2475}: Uniprot, S4R814_PETMA, Supplementary Figure S1b), using on PDB 7N4Y (resolution: 2.61Å) of *Bos Taurus* TG as a reference [https://www.rcsb.org/structure/7N4Y; Kim et al., 2021].

The structural comparison between sea lamprey $TGPM^{576}$ or $TGPM^{1746}$ or $TGPM^{2475}$ monomer (chain B) and human TG monomer (PDB: 6scj, chain A) was conducted using the Matchmaker command in the UCSF ChimeraX program. The sea lamprey $TGPM^{567}$ was aligned with TGHS with an RMSD of 0.200 Å across 396 overlapping atom pairs, with a sequence alignment score: 1067 (Figure 5a1, a2). $TGPM^{567}$ corresponds to region I of TGHS, overlapping amino-acids 266-771. The assembly of sea lamprey $TGPM^{1746}$ and TGHS resulted in an RMSD of 1.076 Å, based on 806 overlapping atom pairs, with a sequence alignment score of 2879.4 (Figura 5b1, b2). Similarly, the assembly of $TGPM^{2475}$ and TGHS yielded an RMSD of 1.219 Å, across 1271

overlapping atom pairs, with a sequence alignment score of 4688.8 (Figure 5c1, c2). A consistent overlap was observed across the four regions (I, II, III, IV) of sea lamprey $TGPM^{1746}$ and $TGPM^{2475}$, aligning with the four regions of TGHS. This confirms the homology between the proteins, extending from the TG type 1-10 module to the carboxy-terminal end in $TGPM^{1746}$ and from the TG type 1-1 module to the first 256 amino acids of the ChEL domain in $TGPM^{2475}$.

Next, we compared the three-dimensional structure of $TGPM^{567}$ (chain A) and $TGPM^{1746}$(chain A), $TGPM^{567}$ (chain A) and $TGPM^{2475}$ (chain B), and $TGPM^{1746}$ (chain A) and $TGPM^{2475}$ (chain A). The Matchmaker tool confirmed the homogeneous overlap of the complexes, with RMSD values of 0.868, 1.176 and 0.295 Å for $TGPM^{567}$ and $TGPM^{1746}$ (42 overlapping atom pairs, sequence alignment score: 163.6, Figure 6a1, a2), for $TGPM^{567}$ and $TGPM^{2475}$ 351 overlapping atom pairs, sequence alignment score: 2407.7, Figure 6b1, b2) and for $TGPM^{1746}$ and $TGPM^{2475}$ (1246 overlapping atom pairs, sequence alignment score: 6249, Figure 6c1, c2) parings, respectively.

Finally, we generated a PDB for TGPM using the Swiss Model program for homology modeling (Figure 7, Supplementary Video 1), using on PDB 7N4Y of *Bos Taurus* TG as a reference. The full-length sea lamprey TG (TGPM, chain B) was then aligned with TGHS (chain A), yielding an RMSD of 1.191 Å across 1449 overlapping atom pairs, with a sequence alignment score: 5212.9) (Figure 7, Supplementary Video 2, Supplementary Video 3). As observed previously with $TGPM^{1746}$ and $TGPM^{2475}$, the alignment between TGPM and TGHS exhibits a homogeneous overlap across the four regions (I, II, III, IV) of sea lamprey TG, corresponding to the four analogous regions in human TG.

Comparative analysis of TG type 1 modules between nidogen-2 and thyroglobulin.

As observed earlier, many proteins incorporate the TG type 1 module in their structure. Among them, nidogen is particularly noteworthy, as it is present at the base of vertebrates evolution. Based

on our comprehensive analysis showing that cyclostomes—such as the sea lamprey—already possess a modular architecture closely resembling that of human TG, we propose that an ancestral nidogen-like protein may have acted as a precursor to TG during its complexation process.

NID-1 and NID-2 are homologous sulfated monomeric glycoproteins found in the basal lamina, already observed in cyclostomes and amphioxus, two extensively studied protovertebrates [Holland and Short, 2010; Novinec et al., 2006; Takagi et al., 2022; Yurchenco and Wadsworth, 2004]. In mammals, both proteins exhibit a modular structure consisting of three globular regions (G1, G2 and G3), separated by a a short flexible linker region between G1 and G2 and a rod-like domain between G2 and G3 [Bechtel et al 2012, Töpfer and Holz, 2024] (Figure 8a). The G1 globular region contains the NIDO domain (nidogen-like domain), while the G2 globular region consists of the nidogen G2 beta-barrel domain and one EGF-like domain (epidermal growth factor-like domain). The rod-like domain is composed of four EGF-like domains and one TG type 1 domain in NID-1 (Uniprot: P14543, human) or two TG type 1 domain in NID-2 (Uniprot: Q14112, human) of mammals (Figure 8a). However, sea lamprey NID-2, which is related to a primitive nidogen, contains only a single TG type 1 module (Uniprot: A0AAJ7X4A4). Notably, invertebrates such as Drosophila and C. elegans do not possess TG type 1 domains. In NID-1, the G3 globular region consists of four LDL-receptor class B repeats and an additional EGF-like domain, whereas in NID-2, it contains five LY modules [Bechtel et al 2012, Töpfer and Holz, 2024]. Figure 8a presents the three-dimensional atomic structure of sea lamprey NID-2 (which contains only one TG type 1 module) and human NID-2 (which possesses two TG type 1 modules). These structures were generated using with AlphaFold server, as their crystallography remains unresolved.

To analyze the homology profile among the of TG type 1 modules of *Petromyzon marinus*, *Mus musculus* and *Homo sapiens*, including NID-1, NID-2, TGPM and TGHS, we performed an alignment using the T-Coffee program and 3D structure modeling using AlphaFold server. The T-

Coffee analysis yielded a high total consistency value of 89 points, indicating strong sequence conservation (Supplementary Figure S15). The six cysteine residues were well conserved across all type 1 TG repeats analyzed. Additionally, AlphaFold modeling of TG type 1-1 and TG type 1-2 modules from nidogens and thyroglobulins in sea lamprey, mouse, and human revealed similar secondary structures, characterized by one α-helix and two β-sheets, with a very high plDDT score (Figure 8b). Figure 9 illustrates the alignment of all TG type 1 modules around the six conserved cysteines. Notably, the structural conservation of these modules is high in both nidogens and TGs. The primary structure alignment, performed using EMBOSS Needle, reveals a sequence divergence between sea lamprey nidogen and human nidogen, with an identity of 37.0% and a similarity of 48.1%—values comparable to those observed in the TG comparison between both species (see above). The SignalP 6.0 program predicts that NID-2 sequence in both sea lamprey and human contains a signal peptide, with a cleavage site between positions 36 and 37 (MGMSRAVARARSPAGYAALPAALLCLAAALLGCARG) and a probability of 0.951473. In human NID-2, the cleavage site is predicted between positions 30 and 31 (MEGDRVAGRPVLSSLPVLLLLPLLMLRAAA) with a probability of 0.943032 (data not shown).

Structural pairing analysis of TG type 1 modules from human nidogen-2 and human thyroglobulin using the Matchmaker tool.

To support the hypothesis that a nidogen-like protein may have served as the ancestral precursor to the earliest form of TG, we compared the 3D structures of TG type 1-1 and TG type 1-2 from human NID-2 and human TG. Since PDB for TG type 1-1 and TG type 1-2 from human NID-2 were not available, we generated them using the Swiss Model program for the homology modeling, based on the human sequence indexada en Uniprot (Q14112) and the 1NPE *mus musculus* PDB model, which is the only available PDB for nidogen.

The structural comparison between TG type 1-1 and TG type 1-2 from human NID-2 monomer (chain A) and human TG monomer (PDB: 6scj, chain A) was conducted using the Matchmaker command of the UCSF ChimeraX program. The TG type 1 modules from human NID-2 were aligned with human TG, resulting in a sequence alignment score of 267.2 (Figure 10). Interestingly, the TG type 1-1 module of NID-2 precisely matches the TG type 1-1 module of human TG (Figure 10). In contrast, the TG type 1-2 module of NID-2 does not align with human TG; likely due to the greater complexity of its alpha helix, which may interfere with structural alignment. A similar result was observed when TG type 1-1 and TG type 1-2 from human NID-2 (chain A)—obtained through homology modeling—were alignd with full human NID-2 (1.cif chain A), generated using AlphaFold server. The alignement resulted in a sequence alignment score of 774.5 (data not shown).

Phylogenetic analysis of thyroglobulin sequences and TG type 1 domains

The TG amino acid sequences were used to construct a maximum likelihood tree encompassing 32 vertebrate species (*Aquarana catesbeiana*, *Astyanax mexicanus*, *Bos Taurus*, *Canis lupus familiaris*, *Chelonia mydas*, *Columba livia*, *Crotalus tigris*, *Cynoglossus semilaevis*, *Danio rerio*, *Gallus gallus*, *Gorilla gorilla,* *Homo sapiens*, *Lampetra fluviatilis*, *Lampetra planeri*, *Larus michahellis*, *Lepisosteus oculatus*, *Macaca fascicularis*, *Macaca mulatta*, *Mus musculus*, *Oryzias latipes*, *Pan troglodytes*, *Panthera leo*, *Petromyzon marinus*, *Pongo pygmaeus*, *Rattus norvegicus*, *Stegastes partitus*, *Struthio camelus*, *Taeniopygia guttata*, *Takifugu flavidus*, *Trichechus manatus latirostris*, *Xenopus tropicalis* and *Xiphophorus maculatus*; Figure 11). The resulting tree revealed the expected overall topology, preserving established phylogenetic relationships among taxa. Most nodes received strong statistical support (bootstrap scores ≥70; Figure 11). *Petromyzon marinus* belongs to the same clade as *Homo sapiens*.

A second phylogenetic analysis focused specifically on TG type 1 domains derived from both TGs and nidogens of *Petromyzon marinus* and *Homo sapiens* (Figure 12). Despite lower bootstrap support values, the resulting tree resolved two primary clades. The first clade included *Homo sapiens* TG type 1-4 and *Petromyzon marinus* TG type 1-6 domains, which clustered with the TG type 1-1, -1-2, -1-7, -1-9, and -1-10 domains from both species. The second clade comprised the *Homo sapiens* TG type 1-6 and *Petromyzon marinus* TG type 1-4 domains, along with the TG type 1-3, -1-5, -1-8, and -1-11 domains from both taxa. Notably, all nidogen-derived TG type 1 domains were consistently grouped within this second clade. Our findings are consistent with those of Belkadi et al. (2012), who investigated the phylogeny of TG type 1 domains through an analysis that included sequences from sea urchin, amphioxus, zebra finches, mice, and humans. Their study also supports a division of TG type 1 domains into two major branches. This phylogenetic clustering suggests that TG type 1 repeats may have originated from an ancestral nidogen-like structure, as indicated by the consistent co-grouping of homologous domains from both protein families within a shared evolutionary lineage.

Evolutionary model of the thyroglobulin complexation process.

In our evolutionary model, we propose that a protein with TG type 1 modules may have gave rise to TG. The TG type 1 domains that phylogenetically cluster with nidogen domains likely represent ancestral segments originally derived from a nidogen-like precursor (Figure 13a). These domains may correspond to the earliest TG type 1 repeats that led to the emergence of a proto-TG containing only a limited number of such domains. On the other hand, the same nidogen-like precursor would give rise to NID-2 and by deletion of a type 1 TG to NID-1 (Figure 13a). A process facilitated by a favorable genomic and genetic rearrangement environment. Among the most significant catalysts for this process was ionizing radiation, particularly ultraviolet (UV) radiation, as early Cambrian atmospheric oxygen levels were estimated to be only about 10% of present levels [Sperling et al., 2015], and the ozone layer was not yet fully formed. Unlike NID-1 and NID-2, TG lacks the three

globular regions (G1, G2 and G3) and none of the associated modules (NIDO, nidogen G2 beta-barrel, EGF-like, LDL-receptor class B). The elimination of the domains in G1 and G2, as well as EGF-like modules, may have resulted from recombination events (Figure 13b). The loss of the G1 domain could have been partial, preserving the signal peptide. The loss of LDL-receptor class B repeats may be explained by molecular mechanisms leading to a premature termination codon before the G3 globular domain (Figure 13c). As a result, the first secretory proto-TG likely emerged, containing five TG type 1 modules (Figure 13d). In addition to the five motifs, the protein appears to have undergone several evolutionary variations, leading to the emergence of specific tyrosine residues. These residues facilitated the organization of iodine, enabling the synthesis of T_4 through the interaction of an acceptor tyrosine and a donor tyrosine (Figure 13e). It is plausible that a short amino-terminal portion of TG—containing a single hormonogenic site and displaying a secretory profile— was sufficient to sustain primitive vertebrates on the earth's surface, exerting physiological effects of HT. This hypothesis is supported by extensive evidence of a mild hypothyroid phenotype in patients homozygous for the p.Arg296Ter variant, which includes only three TG type 1 motifs [Caputo et al., 2007; Citterio et al., 2013; Pardo et al., 2009; Rivolta et al., 2005; Siffo et al., 2018; van de Graaf et al., 1999; Zou et al., 2018]. Further support comes from observations of the p.A340Ter mutant, which retains only four TG type-1 motifs and exhibits an active secretory profile, as described by Lee et al., 2011.

At a certain point, the proto-TG protein likely attained a stable structural and functional configuration comprising five domains—TG type 1-3, -4, -5, -8, and -11—based on those identified in *Petromyzon marinus*. However, under evolutionary pressure, it underwent internal gene duplication events within the TG gene itself, leading to an expansion of TG type 1 domains. These duplications gave rise to additional domains—TG type 1-1, -2, -6, -7, -9, and -10—which now form a distinct phylogenetic clade (Figure 13f). This clade is more divergent from nidogen and indicative of intragenic expansion events specific to the TG lineage. Interestingly, TG type 1-4 and TG type 1-6 occupy an intermediate phylogenetic position between the two major clades,

suggesting that they may represent transitional domains—molecular fingerprints of the divergence process from nidogen-like ancestral sequences toward the fully developed TG type 1 repeats. Their positioning supports the hypothesis of a gradual structural and functional transition, in which certain TG type 1 domains retained ancestral features during the evolutionary remodeling that led to the expansion and specialization of the TG protein in vertebrates.

Subsequent tandem duplication events involving TG type-1 modules—ultimately yielding a protein with 11 repeats—continued to shape the structure of thyroglobulin (TG) (Figure 13f). Insertions of genetic material following modules 4 and 10 are hypothesized to have contributed to the formation of the non-repetitive linker and hinge domains, respectively (Figure 13g). These modifications represent an additional evolutionary milestone in TG's development, progressively refining its architecture toward the definitive form observed in the sea lamprey, a species situated at the base of vertebrate evolution.

Our structural analysis suggests a relationship between the TG type 1 and TG type 3-a modules, as both contain of six cysteines. This finding supports the hypothesis that TG type 3 modules may have originated from the duplication of the TG type 1-11 module (Figure 14h). Through tandem duplications, the TG type 3-a1 module gave rise to the TG type 3-b1, TG type 3-a2 and TG type 3-a3 domains (Figure 14i). Further gene modulations involving partial deletions led to the differentiation of TG type 3-a2 and TG type 3-a3 from TG type 3-a1, while TG type 3-b1 underwent an insertion event, and the addition of two extra cysteines due to the duplication of cysteine 3 and the presence an additional cysteine in the 3-b1 module (Figure 14j). The structural similarity between the TG type 3-b1 and TG type 3-b2 modules is due to the conservation of eight cysteines in both. This suggests that the duplication of the TG type 3-b1 module with its insertion between TG type a-2 and TG type a-3, led to the emergence of the TG type 3-b2 module (Figure 14k). TG type 3-b2 undergoes a deletion, reducing its length to 140 amino acids in sea lamprey TG (Figure 4), compared 172 amino acids in TG type 3-b1 and 134 out of 169 amino acids in

human TG (Figure 4). Thus, TG incorporates the five modules that constitute its region III (Figure 14l). These processes considerably reduced the sequence conservation patterns in region III due to its moderate evolutionary pressure, in contrast to TG type 1 repetitive modules, whose antiproteolytic activity is crucial for protein preservation.

The TG Type 2 modules likely originated from the insertion of a transposable DNA fragment between the hinge and TG Type 1-11 domains (Figure 14m), followed by two duplications events (Figure 14n). To investigate the origins of this module, the TG type 2-1, TG type 2-2 and TG type 2-3 modules were individually compared against the human genome ad other organisms using the BLAST tool (NCBI). The analysis revealed no homology with any genomic elements apart from the TG Type 2 domain itself. However, Lee et al. [2011] and Belkadi et al. [2012] classified it as a distant member of the GCC2/GCC3 domain superfamily. The higher conservation of TG Type 2 modules compared to TG type 1 and TG type 3 modules, as indicated by the sea lamprey-human alignment (Figure 3, Table 1), supports to the hypothesis that the TG Type 2 region emerged after the formation of the TG type 3 modules.

The final—and perhaps most significant—structural modulation event in TG was the incorporation of the ChEL domain and its fusion with regions I, II, and III (Figure 14o and p). The ChEL domain functions as an intramolecular chaperone, enhancing TG's secretory potential [Lee et al., 2008]. The higher conservation of the ChEL domain compared to TG Type 1 and TG Type 3 modules, as indicated by the sea lamprey-human alignment (Table 1), supports the hypothesis that the ChEL domain emerged after the formation of TG Type 1 and TG Type 3 modules. With the incorporation of the ChEL domain, TG is now secreted more efficiently and has an enhanced ability to increase T_4 concentration in the blood. Additionally, TG has acquired the carboxy-terminal hormonogenic site necessary for T_3 formation. The secretion of TH in its active form, T_3, along with the increased levels of T_4, may have enhanced all functions associated with TH. As

previously noted, in the TG of the *Petromyzon marinus*, the complexation process is already fully established.

Discussion

Protein diversity arises through a series of processes, including mutation, fixation, and selection, which can be neutral, deleterious, or advantageous. Domain duplication, gene fusion, and recombination are fundamental evolutionary mechanisms that shape proteins and contribute to proteome complexity [Vogel et al., 2005]. Protein evolution is influenced not only by selection for structure and function but also by the genomic location of coding genes, their expression patterns, their role within biological networks, and their resistance to translation errors [Pál et al., 2006]. Many proteins consist of combinations of multiple domains from diverse origins, enabling the modification of pre-existing proteins or the development of new architectures. Over time, these processes have led to the emergence of novel biological functions, increasing the complexity of multicellular organisms [Patthy, 2003; Vogel et al., 2004].

Interestingly, domain combinations within a protein can be represented as a network, where nodes correspond to functional domains from different superfamilies. The structural organization of TG exemplifies genetic evolution through intragenic duplication events involving TG type 1, TG type 2, and TG type 3 modules, as well as gene fusion, as observed with the ChEL domain (Figures 13 and 14). While the number of possible domain combinations is virtually unlimited, only a select few persist due to strong evolutionary selection. However, the reduction in protein diversity appears to be linked more to the rate of protein production than to overall protein abundance [Pál et al., 2006]. Additionally, radiation—particularly UV radiation—can act as a non-biological catalyst, driving successive cycles of DNA damage and subsequent repair through homologous recombination. This process likely accelerated the rapid emergence and increasing complexity of genomes, leading to the formation of intricate proteins such as TG within a relatively short evolutionary timeframe. All the modulating events proposed in our model occurred randomly, within a brief lapse of time, governed by the fundamental forces of natural selection

and genetic drift, with radiation as persistent physical stressor to DNA that possibly served as a catalyst, accelerating these evolutionary processes.

Using accumulated genomic data, we obtained the full-length cDNA of sea lamprey TG, TGPM (Figure 1), and compared its regions and domains with human TG. Notably, a second lamprey TG transcript, TGPM[1746] (Supplementary Figure S6), was identified in the NCBI database. This transcript begins with the TG type 1-10 module and extends to the carboxy-terminal end, exhibiting high homology with TGPM, except in region III, where the TG type 3 repeats show significant variability, and in spacer 3, they resemble the TG type 3b modules. Additionally, this transcript features an amino-terminal sequence compatible with a signal peptide and was likely generated through alternative splicing from a novel promoter.

Although the primary structures of sea lamprey and human TG diverge, with 35.1% identity and 50.5% similarity, all domains—both repetitive (TG type 1, TG type 2, and TG type 3) and non-repetitive (linker, hinge, and ChEL)—as well as the amino- and carboxy-terminal hormonogenic sites, remain strictly conserved in sea lamprey (Figure 1). Given the lower identity and similarity of TG type 1 modules compared to the ChEL domain, we propose that these two domains emerged at different stages of TG evolution, under different environmental conditions. TG type 1 modules likely arose earlier, in more primitive periods under stressful environmental conditions, which drove more modulated events leading to sequence variability. In contrast, the more conserved amino acid profile of the ChEL domain suggests that its fusion with the proto-TG occurred during a period of reduced genomic rearrangement and lower environmental stress. The structural complexity of genetic traits enabled vertebrates to colonize terrestrial environments and become the dominant species on earth. The fusion of TG with the ChEL domain may have marked a pivotal turning point in vertebrate evolution.

The phylogenetic clustering identified in our analysis offers insight into the evolutionary origins of TG type 1 domains. The consistent co-clustering of nidogen-derived TG type 1 repeats with TG

domains from *Homo sapiens* and *Petromyzon marinus* within a single clade suggests a shared ancestral origin. The presence of homologous TG type 1 repeats in both TG and nidogen—structurally and functionally distinct extracellular proteins—supports the hypothesis that TG type 1 domains may have evolved from an ancestral nidogen-like protein. Following this divergence, the TG gene likely underwent intragenic duplication events, leading to an expanded repertoire of TG type 1 repeats and the emergence of the modular architecture characteristic of modern TG proteins. This evolutionary scenario aligns with the broader model of modular protein evolution, especially among extracellular matrix components, in which domain shuffling, duplication, and specialization have driven functional diversification across protein families. The retention of conserved cysteine-rich motifs within TG type 1 domains further implies selective pressure to preserve structural elements essential for protein stability and interaction. Altogether, these findings support a model wherein nidogen-like proteins acted as evolutionary scaffolds for the emergence of TG type 1 domains, followed by lineage-specific expansions that account for the diversity observed in vertebrate TGs.

The TG type 1 module predates the emergence of TG in the earliest vertebrates and is present in several extracellular proteins, suggesting its likely involvement in the evolution of multicellular organisms by facilitating cell-cell and cell-environment interactions [Novinec et al., 2006]. A total of 170 protein sequences have been identified containing 333 TG type 1 modules [Novinec et al., 2006]. Several proteins with TG type 1 domains, such as nidogen—proposed by us as a precursor to TG type 1 modules—play crucial physiological roles, including partial or complete inhibition of protein degradation [Fowlkes et al., 1997]. The TG type 1 module can also regulate specific degradation events in proteins that incorporate this module into their structure [Fineschi et al., 1995]. The evolutionary success of the TG type 1 module is likely attributed to its ability to reversibly bind proteases, offering protection to proteins with more resilient TG type 1 modules, enabling them to endure and thrive in proteolytic environments during the evolution and

establishment of multicellular life [Molina, 1996; Novinec et al., 2006]. In a multicellular environment, close cellular interactions create a microenvironment more exposed to proteolytic enzymes released by neighboring cells during natural metabolic processes. The process of TH release itself involves the transport of TG to lysosomes [Ericson, 1981; Rousset and Mornex, 1991]. TG degradation is a slow process, requiring several hours [Brix and Herzog, 1994], and involves proteolytic cleavage by thyroid proteases [Dunn, 1991]. The eleven TG type 1 repeats may regulate this process in thyrocytes through selective and reversible inhibition of endosomal and/or lysosomal proteases, playing a vital role in the physiology of TH release [Molina, 1996].

Remarkable, in the TG type 1-3, TG type 1-7, and TG type 1-8 modules, insertions in human TG are highly conserved in sea lamprey TG (Figure 2), likely originating from an intron exonization event, as suggested by Parma et al. [1987] for the central insertion of TG type 1-7, or from random DNA insertion mechanisms. While both possibilities must be considered, it is also plausible that gene rearrangements and recombination-repair processes played a key role in promoting this insertion pattern, alongside tandem duplication and gene modulation events that shaped these modules throughout their evolutionary trajectory. These findings lead us to hypothesize that these TG type 1 repeats underwent extensive intragenic rearrangements, preserving only structural elements essential for their physiological function, such as the six cysteines required for disulfide bond formation. This strongly suggests that these insertions emerged during or prior to lamprey evolution and have remained unchanged in vertebrates.

Proteins require a stable three-dimensional structure to function properly. This stability ensures that a sufficient number of active proteins are available in the cell [Pál et al., 2006]. Each TG monomer contains 122 cysteine residues, accounting for 4.44% of the total amino acids [Citterio et al., 2021; Holzer et al., 2016; Malthiery and Lissitzky, 1987; Mercken et al., 1985; van de Graaf et al., 2001]. Newly synthesized TG folds through the formation of up to 60 intrachain disulfide bonds. As noted above, TG type 1, TG type 2, and TG type 3 domains are cysteine-rich repeat

modules that are covalently linked by intrachain disulfide bonds, playing a crucial role in shaping the tertiary structure of TG. The TG type 1 domains from human and bovine TG are further categorized into two subgroups: type 1A and type 1B. Type 1A repeats (TG type 1–1 to TG type 1–8 and TG type 1–10) contain six cysteines residues, whereas type 1B repeats (TG type 1–9 and TG type 1–11) include only four cysteines residues [Molina et al., 1996]. Conserved intradomain disulfide bonding patterns—Cys1–Cys2, Cys3–Cys4, and Cys5–Cys6—are consistently observed in all type 1A repeats [Molina et al., 1996]. Additionally, TG type 1B, TG type 2, TG type 3 repeats, and the six Cys residues of the ChEL domain also undergo intrachain disulfide bonding. Remarkably, over 90% of the cysteine residues are conserved between sea lamprey and human TG, underscoring their critical role in the structural and functional integrity of TG.

In conclusion, we have identified and characterized the full-length cDNA of sea lamprey TG, analyzing its structural conservation throughout evolution. Additionally, we detected a second TG transcript in this species, likely arising from an alternative splicing process. Based on our findings, we propose a model for the TG complexation process, beginning with the emergence of TG type 1 modules from a nidogen-like precursor, the process continues through intragenic duplication of these modules, followed by the sequential incorporation of TG type 3 and subsequently TG type 2 modules. This process culminates in fusion with the ChEL domain, representing the final step of TG complexation. Finally, our data confirm that the TG complexation process is fully established in the lamprey and has remained unchanged throughout vertebrate evolution.

Methods

Amino acid sequence alignments using Clustal Omega, T-Coffee, EMBOSS Needle and Blat Search Genome.

The TG amino acid sequence homology between multiple sequence were performed using the multiple sequence alignments Clustal Omega with its visualization tool MView (https://www.ebi.ac.uk/Tools/msa/clustalo/) and T-Coffee, (https://tcoffee.crg.eu/). T-Coffee is a versatile multiple sequence alignment tool suitable for aligning various types of biological sequences, including nucleic acids and proteins. More than just an aligner, T-Coffee acts as a framework that integrates different alignment methods with additional structural, evolutionary, or experimental data to enhance accuracy and meaningful alignments. It employs a library generated from a combination of pairwise alignments to construct multiple alignments. The coloured graphical output visually represents the level of consistency between the final alignment and the library used by T-Coffee, where the main score—the overall consistency value—ranges from 0 to 100. A score of 100 indicates complete agreement between the alignment and its reference library. The alignment file uses a color scheme to denote consistency, from blue (low-consistency regions), which are seldom biologically meaningful, to red (high-consistency regions)**,** which reflect strong agreement.

The amino acid sequence homology between sequence pairs was assessed using EMBOSS Needle (https://www.ebi.ac.uk/Tools/psa/emboss_needle/) and Blat Search Genome (https://genome-euro.ucsc.edu/cgi-bin/hgBlat).

SignalP 6.0: signal peptide prediction and cleavage site identification

The SignalP 6.0 server predicts the presence of signal peptides and determines their cleavage sites in proteins (https://services.healthtech.dtu.dk/services/SignalP-6.0/). It can distinguish between five types of signal peptides: (1) Sec/SPI, standard secretory signal peptides transported

by the Sec translocon and cleaved by Signal Peptidase I; (2) Sec/SPII, lipoprotein signal peptides transported by the Sec translocon and cleaved by Signal Peptidase II; (3) Tat/SPI, Tat signal peptides transported by the Tat translocon and cleaved by Signal Peptidase I; (4) Tat/SPII, Tat lipoprotein signal peptides transported by the Tat translocon and cleaved by Signal Peptidase II and (5) Sec/SPIII, Pilin and pilin-like signal peptides transported by the Sec translocon and cleaved by Signal Peptidase III.

Modeling the three-dimensional atomic structure of proteins

The three-dimensional atomic structure of *Homo sapiens* TG has been characterized using a composite cryo-electron microscopy (cryo-EM) density map (PDB: 6SCJ, resolution: 3.60 Å; https://www.rcsb.org/structure/6SCJ; Coscia et al., 2020). Coscia et al., [2020] proposes a new model of the primary structure of the human TG organized in five regions: N-Terminal Domain (NTD), Core, Flap, Arm and C-Terminal Domain (CTD). However, the TG models presented by Adaixo et al. [2022] and Kim et al. [2021] adhere to the general framework of the classical model for the TG monomer. Therefore, we adapted the Coscias et al., [2020] model align with the classical model. Spacers 1, 2, and 3 are located in regions II, III, and IV, respectively. The UCSF Chimera program was used to visualize the 3D model of the *Homo sapiens* TG (Pettersen et al., 2021; 2004; UCSF Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, https://www.cgl.ucsf.edu/chimera/).

The PDB for TG type 1-1 and TG type 1-2 from *Homo sapiens* NID-2, as well as of $TGPM^{2475}$ (Uniprot: S4R814_PETMA), $TGPM^{567}$ (NCBI: XP_032817835.1), $TGPM^{1746}$ (NCBI: XP_032817730.1) and full-lenght assembled TG (TGPM) from *Petromyzon marinus*, were generated using homology modeling with Swiss Model (https://swissmodel.expasy.org/interactive) due to the unresolved crystallographic data.

The three-dimensional structures of NID-2 from *Petromyzon marinus* and *Homo sapiens*, as

well as TG type 1-1 and TG type 1-2 modules from NID-1, NID-2, and TG from *Petromyzon marinus*, *Mus musculus*, and *Homo sapiens*, were generated using the AlphaFold server (https://alphafoldserver.com/). pLDDT (predicted Local Distance Difference Test) provides a per-atom confidence estimate on a 0-100 scale, where a higher value indicate greater reliability. Regions with low pLDDT scores may correspond to unstructured or highly flexible areas.

Structural matching using the Matchmaker command in UCSF ChimeraX.

Structural matching (structural superposition) of different PDBs was performed using the Matchmaker command in UCSF ChimeraX. This command first generates a pairwise sequence alignment and then refines the alignment by optimizing residue pairs at the atomic level. The RMSD (Root-Mean-Square Deviation) value in Matchmaker quantifies the average distance between atoms—typically backbone atoms—of superimposed molecules, measured in Ångströms (Å). In globular protein conformation studies, three-dimensional structural similarity is commonly assessed by calculating the RMSD of Cα atomic coordinates after optimal rigid-body superimposition. A cutoff value of 2 Å is used, where lower RMSD values indicate higher-quality superimposition, with values approaching 0 Å representing near-perfect alignment.

Phylogeny

TG and TG type 1 repetitive phylogeny were based in the amino acid sequence. We obtained the protein sequences available for TG from the NCBI databank (https://www.ncbi.nlm.nih.gov) for thirty species: *Aquarana catesbeiana* (XP_073488945.1), *Astyanax mexicanus* (XP_022529282.2), *Bos Taurus* (NP_776308.1), *Canis lupus familiaris* (NP_001041569.1), *Chelonia mydas* (XP_043394997.1), *Columba livia* (XP_021154537.2), *Crotalus tigris* (XP_039206379.1), *Cynoglossus semilaevis* (XP_008321228.1), *Danio rerio* (NP_001316794.1), *Gallus gallus* (NP_001376406.2), *Homo sapiens* (NP_003226.4), *Lampetra fluviatilis*

(CAL5909345.1), *Lampetra planeri* (CAL5920418.1), *Larus michahellis* (XP_074433977.1), *Lepisosteus oculatus* (XP_015212882.2), *Macaca fascicularis* (XP_045254657.2), *Macaca mulatta* (XP_028708128.1), *Mus musculus* (NP_033401.2), *Oryzias latipes* (XP_011484169.1), *Pan troglodytes* (XP_016815373.4), *Panthera leo* (XP_042780307.1), *Pongo pygmaeus* (XP_054355108.2), *Rattus norvegicus* (NP_112250.2), *Stegastes partitus* (XP_008304814.1), *Struthio camelus* (XP_068789869.1), *Taeniopygia guttata* (XP_072781126.1), *Takifugu flavidus* (TWW79869.1), *Trichechus manatus latirostris* (XP_004373071.1), *Xenopus tropicalis* (NP_001316486.1) and *Xiphophorus maculatus* (XP_023185937.1). We also obtained the protein sequence from Uniprot (https://www.uniprot.org): *Gorilla gorilla* TG (G3QS68), *Homo sapiens* NID-1 (P14543), *Homo sapiens* NID-2 (Q14112), *Petromyzon marinus* NID-1 (A0AAJ7TJ46) and *Petromyzon marinus* NID-2 (A0AAJ7X4A4).

The sequences were aligned using MAFFT tool [Katoh et al., 2013]. Maximum Likelihood (ML) phylogeny was reconstructed using IQ-TREE v1.6.12 [Nguyen et al., 2015], utilizing the best substitution model determined by ModelFinder [Kalyaanamoorthy et al., 2017], and with 2000 ultrafast bootstraps [Hoang et al., 2018]. The phylogenetic tree was visualized with iTOL [Letunic and Bork, 2016].

Acknowledgements

M.G.P. is a research fellow of the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). W.M.S., C.M.R. and H.M.T. are established investigators of the CONICET.

Author contributions

MGP. contributed to bioinformatic analysis and conducted the structural modeling analysis. W.M.S. performed the phylogenetic analisis. CMR. was involved in the bioinformatic analysis, funding acquisition, and contributed to study design. HMT performed the bioinformatics alignement analysis, participated in funding acquisition, conceived and designed the study, and wrote the manuscript. All authors critically reviewed the manuscript, contributed to its revision, and approved the final version.

Funding

This study was funded by grants from the Fondo para la Investigación Científica y Tecnológica (FONCyT-ANPCyT-MINCyT, PICT-2018-02146 to H.M.T.), CONICET (PIP 2021-11220200102976CO to C.M.R.) and Universidad de Buenos Aires (UBACyT 2020-20020190100050BA to C.M.R.).

Data availability

Data and material are available from the authors upon request.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval Not applicable. No procedures requiring ethics approval were performed in this study.

Consent to participate Not applicable.

Consent to publish Not applicable.

References

Adaixo R, Steiner EM, Righetto RD, Schmidt A, Stahlberg H, Taylor NMI (2022) Cryo-EM structure of native human thyroglobulin. Nat Commun 13:61. https://doi.org/10.1038/s41467-021-27693-8

Bechtel M, Keller, MV, Bloch W, Sasaki T, Boukamp P, Zaucke F, Paulsson M, Nischt R (2012) Different domains in nidogen-1 and nidogen-2 drive basement membrane formation in skin organotypic cocultures. FASEB J 26:3637-3648. https://doi.org/10.1096/fj.11-194597

Belkadi A, Jacques C, Savagner F, Malthièry Y (2012) Phylogenetic analysis of human thyroglobulin regions. Thyroid Res 5:3. https://doi.org/10.1186/1756-6614-5-3.

Brix K, Herzog V (1994) Extrathyroidal release of thyroid hormones from thyroglobulin by J774 mouse macrophages. J Clin Invest 93:1388-1396. https://doi.org/10.1172/JCI117115

Caputo M, Rivolta CM, Esperante SA, Gruñeiro-Papendieck L, Chiesa A, Pellizas CG, González-Sarmiento R, Targovnik HM (2007) Congenital hypothyroidism with goitre caused by new mutations in the thyroglobulin gene. Clin Endocrinol 67:351–357. https://doi.org/10.1111/j.1365-2265.2007.02889.x

Citterio CE, Machiavelli GA, Miras MB, Gruñeiro-Papendieck L, Lachlan K, Sobrero G, Chiesa A, Walker J, Muñoz L, Testa G, Belforte FS, Gonzalez-Sarmiento R, Rivolta CM, Targovnik HM (2013) New insights into thyroglobulin gene: molecular analysis of seven novel mutations associated with goiter and hypothyroidism. Mol Cell Endocrinol 365: 277–291. https://doi.org/10.1016/j.mce.2012.11.002

Citterio CE, Rivolta CM, Targovnik HM (2021) Structure and genetic variants of thyroglobulin: pathophysiological implications. Mol Cell Endocrinol 508:111227. https://doi.org/10.1016/j.mce.2021.111227

Coscia F, Taler-Verčič A, Chang VT, Sinn L, O'Reilly FJ, Izoré T, Renko M, Berger I, Rappsilber J, Turk D, Löwe J (2020) The structure of human thyroglobulin. Nature 578:627-630. https://doi.org/10.1038/s41586-020-1995-4

Dunn AD, Cruntchfield HE, Dunn JT (1991) Proteolytic processing of thyroglobulin by extracts of thyroid lysosomes. Endocrinology 128:3073-3080. https://doi.org/10.1210/endo-128-6-3073

Ericson L (1981) Exocytosis and endocytosis in the thyroid follicle cell. Mol Cell Endocrinol 22:1-24. https://doi.org/10.1016/0303-7207(81)90098-8

Fineschi B, Arneson LS, Naujokas MF, Miller J (1995) Proteolysis of major histocompatibility complex class II-associated invariant chain is regulated by the alternatively spliced gene product, p41. Proc Natl Acad Sci USA 92:10257-10261. https://doi.org/10.1073/pnas.92.22.10257

Fowlkes JL, Thrailkill KM, George-Nascimento C, Rosenberg CK, Serra DM (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. https://doi.org/10.1210/endo.138.6.5182

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS (2018) UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol 35:518-522. https://doi.org/10.1093/molbev/msx281.

Holzer G, Morishita Y, Fini J-B, Lorin T, Gillet B, Hughes S, Tohmé M, Deléage G, Demeneix B, Arvan P, Laudet V (2016) Thyroglobulin represents a novel molecular architecture of vertebrates. J Biol Chem 291:16553-16566. https://doi.org/10.1074/jbc.M116.719047

Holland LZ, Short S (2010) Alternative splicing in development and function of chordate endocrine systems: a focus on Pax genes. Integr Comp Biol 50:22-34. https://doi.org/10.1093/icb/icq048

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14:587-589. https://doi.org/10.1038/nmeth.4285.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-80. https://doi.org/10.1093/molbev/mst010.

Kim K, Kopylov M, Bobe D, Kelley K, Eng ET, Arvan P, Clarke OB (2021) The structure of natively iodinated bovine thyroglobulin. Acta Crystallogr D Struct Biol 77:1451-1459. https://doi.org/10.1107/S2059798321010056

Lee J, Di Jeso B, Arvan P (2008) The cholinesterase-like domain of thyroglobulin functions as an intramolecular chaperone. J. Clin. Invest. 118, 2950-2958. https://doi.org/10.1172/JCI35164

Lee J, Di Jeso B, Arvan P (2011) Maturation of thyroglobulin protein región I. J Biol Chem 286:33045-33052. https://doi.org/10.1074/jbc.M111.281337

Letunic I, Bork P (2024) Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. Nucleic Acids Res 52(W1):W78-W82 https://doi.org/10.1093/nar/gkae268.

Malthiéry Y, Lissitzky S (1987) Primary structure of human thyroglobulin deduced from the sequence of its 8448-base complementary DNA. Eur J Biochem 165:491-498. https://doi.org/10.1111/j.1432-1033.1987.tb11466.x

Marlétaz F, Timoshevskaya N, Timoshevskiy VA, Parey E, Simakov O, Gavriouchkina D, Suzuki M, Kubokawa K, Brenner S, Smith JJ, Rokhsar DS (2024) The hagfish genome and the evolution of vertebrates. Nature 627:811-820. https://doi.org/10.1038/s41586-024-07070-3

Mercken L, Simons M-J, Swillens S, Massaer M, Vassart G (1985) Primary structure of bovine thyroglobulin deduced from the sequence of its 8,431-base complementary DNA. Nature 316:647-651. https://doi.org/10.1038/316647a0

Molina F, Bouanani M, Pau B, Granier C (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. https://doi.org/10.1111/j.1432-1033.1996.0125h.x

Monaco F, Andreoli M, La Posta A, Roche J (1978) Thyroglobulin biosynthesis in a larval (ammocoete) and adult freshwater lamprey (*Lampetra planeri Bl*.). Comp Biochem Physiol B 60:87-91. https://doi.org/10.1016/0305-0491(78)90033-0

Novinec M, Kordiš D, Turk V, Lenarčič B (2006) Diversity and evolution of the thyroglobulin type-1 domain superfamily. Mol Biol Evol 23:744-755. https://doi.org/10.1093/molbev/msj082

Ohmiya Y, Suzuki S, Kondo Y (1989) Isolation and characterization of hagfish thyroid iodoprotein by its non-thyroglobulin nature, very high iodine and carbohydrate contents and low hormone/iodine ratio. Eur J Biochem 182:11-8. https://doi.org/10.1111/j.1432-1033.1989.tb14794.x

Pál C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. Nat Rev Genet 7:337-348. https://doi.org/10.1038/nrg1838

Pardo V, Vono-Toniolo J, Rubio IGS, Knobel M, Possato RF, Targovnik HM, Kopp P, Medeiros-Neto G (2009) The p.A2215D thyroglobulin gene mutation leads to deficient synthesis and secretion of the mutated protein and congenital hypothyroidism with wide phenotype variation. J Clin Endocrinol Metab 94:2938–2944. https://doi.org/10.1210/jc.2009-0150

Park Y-n, Arvan P (2004) The acetylcholinesterase homology region is essential for normal conformational maturation and secretion of thyroglobulin. J Biol Chem 279:17085-17089. https://doi.org/10.1074/jbc.M314042200

Parma J, Christophe D, Pohl V, Vassart G (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. https://doi.org/10.1016/0022-2836(87)90403-7

Patthy L (2003) Modular assembly of genes and the evolution of new functions. Genetica 118:217–231. https://doi.org/10.1023/A:1024182432483

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera-a visualization system for exploratory research and analysis. J Comput Chem 25:1605-1612. https://doi.org/10.1002/jcc.20084

Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE (2021) UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci 30:70-82. https://doi.org/10.1002/pro.3943

Putnam NH, Butts T, Ferrier DE, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu JK, Benito-Gutiérrez EL, Dubchak I, Garcia-Fernàndez J, Gibson-Brown JJ, Grigoriev IV, Horton AC, de Jong PJ, Jurka J, Kapitonov VV, Kohara Y, Kuroki Y, Lindquist E, Lucas S, Osoegawa K, Pennacchio LA, Salamov AA, Satou Y, Sauka-Spengler T, Schmutz J, Shin-I T, Toyoda A, Bronner-Fraser M, Fujiyama A, Holland LZ, Holland PW, Satoh N, Rokhsar DS (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064-1072. https://doi.org/10.1038/nature06967

Rivolta CM, Moya CM, Gutnisky VJ, Varela V, Miralles-García JM, González-Sarmiento R, Targovnik HM (2005) A new case of congenital goiter with hypothyroidism due to a homozygous p.R277X mutation in the exon 7 of the thyroglobulin gene: a mutational hot spot could explain the recurrence of this mutation. J Clin Endocrinol Metab 90:3766–3770. https://doi.org/10.1210/jc.2005-0278

Rousset B, Mornex R (1991) The thyroid hormone secretory pathway--current dogmas and alternative hypotheses. Mol Cell Endocrinol 78:C89-C93. https://doi.org/10.1016/0303-7207(91)90176-s

Schumacher M, Camp S, Maulet Y, Newton M, MacPhee-Quigley K, Taylor SS, Friedmann T, Taylor P (1986) Primary structure of *Torpedo californica* acetylcholinesterase deduced from its cDNA sequence. Nature 319:407-409. https://doi.org/ 10.1038/319407a0

Shimeld SM, Donoghue PC (2012) Evolutionary crossroads in developmental biology: cyclostomes (lamprey and hagfish). Development 139:2091-2099. https://doi.org/10.1242/dev.074716

Siffo S, Adrover E, Citterio CE, Miras MB, Balbi VA, Chiesa A, Weill J, Sobrero G, González VG, Papendieck P, Martinez EB, González-Sarmiento R, Rivolta CM, Targovnik HM (2018) Molecular analysis of thyroglobulin mutations found in patients with goiter and hypothyroidism. Mol Cell Endocrinol 473:1–16. https://doi.org/10.1016/j.mce.2017.12.009

Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, Yandell MD, Manousaki T, Meyer A, Bloom OE, Morgan JR, Buxbaum JD, Sachidanandam R, Sims C, Garruss AS, Cook M, Krumlauf R, Wiedemann LM, Sower SA, Decatur WA, Hall JA, Amemiya CT, Saha NR, Buckley KM, Rast JP, Das S, Hirano M, McCurley N, Guo P, Rohner N, Tabin CJ, Piccinelli P, Elgar G, Ruffier M, Aken BL, Searle SM, Muffato M, Pignatelli M, Herrero J, Jones M, Brown CT, Chung-Davidson YW, Nanlohy KG, Libants SV, Yeh CY, McCauley DW, Langeland JA, Pancer Z, Fritzsch B, de Jong PJ, Zhu B, Fulton LL, Theising B, Flicek P, Bronner ME, Warren WC, Clifton SW, Wilson RK, Li W (2013) Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nat Genet 45:415-421, 421e1-2. https://doi.org/10.1038/ng.2568

Sperling EA, Knoll AH, Girguis PR (2015) The ecological physiology of earth's second oxygen revolution. Annu Rev Ecol Evol Syst 46:215-235. https://doi.org/10.1146/annurev-ecolsys-110512-135808

Suzuki S, Kondo Y (1971) Demonstration of thyroglobulin-like iodinated proteins in the branchial sac of tunicates. Gen Comp Endocrinol 17:402-406. https://doi.org/10.1016/0016-6480(71)90149-3

Suzuki S, Gorbman A, Rolland M, Montfort MF, Lissitzky S (1975) Thyroglobulins of cyclostomes and an elasmobranch**.** Gen Comp Endocrinol 26:56-69. https://doi.org/10.1016/0016-6480(75)90215-4

Swillens S, Ludgate M, Mercken L, Dumont JE, Vassart G (1986) Analysis of sequence and structure homologies between thyroglobulin and acetylcholinesterase: possible functional and clinical significance. Biochem Biophys Res Commun 137:142-148. https://doi.org/10.1016/0006-291x(86)91187-3

Takagi W, Sugahara F, Higuchi S, Kusakabe R, Pascual-Anaya J, Sato I, Oisi Y, Ogawa N, Miyanishi H, Adachi N, Hyodo S, Kuratani S (2022) Thyroid and endostyle development in cyclostomes provides new insights into the evolutionary history of vertebrates. BMC Biol 20:76. https://doi.org/10.1186/s12915-022-01282-7

Töpfer U, Holz A (2024) Nidogen in development and disease. Front Cell Dev Biol 12:1380542. https://doi.org/10.3389/fcell.2024.1380542

van de Graaf SAR, Ris-Stalpers C, Veenboer GJM, Cammenga M, Santos C, Targovnik HM, de Vijlder JJM, Medeiros-Neto G (1999) A premature stopcodon in thyroglobulin mRNA results in familial goiter and moderate hypothyroidism. J Clin Endocrinol Metab 84:2537–2542. https://doi.org/10.1210/jcem.84.7.5862

van de Graaf SAR, Ris-Stalpers C, Pauws E, Mendive FM, Targovnik HM, de Vijlder JJM (2001) Up to date with human thyroglobulin. J Endocrinol 170:307-321. https://doi.org/10.1677/joe.0.1700307

Vogel C, Bashton M, Kerrison N D, Chothia C, Teichmann SA (2004) Structure, function and evolution of multidomain proteins. Curr Opin Struct Biol 14:208-216. https://doi.org/10.1016/j.sbi.2004.03.011

Vogel C, Teichmann SA, Pereira-Leal J (2005) The Relationship Between Domain Duplication and Recombination. J Mol Biol 346:355-365. https://doi.org/10.1016/j.jmb.2004.11.050

Xu Y, Zhu S-W, Li Q-W (2016) Lamprey: a model for vertebrate evolutionary Research Zool Res 37:263-269. https://doi.org/10.13918/j.issn.2095-8137.2016.5.263

Yurchenco PD, Wadsworth WG (2004) Assembly and tissue functions of early embryonic laminins and netrins. Curr Opin Cell Biol 16:572-579. https://doi.org/10.1016/j.ceb.2004.07.013

Zou M, Alzahrani AS, Al-Odaib A, Alqahtani MA, Babiker O, Al-Rijjal RA, BinEssa HA, Kattan WE, Al-Enezi AF, Al Qarni A, Al-Faham MSA, Baitei EY, Alsagheir A, Meyer BF, Shi Y (2018) Molecular analysis of congenital hypothyroidism in Saudi Arabia: SLC26A7 mutation is a novel defect in thyroid dyshormonogenesis. J Clin Endocrinol Metab 103:1889–1898. https://doi.org/10.1210/jc.2017-02202

Table 1: **Comparative analysis of the TG type 1, TG type 2, and TG type 3 modules, along with the ChEL domain, of full-length thyroglobulin from *Petromyzon marinus* (TGPM) and *Homo sapiens* (TGHS), performed using the EMBOSS Needle program.**

	Identity %	Similarity %	Identity average %	Similarity average %
TG type 1-1 TGPM /TG type 1-1 TGHS	41.3	57.1	37.94	51.24
TG type 1-2 TGPM /TG type 1-2 TGHS	61.8	70.6		
TG type 1-3 TGPM /TG type 1-3 TGHS	31.4	47.9		
TG type 1-4 TGPM /TG type 1-4 TGHS	44.4	57.1		
TG type 1-5 TGPM /TG type 1-5 TGHS	42.1	57.9		
TG type 1-6 TGPM /TG type 1-6 TGHS	42.6	54.4		
TG type 1-7 TGPM /TG type 1-7 TGHS	34.7	48.8		
TG type 1-8 TGPM /TG type 1-8 TGHS	37.2	53.2		
TG type 1-9 TGPM /TG type 1-9 TGHS	26.3	44.7		
TG type 1-10 TGPM /TG type 1-10 TGHS	38.8	50.7		
TG type 1-11 TGPM /TG type 1-11 TGHS	16.7	21.2		
TG type 2-1 TGPM /TG type 2-1 TGHS	42.9	50.0	45.70	55.86
TG type 2-2 TGPM /TG type 2-2 TGHS	47.1	58.8		
TG type 2-3 TGPM /TG type 2-3 TGHS	47.1	58.8		
TG type 3-a1 TGPM /TG type 3-a1 TGHS	32.5	48.0	30.26	48.58
TG type 3-b1 TGPM /TG type 3-b1 TGHS	31.2	45.7		
TG type 3-a2 TGPM /TG type 3-a2 TGHS	27.8	53.7		
TG type 3-b2 TGPM /TG type 3-b2 TGHS	31.0	41.3		
TG type 3-a3 TGPM /TG type 3-a3 TGHS	28.8	54.2		
ChEL TGPM /ChEL TGHS	41.7	57.4	41.7	57.4

```
                                signal peptide                          TG Type 1-1
TGPM      1 MRTSPLLPATTTLYLVLWIGTISALEYSDSKTSLASACELLRQQAMEDGR    50
            .:.:|...|.|..:.|: :.:..||......| ..|||.|:.|.....
TGHS      1 --MALVLEIFTLLASICWV-SANIFEYQVDAQPL-RPCELQRETAFLKQA    46
                                                                TG Type 1-2
TGPM     51 LHVPQCSPHGAFRPLQCDAAGEPCWCVDAAGEELPGTRRAEGPPPSCLSF   100
            .:||||:..|:|:.:||...|..|||||.|.|.|:.|:|: .|.|.:||||
TGHS     47 DYVPQCAEDGSFQTVQCQNDGRSCWCVGANGSEVLGSRQ-PGRPVACLSF    95

TGPM    101 CQLHRQRVLLSGYLNGTRALHLPQCTQDGRYRPAQVD-SSGQGWCVDADG   149
            |||.:|::|||||:|.|...:||||...|.|.|.|.| ...|.|||||:|
TGHS     96 CQLQKQQILLSGYINSTDTSYLPQCQDSGDYAPVQCDVQQVQCWCVDAEG   145
                                TG Type 1-3
TGPM    150 MEVYGTRQLGAPLSCPSPCQVSVRRAVRASTPGSPPPQCDRDEQRVLATQ   199
            ||||||||||.|..||..|::..||.:......| ||||.. |...:..|
TGHS    146 MEVYGTRQLGRPKRCPRSCEIRNRRLLHGVGDKS-PPQCSA-EGEFMPVQ   193

TGPM    200 CQLIGTAGRSTLSLLDAFSMHPEAFRSLAEFRARLPGVHAYCYCADPQGR   249
            |:.:.|.......|:.:::..|:||.:.:.|:.|.|.|..||:|||.|||
TGHS    194 CKFVNTTDMMIFDLVHSYNRFPDAFVTFSSFQRRFPEVSGYCHCADSQGR   243

TGPM    250 ELPETGDRGVRSFLFVAYFSGHGLNHEVVGSVLGKTRHEKFLKVQDRLPG   299
            ||.|||...:...::...|:|..|.....:.|.:....:||.||..:.|
TGHS    244 ELAETGLELLLDEIYDTIFAGLDLPSTFTETTLYRILQRRFLAVQSVISG   293
                 TG Type 1-4
TGPM    300 HIVRPTVCEVERSLAERLGGRGFAPVCSQDGSYIPTQC--HGPTCWCVDT   347
            ....||.|||||..|... |..:.|.|.::|.|...||  .|| |||||.
TGHS    294 RFRCPTKCEVERFTATSF-GHPYVPSCRRNGDYQAVQCQTEGP-CWCVDA   341
                                Linker
TGPM    348 RGDEVFGSRVRDSVPNCGSSSSPCQLERQLALSHLFLGPSAPFID-DLIS   396
            :|.|:.|:|.:...|:|....| |..|||.|||.|:.|.|..|.. ||.|
TGHS    342 QGKEMHGTRQQGEPPSCAEGQS-CASERQQALSRLYFGTSGYFSQHDLFS   390

TGPM    397 ------ALPRDGSVSFSESPSQEMQHELWLHWRALREQLLDGG-LGEILS   439
                  |.||....:.|..|:              ::|..:|.| |..::.
TGHS    391 SPEKRWASPRVARFATSCPPT-------------IKELFVDSGLLRPMVE   427

TGPM    440 QASRLFPASGQASRLLGDLVGGIFHTQELALTAAGANVQPGRIAEMLFGD   489
            ..|:.|..|   ..||.:.:..||.::.||..|.....|.|:.:.|||.
TGHS    428 GQSQQFSVS---ENLLKEAIRAIFPSRGLARLALQFTTNPKRLQQNLFGG   474

TGPM    490 RFLKNLKFFNFTGAVGGRGTFNFSKVFGQVGLTGMYDGANFEELARLFAP   539
            :||.|:..||.:||:|.|||||||:.|.|:||....:|...|:||:..:.
TGHS    475 KFLVNVGQFNLSGALGTRGTFNFSQFFQQLGLASFLNGGRQEDLAKPLSV   524

TGPM    540 GEDSYLTRGSSNFSRESFNFDQSIDDPFARATNLERNRNVLGFVVTLLED   589
            |.||..:.|:...:::....::.....|....||:.|:|.|.|:.:|||.
TGHS    525 GLDSNSSTGTPEAAKKDGTMNKPTVGSFGFEINLQENQNALKFLASLLEL   574

TGPM    590 ARFY-----------GLVRDITPLIGLSGGDVGLNALTFLRMAEEDRPTQ   628
            ..|.            .:.||:        |||               ..|.
TGHS    575 PEFLLFLQHAISVPEDVARDL--------GDV--------------METV   602
                TG Type 1-5
TGPM    629 NAADAADGPPASRPFVARCGEDGGAYEPVQCHGAACWCVDAHGREIAGTR   678
            .::...:..| .|.||..|..: |:||.|||....||||::.|:|:.|:|
TGHS    603 LSSQTCEQTP-ERLFVPSCTTE-GSYEDVQCFSGECWCVNSWGKELPGSR   650
                        TG Type 1-6
TGPM    679 AVGGAPPRCPSACEGERARALLARRGRPVGTPLFVPECDGAGDYRPVQCS   728
            ..|| .||||:.||.:|||.......:|.|:.||||.|...|.:.||||.
TGHS    651 VRGG-QPRCPTDCEKQRARMQSLMGSQPAGSTLFVPACTSEGHFLPVQCF   699
                                                TG Type 1-7
TGPM    729 GDRCFCVGAGGAEVPGTGRPLGDPVTCPTPCQAAAASELLQQLRHLGPVL   778
            ...|:||.|.|..:|||...:|.|..||||||..:....|:.::.|   |
TGHS    700 NSECYCVDAEGQAIPGTRSAIGKPKKCPTPCQLQSEQAFLRTVQAL---L   746

TGPM    779 RGDAPSSLPPL---YVPRCDARGGWRPVQCDGAGQQIREFVDAWTKDEGN   825
            ...  |.||.|   |:|:|...|.||.|||:|..:|:.|....| :.:..
TGHS    747 SNS--SMLPTLSDTYIPQCSTDGQWRQVQCNGPPEQVFELYQRW-EAQNK   793
```

```
TGPM     826 ASKATLSDLRELLARVRGAGGGGGGGGKGLRSFLSFLYDSGRQDMFPELS    875
             ....|.:.|...:...|.|..|       ....|:..||::|:||:||.||
TGHS     794 GQDLTPAKLLVKIMSYREAASG------NFSLFIQSLYEAGQQDVFPVLS    837

TGPM     876 LYPSLRGPPEVLNFSGSSGR-------FLRNPEVVWKILTENASFVYTGD    918
             .||||:..|    .:...|:        .|..|.:.|:||....| .|.|.
TGHS     838 QYPSLQDVP----LAALEGKRPQPRENILLEPYLFWQILNGQLS-QYPGS    882
                                                   TG Type 1-8
TGPM     919 YAEFSGRYADFESRLCWCVDAAGHEIDGTRTVSPGQLPKCPGACHLAAAD    968
             |::||...|.|:.|.|||||.||.|::|.|: .|.:||.|||:|..|...
TGHS     883 YSDFSTPLAHFDLRNCWCVDEAGQELEGMRS-EPSKLPTCPGSCEEAKLR    931

TGPM     969 VNRYLQQADLLIGSAGAS--SAGEGVAFGRGLAFTEDELLGSPLGLGDVK   1016
             |.:::::.:.::.::.:|  ..||.....:|:....:: ||.| .|...:
TGHS     932 VLQFIRETEEIVSASNSSRFPLGESFLVAKGIRLRNED-LGLP-PLFPPR    979

TGPM    1017 GEVAAVLAGGTGYAVRLAAQAAMHFHWRRRFFGPGTIGEGVFNGFDPYMP   1066
             ...|.....|:.||:|||||:.:.|:.||||....:.|........||||
TGHS     980 EAFAEQFLRGSDYAIRLAAQSTLSFYQRRRFSPDDSAGASALLRSGPYMP   1029
                                                        TG Type 1-9
TGPM    1067 QCTEPGGWEPAQCVVRAGAGFCWCVDAAGEFVAGSLVARPRRRPQCATPC   1116
             ||...|.|||.||  .||.|.|||||..|.|:.|||.||..:.|||.|.|
TGHS    1030 QCDAFGSWEPVQC--HAGTGHCWCVDEKGGFIPGSLTARSLQIPQCPTTC   1077

TGPM    1117 QRARAEALLTGWKSLGSVENGSSIEVLTKHTPACTPSRKYKKCHYLLRPN   1166
             :::|...||:.||...|.||.|..::.   .|||..:.:|.:........
TGHS    1078 EKSRTSGLLSSWKQARSQENPSPKDLF---VPACLETGEYARLQASGAGT   1124
                                TG Type 1-10
TGPM    1167 WWPSPSYSYTPTHPSPTTNAESPSLCELRRRQALLRNAGLGSVPECDSR-   1215
             |...|: |.....|..:::|:.||||.:.:...|.|....|.||.|.:.
TGHS    1125 WCVDPA-SGEELRPGSSSSAQCPSLCNVLKSGVLSRRVSPGYVPACRAED   1173
                                                      HINGE
TGPM    1216 GDYAARQC--ADGTCWCAGAGGEEIPDTRRHAGEPGVVCHEPRCELPQDE   1263
             |.::..||  |.|:|||....|||:|.||...|:|  .|..|||.||.:.
TGHS    1174 GGFSPVQCDQAQGSCWCVMDSGEEVPGTRVTGGQP--ACESPRCPLPFNA   1221

TGPM    1264 RRAGRWALVCEDGAAATAAAGLQRCLLACLRGFAHTLDGAGHDGGAPREF   1313
             .......::||..:..|.:| :|:|.|.|.:|.....        .|...
TGHS    1222 SEVVGGTILCETISGPTGSA-MQQCQLLCRQGSWSVF--------PPGPL   1262

TGPM    1314 QCNATSGEWIGSSPQPDACQEIHPWQAVQLSISFTLRFPEEKMCNPEYEG   1363
             .|:..||.|....|||.|||....||.:|....|.|:.|..|||:.:|..
TGHS    1263 ICSLESGRWESQLPQPRACQRPQLWQTIQTQGHFQLQLPPGKMCSADYAD   1312

TGPM    1364 LLESFNSFLLADLTARGFCHLSNGVGSDAVA--VCDAASVRVRCQDALRL   1411
             ||::|..|:|.:|||||||.:........|:  ||:.:||:|.|....||
TGHS    1313 LLQTFQVFILDELTARGFCQIQVKTFGTLVSIPVCNNSSVQVGCLTRERL   1362

TGPM    1412 LVNVTWRSSLSRLPVQSYPDLHNIERAFVAH--AQRFAQLLDDRGYRLAL   1459
             .|||||:|.|..:||.|.||||:||||.|..  ..||..|:....::|.|
TGHS    1363 GVNVTWKSRLEDIPVASLPDLHDIERALVGKDLLGRFTDLIQSGSFQLHL   1412
                                                      TG Type 2-1
TGPM    1460 NGRDYSSEGAARFRKDELYDTSPAFRVACEPGYVRI--------AQGCAA   1501
             :.:.:.:| ..||.:.:.:.|||.....|..|:.::        ..||..
TGHS    1413 DSKTFPAE-TIRFLQGDHFGTSPRTWFGCSEGFYQVLTSEASQDGLGCVK   1461
                TG Type 2-2          TG Type 2-3          Spacer 1 TG Type 1-11
TGPM    1502 CPRGSLWRDGGCALCPADHYQDEAGQTACTPCPADTGTLTAGAAAAFQCR   1551
             ||.||..:|..|..||...||:||..||.|||....|:|||.:...|.
TGHS    1462 CPEGSYSQDEECIPCPVGFYQEQAGSLACVPCPVGRTTISAGAFSQTHCV   1511

TGPM    1552 TPCER-----EALRRGQLGAGER----LELYC-DPEG----LYEGEKPAK   1587
             |.|:|     :..:.||..|.::    .:.:| |.||    .:|.|.|.:
TGHS    1512 TDCQRNEAGLQCDQNGQYRASQKDRGSGKAFCVDGEGRRLPWWETEAPLE   1561
                Spacer 2                                TG Type 3-a1
TGPM    1588 GGQDEVRRMFEEVPSEDLVLGAVDAVALRIRTETG--PVQEVEPRCLQEC   1635
             ..|..:.:.||:||...:..|...||:|.:....  ||.:    ||.:|
TGHS    1562 DSQCLMMQKFEKVPESKVIFDANAPVAVRSKVPDSEFPVMQ----CLTDC   1607
```

```
                    - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
TGPM        1636 LKEERCDFFVFSTQGDVSSCHFYNGSRESYQCEKAPEEKSALGISLCVVF   1685
                 .::|.|.||..||.....||.||..:.::..|..:.:::.|||.|....|
TGHS        1608 TEDEACSFFTVSTTEPEISCDFYAWTSDNVACMTSDQKRDALGNSKATSF   1657
                 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
TGPM        1686 IAIKCFVRGVGFHVSSASGRFYFHAGHELPSWDGQLYRPTGFGNAASGAY   1735
                 .:::|.|: |..| ...|...|...|....:...:.:.||||.|..||.|
TGHS        1658 GSLRCQVK-VRSH-GQDSPAVYLKKGQGSTTTLQKRFEPTGFQNMLSGLY   1705
                                          TG Type 3-b1
                 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
TGPM        1736 RTLALPAGTTTLPDAHLYCRSACSAQACCDGFLLKELPPGVLVCALLSGP   1785
                 ..:...|....|.||||:|..||....|||||:|.:::..|.::|.|||.|
TGHS        1706 NPIVFSASGANLTDAHLFCLLACDRDLCCDGFVLTQVQGGAIICGLLSSP   1755
                 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
TGPM        1786 TVLTCRAPWWPRTADEYGDGECSGLRVHKPSRTFGFSLGGVHYNASNPPS   1835
                 :||.|....|...::.:.:..|.|:...:.|......||...:..|..|
TGHS        1756 SVLLCNVKDWMDPSEAWANATCPGVTYDQESHQVILRLGDQEFIKSLTP-   1804
                 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
TGPM        1836 AVASTAAAAAAASQVIYLWAGSEAGSRDVCLHTKAYT-PQPGEPAGAADV   1884
                  :..|........|| |||..|:.|||...:..:..| |:|..|..|...
TGHS        1805 -LEGTQDTFTNFQQV-YLWKDSDMGSRPESMGCRKDTVPRPASPTEAGLT   1852
                                                          TG Type 3-a2
                 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
TGPM        1885 SGRFSTLRPEEVTLQPEMRVAQQNFWLFRRAFSPPQQATSWCLRRC-RRD   1933
                 :..||.:...:|.:.....::.|..|||:..|| .|||..|||.|| :..
TGHS        1853 TELFSPVDLNQVIVNGNQSLSSQKHWLFKHLFS-AQQANLWCLSRCVQEH   1901
                 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
TGPM        1934 ALCKAAAVGDGPAGWLECLLYPDTQSCGTSQELLERPARAPSCASLLPRI   1983
                 :.|:.|.:.:..:.:..|.|||:.|.|   .:::|  :.|..|..:||::
TGHS        1902 SFCQLAEITESASLYFTCTLYPEAQVC---DDIME--SNAQGCRLILPQM   1946
                 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
                                                                   TG Type 3-b2
TGPM        1984 DGTLYRKRDGAEGPVRRLYRRTAFRSEIGVLLRSVANLTGMALSDGFSRC   2033
                 ...|:||:...|..|:..|.|..|:..:|:.:|:...::..::|:||..|
TGHS        1947 PKALFRKKVILEDKVKNFYTRLPFQKLMGISIRNKVPMSEKSISNGFFEC   1996
                 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
TGPM        2034 ERTCDEDPCCEGFAFLRRALSSAGEALLQCATHTGPGVQLCGRESPMGA-   2082
                 ||.||.||||.||.||..:....||  :.|.|....|:|:|..|:. ||
TGHS        1997 ERRCDADPCCTGFGFLNVSQLKGGE--VTCLTLNSLGIQMCSEENG-GAW   2043
                 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
TGPM        2083 -TFGCPPPGVDTASSLFGWYRAQARSSPQTAQLCPAVTLPPAPTKGAADG   2131
                  ...|..|.::..:..||||:.....: .....||.|.||....|.:.|.
TGHS        2044 RILDCGSPDIEVHTYPFGWYQKPIAQN-NAPSFCPLVVLPSLTEKVSLDS   2092
                                                                TG Type 3-a3
                 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
TGPM        2132 FQSLDVSTAAVDPTISAFDVVVLLGPGAGGSPRAAPPSD-----EWCLAV   2176
                 :|||.:|:..|||:|..|||..:         ..|..|:     :.||:.
TGHS        2093 WQSLALSSVVVDPSIRHFDVAHV---------STAATSNFSAVRDLCLSE   2133
                 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
TGPM        2177 CQRSPWCASVGVRHLGDSTRCVLHAETLACGYGARGGRHCRLGVLEPPAR   2226
                 |.:...|....::....:.||:.:|:|.:|.:..: |::|||.:.|....
TGHS        2134 CSQHEACLITTLQTQPGAVRCMFYADTQSCTHSLQ-GQNCRLLLREEATH   2182
                                Spacer 3                 ChEL
                 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
TGPM        2227 LYRKKAPSPGGGQGPVRPVVPLGS-GQLQGASRTVEVDSQVKVIDVFLGV   2275
                 :|||...|....:..| |.||:.: |:|.|.|:.::|.:..|.:|.||||
TGHS        2183 IYRKPGISLLSYEASV-PSVPISTHGRLLGRSQAIQVGTSWKQVDQFLGV   2231
                 ================================================================
TGPM        2276 PYAAPPLDANRFSGPQPPRPLAVNETWDADRYKTRPDCLQPDGQRGTS--   2323
                 |||||||...||   |.|.||....:|||.  |.|..|.||..:..||
TGHS        2232 PYAAPPLAERRF---QAPEPLNWTGSWDAS--KPRASCWQPGTRTSTSPG   2276
                 ================================================================
TGPM        2324 LSEDCLYLNIFVPKVKPHNASVLVFFHG---GDNSFGGSAQGPLDPSYLA   2370
                 :||||||||:|:|:....|||||||||.   .:.|.|..|   :|.|:||
TGHS        2277 VSEDCLYLNVFIPQNVAPNASVLVFFHNTMDREESEGWPA---IDGSFLA   2323
                 ================================================================
TGPM        2371 ALGDIIVVTANYRLGLFGFLSTGDEAAAGNWGLLDQQAALRWVRDHAALF   2420
                 |:|::|||||:||:|:|||||:|....:||||||||.|||.||:.|...|
TGHS        2324 AVGNLIVVTASYRVGVFGFLSSGSGEVSGNWGLLDQVAALTWVQTHIRGF   2373
                 ================================================================
TGPM        2421 GGSAGAVTVAGDRSSADNVGLHLV---APGSRGLFQRAILMGGSVLSPSA   2467
                 ||....|::|.||..||...:||:   |..|: ||:||:|||||.|||:|
TGHS        2374 GGDPRRVSLAADRGGADVASIHLLTARATNSQ-LFRRAVLMGGSALSPAA   2422
```

```
TGPM        2468 VQLDSATARSQAASLAQLVGCGHSSSEYLVRCLRDRSALALNAAQAQLL-   2516
                 | :....|:.||.:||:.|.|..|||:.:|.|||.:.|..||.||.:||
TGHS        2423 V-ISHERAQQQAIALAKEVSCPMSSSQEVVSCLRQKPANVLNDAQTKLLA   2471

TGPM        2517 -SGEYLQRWAAVVDGHFLRDIPAEAAARGALADVDIVLGATEDDGALGRS   2565
                  ||.: ..|..|:||||||:.||.|..|....:||:::|:::|||.:.|:
TGHS        2472 VSGPF-HYWGPVIDGHFLREPPARALKRSLWVEVDLLIGSSQDDGLINRA   2520

TGPM        2566 RLPQVFPRFTSMSYSSAGFDVALQKVLRGAGASSFARDAVRWFYKRGDGA   2615
                 :..:.|......:.|...|..|||..|.|..:.:....|..|:|.....
TGHS        2521 KAVKQFEESRGRTSSKTAFYQALQNSLGGEDSDARVEAAATWYYSLEHS-   2569

TGPM        2616 NPPPLADDE---SWLLSNVSRDYDVICPAVHMAEMWAARGKANAFLYYVP   2662
                      .||.   |..|.|.:|||.:|||.:.||..||.|.:.|.|:|:.|
TGHS        2570 -----TDDYASFSRALENATRDYFIICPIIDMASAWAKRARGNVFMYHAP   2614

TGPM        2663 TQQSWNTLEWEPHSDTQLAFGVPLHPERKGRSCVGEQQLSRHFIGYLANF   2712
                 ......:||.  .:|.|.|.|:|.:|..:|:..:.|:.||...:.|.::|
TGHS        2615 ENYGHGSLEL--LADVQFALGLPFYPAYEGQFSLEEKSLSLKIMQYFSHF   2662

TGPM        2713 VKSGDPSFPNRHARGAGSRLLPAWPRFSLNEAGGFYKEVRAGMRNQRGLK   2762
                 ::||:|::|...:|...:...| ||.|.....|..|||....:.|::|||
TGHS        2663 IRSGNPNYPYEFSRKVPTFATP-WPDFVPRAGGENYKEFSELLPNRQGLK   2711

TGPM        2763 MRECSFWKDYVQPLMAATGGI--GEVERQWREDYNAWRQEALLEWRVQMS   2810
                 ..:||||..|:..|..:..|.  |:.....|:..|.            |
TGHS        2712 KADCSFWSKYISSLKTSADGAKGGQSAESEEEELTAG------------S   2749

TGPM        2811 NFRAAVTGSMATERPAAAPYV-   2831
                 ..|..:   ::.:.|.:..|.
TGHS        2750 GLREDL---LSLQEPGSKTYSK   2768
```

Figure 1. Comparative analysis of full-length thyroglobulins. A comparative analysis of the full-length thyroglobulin from *Petromyzon marinus* (TGPM, 2831 amino acids) and *Homo sapiens* (TGHS, 2768 amino acids) was conducted using the EMBOSS Needle program. Amino acids are represented using single-letter codes, with signal peptides enclosed in boxes. The linker, hinge and ChEL domain, along with eleven TG type 1 modules, three TG type 2 modules, five TG type 3 modules, and spacers 1, 2 and 3 are displayed. The cysteines are highlighted in yellow, while the amino-terminal hormonogenic sites (TGPM: Y^{27} and Y^{153}, TGHS: Y^{24} and Y^{149}) and carboxyl-terminal hormonogenic sites (TGPM: Y^{2830}, TGHS: Y^{2766}) are marked in red. Additionally, the 202 amino acids missing in the GL476337 scaffold are shaded gray.

Petromyzon marinus	TG11PM	ASA	CE	12aa	HVPQC	3aa	GAFRPLQC	6aa	CWCVD	2aa	GEE	2aa	GTR	8aa	C	
	TG12PM	LSF	CQ	18aa	HLPQC	3aa	GRYRPAQV	5aa	GWCVD	2aa	GME	2aa	GTR	7aa	C	
	TG13PM	PSP	CQ	14aa	PPPQC	4aa	QRVLATQC	40aa	CYCAD	2aa	GRE	4aa	GDR	46aa		←
	TG14PM	PTV	CE	13aa	FAPVC	3aa	GSYIPTQC	4aa	CWCVD	2aa	GDE	2aa	GSR	7aa	C	
	TG15PM	ADA		9aa	FVARC	4aa	GAYEPVQC	4aa	CWCVD	2aa	GRE	2aa	GTR	8aa	C	
	TG16PM	PSA	CE	19aa	FVPEC	3aa	GDYRPVQC	4aa	CFCVG	2aa	GAE	2aa	GTG	8aa	C	
	TG17PM	PTP	CQ	29aa	YVPRC	3aa	GGWRPVQC	128aa	CWCVD	2aa	GHE	2aa	GTR	9aa	C	←
	TG18PM	PGA	CH	100aa	YMPQC	3aa	GGWEPAQC	8aa	CWCVD	2aa	GEF	2aa	GSL	9aa	C	←
	TG19PM	ATP	CQ	28aa	HTPAC	1aa	PSRKYKKC	29aa								
	TG110PM	PSL	CE	14aa	SVPEC	3aa	GDYAARQC	4aa	CWCAG	2aa	GEE	2aa	DTR	9aa	C	
	TG111PM	RTP	CE	14aa	LELYC	3aa	GLYEGEKP	6aa								
Homo sapiens	TG11HS	LRP	CE	12aa	YVPQC	3aa	GSFQTVQC	6aa	CWCVG	2aa	GSE	2aa	GSR	7aa	C	
	TG12HS	LSF	CQ	18aa	YLPQC	3aa	GDYAPVQC	6aa	CWCVD	2aa	GME	2aa	GTR	7aa	C	
	TG13HS	PRS	CE	13aa	SPPQC	3aa	GEFMPVQC	40aa	CHCAD	2aa	GRE	4aa	GLE	45aa	C	←
	TG14HS	PTK	CE	12aa	YVPSC	3aa	GDYQAVQC	5aa	CWCVD	2aa	GKE	2aa	GTR	7aa	C	
	TG15HS	SQT	CE	6aa	FVPSC	3aa	GSYEDVQC	4aa	CWCVN	2aa	GKE	2aa	GSR	7aa	C	
	TG16HS	PTD	CE	19aa	FVPAC	3aa	GHFLPVQC	4aa	CYCVD	2aa	GQA	2aa	GTR	8aa	C	
	TG17HS	PTP	CQ	27aa	YIPQC	3aa	GQWRQVQC	123aa	CWCVD	2aa	GQE	2aa	GMR	8aa	C	←
	TG18HS	PGS	CE	100aa	YMPQC	3aa	GSWEPVQC	6aa	CWCVD	2aa	GGF	2aa	GSL	9aa	C	←
	TG19HS	PTT	CE	25aa	FVPAC	3aa	GEYARLQA	4aa	TWCVD	2aa	SGE	2aa	RPG	6aa	C	
	TG110HS	PSL	CN	14aa	YVPAC	4aa	GGFSPVQC	6aa	CWCVM	2aa	GEE	2aa	GTR	7aa	C	
	TG111HS	VTD	CQ	3aa	AGLQC	3aa	GQYRASQK	6aa	AFCVD	2aa	GRR	2aa	WWE	9aa	C	

Figure 2. Alignment of eleven TG type 1 modules. The alignment of the eleven TG type 1 modules of full-length thyroglobulin from *Petromyzon marinus* (TGPM) and *Homo sapiens* (TGHS) was conducted according to Parma y col. [1987], Malthíery y col. [1987], Molina y col. [1996], van Graaf y col., [2001]**.** The six cysteines are conserved across all modules, except in TG type 1-2, TG type 1-3, TG type 1-5, TG type 1-9 and TG type 1-11 in TGPM, and in TG type 1-9 and TG type 1-11 in TGHS. Additionally, TG type 1-3, TG type 1-7 and TG type 1-8 contain long inserts, which are underlined in the alignment.

TG11PM: TG type 1-1 TGPM, TG12PM: TG type 1-2 TGPM, TG13PM: TG type 1-3 TGPM, TG14PM: TG type 1-4 TGPM, TG15PM: TG type 1-5 TGPM, TG16PM: TG type 1-6 TGPM, TG17PM: TG type 1-7 TGPM, TG18PM: TG type 1-8 TGPM, TG19PM: TG type 1-9 TGPM, TG110PM: TG type 1-10 TGPM, TG111PM: TG type 1-11 TGPM.

TG11HS: TG type 1-1 TGHS, TG12HS: TG type 1-2 TGHS, TG13HS: TG type 1-3 TGHS, TG14HS: TG type 1-4 TGHS, TG15HS: TG type 1-5 TGHS, TG16HS: TG type 1-6 TGHS, TG17HS: TG type 1-7 TGHS, TG18HS: TG type 1-8 TGHS, TG19HS: TG type 1-9 TGHS, TG110HS: TG type 1-10 TGHS, TG111HS: TG type 1-11 TGHS.

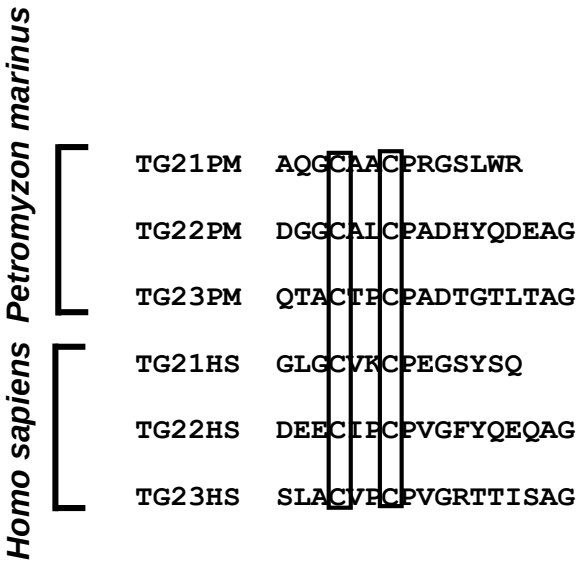


Figure 3. Alignment of three TG type 2 modules. The alignment of the three TG type 2 modules of full-length thyroglobulin from *Petromyzon marinus* (TGPM) and *Homo sapiens* (TGHS) was conducted according to Parma et al. (1987), Malthiéry et al. (1987), Molina et al. (1996), and van Graaf et al. (2001). The two cysteines are conserved across all modules.
TG21PM: TG type 2-1 TGPM, TG22PM: TG type 2-2 TGPM, TG23PM: TG type 2-3 TGPM.
TG21HS: TG type 2-1 TGHS, TG22HS: TG type 2-2 TGHS, TG23HS: TG type 2-3 TGHS.

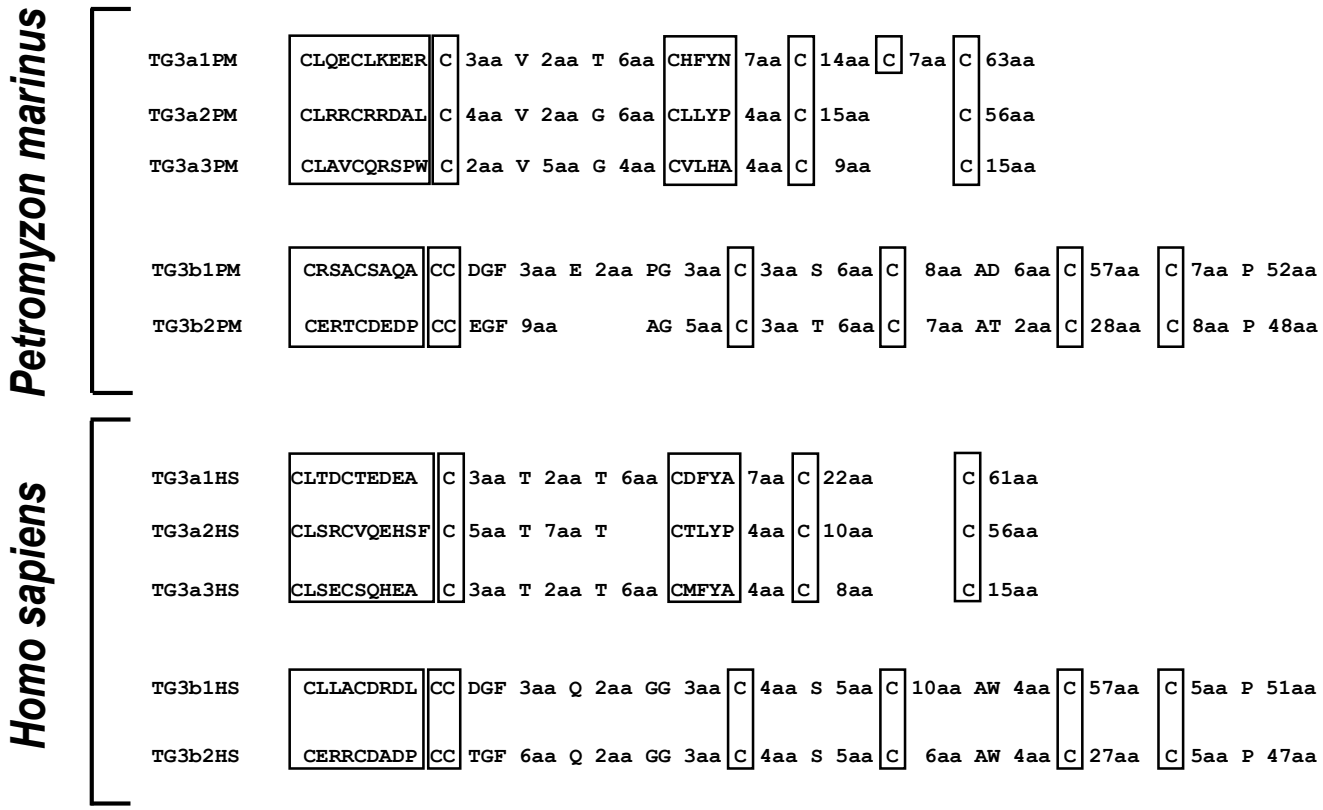


Figure 4. Alignment of five TG type 3 modules. The alignment of the five TG type 3 modules of full-length thyroglobulin from *Petromyzon marinus* (TGPM)and *Homo sapiens* (TGHS) was conducted according to Parma y col. [1987], Malthíery y col. [1987], Molina y col. [1996], van Graaf y col., [2001]. All six cysteine are conserved in the TG type 3-a modules, while all eight cysteines are preserved in the TG type 3-b modules of TGPM and TGHS. TG type 3-a1 in TGPM contains an extra cysteine.

TG3a1PM: TG type 3-a1 TGPM, TG3b1PM: TG type 3-b1 TGPM, TG3a2PM: TG type 3-a2 TGPM, TG3b1PM: TG type 3-b2 TGPM, TG3a3PM: TG type 3-a3 TGPM.

TG3a1HS: TG type 3-a1 TGHS, TG3b1HS: TG type 3-b1 TGHS, TG3a2HS: TG type 3-a2 TGHS, TG3b2HS: TG type 3-b2 TGHS, TG3a3HS: TG type 3-a3 TGHS.

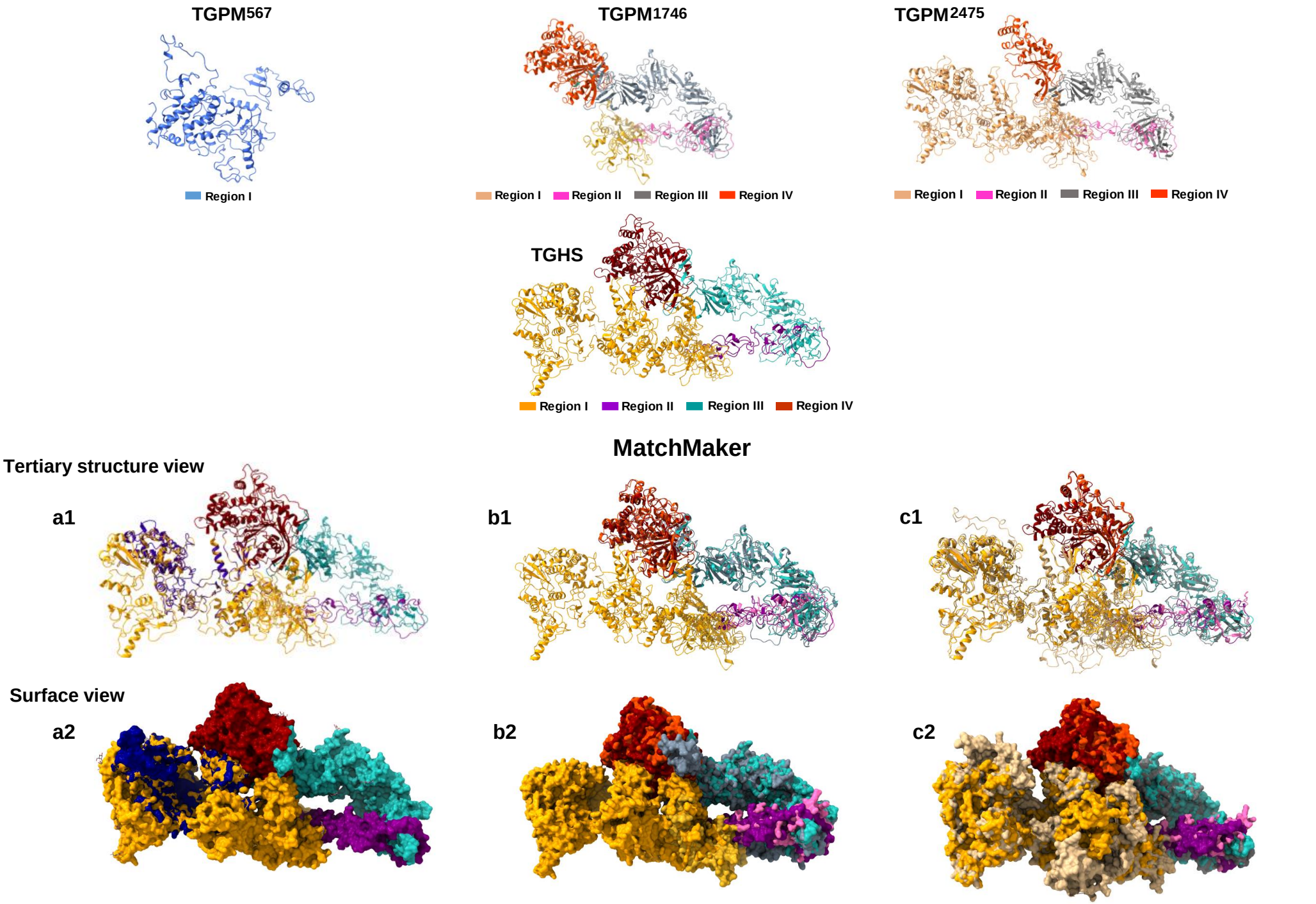

TGPM567
Region I
TGPM1746
Region I
Region II
Region III
Region IV
TGPM2475
Region I
Region II
Region III
Region IV
TGHS
Region I
Region II
Region III
Region IV
MatchMaker
Tertiary structure view
a1
b1
c1
Surface view
a2
b2
c2

Figure 5. Structural alignment of thyroglobulins from *Petromyzon marinus* and *Homo sapiens.* The structural alignment of **thyroglobulins** $TGPM^{567}$, $TGPM^{1747}$, and $TGPM^{2475}$ from *Petromyzon marinus* with **Homo sapiens thyroglobulin (TGHS)** is presented. The **top section** displays the **homology modeling** of $TGPM^{567}$, $TGPM^{1747}$, and $TGPM^{2475}$. The **middle section** shows the **TGHS model**. The **bottom section** illustrates the **superposition of thyroglobulin monomers** ($TGPM^{567}$ **with TGHS,** $TGPM^{1747}$ **with TGHS, and** $TGPM^{2475}$ **with TGHS**) across the **four regions (I, II, III, IV)** of the **classical model,** using the **Matchmaker command** in **UCSF ChimeraX** (https://www.cgl.ucsf.edu/chimera/). Additionally, in **a1, b1, and c1**, the **superposition of monomers** is displayed **without the surface**, while in **a2, b2, and c2**, it is shown **with the surface**.

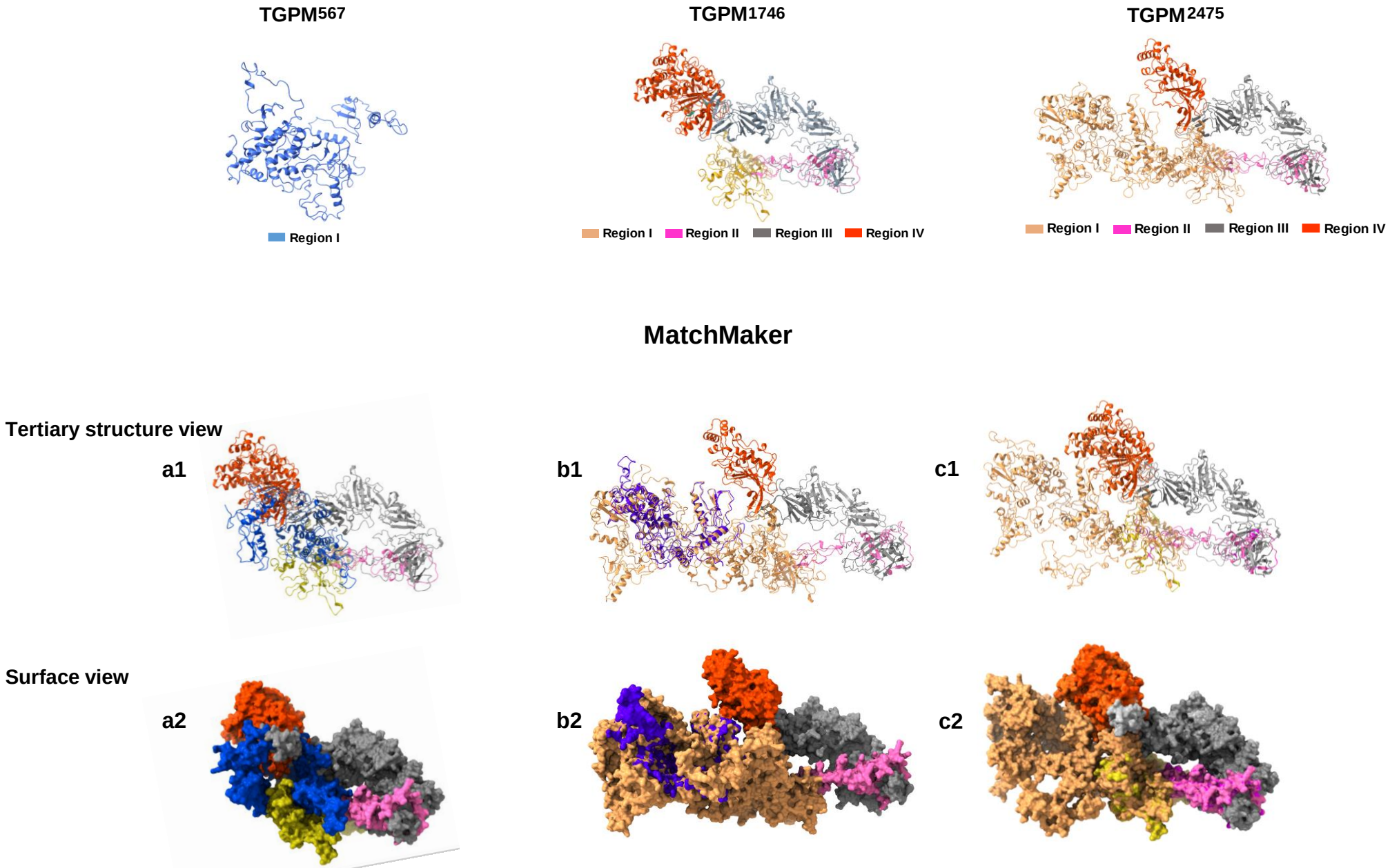


Figure 6. Structural alignment of thyroglobulins from *Petromyzon marinus*. The structural alignment of thyroglobulins TGPM^{567}, TGPM^{1747}, and TGPM^{2475} from *Petromyzon marinus* is presented. The top section displays the homology modeling of TGPM^{567}, TGPM^{1747}, and TGPM^{2475}. The middle and bottom sections illustrate the superposition of thyroglobulin monomers, including TGPM^{567} with TGPM^{1747}, TGPM^{567} with TGPM^{2475}, and TGPM^{1747} with TGPM^{2475}, across the four regions (I, II, III, IV) of the classical model, using the Matchmaker command in UCSF ChimeraX (https://www.cgl.ucsf.edu/chimera/). Additionally, in a1, b1, and c1, the superposition of monomers is displayed without the surface, while in a2, b2, and c2, it is shown with the surface.

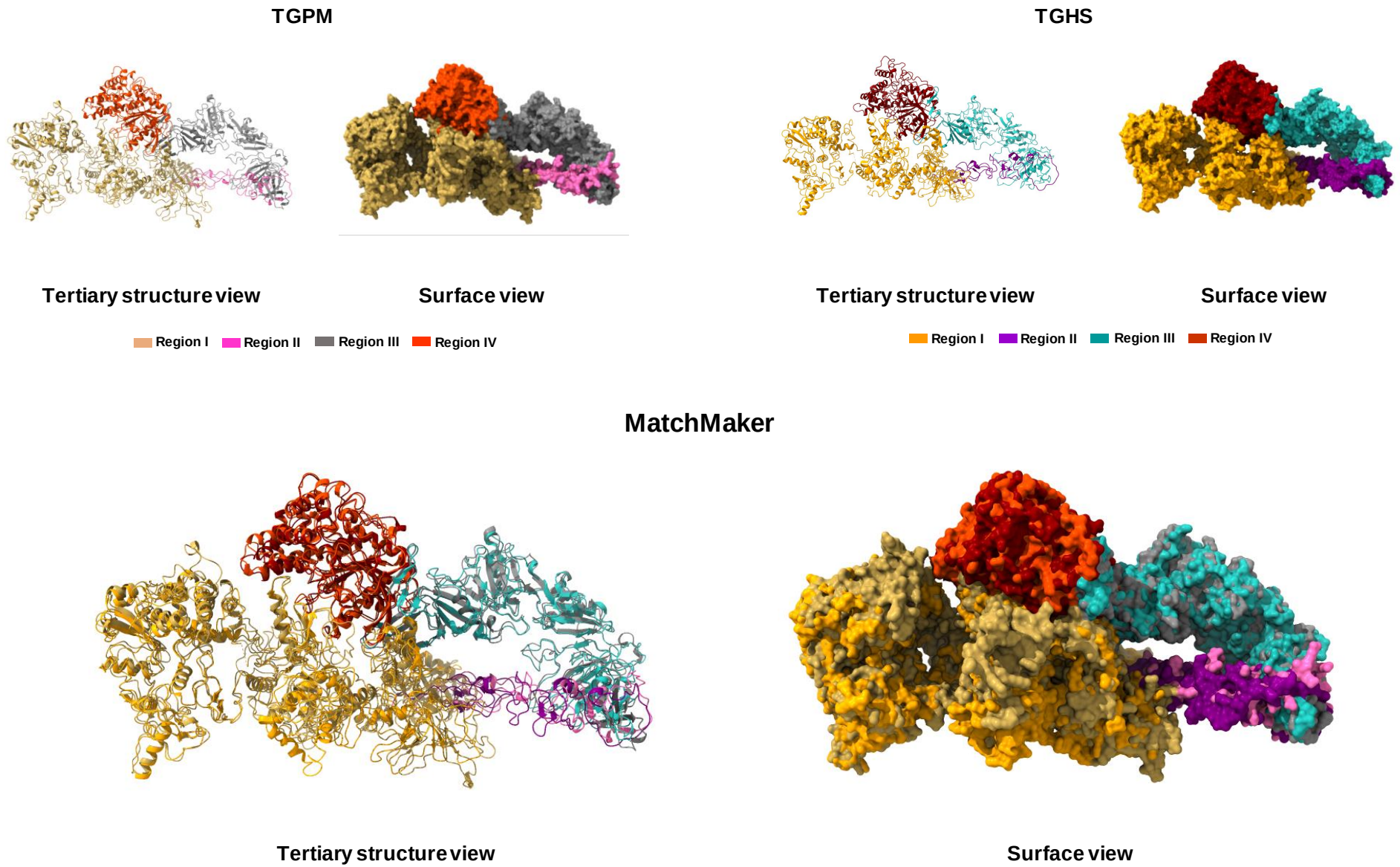
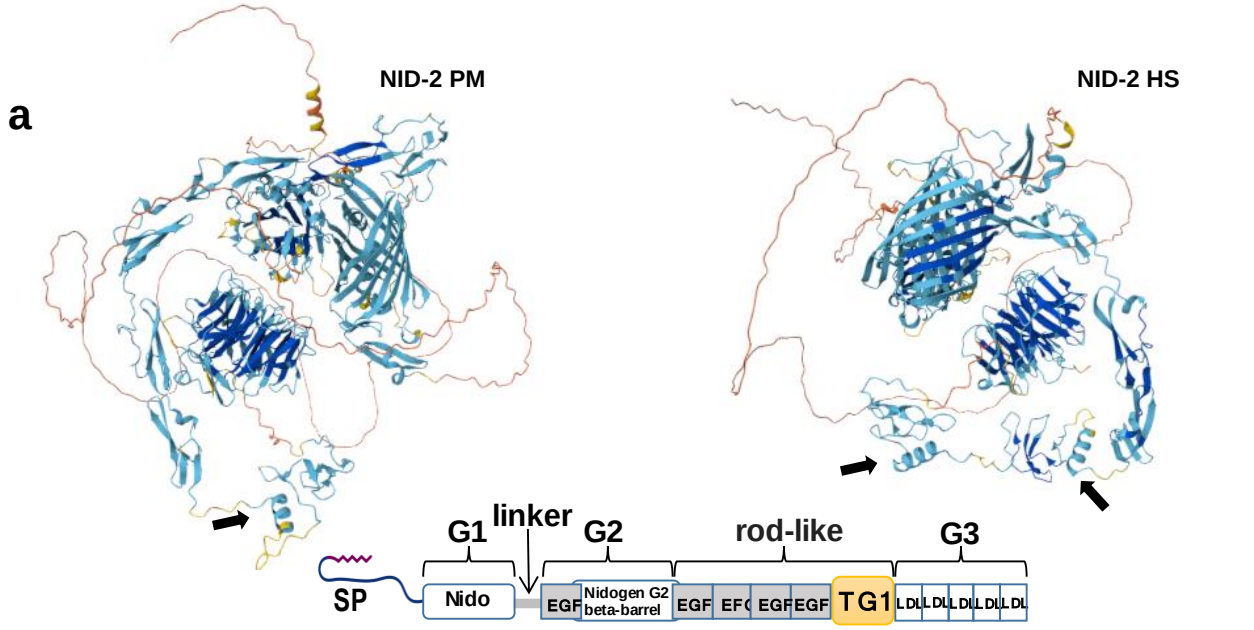
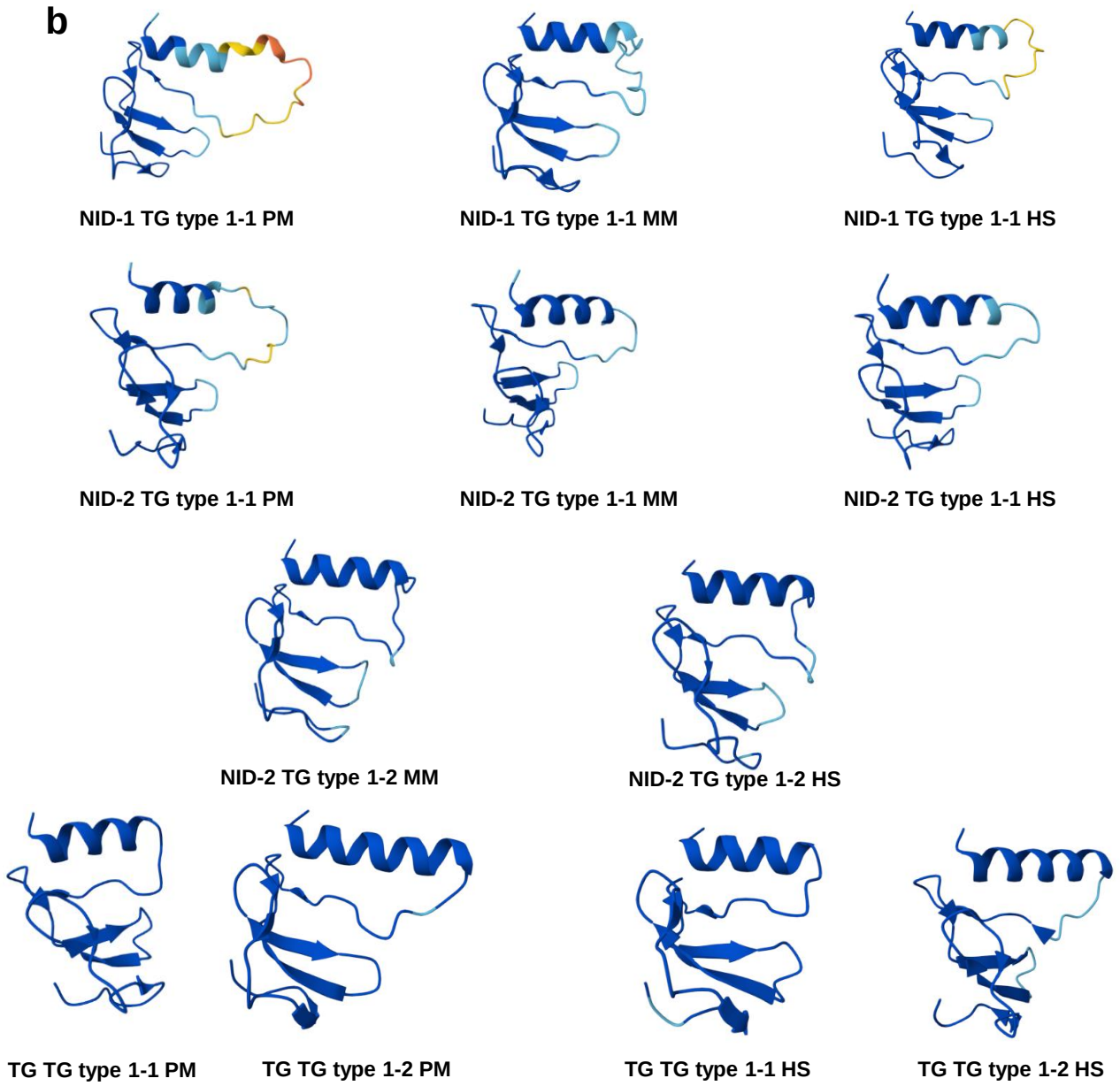


Figure 7. Structural alignment of full-length thyroglobulins from *Petromyzon marinus* and *Homo sapiens.* The structural alignment of full-length **thyroglobulins** from *Petromyzon marinus* **(TGPM)** and *Homo sapiens* **(TGHS)** is presented. The **top section** displays the **homology modeling** of **TGPM and TGHS.** The **bottom section** illustrates the **superposition of thyroglobulin monomers (TGPM with TGHS)** across the **four regions (I, II, III, IV)** of the **classical model**, using the **Matchmaker command** in **UCSF ChimeraX** (https://www.cgl.ucsf.edu/chimera/). Additionally, the **left side** shows the **superposition of monomers without the surface**, while the **right side** displays the **superposition with the surface**.





Very high (pIDDT>90) Confident (90>pIDDT) Low (70>pIDDT<50) Very low (pIDDT<50)

Figure 8. Three-dimensional atomic structure of nidogen 2, and TG type 1 modules using AlphaFold server. (a) The 3D structure of nidogen 2 (NID-2) from *Petromyzon marinus* and *Homo sapiens*. Arrows indicate the TG type 1 modules. (b) The modeled structure of TG type 1-1 and TG type 1-2 modules of nidogen 1 (NID-1), nidogen 2 (NID-2), and thyroglobulin (TG) from *Petromyzon marinus* (PM), *Mus musculus* (MM), and *Homo sapiens* (HS). The pLDDT (predicted Local Distance Difference Test) provides a per-atom confidence estimate on a 0-100 scale, where a higher value indicate greater reliability. Regions with low pLDDT scores may correspond to unstructured or highly flexible areas.

TG11N1PM	KST	CE	25aa	YVPQC	3aa	GTFLPVQC	6aa	CWCVD	2aa	GQE	2aa	GTR	8aa	C
TG11N1MM	KTR	CQ	23aa	FVPQC	3aa	GHYVPTQC	6aa	CWCVD	2aa	GRE	2aa	GSR	8aa	C
TG11N1HS	KTR	CQ	23aa	FVPEC	3aa	GHYAPTQC	6aa	CWCVD	2aa	GRE	2aa	GTR	8aa	C
TG11N2PM	RTP	CE	22aa	YVPQC	3aa	GAFRPVQC	6aa	CWCVT	2aa	GQE	2aa	GTR	8aa	C
TG11N2MM	LKP	CE	17aa	HIPQC	3aa	GNFVPLQC	6aa	CWCVD	2aa	GHE	2aa	GTQ	9aa	C
TG11N2HS	LTP	CE	17aa	HIPQC	3aa	GNFLPLQC	6aa	CWCVD	2aa	GHE	2aa	GTQ	9aa	C
TG11PM	ASA	CE	12aa	HVPQC	3aa	GAFRPLQC	6aa	CWCVD	2aa	GEE	2aa	GTR	8aa	C
TG11HS	LRP	CE	12aa	YVPQC	3aa	GSFQTVQC	6aa	CWCVG	2aa	GSE	2aa	GSR	7aa	C
TG12N2MM	RTV	CE	18aa	YVPQC	3aa	GHFIPLQC	6aa	CWCVD	2aa	GRE	2aa	GTR	8aa	C
TG12N2HS	PTI	CE	18aa	YVPQC	3aa	GHFIPLQC	6aa	CWCVD	2aa	GRE	2aa	GTR	8aa	C
TG12PM	LSF	CQ	18aa	HLPQC	3aa	GRYRPAQV	5aa	GWCVD	2aa	GME	2aa	GTR	7aa	C
TG12HS	LSF	CQ	18aa	YLPQC	3aa	GDYAPVQC	6aa	CWCVD	2aa	GME	2aa	GTR	7aa	C

Figure 9. Alignment of TG type 1 modules across species. The alignment of TG type 1 modules from *Petromyzon marinus* (PM), *Mus musculus* (MM), and *Homo sapiens* (HS)—including nidogen 1 (NID-1) and nidogen 2 (NID-2), as well as full-length thyroglobulins from *Petromyzon marinus* (TGPM) and *Homo sapiens* (TGHS)— was conducted according to Parma et al. (1987), Malthiéry et al. (1987), Molina et al. (1996), and van Graaf et al. (2001). The six cysteines are conserved across all modules.
TG11N1PM: PM NID-1 TG type 1-1, TG11N1MM: MM NID-1 TG type 1-1, TG11N1HS: HS NID-1 TG type 1-1, TG11N2PM: PM NID-2 TG type 1-1, TG11N2MM: MM NID-2 TG type 1-1, TG11N2HS: HS NID-2 TG type 1-1, TG11PM: TGPM TG type 1-1, TG11HS: TGHS TG type 1-1, TG12N2MM: MM NID-2 TG type 1-2, TG12N2HS: HS NID-2, TG type 1-2, TG12PM: TGPM TG type 1-2, TG12HS: TGHS TG type 1-2.

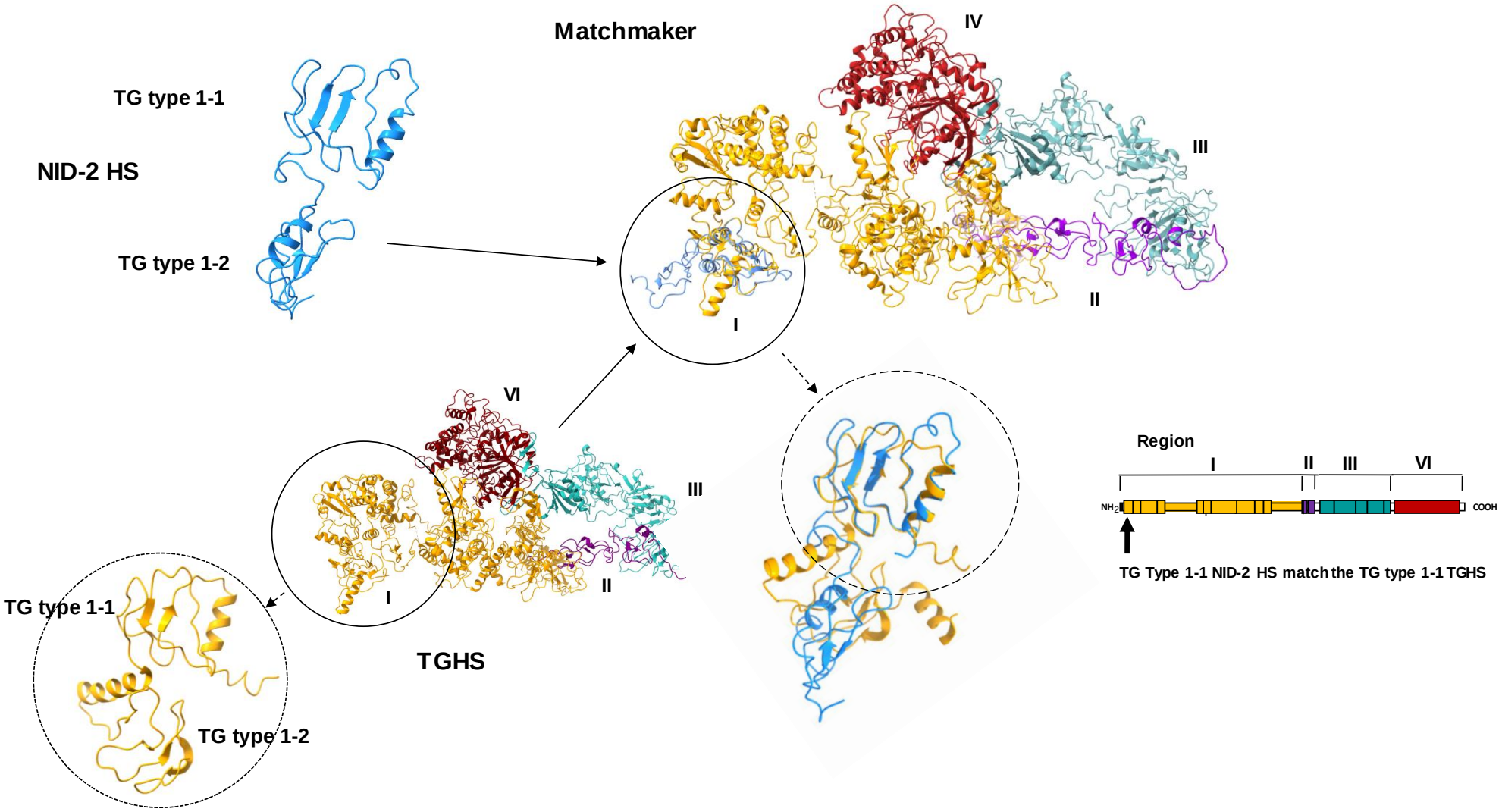


Figure 10. Structural alignment of TG type 1-1 modules. The structural alignment of TG type 1-1 module from *Homo sapiens* nidogen 2 (NID-2 HS) and thyroglobulin (TGHS) was performed using the Matchmaker command in UCSF ChimeraX (https://www.cgl.ucsf.edu/chimera/). Interestingly, the TG type 1-1 module of NID-2 HS exactly matches the TG type 1-1 module of TGHS (dotted circle). In contrast, the TG type 1-2 module of NID-2 HS does not align with TG type 1-2 module of TGHS. The four regions (I, II, III, IV) of the classical model of TGHS are displayed.

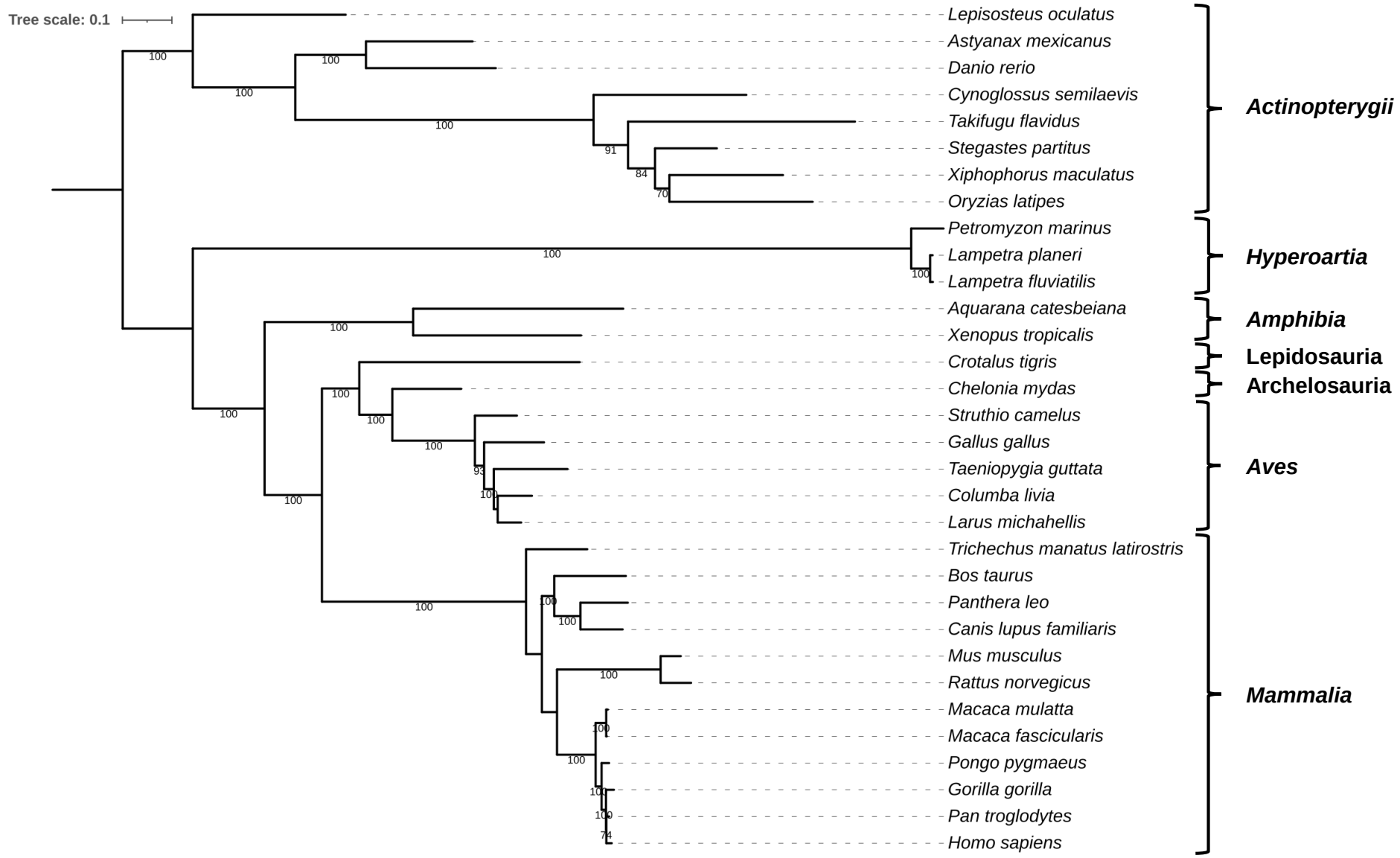


Figure 11. Phylogenetic analysis of thyroglobulins from 32 species. Maximum-likelihood phylogeny generated from alignment of thirty TG amino acid sequences shows the phylogenetic relationship among the species. Numbers over the nodes represent the ultrafast bootstrap values ≥70 computed by 2000 replicates. Actinopterygii, Hyperoartia, Amphibia, Aves, and Mammalia are taxonomic classes, whereas Lepidosauria and Archelosauria are taxonomic clades.

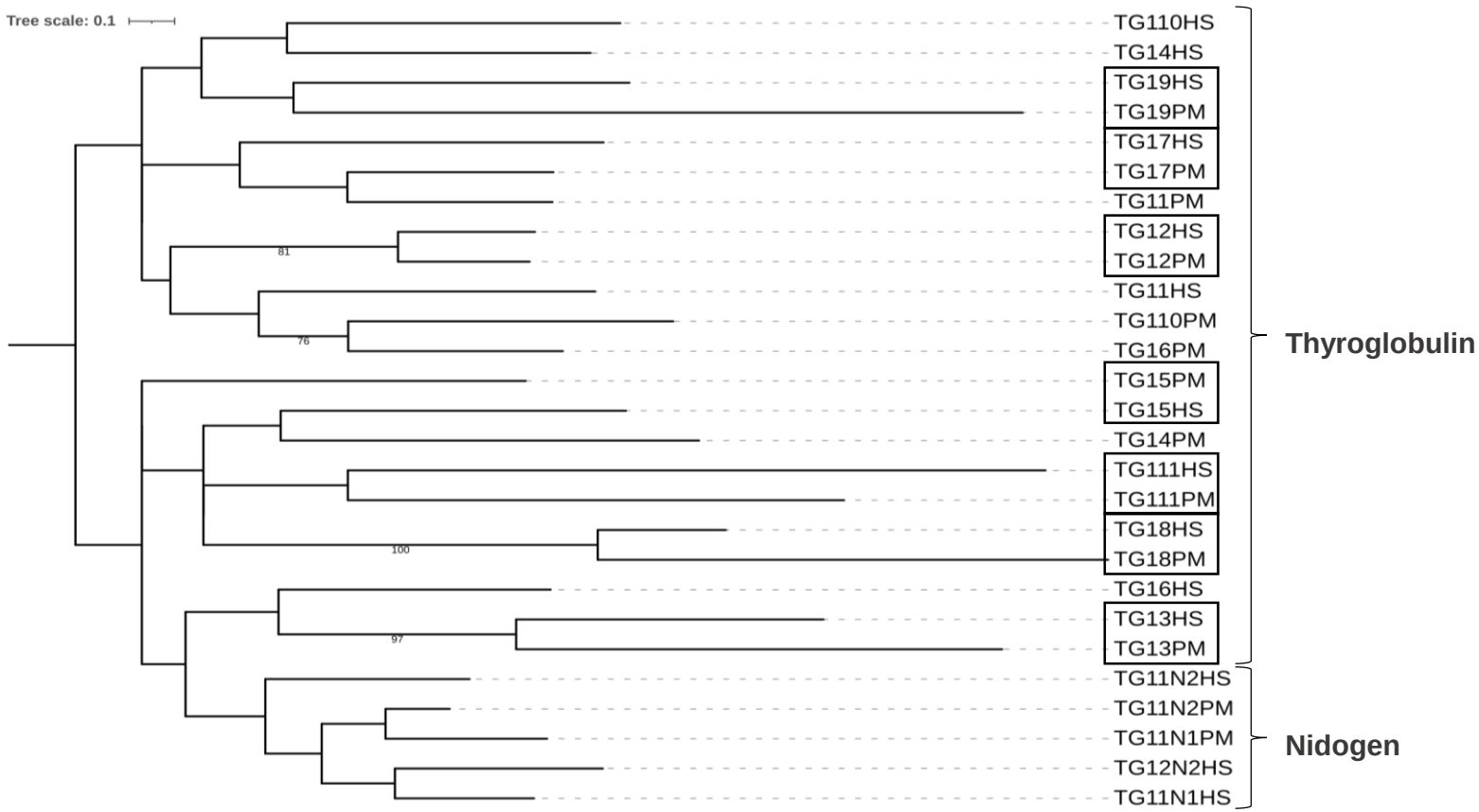


Figure 12. Phylogenetic analysis of TG type 1 domains for *Homo sapiens* and *Petromyzon marinus*. Maximum-likelihood phylogeny generated from alignment of TG type 1 domains shows the phylogenetic relationship among the species. Numbers over the nodes represent the ultrafast bootstrap values ≥70 computed by 2000 replicates. *Homo sapiens* (HS) and *Petromyzon marinus* (PM).

TG11PM: TG type 1-1 TGPM, TG12PM: TG type 1-2 TGPM, TG13PM: TG type 1-3 TGPM, TG14PM: TG type 1-4 TGPM, TG15PM: TG type 1-5 TGPM, TG16PM: TG type 1-6 TGPM, TG17PM: TG type 1-7 TGPM, TG18PM: TG type 1-8 TGPM, TG19PM: TG type 1-9 TGPM, TG110PM: TG type 1-10 TGPM, TG111PM: TG type 1-11 TGPM.

TG11HS: TG type 1-1 TGHS, TG12HS: TG type 1-2 TGHS, TG13HS: TG type 1-3 TGHS, TG14HS: TG type 1-4 TGHS, TG15HS: TG type 1-5 TGHS, TG16HS: TG type 1-6 TGHS, TG17HS: TG type 1-7 TGHS, TG18HS: TG type 1-8 TGHS, TG19HS: TG type 1-9 TGHS, TG110HS: TG type 1-10 TGHS, TG111HS: TG type 1-11 TGHS.

TG11N1PM: PM NID-1 TG type 1-1, TG11N2PM: PM NID-2 TG type 1-1, TG11N1HS: HS NID-1 TG type 1-1, TG11N2HS: HS NID-2 TG type 1-1, TG12N2HS: HS NID-2 TG type 1-2.

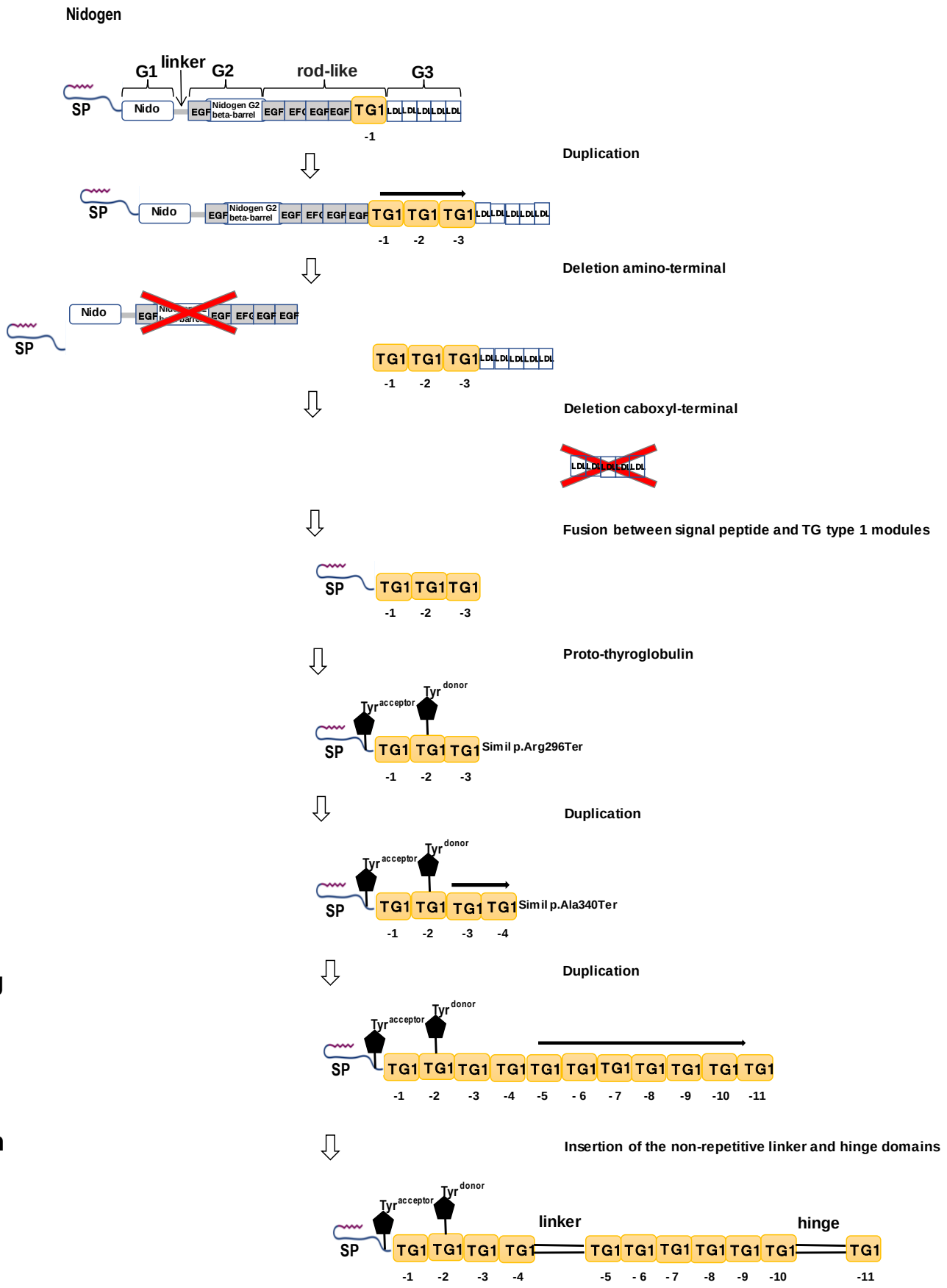


Figure 13. Flow chart of thyroglobulin complexation (stages a, b, c, d, e, f, g, and h). The flow chart illustrates the complexation of thyroglobulin, tracing its progression from nidogen to the acquisition of eleven TG type 1 modules and the insertion of non-repetitive linker and hinge domains.

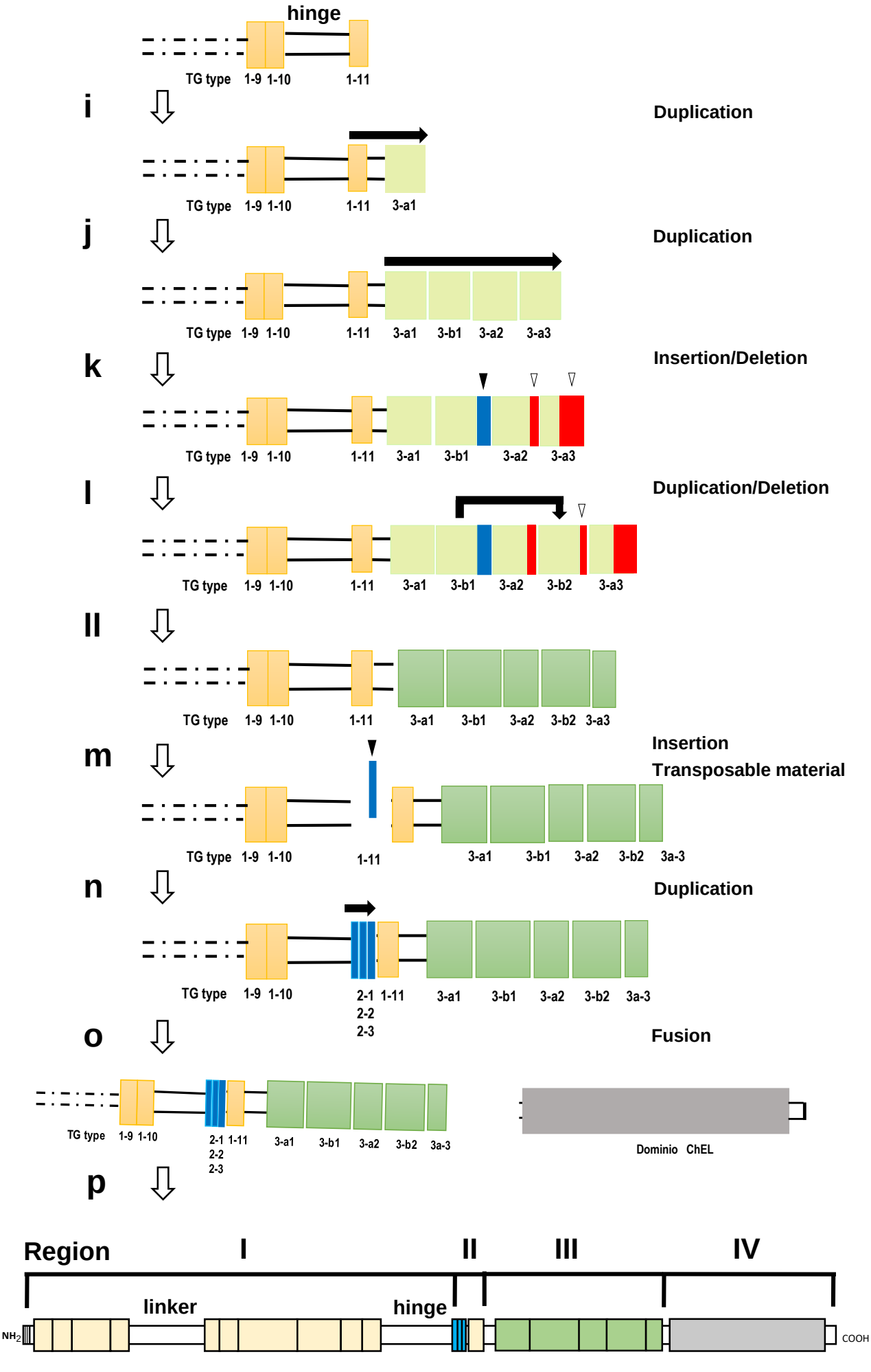

hinge
TG type 1-9 1-10 1-11
i
Duplication
TG type 1-9 1-10 1-11 3-a1
j
Duplication
TG type 1-9 1-10 1-11 3-a1 3-b1 3-a2 3-a3
k
Insertion/Deletion
TG type 1-9 1-10 1-11 3-a1 3-b1 3-a2 3-a3
l
Duplication/Deletion
TG type 1-9 1-10 1-11 3-a1 3-b1 3-a2 3-b2 3-a3
II
TG type 1-9 1-10 1-11 3-a1 3-b1 3-a2 3-b2 3-a3
m
Insertion
Transposable material
TG type 1-9 1-10 1-11 3-a1 3-b1 3-a2 3-b2 3a-3
n
Duplication
TG type 1-9 1-10 2-1 2-2 2-3 1-11 3-a1 3-b1 3-a2 3-b2 3a-3
o
Fusion
TG type 1-9 1-10 2-1 2-2 2-3 1-11 3-a1 3-b1 3-a2 3-b2 3a-3
Dominio ChEL
p
Region I II III IV
linker
hinge
NH2
COOH

Figure 14. Flow chart of thyroglobulin complexation (stages i, j, k, d, l, ll, m, n and p). The flow chart illustrates the complexation of thyroglobulin, tracing its progression from the acquisition of five TG type 3 modules and three TG type 2 modules to the fusion of the ChEL domain. The four regions (I, II, III, IV) of the classical model of TGHS are displayed in botton.